Abstraction-Dependence in Partially Schematic Constructions: Named-Entity Slots in Spoken English

Yoichiro Hasebe
Doshisha University
yhasebe@mail.doshisha.ac.jp

Abstract

Partially schematic constructions—fixed lexical anchors paired with open slots—are difficult to retrieve from surface-indexed corpora without severe coverage loss. To overcome this bottleneck, this study integrates syntactic noun-chunking, grammatical category tagging, and named-entity typing within the TED Corpus Search Engine (TCSE). Using a 13-million-token snapshot of TED Talks, the paper introduces a quantitative metric—*abstraction-dependence*—defined as the ratio of a construction's total frequency to that of its single most frequent surface realization. Conventional surface queries retrieve a median of only 8.2% of instances across 32 single-slot prepositional frames, dropping as low as 0.6% in highly productive open slots; the proposed multi-layer queries retrieve these constructions systematically. For complex argument structures like the ditransitive, whose recipient slot blends pronouns, proper names, and full noun phrases, a single typed-noun-chunk query retrieves all 12,301 instances of the pattern across five prototypical verbs, and head-category conditions on the recipient slot partition them automatically into pronominal (85.5%), common-noun (13.3%), and proper-noun (1.2%) realizations—a classification corroborated by dependency parsing. These findings convert qualitative claims about slot productivity into a precise corpus-based metric, with applications extending from descriptive grammar to the evaluation of abstract structural generalization in language models.

Keywords: Construction Grammar, Corpus Linguistics, Named Entity Recognition, Slot Productivity, TED Corpus Search Engine

1. Introduction

A central premise of usage-based linguistics is that grammar consists of a structured inventory of constructions—form–meaning pairings that range from fully specified lexical items to fully abstract phrasal schemas (Langacker, 1987, 2008; Goldberg, 1995, 2006; Kay and Fillmore, 1999; Croft, 2001). Between these two extremes lie partially schematic constructions: patterns that combine fixed lexical material with one or more open slots, such as single-slot prepositional frames (`in %DATE`, `by %PERSON`)[1] or complex argument structures like the ditransitive (`[verb] _ _`), whose recipient slot accommodates a heterogeneous mix of pronouns, proper names, and full noun phrases. In principle, a construction's quantitative distribution across a corpus should illuminate its productivity, its semantic constraints, and its degree of schematization (Bybee, 2010; Goldberg, 2019; Schmid, 2020).

However, in empirical practice, a significant methodological bottleneck persists: surface-frequency-based corpus retrieval methods inherently fragment partially schematic constructions. Standard corpus search tools index surface forms or single-token POS tags. Even when query languages allow for numerical distance constraints or variable-length wildcards (as in COCA), surface methods struggle to delimit slot boundaries precisely—for instance, distinguishing the scope of a one-word pronominal recipient from that of a multi-word noun phrase recipient in ditransitive constructions.

Consequently, researchers seeking to quantify open constructional slots have been forced into a binary choice: limit retrieval to a curated subset of surface realizations (e.g., *in 2015*, *give him*), as when Goldberg (2011) deliberately restricts COCA queries for the ditransitive to pronominal recipients,[2] thereby forfeiting an unknown proportion of the construction's instances, or attempt structural retrieval on small, fully parsed treebanks (such as ICE-GB; e.g., Gries and Stefanowitsch, 2004), thereby forfeiting statistical scale and generalizability. Because open slots accommodate diverse filler types and variable constituent spans, surface queries scatter a single schematic construction into hundreds of isolated string frequencies. While this fragmentation is widely acknowledged in usage-based work, the extent of retrieval loss is rarely quantified systematically across different constructional frames.

This paper addresses this methodological challenge by demonstrating how explicitly typed annotation layers—combining syntactic noun-chunking, part-of-speech tagging, and named-entity recognition (NER)—can be integrated within a unified query engine to retrieve partially schematic constructions. Utilizing the TED Corpus Search Engine (TCSE; Hasebe, 2015, 2018), which indexes a 13-million-token snapshot of TED Talks across multiple annotation layers, this study shows that typed abstraction enables researchers to query open constructional slots directly as primary search terms. This multi-layer architecture bridges the gap between surface string matching and full dependency parsing, rendering complex constructional slots systematically countable at scale.

To evaluate this framework empirically, this study addresses four specific research questions:

1. To what extent do single-slot prepositional frames depend on type abstraction for their retrieval, and how severely do top surface queries undercount them?
2. How can this retrieval dependence be quantified through a straightforward corpus-level metric across target frames?

[1] Here and throughout, expressions in monospaced type are TCSE search queries, executable verbatim in its freely accessible web interface; see https://yohasebe.github.io/tcse-doc/ for the full query syntax. The sole abbreviation is `[verb]`, which stands for the concrete verb disjunction `[give|send|tell|show|offer]` used in the ditransitive analysis (Section 3.3).

[2] Goldberg's (2011) ditransitive counts rest on the COCA pattern "[v] [definite pronoun] [lexical NP]", which fixes the recipient as a definite pronoun and the theme as a lexical noun phrase; the typed-chunk queries introduced below retrieve both slots without such surface commitments (Section 3.3).

Type	Description
%PERSON	People, including fictional
%NORP	Nationalities, religious and political groups
%FAC	Buildings, airports, highways, bridges
%ORG	Companies, agencies, institutions
%GPE	Countries, cities, states
%LOC	Non-GPE locations: mountain ranges, bodies of water
%PRODUCT	Objects, vehicles, foods (not services)
%EVENT	Named hurricanes, battles, wars, sports events
%WORK_OF_ART	Titles of books, songs, etc.
%LAW	Named documents made into laws
%LANGUAGE	Any named language
%DATE	Absolute or relative dates or periods
%TIME	Times smaller than a day
%PERCENT	Percentage expressions
%MONEY	Monetary values
%QUANTITY	Measurements of weight, distance, etc.
%ORDINAL	*first*, *second*, etc.
%CARDINAL	Numerals not covered by another type

Table 1: The 18 entity types of spaCy's en_core_web_lg model, as indexed in TCSE (descriptions abridged from the spaCy documentation; see https://spacy.io/models/en).

3. What entity-type preferences do open prepositional slots exhibit in spoken monologue data?
4. How can multiple abstraction layers be composed to retrieve complex argument structures like the ditransitive and other multi-slot patterns?

2. Data and Infrastructure

2.1 Dataset and System Architecture

The empirical foundation of this study is the TED Corpus Search Engine (TCSE; Hasebe, 2015, 2018), a freely available web-based corpus system built on transcripts of TED Talks (no login required). The dataset analyzed here comprises a frozen snapshot (retrieved August 10, 2026) containing 6,419 talks, totaling 13,017,589 tokens across 677,487 sentence segments. The corpus represents planned spoken monologues delivered in a controlled presentation format. While the dataset forms a single specialized genre, Hasebe (2018) established, on an earlier snapshot of the corpus, that its lemma frequencies correlate slightly more strongly with COCA's popular magazine register (Kendall's $\tau = 0.61$) than with its spoken register (0.59)—a pattern consistent with structured, audience-directed spoken delivery.

All text in TCSE is processed through an automated NLP pipeline based on spaCy's large general-purpose English model (en_core_web_lg, version 3.8). The pipeline adds structural annotations directly over the token stream across multiple distinct layers:

1. **Surface Layer**: Token forms and normalized lowercase forms.
2. **Lemma Layer**: Canonical lemmas.
3. **Part-of-Speech Layer**: Coarse Universal POS tags and fine-grained Penn Treebank tags.
4. **Syntactic Chunk Layer**: Non-recursive noun-chunk spans (_), delimiting basic noun-phrase boundaries.
5. **Named-Entity Layer**: 18 standard spaCy entity categories (Table 1), tagged over entity spans.
6. **Dependency Layer**: Syntactic head-dependent relations and grammatical functions (e.g., nsubj, dative, dobj).

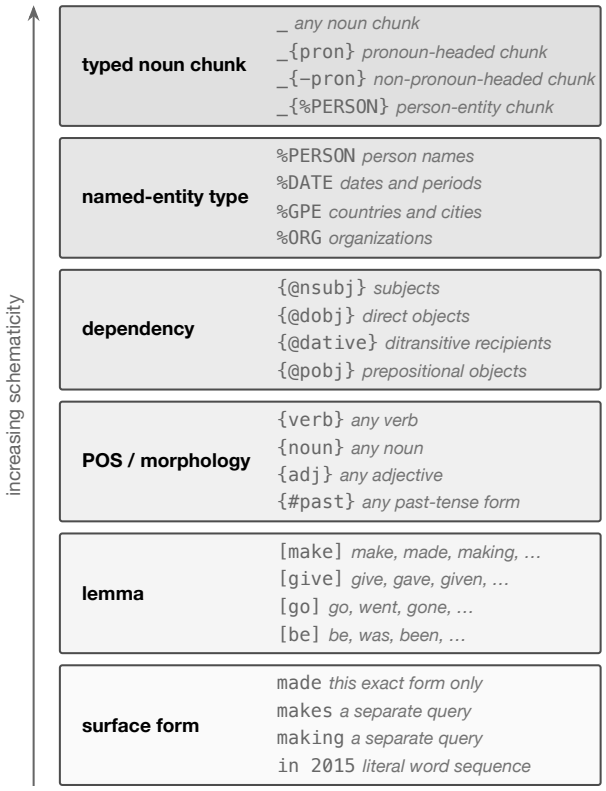

Figure 1: The TCSE annotation layer stack, ordered by increasing schematicity, with query-syntax examples at each level.

TCSE indexes all these annotation layers simultaneously (Figure 1), allowing queries to constrain surface forms, POS tags, chunk boundaries, and entity types within a single expression.

2.2 The Abstraction-Dependence Metric

To measure how heavily a construction relies on type abstraction for its retrieval, this paper introduces a quantitative metric termed *abstraction-dependence*. Let a construction be realized by a set of surface strings, each combining the construction's fixed lexical material with a concrete slot filler. Its abstraction-dependence is defined as the total frequency of the construction divided by the frequency of its single most frequent surface realization:

$$\text{abstraction-dependence} = \frac{f(\text{construction})}{\max f(\text{string})}$$

where f denotes corpus frequency and the maximum is taken over that set of surface strings.

A construction whose frequency is carried almost entirely by a single surface string yields a ratio close to 1.0, indicating that it is easily discoverable through surface queries alone. Conversely, a construction whose instances are widely distributed across diverse surface fillers exhibits a high ratio, rendering it effectively invisible to surface methods. The reciprocal of this ratio carries a direct methodological interpretation: it quantifies the proportion of a construction's instances that the single best surface query can retrieve (defined here as its *coverage*). Abstraction-dependence itself, rather than coverage, serves as the pri-

mary index: it rank-orders constructions on an ascending scale of dependence (Section 3.1), in the way collostructional analysis ranks collexemes by association strength (Stefanowitsch and Gries, 2003), and it spreads out precisely the most dependent frames, which coverage compresses into differences of well under a percentage point.

The selection of the single most frequent surface string requires methodological justification. First, it represents the strongest possible position a surface-only analyst can adopt, establishing a conservative upper bound on what a single surface query can retrieve. Realistic surface workflows, which must select strings in advance, necessarily yield lower coverage. Second, this criterion is parameter-free, avoiding arbitrary thresholds regarding how many strings to include. Third, the low coverage of surface queries persists regardless of list length: as Section 3.1 demonstrates, even combining the ten most frequent surface strings of an open frame accounts for only a small fraction of its total instances (e.g., 10.5% for `in %DATE`).

While collostructional methods quantify the mutual attraction between constructions and their fillers (Stefanowitsch and Gries, 2003), hapax-based productivity measures gauge the growth rate of a filler inventory (Baayen, 1992), and entropy or Zipfian metrics offer granular insights into filler diversity, the present ratio prioritizes operational simplicity and upper-bound estimation.

The metric operates at a specific level within the layer stack. To make every step of the calculation transparent, we consider a deliberately minimal example. Counting occurrences of the surface string *in 2015* yields 159 instances, whereas counting the entity-typed frame `in %DATE` (with the identical fixed lexical material and an open slot) yields 13,116 instances. Both counts describe the same domain of the grammar and are empirically valid; they differ because they are evaluated at distinct abstraction layers. The frame's abstraction-dependence is calculated as 13,116 / 159 = 82.5, indicating that the best surface string retrieves only 1.2% of the construction's instances.

2.3 Counting Protocols and Reproducibility

The empirical analysis in this study relies on two distinct counting protocols, chosen to address different structural levels of constructional schematicity.

The first protocol targets single-slot prepositional frames via exhaustive corpus aggregation. Its purpose is to measure abstraction-dependence and surface query retrieval loss directly within basic, highly frequent constructional environments where a fixed lexical anchor pairs with a single open entity slot. A target instance is identified when a named-entity span (encompassing both single- and multi-word entity expressions) occurs immediately following one of nine targeted prepositions (*in, at, by, from, to, of, for, with, on*). Non-talk sentences, such as production credits and on-screen captions, are excluded from this aggregation, so that the counts reflect talk content proper.[3]

To focus the analysis on well-attested, statistically robust patterns while filtering out low-frequency noise, a frame is included if its total constructional frequency exceeds 800 instances; this criterion identifies 32 distinct frames.[4] The frame set is robust to this cutoff.[5]

Whereas the first protocol relies on exhaustive aggregation over entity spans, the second protocol quantifies multi-slot argument structures and relational patterns by using composed query expressions as direct measuring instruments. Its purpose is to evaluate how multiple abstraction layers (combining syntactic chunking, POS conditions, and entity typing) can be composed to retrieve complex constructions whose open slots are structurally heterogeneous or interact across constituent boundaries. For example, in analyzing the ditransitive construction, the typed noun-chunk operator `_{COND}` —where COND is a placeholder for a condition on the chunk's head—constrains constituent boundaries while specifying the internal head category. The query `[verb] _{pron} _` isolates pronominal recipients, whereas the complement operator `[verb] _{-pron} _` retrieves all remaining recipient structures.

All numerical values reported in this paper were extracted from the frozen data snapshot (retrieved August 10, 2026). While the frame-level aggregation tables (including the boilerplate exclusion described above) were generated via dedicated backend scripts, all query-based counts reported throughout the paper are fully verifiable and reproducible by any user via TCSE's freely accessible advanced search interface using the query specifications provided.

3. Results

3.1 Abstraction-Dependence in Prepositional Frames

Table 2 reports key empirical measurements for all 32 single-slot frames meeting the 800-instance frequency floor, ordered by descending abstraction-dependence; Figure 2 plots the same ranking on a logarithmic scale.

The quantitative distribution reveals the scale of the retrieval loss: across all 32 frames, the median abstraction-dependence is 12.2 (coverage 8.2%). For the most productive open slots, top surface strings capture barely 1% to 4% of total constructional instances; in the most dependent frame, `by %PERSON`, 964 instances are spread across 828 distinct fillers, so that even the best surface string (*by TED*, 6 instances) covers 0.6%. The frame `in %DATE` illustrates how this dependence operates: a surface-only query for its top realization (*in 2015*, 159 instances) retrieves a coverage of 1.2% of the frame's total usage domain (13,116 instances; calculated as 159 / 13,116). The proposed abstraction-dependence metric is the direct reciprocal of this coverage (13,116 / 159 = 82.5), quantifying how the total constructional domain expands to 82.5 times the size of its single most frequent string.

[3] Operationally, a sentence is excluded if its verbatim text recurs in three or more talks, if it matches standard production-credit phrases (e.g., "The show is produced by …", "original music by …"), or if it consists of a bracketed line, optionally followed by a parenthetical (e.g., "(Music)"), such as on-screen labels and captions.

[4] Treating these partially schematic frames as constructions follows Goldberg's (2006: 5) frequency-based criterion, under which patterns are stored as constructions even when fully predictable, provided they occur with sufficient frequency. The 800-instance floor operationalizes this criterion at corpus scale.

[5] Halving the threshold to 400 instances expands the frame set to 44, while the median dependence shifts only slightly from 12.2 to 10.2 and extreme values remain identical.

Construction	Total	Top Surface String	AD	Cov.
by %PERSON	964	by TED (6)	160.7	0.6%
in %DATE	13,116	in 2015 (159)	82.5	1.2%
with %PERSON	995	with TED (21)	47.4	2.1%
to %ORG	1,396	to AI (37)	37.7	2.7%
from %ORG	1,022	from Facebook (28)	36.5	2.7%
from %GPE	1,934	from India (67)	28.9	3.5%
of %PERSON	1,827	of TED (68)	26.9	3.7%
in %ORG	1,174	in Congress (58)	20.2	4.9%
to %GPE	2,877	to the United States (145)	19.8	5.0%
at %ORG	2,540	at MIT (131)	19.4	5.2%
on %DATE	1,391	on the day (72)	19.3	5.2%
of %GPE	4,171	of the United States (224)	18.6	5.4%
of %ORG	3,236	of AI (196)	16.5	6.1%
in %GPE	15,134	in the United States (1,090)	13.9	7.2%
for %DATE	5,661	for years (442)	12.8	7.8%
of %DATE	1,978	of today (158)	12.5	8.0%
to %PERSON	1,245	to TED (104)	12.0	8.4%
with %ORG	990	with AI (117)	8.5	11.8%
of %NORP	1,332	of Americans (159)	8.4	11.9%
by %DATE	1,295	by 2050 (161)	8.0	12.4%
on %ORG	984	on Facebook (140)	7.0	14.2%
of %CARDINAL	3,418	of one (511)	6.7	15.0%
at %DATE	1,164	at the end of the day (188)	6.2	16.2%
by %CARDINAL	941	by one (155)	6.1	16.5%
of %LOC	1,231	of Africa (208)	5.9	16.9%
to %CARDINAL	1,610	to one (275)	5.9	17.1%
with %CARDINAL	1,639	with one (375)	4.4	22.9%
in %LOC	2,530	in Africa (643)	3.9	25.4%
for %CARDINAL	1,013	for one (303)	3.3	29.9%
in %CARDINAL	2,757	in one (903)	3.1	32.8%
from %CARDINAL	945	from one (404)	2.3	42.8%
on %LOC	1,267	on Earth (795)	1.6	62.7%

Table 2: Abstraction-dependence (AD) and coverage (Cov.) of the top surface string across 32 single-slot prepositional frames (top-string frequencies in parentheses). Some top strings reflect systematic tagging idiosyncrasies of the entity layer (e.g., *TED* tagged as %PERSON, *AI* as %ORG); see Sections 3.2 and 4.4.

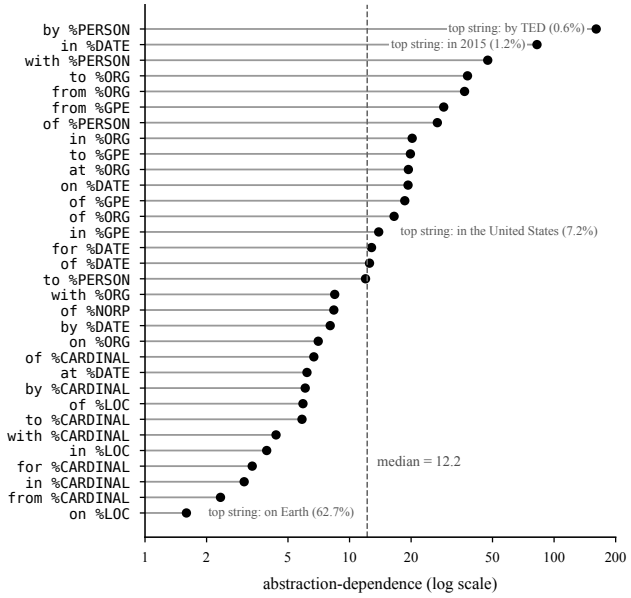

Figure 2: Abstraction-dependence of the 32 prepositional frames (log scale).

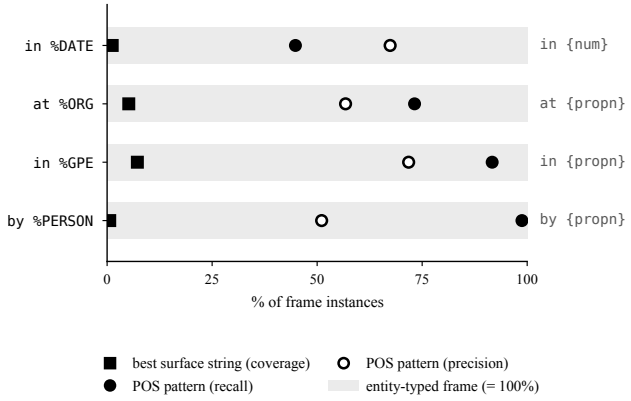

Figure 3: Coverage of the best surface string, and recall and precision of a single-token POS pattern, for four representative frames, relative to the entity-typed frame (= 100%).

Even expanding surface queries to the top-ten most frequent strings captures only a small fraction of open frames (e.g., top-10 strings for in %DATE account for only 10.5% of total instances). Only frames dominated by a single culturally salient proper noun—such as on %LOC dominated by *on Earth* (coverage 62.7%)—exhibit high surface coverage. For the vast majority of schematic frames, surface-frequency retrieval methods fail to capture the construction.

Single-token POS patterns occupy an intermediate level between these two layers. Figure 3 measures this level for four representative frames, chosen to span the two relevant single-token patterns ({num} and {propn}) and four distinct entity types: patterns such as by {propn} recover far more than any surface string (recall between 44.8% and 98.8%), but they both undergenerate and overgenerate—multi-word spans like *in the 1990s* escape a single-token pattern, and only 51.1% to 71.8% of the pattern's matches actually belong to the targeted frame. Entity-typed frames therefore do more than add recall: they delimit the constructional slot that POS patterns can only approximate.

3.2 Entity-Type Profiles Across Prepositions

To examine how entity types associate with prepositional slots, Figure 4 reports the within-preposition proportions of entity spans across all nine target prepositions without applying a frequency floor. Seven dominant categories are broken down individually; the remaining 11 (such as %MONEY, %NORP, and %EVENT) are aggregated under *Other* (see Table 1 for the full inventory).

The raw co-occurrence counts demonstrate that prepositions possess highly distinct, systemic entity-type preferences. The spatial prepositions *in, from, to* attract high proportions of %GPE (29.4% to 40.3%). In contrast, *with* and *by* show strong preferences for %PERSON (20.9% and 19.6%, respectively), reflecting comitative/instrumental and agentive roles. Temporal reference likewise patterns by preposition: *for* and *in* favor %DATE (54.9% and 34.9%), whereas %TIME, a much rarer category overall, reaches its highest shares under *at* and *for* (9.0% and 9.3%). *At* is instead dominated by %ORG (34.0%, as in *at MIT*), and *on* stands out for %LOC (22.6%, as in *on Earth*). Read constructionally, these profiles amount to a quantitative semantic signature of each frame: what a qualitative description would call the semantic restrictions of a slot appears here as a measurable distribution of entity types, obtained directly from raw co-occurrence counts.

Furthermore, automatic annotation yields systematic classification patterns (for instance, the conference name *TED* is tagged as %PERSON, appearing as the top filler of of %PERSON with 68 instances). Rather than representing error-free natural language, these reported metrics faithfully capture the output of the annotated corpus layer. However, the tagging

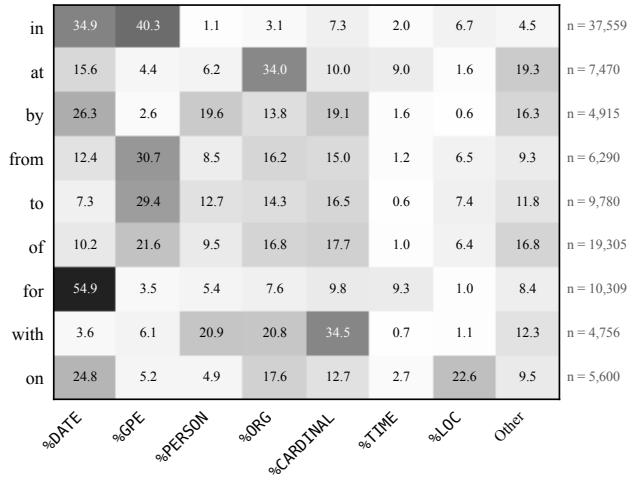

Figure 4: Within-preposition entity-type profiles across the nine target prepositions (row shares in %; row totals at right).

Recipient Head	Composed Query	Dependency Layer	Entity Typing
Total instances	12,301	14,152	—	
Pronominal (`_{pron}`)	10,521 (85.5%)	12,691 (89.7%)	—	
Proper name (`_{propn}`)	147 (1.2%)	155 (1.1%)		101 (0.8%)
Common noun (`_{noun}`)	1,630 (13.3%)	1,306 (9.2%)	—	

Table 3: Distribution of recipient slot types across three analytical instruments: the composed query `[verb] _ _`, the dependency layer (`dative` relation), and the named-entity intersection `_{%PERSON|%GPE|%ORG}`.

idiosyncrasies identified here are far too small to reverse the broad contrasts observed across prepositions (such as the 20% to 21% person preference for *with* and *by* versus the 40.3% GPE preference for *in*).

3.3 Composed Abstraction in Ditransitive Constructions

Whereas the single-slot analysis demonstrates the high retrieval loss of surface queries within a single open entity slot, multi-slot argument-structure constructions introduce constituent-boundary ambiguity and internal structural heterogeneity. To evaluate how multiple abstraction layers compose to retrieve complex patterns, Table 3 examines the ditransitive construction across five prototypical ditransitive verbs (*give, send, tell, show, offer*).

The typed noun-chunk slot `_{COND}` provides a straightforward mechanism for partitioning recipient types within a single query: head-category conditions divide the 12,301 ditransitive instances into pronominal (`_{pron}`, 85.5%), proper-name (`_{propn}`, 1.2%), and common-noun (`_{noun}`, 13.3%) recipients, as shown in the Composed Query column of Table 3.

Crucially, three distinct analytical instruments within TCSE corroborate the small size of the proper-name recipient category—providing clear quantitative support for prior corpus findings on the dative alternation in spoken English (such as Bresnan et al., 2007; Szmrecsanyi et al., 2017; Engel et al., 2022), which establish that pronominal forms heavily dominate the recipient position:

- **Query-based complement count**: The complement operator `_{-pron}` retrieves 1,780 total non-pronominal recipients: 1,630 common-noun and 147 proper-name instances (1.2%), with the remaining 3 headed by numerals or unclassified tags.
- **Dependency-layer extraction**: Syntactic dependency parsing (`dative` relation) yields 14,152 ditransitive instances with pronominal or nominal recipient heads (excluding prepositional *to*/*for* variants and minor unclassifiable heads), of which 155 are tagged as proper-noun recipients (1.1%).
- **Named-entity layer tagging**: Intersecting the recipient slot with entity typing (`_{%PERSON|%GPE|%ORG}`) yields 101 tagged instances (0.8%).

All three measurement approaches consistently demonstrate that proper names constitute a small minority (around 1%) of ditransitive recipients in spoken monologues, with pronouns dominating the slot.

4. Discussion and Implications

4.1 Theoretical and Methodological Implications

These empirical findings convert the general problem of retrieval loss into a precise, measurable parameter across construction types. Abstraction-dependence assigns a numerical index to each construction, and its reciprocal carries a direct methodological interpretation: top surface queries retrieve a coverage of barely 0.6% for the most open frame evaluated and a median coverage of 8.2% across single-slot frames (Section 3.1). Meanwhile, complex argument structures introduce heterogeneous slots, such as a recipient slot that is 85.5% pronominal yet structurally diverse (Section 3.3). Thus, the methodological limitation of surface retrieval is expressed as an empirically measurable metric that can be computed, compared, and tracked systematically. Slot productivity thereby gains an additional, straightforward corpus-based diagnostic.

These results also cast named-entity recognition in a practical, complementary role. Entity typing is conventionally treated as an end in itself—a computational task designed to extract individuals, locations, and dates from unstructured text. In the present framework, it functions instead as an essential instrument for construction discovery: the entity layer serves to make the open constructional slots accommodating those entities systematically countable at scale. This pipeline also exposes its own classification errors (Section 3.2), but these errors are systematic rather than random: the same misclassification recurs deterministically wherever the same configuration appears. This stability reinforces the present perspective: what is essential for corpus-based construction grammar research is not error-free extraction as such, but rather a consistent, queryable level of structural abstraction.

This architecture thereby demonstrates that grammatical categories and named-entity types function not as competing alternatives, but as orthogonal abstraction layers whose structured composition enables systematic constructional retrieval. Grammatical categories specify syntactic slot boundaries and part-of-speech constraints, while entity types specify semantic category constraints, within the accuracy limits of automatic annotation (Section 4.4). Combining these orthogonal layers inside a single multi-layer query allows researchers to move between string-level concreteness and schematic abstractness.

4.2 Extending Composed Abstraction across Construction Types

The utility of composed multi-layer abstraction is not restricted to single-slot prepositions or the ditransitive construction. The same abstraction layers underpin TCSE's curated inventory of 1,167 constructional patterns, of which 131 rely on noun-chunk slots and 39 on entity types. When evaluated against traditional query designs that constrain slots to single tokens, the necessity of multi-word chunk abstraction becomes clear: such traditional single-word approximations can substantially undercount: ditransitive frequency drops to 55% (6,781 of 12,301 instances) when slots are restricted to single-word noun phrases.

Composed queries similarly capture other canonical argument-structure constructions in the constructionist literature:

- **Caused-motion construction**: The query `[put] _ {adp} _` retrieves 3,924 instances (e.g., *put embedded nanoparticles into that material*), and its theme slot exhibits a profile distinct from the ditransitive recipient: 37% pronominal and 61% common-noun fillers.
- **Resultative construction**: The query `[make] _ {adj}` retrieves 3,445 instances (e.g., *make current data centers flexible*), occupying an intermediate position at 63% pronominal.
- **Cognate object construction**: A lemma condition on the chunk head delimits the entire object noun phrase, verb by verb (`[live]{verb} _{[life]}`: 415; `[sing]{verb} _{[song]}`: 85; `[dream]{verb} _{[dream]}`: 10), capturing multi-word objects such as *live a good life* in a single slot.

Furthermore, the scope of typed multi-layer abstraction extends beyond fixed slot conditions to relational constraints across constituent slots. The lemma condition just applied to cognate objects has a variable counterpart: the lemma-echo mechanism requires a chunk head to replicate a lemma bound elsewhere in the query, so the fully schematic `{verb:1} _{=1}` retrieves the identical-lemma core of the cognate-object family in one query (132 instances: *walk the walk*, *dream the impossible dream*), with no lexical material specified at all.

A distinct family of repetitive and cumulative patterns—including nominal binomials (*day by day*, *side by side*) and repetitive adverbial constructions (*on and on*, *round and round*)—is defined by a structural identity relation across slots. Such cross-slot identity cannot be formulated by single-slot queries, and surface search can only approximate it by manually enumerating candidate word pairs. The same mechanism formulates this directly: `{noun:1} by _{=1}` binds the lemma of the initial noun and requires the subsequent chunk to replicate it, retrieving 450 instances (*year by year*, *side by side*), while `{noun:1} after _{=1}` retrieves 416 instances (*day after day*). A concrete application of such repetitive patterns is demonstrated by Kuryu (2025), who investigated the multimodal collostructional properties of repetitive adverbial constructions ([ADV and ADV]) and co-speech gestures using enumerated surface queries prior to the availability of automatic echo operators. The present lemma-echo mechanism supports such inquiries directly, making cross-slot repetitive patterns countable within a single query.

4.3 Infrastructure, Pedagogy, and AI Diagnostic Benchmarking

For corpus linguistics infrastructure, the implications are direct. Querying complex constructional slots has so far required giving up corpus size, relying either on fully parsed treebanks such as ICE-GB (approx. 1 million words) or on Universal Dependencies treebanks (approx. 250,000 tokens per corpus). The measurements presented here come from a corpus more than ten times larger than ICE-GB and roughly fifty times larger than standard UD treebanks, and they were computed on a publicly available query engine where abstract slots themselves serve as primary query terms. The limiting factor is thus no longer automatic annotation, which modern NLP pipelines provide, but multi-layer indexing: corpus systems that index entity and chunk layers alongside surface forms make this kind of constructional inquiry a routine matter.

Practical applications extend to descriptive grammar and applied linguistics. The slot profiles of Section 3.2 and the recipient partition of Section 3.3 provide empirical evidence that previously required labor-intensive manual coding on small specialized datasets. By making abstract constructional slots queryable at corpus scale within authentic discourse contexts, the present framework enables researchers and learners to systematically observe how schematic templates interact with lexical and semantic constraints in natural spoken language.

Furthermore, the tension between surface patterns and abstract representations has recently emerged in artificial intelligence research. Mackintosh et al. (2025) demonstrate that large language models (LLMs) rely heavily on surface-level statistical cues, exhibiting brittle generalization over abstract constructional structures. The quantitative framework introduced here renders this divide explicit and measurable for corpus analysis. Beyond corpus queries, the composed abstraction framework offers a basis for constructing diagnostic benchmarks to evaluate whether neural language models capture abstract constructional representations or merely exploit surface-level statistical associations.

4.4 Methodological Limitations

Four methodological limitations qualify these findings. First, all structural annotations are generated automatically, allowing parser errors to propagate across abstraction layers (such as when the proper name of the conference *TED* is misclassified as the top filler of `of %PERSON` in Section 3.2). Additionally, automatic noun-chunking introduces boundary-delimitation noise (such as complex post-modified noun phrases being truncated), which slightly affects the exact span boundaries of extracted fillers. The reported metrics reflect the properties of the annotated corpus layer rather than pure natural language distributions. Without manual validation, the total scale of annotation error cannot be bounded, and any individual count may include unidentified misclassifications; counts resting on small frequencies, such as the top-filler frequencies underlying abstraction-dependence, are the most sensitive. The conclusions in Section 3, however, concern aggregate patterns over hundreds or thousands of instances and are designed not to hinge on particular classification decisions.

Second, the dataset comprises a single specialized genre (planned spoken monologues delivered within TED's struc-

tured presentation format), meaning slot productivity profiles may vary in casual conversation or written registers. In written registers where noun phrases exhibit higher structural complexity and modifier density, the reliance on surface queries is expected to result in even more severe coverage loss than that observed in TED monologues. This genre hybridity is empirically measurable rather than impressionistic: as detailed in Section 2.1, the corpus's lemma frequencies correlate slightly more strongly with COCA's popular magazine register than with its spoken register (Hasebe, 2018).[6]

Third, TED speakers represent a linguistically heterogeneous population, including numerous non-native English speakers; the corpus thus reflects a blend of native and non-native usage, which researchers analyzing native-speaker grammars should take into account.

Fourth, all annotation layers depend on a specific spaCy model and parse version; while the versioning and traceability protocols detailed in Sections 2.1 and 2.3 mitigate this dependency, they do not eliminate reliance on a single automated pipeline.

5. Conclusion

This study investigated the broader relationship between type abstraction and construction discovery in corpus-based construction grammar. Specifically, it evaluated the extent to which open constructional slots—analyzed across 32 single-slot prepositional frames and the ditransitive construction—depend on multi-layer abstraction, how that dependence can be quantified, and how entity types associate with those slots.

Within this empirical scope, the findings provide clear answers. First, surface query coverage (ranging from 63% down to below 1% across single-slot frames) demonstrates the high retrieval loss of surface methods, and its reciprocal, abstraction-dependence, provides a direct metric quantifying how heavily constructions rely on type abstraction. Second, constructional slots exhibit distinct, systemic entity-type preferences that appear clearly in raw co-occurrence counts and possess a composite character: each frame pairs its fixed lexical anchor with a characteristic distribution of entity types in the open slot. Third, where slots are structurally heterogeneous, multiple abstraction layers can be composed effectively: a single advanced query retrieves the ditransitive construction, automatically partitioning its recipient slot into pronouns, proper names, and full noun phrases (with three distinct analytical instruments confirming the proper-name minority), while a lemma-echo mechanism makes cross-slot relational patterns visible.

This methodology places minimal demands on data beyond what modern automatic annotation pipelines already provide (namely, indexed entity types and noun chunks), allowing high transferability to diverse genres and linguistic corpora. What surface-frequency methods scatter, typed multi-layer abstraction unifies, rendering the structural layers of grammar between surface strings and abstract constructions systematically countable within a single automatically annotated corpus.

[6] All five COCA registers: magazines .61, spoken .59, academic .57, newspapers .57, fiction .51 (Kendall's τ, 28,180 shared lemmas, $p < .001$; Hasebe, 2018).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP26K15571 and JP26K00003.

References

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Geert Booij and Jaap van Marle (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Kluwer.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer and Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: KNAW.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Engel, Alexandra, Jason Grafmiller, Laura Rosseel and Benedikt Szmrecsanyi. 2022. Assessing the complexity of lectal competence: The register-specificity of the dative alternation after *give*. *Cognitive Linguistics* 33(4). 727–766.

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.

Goldberg, Adele E. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics* 22(1). 131–153.

Goldberg, Adele E. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton: Princeton University Press.

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

Hasebe, Yoichiro. 2015. Design and implementation of an online corpus of presentation transcripts of TED Talks. *Procedia - Social and Behavioral Sciences* 198. 174–182.

Hasebe, Yoichiro. 2018. TED Corpus Search Engine: TED Talks o kenkyū to kyōiku ni katsuyō suru tame no purattofōmu [TED Corpus Search Engine: A platform for using TED Talks in research and education]. *English Corpus Studies* 25. 159–172. (In Japanese.)

Kay, Paul and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language* 75(1). 1–33.

Kuryu, Daiya. 2025. Crossmodal collostructional analysis of English [ADV and ADV] constructions: Multimodal constructions or crossmodal collostructions? *Language and Cognition* 17. e39, 1–28.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar, vol. 1: Theoretical prerequisites*. Stanford: Stanford University Press.

Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Mackintosh, Tom, Harish Tayyar Madabushi and Claire Bonial. 2025. Evaluating CxG generalisation in LLMs via construction-based NLI fine tuning. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, 180–189. Düsseldorf: Association for Computational Linguistics.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford: Oxford University Press.

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

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte and Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: A Journal of General Linguistics* 2(1). 86, 1–27.

Language Resource References

The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

ICE-GB: The British Component of the International Corpus of English. https://www.ucl.ac.uk/arts-humanities/research-projects/1998/sep/ice-gb

spaCy: Industrial-strength natural language processing in Python. https://spacy.io/

TED Corpus Search Engine (TCSE). https://yohasebe.com/tcse/

Universal Dependencies. https://universaldependencies.org/