Article

Vladimir Buskin*

Modelling object omission availability in English transitive verbs with corpus and psycholinguistic data

https://doi.org/10.1515/cllt-2025-0004
Received January 10, 2025; accepted June 20, 2026; published online July 6, 2026

Abstract: This data-driven study investigates the licensing factors of object omission in English transitive verbs (e.g., *I'm eating* Ø vs. **I'm devouring* Ø). Using corpus data from the British component of the International Corpus of English and psycholinguistic measurements from the South Carolina Psycholinguistic Metabase (SCOPE), generalised additive models were trained on distributional and semantic features of 972 verbs to examine linear and non-linear effects on omission availability. In a complementary analysis, linear discriminant analysis was applied to `fastText` word embeddings to assess the predictive value of distributional co-occurrence information. Discriminant scores emerged as the strongest predictor, followed by frequency and dispersion, while purely semantic predictors contributed comparatively little once distributional measures were controlled for. Object-omitting verbs cluster towards the stative and communicative end of the verb space. The results challenge earlier monofactorial accounts and support usage-based explanations of argument realisation, though the overlap between distributional and semantic measures suggests that future research should address potential confounds and consider semantic frames as a key missing predictor.

Keywords: object omission; null instantiation; syntax; corpus linguistics; frame semantics; machine learning

1 Introduction

In English, a large but not clearly delineated class of transitive verbs licenses object omission under conditions that remain poorly understood. This alternation and its

***Corresponding author: Vladimir Buskin**, Department of English Language and Linguistics, Catholic University of Eichstätt-Ingolstadt, Universitätsallee 1, 85072 Eichstätt, Germany, E-mail: VBuskin@ku.de. https://orcid.org/0009-0005-5824-1012

range of subtypes are subsumed in the literature under various umbrella terms, including 'understood object alternations' (Levin 1993), 'objectless transitives' (Lemmens 2006) and, in more recent cognitive approaches, 'object deprofiling' (Goldberg 2001, 2005), 'implicit object constructions' (Gavruseva 2024), and 'null instantiation' (e.g., Buskin 2025; Chaves et al. 2025; Fillmore and Kay 1995: §7; Lambrecht and Lemoine 2005; Ruppenhofer and Michaelis 2010). Despite decades of research, linguists have yet to reach a consensus on why some verbs participate in this alternation, whereas others do not (cf. 1a vs. 1b; see Rice 1988). Crucially, this question of alternation availability must be distinguished from the more widely studied question of what governs the choice between two syntactic alternants, as in examples (2a) and (2b).

(1) a. *I haven't **eaten** Ø yet.*
b. **I haven't **devoured** Ø yet.*

(2) a. *She **lost** the game.*
b. *She **lost** Ø.*

In frame-semantic terms, *eat* and *devour* both evoke an Ingestor frame element within the Ingestion event.[1] Yet only *eat* permits null instantiation of this core frame element. It is puzzling why verbs with the same schematic representation of events should differ so markedly in their argument instantiation behaviour. This apparently contradicts Ruppenhofer (2004: 467), who predicts that frames group verbs with similar omission patterns. Even though there is ample evidence in favour of this generalisation, frame membership alone is insufficient to capture the full complexity of the phenomenon. Additional factors that have been proposed to contribute to omission behaviour include the verb lemma, its frequency of occurrence, its selectional preferences, degree of polysemy, and specificity (Fillmore and Kay 1995: §7; Goldberg 2001: 507; Rice 1988; Resnik 1996; Ruppenhofer 2004: 457–464). In short, the selective availability of object omission likely reflects an interaction of distributional and semantic properties at the level of individual verbs.

This article investigates the extent to which these distributional and semantic predictors can account for variation in object omissibility. To this end, this study draws on English transitive verbs sampled from the British component of the International Corpus of English (ICE; Greenbaum 1996). Features that could not be derived from corpus data were approximated using measures from the South Carolina Psycholinguistic Metabase (SCOPE; Gao et al. 2022). Sets of alternating and non-alternating verbs were established through a combination of manual and automatic annotation, after which non-linear regression models were trained to evaluate the

1 Cf. https://framenet.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Ingestion [Last accessed: September 23, 2025].

explanatory contribution of frequency, dispersion, polysemy, concreteness, specificity, taxonomic neighbourhood density, and selectional preference strength. Additionally, given recent advances in distributional semantics (e.g., Perek 2016), a linear discriminant analysis was conducted on `fastText` word vectors to assess the predictive value of co-occurrence information.

Structurally, Section 2 outlines the properties and current explanations of object null instantiation relevant to its identification and classification in natural language data, establishing the conceptual foundation for the empirical analysis. Section 3 formalises these notions, followed by a description of the corpus and psycholinguistic data, the sampling procedure, and the statistical methods. Section 4 presents the results. Section 5 recapitulates the core findings and concludes with a discussion of methodological limitations, theoretical implications, and directions for further research.

2 Argument omission

Object omission, and argument realisation more broadly, are examined through the lens of Cognitive Linguistics and specifically Frame Semantics (Boas 2020; Boas et al. 2024; Fillmore 1986). In line with usage-based work in this field (Buskin 2025; Diessel 2019; Hoffmann 2022), generalisations in language, including those concerning argument structure, are assumed to emerge gradually from exposure to rich linguistic exemplars (Bybee 2006; Perek 2015); in other words, the factors licensing argument omission are expected to arise from speakers' experience with language, as opposed to configurations of an innate Universal Grammar (Chomsky 1995; Radford 2005). The discussion of the phenomenon begins with a review of the predominantly qualitative and monofactorial research.

2.1 Survey of subtypes

The omission of obligatory verbal complements that are integral to the interpretation of the semantic frame evoked by the verb, i.e., core frame elements, will henceforth be referred to as Null Instantiation (NI), following the working definition proposed by Fillmore and Kay (1995: §7–3). Null-instantiated complements are always linked to interpretive strategies. The most frequently discussed construals include Indefinite Null Instantiation (INI henceforth; cf. 3a) and Definite Null Instantiation (DNI henceforth; cf. 3b). Fillmore and Kay (1995: §7) propose a third one,

Free Null Instantiation (FNI), which may be regarded as a residual category subsuming "null elements that could be either definite or indefinite depending on the context" (Lyngfelt 2012: 2; cf. 3c).

(3) a. *I haven't **eaten** $\emptyset_{\text{Ingestibles}}$ yet.*
b. *I **won** $\emptyset_{\text{Game}}$.*
c. *The packages were **delivered** on time $\emptyset_{\text{Deliverer}}$.*

INI construals correspond to indefinite readings of the unrealised core frame element. Accordingly, it is neither necessary nor relevant whether the referent is salient or recoverable from the pre-text or situational context, respectively (Fillmore 1986: 96; Fillmore and Kay 1995: §7–4). These readings are alternatively referred to as Indefinite Deletion (Allerton 1975: 214) or the Unspecified Object Alternation (Levin 1993: 33) in other works. The primary function of this backgrounding strategy appears to be the shifting of the current center of attention onto the verbal action, thereby presenting it as an atelic activity rather than as an accomplishment or another of Vendler's verb classes (Lambrecht and Lemoine 2005: 20; Rappaport Hovav 2008: 13; Rappaport Hovav and Levin 1998: 98). Mainstream accounts of INI treat it as an arbitrary lexical affordance, yet Glass (2021) provides new evidence that highlights the influence of extra-linguistic factors on the availability of this reading.

DNI construals rest on the assumption that the core frame element is identifiable in the widest sense. Contrary to INI, there are several additional layers to what can 'trigger' DNI. In addition to being idiosyncratic to certain verb lemmas, grammatical constructions can also impose DNI readings of certain argument slots. Notably, seemingly subjectless imperative clauses invariably receive a default reading of the 2.p.sg. as the subject, overriding any potential lexical restrictions. Within the realm of object omission, instructional imperatives (e.g., *cool Ø briefly, then eat Ø warm*) and labelese (on a bottle: *Ø Contains alcohol*) are particularly noteworthy because they illustrate the influence of genre on argument realisation (Ruppenhofer and Michaelis 2010). Furthermore, DNI readings vary with respect to the accessibility of the referent. If the unmentioned referent is both highly accessible and predictable in the discourse context, while also being the 'target' of the current proposition, then the null-instantiated argument is "anaphoric and can in principle also appear in the form of unaccented pronouns, indicating ratified-topic status of the denoted referents" (Lambrecht and Lemoine 2005: 31).

Finally, passive clauses exemplify a constructionally-licensed subtype of FNI, often involving the backgrounding of the *by*-agent. Since the agent, which is more often than not a core frame element, shows considerable variation in information

status, Lyngfelt (2012: 4) proposes treating its construal as 'unspecified NI'. In fact, unspecified NI is only attested in passive clauses (Lyngfelt 2012: 4), thus indicating a clear constructional bias towards agent omission.[2]

2.2 Potential predictors of omission availability

In one of the earliest theoretical accounts of this phenomenon, Rice (1988) observes that "[verbs] that conflate action and manner tend to resist omission while synonymous yet more neutral verbs tend to allow it" (1988: 204). In other words, the specificity of a verb's meaning appears to correlate with omission likelihood. Therefore, as events become more specific, the likelihood of object omission should decrease (1988: 204). Under this line of reasoning, seemingly 'neutral' verbs such as *eat*, *write* and *study* should be more prone to object deprofiling than their hyponyms *devour*, *inscribe* or *review*.

However, not only verb specificity appears to play a role but also that of the object argument. Generally speaking, Rice suggests that verb and object specificity display a fairly linear relationship, yet their effect on the probability of object omission appears to be very much non-linear: "It is the combination of a semantically basic verb and a relatively contingent object […] that tends towards omission" (1988: 207). That is, if lexical specificity is treated as a continuous cline, only certain ranges of verb or object specificity are likely to influence omission probability.

Rice (1988: 207) further argues that lexical specificity interacts with a verb's range of possible complements. Even if a verb is highly general and might, for this reason, be expected to license null objects more readily, low variability in the semantic classes of objects selected by the verb may in fact counteract this tendency. In sum, Rice points out a number of closely related semantic variables that could produce non-linear effects. While intuitively plausible, her claims remain challenging to verify empirically, as the notion of 'specificity' is difficult to operationalise precisely; nevertheless, Rice's account remains a valuable and informative starting point.

Resnik (1993, 1996) has developed an information-theoretic approach that captures "the effect that the predicate has on what appears in an argument position" (Resnik 1996: 134). It describes the probabilistic relationship between a verb *v* and the WordNet noun class *c* of the co-occurring object noun *n*. More concretely, a verb such as *eat* may preferably co-occur with nouns relating to the semantic domain of Food rather than Furniture or People (Manning and Schütze 1999: 289–290). If the usage patterns of a verb are known with sufficient precision, it is possible to compute "the

2 For the methodological implications, see Section 3.2.

amount of information it carries about its argument" (Resnik 1996: 136) by means of Kullback-Leibler divergence D_{KL}.

D_{KL}, also known as relative entropy, is bounded in [0, ∞) and quantifies the divergence between two probability distributions $p(x)$ and $q(x)$ for a discrete random variable X (cf. Equation (1)). Note that $D_{KL} = \infty$ if $q(x) = 0$ for any x with $p(x) > 0$.

$$D_{KL}(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

Applying this notion to argument realisation, Resnik formalises selectional preference strength $S(v_i)$ using the Kullback-Leibler divergence between two probability distributions $P(c)$ and $P(c|v_i)$. $P(c)$ denotes the prior noun class distribution regardless of the predicate under consideration, and $P(c|v_i)$ is the posterior (or true) distribution of a noun class given a specific predicate (Resnik 1996: 135–136; see Equation (2)). He reports a significant association between object omissibility and selectional preference, concluding that "verbs participating in implicit object alternations select more strongly for the direct objects than verbs that do not" (Resnik 1996: 148–149); however, this test is only based on 34 verbs, which severely limits both the statistical power of the analysis and the generalisability of the finding.[3] That said, working with a small, carefully curated sample does allow for exhaustive qualitative control that is difficult to maintain at scale. Important outliers include verbs such as *wear*, which does not seem to license object drop despite strongly selecting for its object (Resnik 1996: 150).

$$S(v_i) = D_{KL}(P(c \mid v_i) \parallel P(c)) = \sum_{c} P(c \mid v_i) \ \log \frac{P(c \mid v_i)}{P(c)} \tag{2}$$

Ruppenhofer (2004) systematically evaluates a range of explanatory strategies, and while he finds limited evidence for most of them, his analysis usefully foregrounds the role of semantic frames. Regarding verb frequency, he observes "no interactions at all between the frequency of a lemma and the ability to license omission nor between lemma frequency and the type of omission" (2004: 441). It is indeed true that in his small sample of 34 verbs there is little evidence for a linear relationship between these two variables. However, this does not imply that frequency information is altogether irrelevant: it may only contribute a small part of the puzzle rather than the full picture in a way that is perhaps not immediately apparent owing, for instance, to non-linear rather than linear effects, or even interactions with other variables. Moreover, the frequency of a lemma should never be interpreted in

3 Resnik's sample includes the following verbs (Resnik 1996: 138): *pour, drink, pack, sing, steal, eat, hang, wear, open, push, say, pull, like, write, play, hit, catch, explain, read, watch, do, hear, call, want, show, bring, put, see, find, take, get, give, make, have.*

isolation from its dispersion across a corpus (Gries 2020); otherwise, the frequency data offer only a vague representation of a lemma's usage profile (2020: 101).

Ruppenhofer (2004: 441–444) arrives at similar conclusions about the polysemy of the lemma and the definiteness of the object argument; similarly questionable is the effect of object animacy. While he assigns greater explanatory value to semantic taxonomies and Resnik's selectional preference strength outlined above, he acknowledges both the limitations of these approaches and the absence of consistent patterns (2004: 447–452). In this respect, the study by Olsen and Resnik (1997), which builds on Resnik (1996), is particularly insightful as it integrates a gradual notion of transitivity into its account of object omission (Hopper and Thompson 1980). Despite Ruppenhofer's justified criticism of the missing outlier analysis, the authors have still identified an important probabilistic tendency, namely that "implicit object constructions are less likely than their counterparts with overt objects to have telic interpretations, or semantically individuated objects" (Olsen and Resnik 1997: 9).

Finally, Ruppenhofer (2004: 457–471) proposes semantic frames[4] as a more informative analytical category because they are intended to group verbs with similar semantic properties, aspectual structures, selectional preferences, and the overall availability of omission in addition to syntactic features of the null-instantiated frame elements. Empirical validation of this claim would necessitate large-scale frame semantic annotation, with several studies offering automatised solutions to this problem (Chanin 2023; Kabbach et al. 2018; Markard and Klie 2020; Ringgaard et al. 2017; Swayamdipta et al. 2017); however, their ability to extract arguments still remains relatively modest, only marginally exceeding the random baseline.

In one of the most recent corpus-based studies of object omission, Glass (2021) argues that the notion of social routines contributes meaningfully to the explanation of omission patterns. In a series of experiments, participants were asked to rate the acceptability of sentences where object omission occurred in low-routine or high-routine contexts. The ratings were moreover conditioned on selectional preference strength and frequency, both coded dichotomously ('low' vs. 'high'). Glass finds that verbs denoting activities highly entrenched in a speech community (e.g., *eat* or *lift*) are significantly more likely to display some form of NI reading than less routinised actions (e.g., *devour*) (2021: 63–67). Routinisation is explained as the cause for a verb's

4 His working definition of a semantic frame is: "semantic frames are schematic representations of situations involving various participants, props, and other conceptual roles, each of which is a frame element. The semantic arguments of predicates – which can belong to any part of speech – correspond to the frame element of the frame (or frames) associated with those words" (Ruppenhofer 2004: 464–465).

increased frequency and selectional preferences in a given speech community (2021: 60–61; 69).

While the author appears to identify a genuine difference between alternating and non-alternating verbs, the experiments only take into account 16 verbs in total. This raises concerns regarding the power of the statistical analyses, especially considering the absence of effect size estimates. What is more, the fact that the subjects had been recruited via Amazon's Mechanical Turk (Glass 2021: 64), which is notorious for supplying low-quality research data (cf. Douglas et al. 2023: 14; Peer et al. 2022: 1657), may constitute a non-negligible source of bias. Finally, the experimental design does not permit causal inference, as this would require rigorous control for confounders, colliders, and mediators in addition to establishing temporal precedence (Heumann et al. 2022: 376–382). These observations are not intended as a dismissal of Glass's important contribution, which remains the most direct empirical predecessor of the present study. Rather, they point towards methodological refinements that the present study seeks to address and that future work should continue to develop.

3 Methodology

Before investigating the mechanisms of null object availability, several iterations of data sampling and preprocessing were required. First and foremost, this involved defining alternating and non-alternating transitive verbs and subsequently obtaining them from a representative corpus of English. The British component of the International Corpus of English (ICE-GB; Greenbaum and Nelson 1996) was deemed suitable for this task. It provides a comprehensive profile of English usage in both speech and writing across 32 text categories, which can be grouped into 12 overarching registers that vary in formality, interactivity and target audience. Comprising 500 texts of 2,000 words each, this 1-million-word corpus allows for a complete qualitative inspection of all available alternating verb patterns, without the need for further downsampling. Currently, there are 15 available ICE-components[5] that adhere to the same notational guidelines, providing an excellent resource for future comparisons with other L1 and L2 varieties of English.

3.1 Formalisation of object omission

In order to facilitate data collection and statistical modelling, all relevant linguistic entities will be defined in set-theoretic terms. The bottom line forms the set V of

5 Cf. https://www.ice-corpora.uzh.ch/en.html [Last accessed: September 13, 2024].

distinct English verbs $\{v_1, v_2, \ldots, v_n\}$ which evoke semantic frames, and more importantly, frame elements $E = \{e_1, e_2, \ldots, e_m\}$, where $n, m \in \mathbb{N}$. Consequently, each verb v is associated with a subset of frame elements E_v. Those elements which are necessary to the central meaning of the frame (Fillmore 2007: 133) are the core elements $C_v \subseteq E_v$. The semantic features as well as the number of core elements evoked are relevant for the distinction between transitive verbs $V_{\text{trans}} \subseteq V$ and intransitive verbs $V_{\neg\text{trans}} \subseteq V$; for simplicity, these sets are assumed to be disjoint. In this approach, transitivity is defined in terms of semantic frame structures. Being rooted in theory, the applicability of these definitions can only be assessed via qualitative inspection and reference to the FrameNet[6] database.[7]

Definition 3.1. (Transitive verb). A verb $v \in V$ is considered transitive if it evokes at least two core frame elements c and c' with $c, c' \in C_v$, where one of them can function as the subject and at least one other can function as the object of the clause.

Definition 3.2. (Intransitive verb). A verb $v \in V$ is considered intransitive if it does not evoke any core elements that can fulfil the function of the object.

Object omission availability appears to be lexically restricted, thus necessitating a further differentiation between alternating transitive verbs V_{alt} (e.g., *eat*) and non-alternating transitive verbs $V_{\neg\text{alt}}$ (e.g., *devour*; cf. Resnik 1996: 148–149). This differentiation will be treated as a statistical one. Define $I(\alpha)$ as an indicator function that returns 1 if a proposition α is true, and 0 otherwise. Given observations $i \in \{1, \ldots, N\}$ of a single transitive verb in a sample of size N, let the indicator function

$$I(\text{object}_i) = \begin{cases} 1 & \text{if a core element is a direct object in observation } i. \\ 0 & \text{if no core element is a direct object in observation } i. \end{cases}$$

The objects of non-alternating verbs should be either realised without exception or not at all; that is, the sum of all object attestations $\sum_{i=1}^{N} I(\text{object}_i)$ must be 0 or equivalent to the size N of the verb sample: if a non-alternating transitive verb occurs 100 times, one would expect either 100 direct objects or none at all. In other words, the probability that an object occurs should be 1 or 0. I will denote the corresponding probability of overt object realisation in the population as π_{obj}.

By the same token, one can define:

6 Cf. https://framenet.icsi.berkeley.edu [Last accessed: August 25, 2025].

7 In the same vein, I acknowledge that object annotation necessarily involves interpretive judgements guided by theory and corpus evidence, as is standard in linguistic annotation tasks.

Definition 3.3. (Non-alternating verb). A verb is non-alternating if π_{obj} is equal to 0 or 1.

Definition 3.4. (Alternating verb). Verbs are alternating if the proportion of realised objects falls in the range $0 < \pi_{\text{obj}} < 1$.

In the corpus sample, these population proportions will be approximated using the maximum likelihood estimator

$$\widehat{\pi}_{\text{obj}} = \frac{\sum_{i=1}^{N} I(\text{object}_i)}{N}. \tag{3}$$

The problem with Definition 3.4 is that in a sample of 1,000 observations a verb with only one omitted object (or 999 realised ones, respectively) would be classified as alternating. Rather than being instances of linguistically meaningful patterns, these minor fluctuations may have arisen due to random noise, such as sporadic transcription errors in the corpus data, speaker dysfluencies, false starts, or errors during data annotation. Therefore, in order to control for accidental (non)-alternating behaviour, the interval should be constrained by a threshold parameter τ at both extremes, such that $\tau < \widehat{\pi}_{\text{obj}} < 1 - \tau$. I will employ a value of $\tau = 0.05$ to control for the most conspicuous borderline cases.[8]

While offering a convenient operationalisation, the present view of omission availability raises the notorious 'negative data' problem in corpus linguistics: "just because a construction does not surface in a corpus, it does not follow that it is ungrammatical" (Hoffmann 2011: 10–11). If a verb never drops its objects in a sample, it does not imply that it categorically rejects omission; the same logic applies to verbs that seem to always drop their objects. Undoubtedly, the nature of the sample can lead to a systematic over- or underestimation of π_{obj}. However, the present study mitigates this concern in two ways. First, rather than working with downsampled data, all attestations of target verbs in ICE-GB are examined, providing complete coverage of their usage patterns across the corpus's 12 registers. Second, this comprehensive register differentiation means that categorical (non-)occurrence likely reflects genuine linguistic constraints rather than sampling artifacts or genre-specific biases. In a balanced 1-million-word corpus, consistent omission behavior across diverse registers provides stronger evidence that the true population proportion approaches 0 or 1 as N tends to infinity. The threshold parameter τ further reduces the likelihood that borderline cases ($\pi_{\text{obj}} \approx 0.01$ or 0.99) are misclassified due to residual sampling variation or annotation errors.

8 The supplementary script includes a function that allows readers to explore the effects of different threshold values.

3.2 The verb sample

In order to investigate null instantiation availability, three sets need to be established: V_{trans}, V_{alt}, and $V_{\neg\text{alt}}$. The overarching set of transitive verbs can be approximated based on the extensive, yet not exhaustive list of transitive verbs (n = 1650) compiled by Glass (2021), which has been derived from Levin (1993).[9] To this base list, an additional 72 transitive verbs identified through the literature review described below were added, yielding a total sample of 1,722 transitive verbs. Although Glass acknowledges in her script that not all verbs in her list may be prototypically transitive, it is reasonable to assume that a qualitatively collated database of this size is likely to provide representative insights into the characteristics of transitive verbs. When queried against ICE-GB, only 972 of these 1,722 transitive verbs yielded attestations in the corpus.

Identifying V_{alt} has posed a considerable challenge. As a point of departure, putative alternating verbs have been systematically compiled from more or less recent papers published on the Understood Object Alternation, i.e., Glass (2021), Ruppenhofer and Michaelis (2010), Ruppenhofer (2004), Fillmore and Kay (1995), Goldberg (1995), Levin (1993), Rice (1988), Fillmore (1986), and Allerton (1975). Besides those, the valency dictionary by Herbst et al. (2004) was searched for any additional attestations.[10] Having compiled more than 200 lexical items, each verb was then inspected for its complementation patterns in Herbst et al. (2004), FrameNet,[11] and in the Unified Verb Index[12] to ascertain whether they could be used transitively or ditransitively and entailed a participant role that could be realised as the direct object. The syntactic forms of the objects were treated quite liberally, including both NPs and (non-)finite clauses, such as the typical objects of *eat*, *drink*, *forget*, or *understand*. Conversely, this excludes predicates such as *arrive*, *swim*, *stay* etc. whose core frame elements are, aside from the subject, oblique in nature, i.e., mostly PPs or NPs that cannot be conceptualised as objects in the traditional sense (cf. Hopper and Thompson 1980).

Given the usage-based orientation of the present approach, the alternating status of these putative object-omitting verbs had to be verified empirically. This was accomplished through a combination of manual annotation and automatic POS and dependency parsing in R (RStudio Team 2020) as well as Python (Van Rossum and

9 This list ('trans_verbs.csv') can be accessed via Lelia Glass's OSF repository (cf. https://osf.io/kxe3u/ [Last accessed: September 16, 2024]).

10 Of primary interest were those lemmas with complementation patterns marked as 'only if clear from context'. Numerous additional transitive verbs were scouted for examples that hint at object null instantiation. This list is intended to be comprehensive, but by no means exhaustive.

11 Cf. https://framenet.icsi.berkeley.edu [Last accessed: September 16, 2024].

12 Cf. https://uvi.colorado.edu [Last accessed: November 28, 2024].

Drake 2009).[13] Following the initial verb screening, all putative alternating verbs were queried in the British ICE-corpus and annotated manually for their object realisation patterns, yielding approximately 16,000 observations.

The corpus query returned a wide array of instances that could not plausibly be regarded as genuinely variable contexts. I excluded observations if they met at least one of following three criteria:

(i) The utterance is unintelligible.
(ii) The utterance shows an invariant pattern.
(iii) The utterance instantiates a formally similar yet functionally distinct transitivity alternation.

Case (i) was taken to apply if the corpus-annotation marked an utterance as containing unclear syllables or words in the immediate vicinity of the keyword (approx. within a five word window). Case (ii) concerns strongly lexicalised and hence syntactically invariable patterns (e.g., formulaic *to begin with*, *take care*, *mind you*, but also discourse markers such as *you know* or *you see*) as well as grammatical structures that consistently force a specific object realisation pattern (e.g., the absence of syntactic objects in passives and causatives). Another special case of invariance involves verbs which, according to the literature, should allow object null instantiation but do not display any such occurrences in the corpus data.[14]

Case (iii) refers to closely related transitivity alternations that are easily confused with object omission, such as the causative/inchoative (or ergative) alternation (Levin 1993: 27–30).[15] It involves syntactic subjects in active clauses realising thematic roles that are more typical of objects, such as the PATIENT or THEME (cf. 4–5). To lower the chance of false positives, verbs participating in this alternation were systematically excluded from the analysis, including *begin*, *taste*, *change*, *divide*, *clean*, *close*, and *drive* (which exemplifies the related 'induced action alternation').

(4) a. *[Janet]*$_{\text{AGENT}}$ ***broke*** *[the cup]*$_{\text{PATIENT}}$.
b. *[The cup]*$_{\text{PATIENT}}$ ***broke***.

(5) a. *[…] the United States and the Soviet Union* ***began*** *a menacing nuclear-tipped arms race* (S2B-034#4:1:A)

13 The main libraries leveraged to this end were `quanteda` (Benoit et al. 2018) in R and `spacy` (Honnibal and Montani 2017) in Python.

14 If the verb is relatively infrequent, this observed invariance may indicate that there are too few attestations to accurately represent its complementation patterns. At the same time, invariance in more frequent verbs raises the possibility of true object omission unavailability and calls into question earlier judgements in the literature.

15 I thank the anonymous reviewer who has brought these cases to my attention.

b. *Our relationship is just **beginning** – growing pains will be part of its growth.* (W1B-001#52)

After excluding as many ineligible items as possible during the manual annotation phase, the status of a verb as alternating or non-alternating was determined in accordance with Definition 3.4 and a noise threshold $\tau = 0.05$, returning 133 confirmed alternating verbs.

Finally, since $V_{\text{alt}} \cup V_{\neg\text{alt}} = V_{\text{trans}}$, the non-alternating verbs can be inferred straightforwardly by subtracting the empirically verified alternating verbs from Glass's set of transitive verbs. In concrete terms, the ICE-GB data is expected to contain $972 - 133 = 839$ non-alternating verbs. Naturally, this procedure introduces some degree of selection bias. The literature is unlikely to have documented all existing alternating verbs, but it is reasonable to assume that over a 50-year period the most salient ones have been identified. Crucially, verbs that systematically license object omission are likely to be frequent enough to attract scholarly attention across multiple studies and theoretical frameworks, as evidenced by verbs like *eat*, *write*, and *read* appearing consistently in the literature. Conversely, truly rare alternating verbs would likely contribute little to the statistical patterns under investigation.[16]

This set-theoretic shortcut is motivated by practical considerations as well: the 133 alternating verbs generated over 50,000 raw corpus hits, which after applying exclusion criteria (i–iii) yielded approximately 16,000 eligible observations for analysis. Extending this manual annotation process to all 972 verbs would require the examination of an estimated 200,000+ raw observations. Such a scale would exceed current resource constraints and could yield diminishing analytical returns, given the frequency argument above.

3.3 Feature annotation

The theoretical accounts reviewed in Section 2.1 have identified several dimensions likely to influence argument omission: lexical specificity, selectional preference strength, frequency and dispersion effects, and social routines. However, the predominantly qualitative nature of previous research has made many of these factors difficult to test empirically. While distributional features such as frequency and dispersion can be calculated directly from corpus data,

16 A more formal perspective on selection bias in the verb data is available as an online appendix in this study's OSF repository.

semantic properties, such as lexical specificity, can only be operationalised indirectly through psycholinguistic measurements. To address this challenge, variables that cannot easily be calculated are derived from the South Carolina Psycholinguistic Metabase (SCOPE; Gao et al. 2022) or from pre-trained semantic vector space models.[17] The predictor variables are presented along with their measures of central tendency and moments in Table 1.

Frequency and dispersion measures are calculated directly from the corpus data following established protocols (Gries 2020). Since the distribution of the raw frequencies shows the characteristic positive skew associated with influential outliers, the counts were log-transformed to render them more symmetric around the mean. Dispersion and Resnik's selectional preference strength were both computed using relative entropy.[18] Dispersion values close to 0 signify that a lexical item's observed distribution matches the expected distribution based on corpus structure; conversely, higher deviation scores indicate specialisation in particular genres.[19] Analogously, verbs with a selectional preference strength of 0 do not prefer any particular noun classes, whereas higher scores indicate stronger selectivity. Although it was only possible to compute meaningful Resnik scores for 399 verbs, this verb sample still exceeds that of Resnik (1996) by nearly a factor of 12. Social routine ratings were excluded due to the problems outlined at the end of Section 2.2.

Rice's particular notion of specificity, which should distinguish general verbs such as *eat* or *write* from seemingly more specific alternatives like *devour* or *inscribe*, is rather vague: it is not clear whether it represents "category inclusiveness" (Villani et al. 2024: 870), or concreteness, with which it is typically confused. Accordingly,

---

17 A full breakdown of all variables in the SCOPE database is given in Gao et al. (2022: 2855–2864). See also the online search interface at https://sc.edu/study/colleges_schools/artsandsciences/psychology/research_clinical_facilities/scope/search.php [Last accessed: June 13, 2025].

18 The calculation of selectional preference strength follows Resnik's (1993) equations (cf. Section 2.2) and has been implemented in close accordance with Glass's (2021) Python script. To validate the output, the measures were compared to the published values in Resnik (1996: 138), which had been derived from the Brown, CHILDES, and Norms datasets. Across the 25 verbs present both in the current and in Resnik's dataset, the calculated measures show a strong positive correlation with the Brown corpus values (Pearson's $r = 0.852$, 95 % CI [0.689, 0.933], $t(23) = 7.81$, $p < 0.001$), a moderate one with the CHILDES data (Pearson's $r = 0.535$, 95 % CI [0.177, 0.768], $t(23) = 3.04$, $p < 0.01$) and no significant one with the Norms data (Pearson's $r = 0.380$, 95 % CI [-0.018, 0.674], $t(23) = 1.97$, $p = 0.06$). Therefore, the selectional preferences measured in this study closely mirror Resnik's measurements in the Brown and CHILDES data, supporting their validity and further analysis.

19 More precisely, for the 12 ICE text categories with 500 files in total, the expected proportions are: S1A (0.2), S1B (0.16), S2A (0.14), S2B (0.1), W1A (0.04), W1B (0.06), W2A (0.08), W2B (0.08), W2C (0.04), W2D (0.04), W2E (0.02), W2F (0.04); see also https://www.ice-corpora.uzh.ch/en/design.html [Last accessed: September 23, 2025].

Table 1: Summary statistics on 972 transitive verbs.

Variable	Missing *n* (%)	Mean	SD	Median	MAD	Min	Max	Skewness	Kurtosis
Log_frequency	0.00	1.96	1.71	1.61	1.53	0.00	8.15	0.87	0.13
Dispersion	0.00	2.08	1.35	1.98	1.50	0.01	5.64	0.47	−0.64
Concreteness	2.67	3.39	0.91	3.39	1.05	1.42	5.00	−0.08	−1.01
Number_meanings	1.44	8.88	8.18	7	5.93	1.00	75.00	2.71	11.61
Number_features	45.27	18.04	14.23	13	5.93	5.00	143.00	3.11	5.78
Neighbourhood_density	0.41	93.70	53.88	83.9	52.09	6.87	332.19	0.98	1.10
Selectional_preference_strength	58.33	4.03	2.06	3.77	2.15	0.31	11.87	0.73	0.69

Rice's specificity will be quantified via several complementary measures, intended to distinguish between 'true' specificity and concreteness.

The SCOPE database lists four concreteness metrics in the 'Concreteness/Imageability' group, all of which exhibit strong intercorrelations (Spearman's $\rho > 0.77$ for all pairwise comparisons). Given this high degree of redundancy, 'Conc_Brys' (Brysbaert et al. 2014) has been selected as it exhibits the lowest proportion of missing values (2.70 % vs. > 29 % for alternative measures).[20] These Likert-scale ratings range from 1 (abstract) to 5 (concrete) and reflect how word meanings tend to be acquired, with concrete meanings being more likely to be learnt through bodily experience. Furthermore, lemma polysemy is additionally controlled for because it can confound the effect of concreteness (Reijnierse et al. 2019). I rely on 'NSenses_WordNet' (G. A. Miller 1995), which counts distinct word senses in the WordNet database. This measure correlates highly with alternative polysemy measures (e.g., 'Nsenses_Webster', $\rho = 0.49$) while also providing nearly complete coverage for all target verbs.

Additionally, SCOPE's semantic neighbourhood density measure is included to capture certain aspects of specificity. This measure places greater emphasis on taxonomic relations (e.g., *cow – bull*) over associative ones (e.g., *cow – milk*), with empirical support from the finding that "abstract words tend to have more sparse neighborhoods than concrete words" (Reilly and Desai 2017: 51). The 'NFeatures' norms (Buchanan et al. 2019), which display substantial data sparsity (45 % missing values), collect the number of distinct semantic features associated with a lexical item; for example, the cue word *duck* could trigger responses such as *is a bird*, *quacks*, or *has wings* (Buchanan et al. 2019: 1852). Given their taxonomic orientation, neighbourhood density and a word's number of features are probably most suitable for quantifying verb specificity (cf. Villani et al. 2024: 869).

As a diagnostic precaution, it is worth inspecting the pairwise correlations of the predictors to gauge potential (multi-)collinearity issues. In fact, Table 2 reveals a multitude of significant correlations that may create problems with parameter estimation. Frequency and neighbourhood density are positively correlated with each other, and both are negatively correlated with dispersion and selectional preference strength, suggesting that higher verb frequencies are associated with more dense taxonomic neighbourhoods, more even dispersion and lower selectional preferences. By contrast, uneven dispersion is linked to lower polysemy and more sparse semantic

20 The low NA counts of concreteness, the number of meanings, and neighbourhood density rendered data imputation a plausible option. Missing values were reconstructed via the `imputePCA()` function from the `missMDA` package in R (Josse and Husson 2016).

neighbourhoods. Concreteness and the number of features show the weakest pairwise correlations with the remaining predictors.[21] Computing variance inflation factors[22] returns elevated values for dispersion (3.30), selectional preferences (5.15) and log frequency (6.19), while the remaining features stay below 2. The VIFs fall below 5 if either log frequency or Resnik's strength is removed. In sum, the correlated features point to partial mutual redundancies among the predictors that are likely to affect model performance and interpretation.

To complement these distributional and psycholinguistic measures, pre-trained word vectors from the R library `fastText` (specifically the `cc.en.300.bin` model) are included for all 972 verb lemmas (Mikolov et al. 2018; Mouselimis 2024).[23] Semantic vector space models capture lexical semantic properties through high-dimensional co-occurrence matrices, thereby yielding fine-grained representations of contextual word use. These models are grounded in the Distributional Hypothesis: "The degree of semantic similarity between two linguistic expressions *A* and *B* is a function of the similarity of the linguistic contexts in which *A* and *B* can appear" (Lenci 2008: 3).[24] Considering that vector space models have demonstrated effectiveness in modelling lexical change (Hamilton et al. 2016), inflectional morphology (Shafaei-Bajestan et al. 2022), and syntactic productivity patterns (Perek 2016), they are likely to capture at least some of the differences and similarities between alternating and non-alternating verbs.

3.4 Statistical methods

Following the definitions set forth in Section 3.1, object omission availability is modelled as a dichotomous random variable:

21 The moderate, yet statistically significant pairwise correlations involving concreteness, number of features, and neighbourhood density mirror the weak relationship between concreteness and specificity reported in the literature (Villani et al. 2024: 869). What is noteworthy, however, is the weak association between 'Number_features' and 'Neighbourhood_density', as these features were expected to capture the same latent specificity construct. It may be more plausible to assume that they are capturing distinct aspects of specificity, or semantic representation altogether. Note also the expected correlation between concreteness and polysemy (Reijnierse et al. 2019) as well as between polysemy and frequency (cf. Zipf 1949: 27–31).

22 These correspond to the diagonal elements of the inverted feature correlation matrix, i.e., $\mathrm{diag}(\mathbf{R}^{-1})$.

23 The exact model generating these vectors is described in Bojanowski et al. (2017).

24 For an overview of the conceptual and mathematical foundations of semantic vector space models, see Turney and Patrick (2010).

Table 2: Feature correlation matrix based on Spearman rank correlation coefficient ρ; p-values were adjusted using Holm's method.

	Predictors						
	Log_frequency	**Dispersion**	**Concreteness**	**Number_meanings**	**Number_features**	**Neighbourhood_density**	**Selectional_preference_strength**
Log_frequency	1.000						
Dispersion	−0.893***	1.000					
Concreteness	−0.315***	0.293***	1.000				
Number_meanings	0.383***	−0.398***	0.319***	1.000			
Number_features	0.235***	−0.226***	0.190*	0.139	1.000		
Neighbourhood_density	0.651***	−0.645***	−0.150	0.430***	0.146	1.000	
Selectional_preference_strength	−0.930***	0.870***	0.353***	−0.337***	−0.222***	−0.615***	1.000

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

$$\text{Object omissible} = \begin{cases} \text{yes} & \text{if } v \in V_{\text{alt}}. \\ \text{no} & \text{if } v \notin V_{\text{alt}}. \end{cases} \quad (4)$$

The aim is to estimate the probability that a verb licenses object omission conditional on the independent variables. Formally, this conditional probability can be written as $P(\text{Object omissible} = \text{yes}|\mathbf{X})$, where $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ is the $N \times p$ matrix containing the explanatory variables.

3.4.1 Generalised additive models

Generalised additive models (GAMs henceforth, Hastie and Tibshirani 1991; Wood 2017) provide a flexible framework for modelling this probability while accommodating non-linear relationships between the response variable and the predictors.[25] The assumption of linearity is unlikely to hold here, as argued on theoretical grounds in Section 2.2, rendering GAMs more appropriate than strictly linear models. Concretely, the log odds of object omissibility is modelled as a sum of smooth functions, one per predictor, linked to the response via a logistic function.

Each smooth term is constructed from a set of flexible basis functions whose shape is estimated from the data. To prevent overfitting, a smoothness penalty controls the curvature of each function; smoother fits are preferred unless the data strongly support greater complexity. This trade-off is governed by a smoothing parameter estimated from the data via REML (Wood 2017). The residual flexibility of a smooth term after penalisation is summarised by its effective degrees of freedom (edf): values near 1 indicate a nearly linear effect, whereas higher values indicate increasing non-linearity. Smooth terms with edf ≈1 are additionally reported as linear coefficients $\widehat{\beta}$ with effect sizes expressed as odds ratios $e^{\widehat{\beta}}$. Based on the literature review in Section 2.2, non-zero effects are expected for all predictors included in the models. In the non-linear case, p-values are derived by testing whether a given smooth term has no effect on the response (D. L. Miller 2025: 450), using a χ^2-distributed test statistic (Wood 2013: 224–225).

Model fit is assessed via the proportion of deviance explained. To also assess the predictors' individual contributions to the explained residual deviance, relative feature importance is estimated with the `gam.hp` package in R (Lai et al. 2024). Since some predictors are strongly correlated (e.g., frequency and dispersion), the package accounts for concurvity (i.e., multicollinearity in the non-linear case) and the "overlapping variance" it introduces (Lai et al. 2024: 543). Further technical details on the model specification, measures of fit, and the individual variable importance are given in Appendix B.

25 The code implementation relies on Simon Wood's `mgcv` library in R (Wood 2017).

3.4.2 Linear discriminant analysis

Linear Discriminant Analysis (LDA; Hastie et al. 2017: 106–112; James et al. 2021: 141–152) is used to analyse the high-dimensional semantic vector space data. LDA is computationally efficient, has a closed-form solution requiring no iterative optimisation, and is well suited to high-dimensional datasets with continuous predictors (Graf et al. 2024; Hastie et al. 2017: 438–439). If Y is a discrete response variable with K classes, both LDA and logistic regression model the conditional probability $P(Y = k|\mathbf{X} = \mathbf{x})$, but differ in how they arrive at it. Rather than maximising a likelihood function, LDA models the class-conditional densities $P(\mathbf{X} = \mathbf{x}|Y = k)$ and applies Bayes' theorem to obtain posterior class probabilities. The resulting discriminant functions and the derivation of the classification rule are given in Appendix C. Drawing on pre-trained `fastText` embeddings, discriminant scores were estimated for all 972 transitive verbs and included in the GAM as an additional predictor.[26]

3.4.3 Predictive performance

Once the probabilities have been estimated, the models can be evaluated in terms of their classification performance, i.e., their ability to correctly identify alternating and non-alternating verbs in the data. Conventionally, the model predictions and true labels are cross-classified in a confusion matrix, where the cells identify the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In the verb data, TPs are alternating verbs which the model also predicts to be alternating; conversely, FPs are non-alternating verbs erroneously classified as alternating. TNs and FNs represent the analogous cases for non-alternating verbs.

These four quantities permit the computation of several evaluation metrics, each capturing a different aspect of classification performance. In addition to relative model fit, the analyses report the cross-validated accuracy, precision, recall, F1 as well as Matthew's correlation coefficient of each model examined (MCC; cf. Equation (5)). The MCC has been argued to be more informative than metrics such as precision or recall alone, as it balances all four confusion matrix rates (sensitivity, specificity, precision, and negative predictive value; cf. Chicco and Jurman 2023: 4). The coefficent ranges from −1 (every observation was misclassified) to +1 (every observation was classified correctly), with 0 corresponding to random predictions (Chicco and Jurman 2020: 5).

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}}. \tag{5}$$

26 Since the scores are constructed independently for the training and test data, there should not be any data leakage involved.

4 Results

After applying threshold control as outlined in Section 3.1, the number of alternating verbs included in the final sample was reduced from 142 to 133. The violin plots in Figure 1 reveal systematic differences between verbs that allow object omission and those that do not across seven dimensions. Six of the seven predictors show statistically significant differences ($p < 0.001$) according to unpaired Wilcoxon signed rank tests (see Table 3). Verbs with omissible objects demonstrate consistently higher values across most measured variables: greater log-frequency (Panel A), more even dispersion (Panel B), more meanings (Panel D), more semantic features (Panel E), and sparser semantic neighbourhoods (Panel F), but, notably, lower selectional preference strength (Panel G). The direction of the difference in selectional preferences is unexpected, as Resnik (1996) had observed the opposite tendency.[27]

The effect sizes, as measured by $z/\sqrt{N}$, range from moderate (0.19 for number of features) to strong (0.43 for selectional preference strength). Notably, concreteness (Panel C) is the only variable that did not show a significant difference between the two groups ($p = 0.82$). They display nearly identical distributions, indicating that the mode of semantic acquisition (experience-based vs. language-based) does not influence object omissibility patterns in isolation.

4.1 Vector space analysis

Having retrieved 300 word vectors for each of the 972 verbs in the sample, this high-dimensional space was projected onto a two-dimensional plane using *t*-distributed Stochastic Neighbour Embedding (*t*-SNE; van der Maaten and Hinton 2008) as implemented in the R package `Rtsne` (Krijthe 2015). This unsupervised technique is well-suited for revealing cluster structure in high-dimensional data. The *t*-SNE visualisations in Figure 2 reveal the verbs' semantic organisation, with alternating verbs and non-alternating verbs colour-coded for comparison. While the two classes are not cleanly separable, the distribution of alternating verbs does not seem random.

Inspecting these regions reveals several semantically coherent neighbourhoods. In the upper-centre and upper-left, alternating verbs cluster around transactional and exchange events: *sell*, *buy*, *pay*, *borrow*, and *lease* suggest commercial transactions. A similar, yet more markedly formal transactional theme is represented in

27 Resnik (1996) took into account 34 transitive verbs, 15 of which he deemed alternating, whereas this study examines more than a hundred. That this tendency is reversed with a larger sample is completely within the realm of the possible.

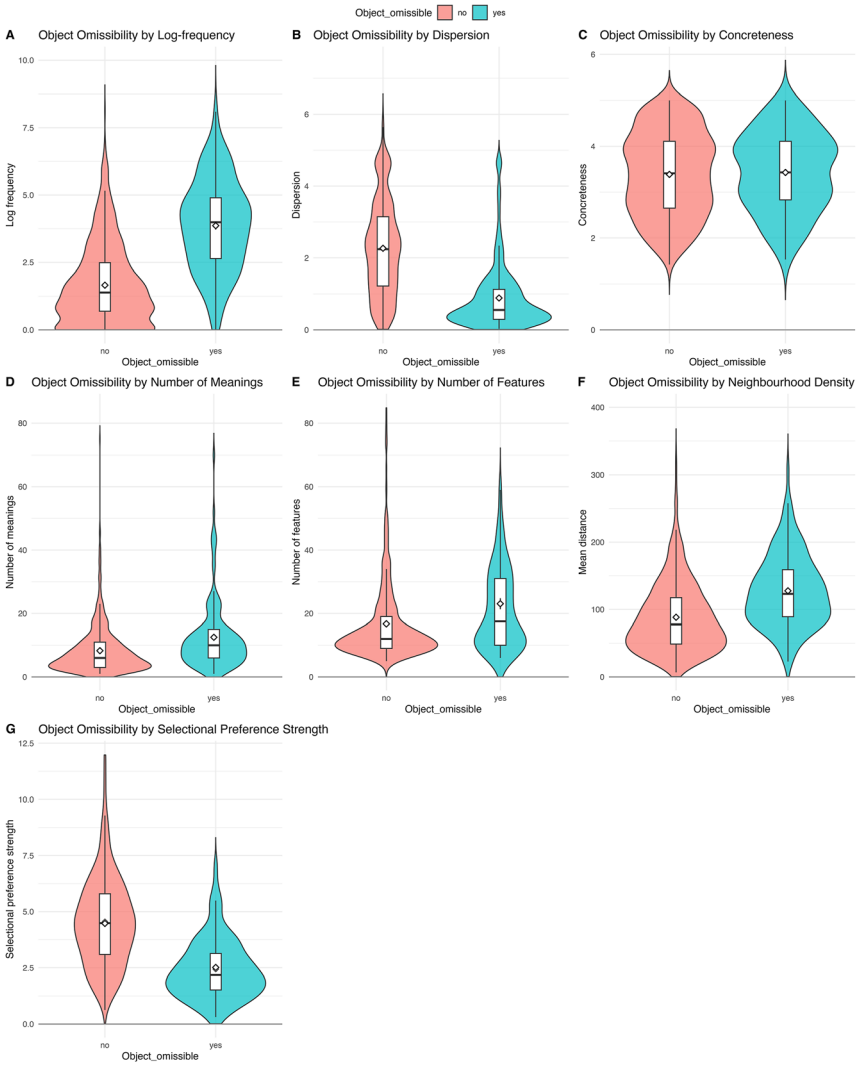

Figure 1: Descriptive statistics on object omissibility.

the centre by *withdraw*, *serve*, *share*, *contribute*, and *permit*, among others. The most dense neighbourhood in the upper-right of the figure involves cognitive processes (*imagine*, *remember*, *believe*, *understand*, *forget*), gradually transitioning into a communication cluster (*explain*, *answer*, *oppose*, *insist*, *advise*, *warn*, *sing*). The uppermost subgroup is more heterogeneous, encompassing interactions with people and objects in the broadest sense (*enter*, *leave*, *try*, *ask*, *refuse*, *help*, *join*), partially

Table 3: Wilcoxon signed-rank tests with Holm-adjusted *p*-values (p_{adj}).

Feature	Test statistic *W*	p_{adj}	Effect size
Log frequency	18,358	<0.001	0.40
Dispersion	91,686	<0.001	0.38
Concreteness	52,634	0.82	0.63
Number of features	17,063	<0.001	0.19
Number of meanings	37,912	<0.001	0.19
Neighbourhood density	31,174	<0.001	0.26
Selectional preference strength	22,699	<0.001	0.43

overlapping with the aforementioned stative verbs (*see*, *know*, *hear*). What these clusters share is that the events they describe are mundane yet essential to communal life.

Examples (6–11) illustrate variable object realisation patterns for a subset of the verbs above. Their annotation includes the semantic frame realised (e.g., COMMERCE_SELL), the thematic role of interest (e.g., GOODS), and, in case of omission, the interpretive pattern of the frame element (INI or DNI), following the labelling conventions of FrameNet.[28] For every utterance, the corresponding text category, file number, and speaker turn in the ICE corpus is indicated in parentheses; the corresponding feature values in the dataset are reported in Table 7 in Appendix A.

(6) COMMERCE_SELL
   a. *They do actually **sell** [books]*$_{\text{GOODS}}$ (S1A-066#127:1:B)
   b. *Philips were not too keen to **sell** Ø*$_{\text{INI:GOODS}}$ *to a competitor but were aware what this sale could do for jobs in the area.* (W2B-012#28:1)

(7) SIGN_AGREEMENT
   a. *You will be asked to **sign** [an acknowledgement*$_{\text{AGREEMENT}}$*] that you have been given your prescription.* (W2D-001#39:1)
   b. *Just put your name address and **sign** Ø*$_{\text{AGREEMENT}}$ *at the bottom here* (S1A-051#237:3:A)

(8) DECIDING
   a. *But it will be difficult for Rajiv Gandhi to **decide** [how to take advantage of this]*$_{\text{DECISION}}$ (S2B-007#57:1)

28 If the frame patterns of a certain verb are not documented in FrameNet, such as *milk* (cf. 18), then the participant role labels were borrowed from closely related frames denoting a similar event structure. For example, a plausible frame instantiated by *milk* could be EMPTYING, in which an AGENT removes a THEME from a SOURCE.

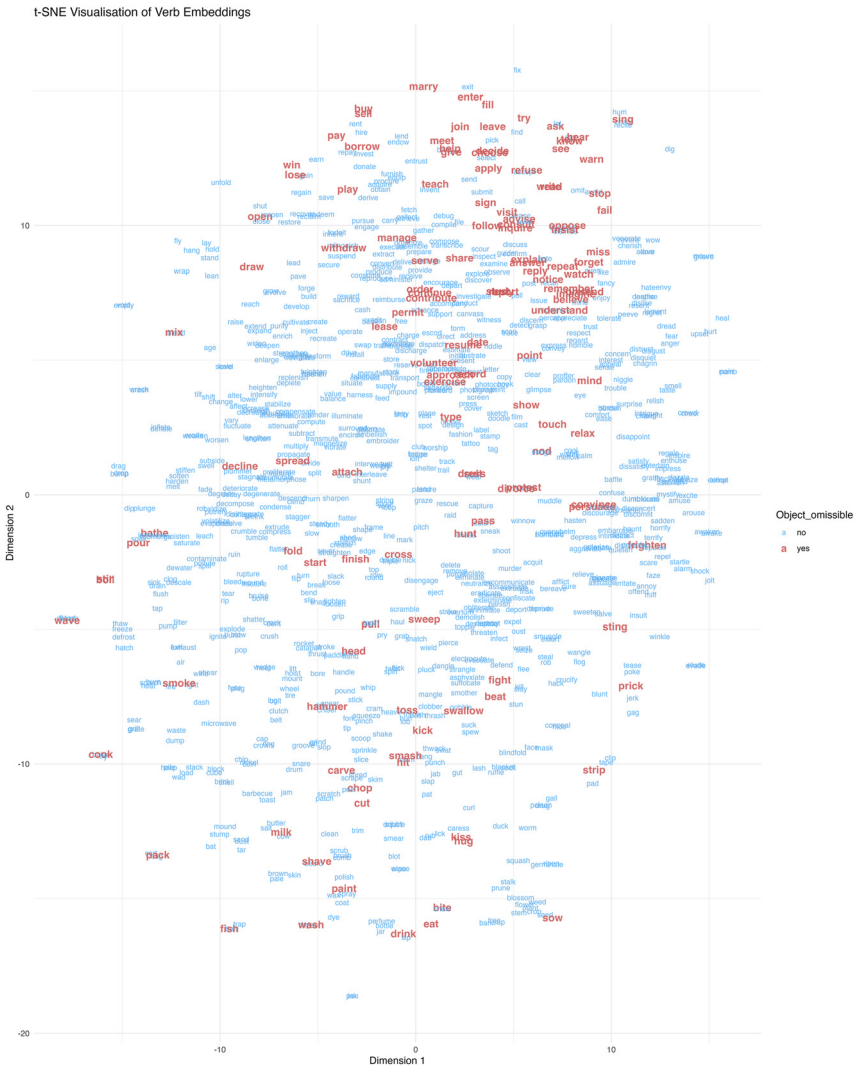

Figure 2: *t*-SNE visualisation of (non)-alternating verbs in vector space.

b. *Your Lordship has to **decide** Ø*$_{\text{INI:DECISION}}$ (S2A-068#41:1:A)

(9) STATEMENT

a. *That would **explain** [a few things*$_{\text{MESSAGE}}$*]* (S1A-010#77:1:B)

b. *I'll **explain** Ø*$_{\text{DNI:MESSAGE}}$ (S1A-096#2:1:B)

(10) MEMORY
  a. *So glad to* ***remember*** *[the way you speak]*$_{\text{CONTENT}}$ (W1B-008#64:3)
  b. *I'll* ***remember*** $\emptyset_{\text{DNI:CONTENT}}$ *as soon as I see your face* (W2F-020#204:1)

(11) ASSISTANCE
  a. *How can I* ***help*** *[you]*$_{\text{BENEFITTED\_PARTY}}$ (S1A-066#24:1:A)
  b. *But it might well* ***help*** $\emptyset_{\text{DNI:BENEFITTED\_PARTY}}$ *if you've uhm actually done some [phonetics]* (S1A-008#17:1:B)

Moving downward along Figure 2, alternating verbs become sparser; simultaneously, causative and force-dynamic senses predominate (e.g., *touch*, *nod*, *hunt*, *frighten*, *sting*), with their relative intensity somewhat increasing towards the bottom (*fight*, *beat*, *toss*, *smash*). This region is where prototypical alternating verbs such as *kiss*, *eat* or *drink* are found. Their neighbourhoods are sparse, with few transitive verbs occupying this region in the vector space overall. The bottom-left corner captures substance manipulation, grooming, and the procurement of food: *paint*, *carve*, *chop*, *cut*, *wash*, *fish*, and *milk*, among others. The left-most region is almost devoid of alternating verbs, but the few existing ones fit the general themes of grooming, food, and consumption: *bathe*, *stir*, *smoke*, *cook*. This pattern continues to support the observation regarding the semantic basicness of alternating verbs; further examples are given in (12–14).

(12) TEXT_CREATION
  a. *At last I put pen to paper and actually* ***write*** *[a letter]*$_{\text{TEXT}}$ *to you!* (W1B-002#3:1)
  b. *and then as you're* ***writing*** $\emptyset_{\text{INI:TEXT}}$ *the inspiration comes* (S1B-048#13:1)

(13). HUNTING
  a. *[…] you might as well be* ***fishing*** *[simple maggot or bread]*$_{\text{FOOD}}$ (W2D-017#20:1)
  b. *If you* ***fish*** $\emptyset_{\text{INI:FOOD}}$ *with a quality bait in the right place […]* (W2D-017#25:1)

(14) CUTTING
  a. ***Cut*** *[the lamb]*$_{\text{ITEM}}$ *into cubes and put it into the bowl with all the other ingredients.* (W2D-020#81:1)
  b. *Add the meat,* ***cut*** $\emptyset_{\text{DNI:ITEM}}$ *into neat cubes with the fat and gristle removed, and process for another 10-15 seconds until thoroughly blended.* (W2D-020#13:1)

The remaining regions, especially those in the left and right peripheries, are dominated overwhelmingly by non-alternating transitive verbs, which tend to describe more abstract and specific actions with causative semantics (*develop*, *inject*, *acquit*,

replenish, *subtract*, *ruin*, *filter*) and potentially less basic social acts and punishments (*banish*, *excommunicate*, *destroy*, *ruin*, *steal*, *rob*, *sever*). This recurrent negative valence among non-alternating verbs is striking and warrants further dedicated analysis. On the whole, this qualitative exploration suggests that alternating verbs tend to describe functionally generic events that are likely to be strongly routinised (Glass 2021).

Next, a linear discriminant analysis was conducted on the full feature space with 300 word vectors. After performing leave-one-out cross-validation, the LDA model exhibits decent accuracy (81 %), high precision (0.90), yet poor recall (0.32), thereby only capturing approximately a third of all alternating verbs (49 TPs and 105 FNs). Nevertheless, the MCC lies above the random baseline (0.23). Figure 3 illustrates how the posterior probabilities are obtained from the discriminant scores after a logistic transformation: higher scores correspond to higher posterior probabilities of object omission availability.[29] Some alternating verbs allow for clean identification in vector space, most of which are part of the dense neighbourhoods identified above (e.g., *buy*, *try*, *win*, or *ask*). Non-alternating verbs like *say* or *let* constitute exceptions. Conversely, verbs with low discriminant scores are more likely to be non-alternating, with few alternating verbs interspersed between them. These scores are treated as an independent variable henceforth, serving as a proxy for the information contributed by the semantic vector space.

4.2 (Non-)linear patterns in object omissibility

Because predictors differ in their numbers of complete observations, GAMs were fitted to three nested subsets of the full dataset (Figure 4). Note that `s()` represents smooth terms and `ti()` non-linear interaction terms. Model A is trained only on those observations where none of the predictors display missing values; as such, it will inevitably show the lowest power and provide the most conservative estimates of the effects. Model B excludes selectional preference strength, and Model C additionally disregards 'Number_features', thereby increasing the number of complete observations to 532 and 972, respectively, the last of which corresponds to the full dataset. Each model was estimated via restricted maximum likelihood (REML) and a `gamma` penalty of 1.4, which can improve the test performance (Wood 2017: 231).

Table 4 summarises the model statistics of Models A, B, and C, respectively. Note that the 'no:yes' row captures the ratio of non-alternating verbs to alternating ones, summing up to *n* for each model. To evaluate their predictive performance, all

29 Note that since the target variable is dichotomous, there is only one linear discriminant axis (LD1).

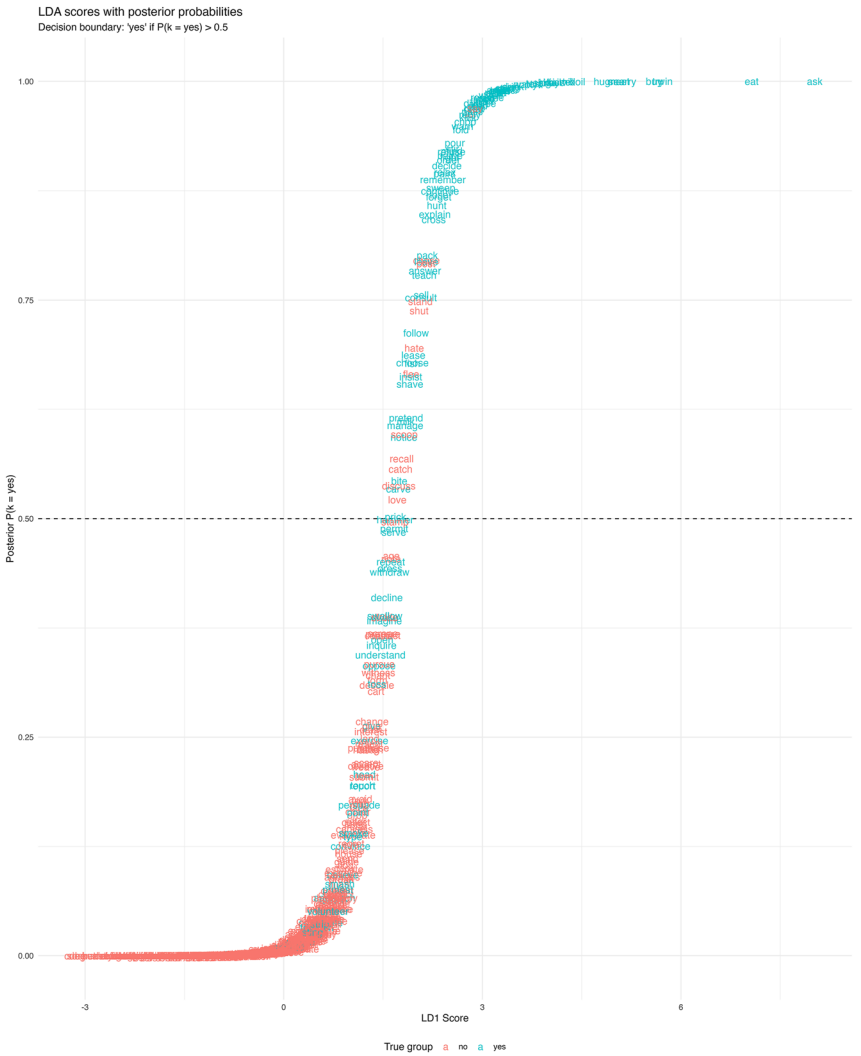

Figure 3: Linear discriminant (LD1) scores by posterior class probabilities and observed omissibility group with default classification boundary at 0.5.

relevant classification metrics, including the underlying confusion matrices, are reported in Table 5.

Model A, which was trained on complete cases (n = 229), displays excellent fit with approximately 67 % explained model deviance and very balanced predictive performance throughout (MCC = 0.783), capturing a substantial portion of true

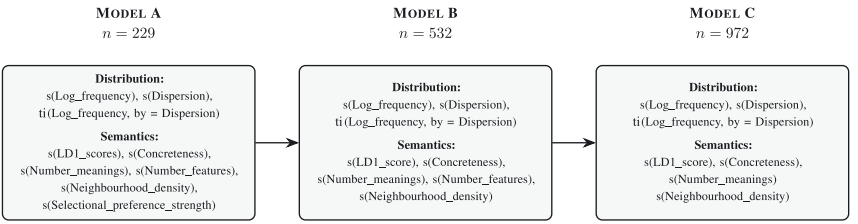

Figure 4: Summary of modelling steps.

positives (Recall = 0.855). The edfs indicate some non-linearity in the effects of neighbourhood density and LD1 scores, yet only the latter reaches statistical significance. Treating predictors with edf = 1 as linear terms, none reaches significance. The concrete contributions of the features to the proportion of explained deviance follow a clear hierarchy: LD scores > distributional features > semantic features. It is worth noting that verb frequency, dispersion, their interaction, and selectional preference strength exhibit substantial concurvity, each approaching non-identifiability. This means that to a large extent these terms are mutually redundant, capturing virtually the same information about object omissibility.

Model B was trained on a larger subset of the input data (n = 532), excluding selectional preference strength and thus increasing statistical power. This gain in power comes at virtually no cost, as selectional preference strength contributes negligible information once frequency and dispersion are included. Regardless of the more pronounced class imbalance in the dependent variable, both model fit and predictive accuracy remain strong. Once more, except for the LD1 scores, none of the model predictors exhibit meaningful (non)-linear effects. The feature importance ranking of Model A applies to Model B as well.

Model C has been fitted to the largest dataset, albeit with the fewest predictors and the greatest class imbalance with a ratio of almost 6:1 in favour of non-alternating verbs. In terms of fit and predictive performance, it is comparable to Models A and B in all respects. One noteworthy weakness is the mild decrease in recall (0.714) compared to the other models. Considering that the dataset underlying Model C has an additional 417 non-alternating and 23 alternating verbs compared to Model B, this decrease is negligible. Instead, it highlights the robustness of Model C in light of strong class imbalance, suggesting that the model has captured the essential determinants of object omissibility. In this final model, the LD1 scores account for 48.1 % of the deviance explained, followed by the distributional predictors, which jointly contribute approximately 16.1 %, and the semantic ones (≈6.6 %). The larger sample size also yields statistically significant smooth terms for dispersion and its interaction with log-frequency.

Table 4: Generalized additive model results for three nested models; final column reports individual contribution to explained deviance.

Term	Model A					Model B					Model C				
	edf	Ref.df	χ^2	*p*	Dev. expl.	edf	Ref.df	χ^2	*p*	Dev. expl.	edf	Ref.df	χ^2	*p*	Dev. expl.
s(Log_frequency)	1.00	1.00	0.06	0.80	0.041	1.63	2.07	1.43	0.49	0.060	2.25	2.96	3.50	0.30	0.068
s(Dispersion)	1.00	1.00	0.04	0.85	0.033	2.00	2.56	3.10	0.23	0.051	1.87	2.42	2.47	0.31	0.059
ti(Log_frequency, Dispersion)	1.00	1.00	3.29	0.27	0.030	1.00	1.00	0.17	0.68	0.025	1.00	1.00	0.038	0.85	0.034
s(LD1_score)	1.00	1.00	46.41	<0.001***	0.486	1.00	1.00	81.93	<0.001***	0.486	1.00	1.00	114.04	<0.001***	0.481
s(Concreteness)	1.00	1.00	3.27	0.07	0.010	1.00	1.00	1.32	0.25	0.005	1.00	1.00	0.04	0.83	0.044
s(Number_meanings)	1.00	1.00	0.00	0.98	0.006	1.00	1.00	0.05	0.83	0.004	1.00	1.00	0.00	0.98	0.009
s(Number_features)	1.40	1.69	2.49	0.15	0.019	1.00	1.00	2.78	0.10	0.019					–
s(Neighbourhood_density)	1.93	2.45	2.34	0.39	0.012	1.00	1.00	0.22	0.64	0.016	1.00	1.00	0.05	0.93	0.013
s(Selectional_preference_strength)	1.00	1.00	0.01	0.92	0.034					–					–
Deviance explained					0.670					0.662					0.674
n					229					532					972
no:yes					153:76					422:110					839:133

Table 5: Confusion matrices for Models A–C (rows = actual, columns = predicted), with cross-validated performance metrics.

Actual\Predicted	Model A		Model B		Model C	
	no	yes	no	yes	no	yes
no	142	11	408	14	822	17
yes	11	65	29	81	38	95
Accuracy		0.904		0.919		0.943
Precision		0.855		0.853		0.848
Recall		0.855		0.736		0.714
F1		0.855		0.790		0.776
AUC		0.941		0.958		0.969
MCC		0.783		0.743		0.747

The smooths of Model C are plotted in Figure 5. The only predictor that seems to exhibit some non-linearity is dispersion, yet it is insufficient to reject the null hypothesis that there is no non-linear effect. Although the surface captured by the term 'ti(Log_frequency, Dispersion)' suggests some minor interaction effects in the contour plot (see Figure 5, top-right), it does not reach statistical significance. None of concreteness, number of meanings, or neighbourhood density displays a statistically significant effect. Crucially, Model C reveals a linear effect of LD1 scores on object omissibility that is highly significant and strong ($\widehat{\beta} = 2.39$, 95 % CI [1.95, 2.82], $z = 10.68$, $p < 0.001$, $e^{\widehat{\beta}} = 10.86$). A one-unit increase in a verb's LD1 scores thus increases the odds of object omissibility by a factor of approximately 11.

4.3 Word vectors revisited

Word vectors consistently account for the bulk of deviance in all three GAMs. Consequently, word vectors correlating positively and strongly with LD1 tend to characterise contexts favouring object omission, whereas those correlating negatively characterise contexts favouring object retention. Word vectors with non-negligible contributions to the discriminant scores are shown in Figure 6. While most of these vectors are nearly orthogonal, there is a small subset with moderate dependencies (e.g., V137 and V145, $r = -0.442$).

The predictive power of word vectors raises the question of what semantic dimensions they actually capture. Hollis and Westbury (2016), in an attempt "to see whether coherent semantic dimensions can be reverse-engineered from the vector

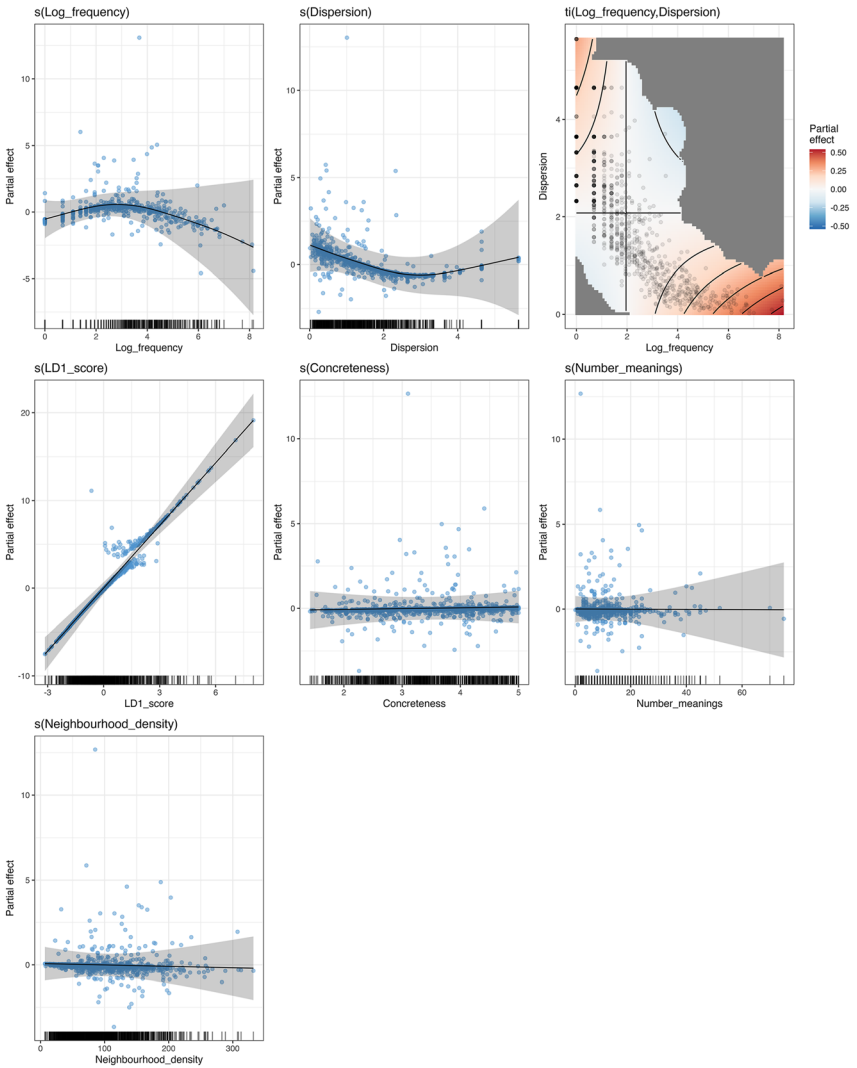

Figure 5: Smooth effects for distributional and semantic predictors with 95 % credible intervals.

representations of meaning" (Hollis and Westbury 2016: 1745), find that in the specific embedding space they investigated, there were noteworthy correlations between word vectors and psycholinguistic features such as concreteness, specificity, or emotional valence, among others (Hollis and Westbury 2016: 1753–1754), none of

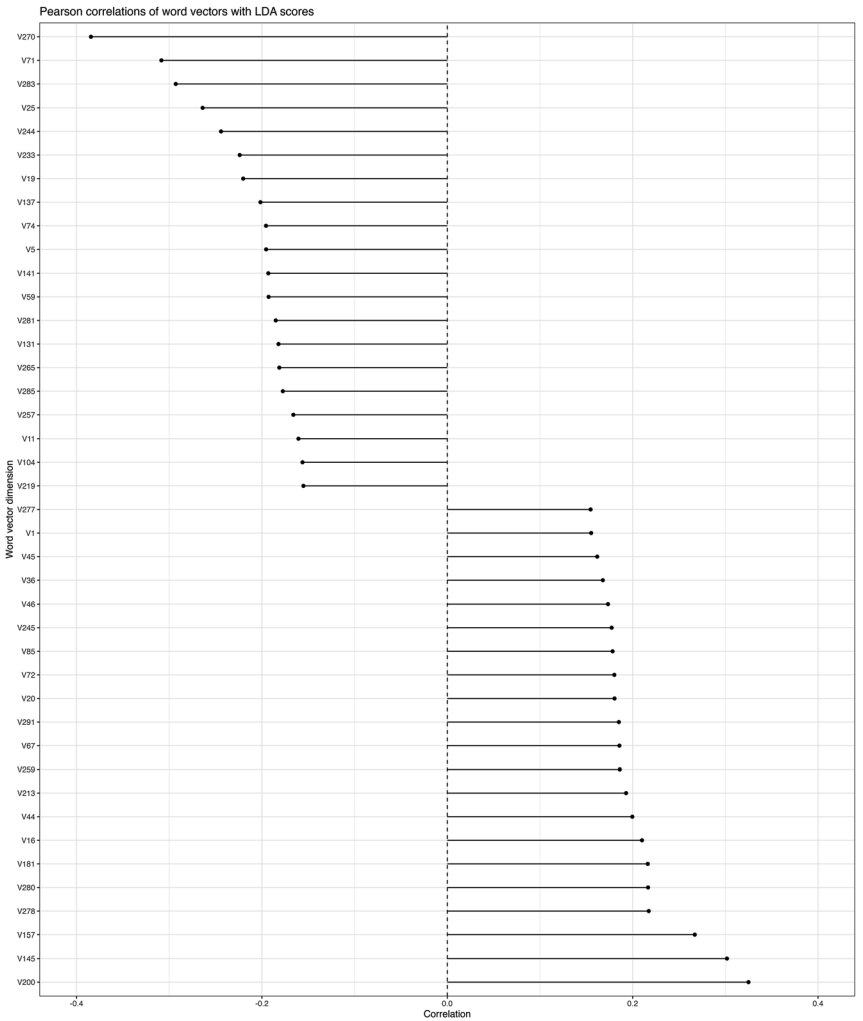

Figure 6: Pearson correlations of word vector dimensions with discriminant scores, only showing coefficients $|r| \geq 0.15$.

which had been supplied during model training.[30] Analogous patterns emerge for the word vectors most strongly associated with object omissibility.

30 Strictly speaking, the correlations concern principal components of the embedding space rather than individual word vectors in their study.

Restricting the scope to the most influential word vectors (V200, V145, V157, V270, V71, V283), their pairwise correlations with the distributional and semantic predictors offer some initial clues (see Figure 6 and Table 6). On the whole, the correlations tend to be rather weak, with a small set of moderate ones. This suggests that the vectors capture aspects of object omissibility not already encoded by the remaining model predictors, though the stronger correlations hint at some overlapping variance (cf. V200 and concreteness; $r = -0.407$). Clearer patterns emerge when the verbs are plotted against their vector coordinates (see Figures 7 and 8).

The word vectors in the panels of Figure 7 positively contribute to the discriminative potential of the LDA and GAM models. This means that higher values of these vectors are associated with clearer class separation and a higher likelihood of object omissibility. Focusing on Panel A, transitive verbs scoring highly on V200 tend to be stative verbs describing psychological processes and perception (*mind*, *know*, *see*, *hear*, *forget*, *imagine*) as well as some communicative acts (*say*, *ask*, *tell*). Strikingly, most of these verbs are alternating verbs, consistent with the cluster formation observed above. Following Table 6, verbs with high V200 scores also tend to be more frequent and more evenly distributed across corpus registers. The negative correlation between V200 and concreteness supports the rather abstract nature of the alternating verbs in Panel A. Conversely, verbs with low V200 scores denote physical acts more likely to be acquired in an embodied fashion, such as *jam*, *cap*, *pad*, *wield*, *drag*, or *rub*. In essence, V200 accounts well for the dense cluster of alternating verbs observed in the *t*-SNE visualisation earlier. Two representative examples are given in (15) and (16).

(15) PERCEPTION_EXPERIENCE

a. *You've **seen** [it]*$_{\text{PHENOMENON}}$ *haven't you* (S1A-073#109:1:B)

b. *You went round there and **saw** Ø*$_{\text{DNI:PHENOMENON}}$ (S1A-073#110:1:B)

Table 6: Pearson correlations between selected embedding dimensions and lexical-semantic variables.

Dim	Log frequency	Dispersion	Concreteness	N_{meanings}	Neigh. density	N_{features}	Selectional strength
V200	0.323	−0.244	−0.407	0.025	0.139	−0.025	−0.354
V145	0.270	−0.193	−0.305	−0.083	0.078	0.043	−0.303
V157	0.216	−0.221	0.223	0.268	0.259	0.082	−0.301
V270	−0.382	0.303	0.214	−0.123	−0.136	−0.061	0.414
V71	−0.312	0.234	0.124	−0.130	−0.160	−0.087	0.310
V283	−0.248	0.201	−0.195	−0.244	−0.179	−0.071	0.406

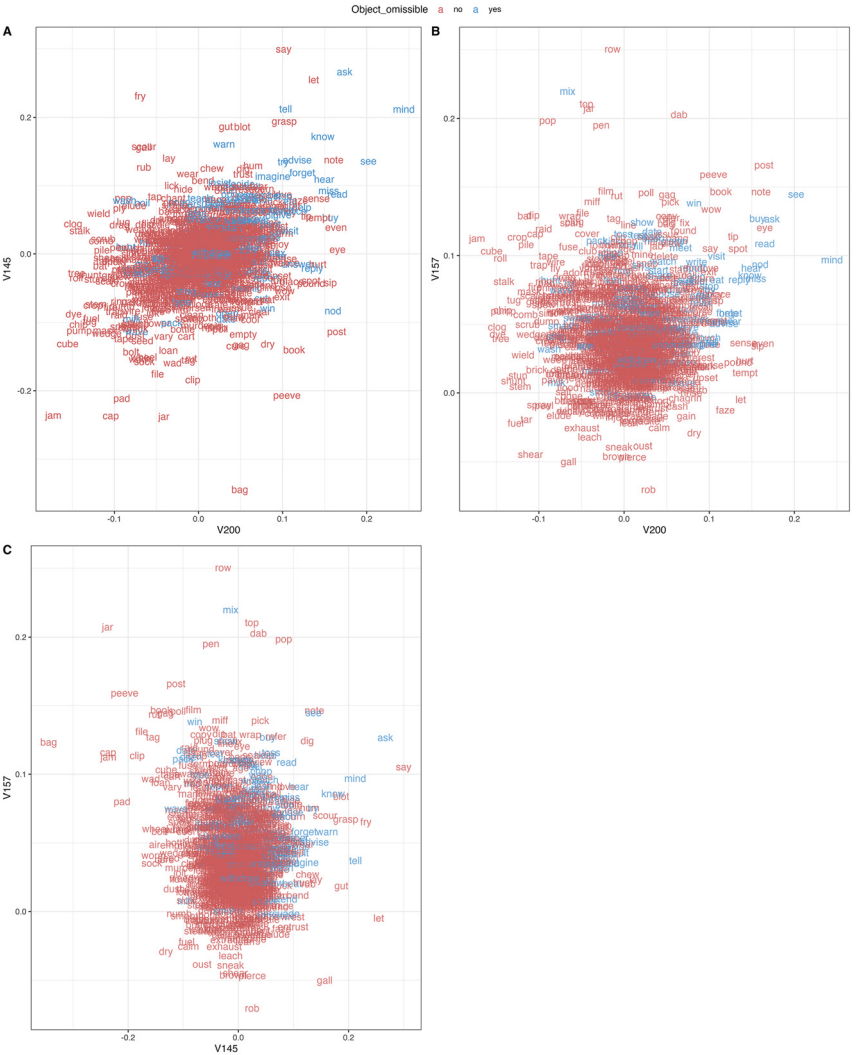

Figure 7: Visualisation of word vectors with strongest positive correlations with discriminant scores.

(16) PERCEPTION_EXPERIENCE

a. *Do you want to **hear** [my joke]*$_{\text{PHENOMENON}}$ (S1A-041#297:1:B)

b. *I didn't **hear** Ø*$_{\text{DNI:PHENOMENON}}$ (S1A-041#400:1:B)

Dimension V145 shares a nearly identical correlational profile with V200. Higher scores are associated chiefly with abstract communication verbs (*tell*, *say*, *let*, *ask*, *warn*, *advise*), though there are some inconsistencies (e.g., *fry*, *chew*), and the lower

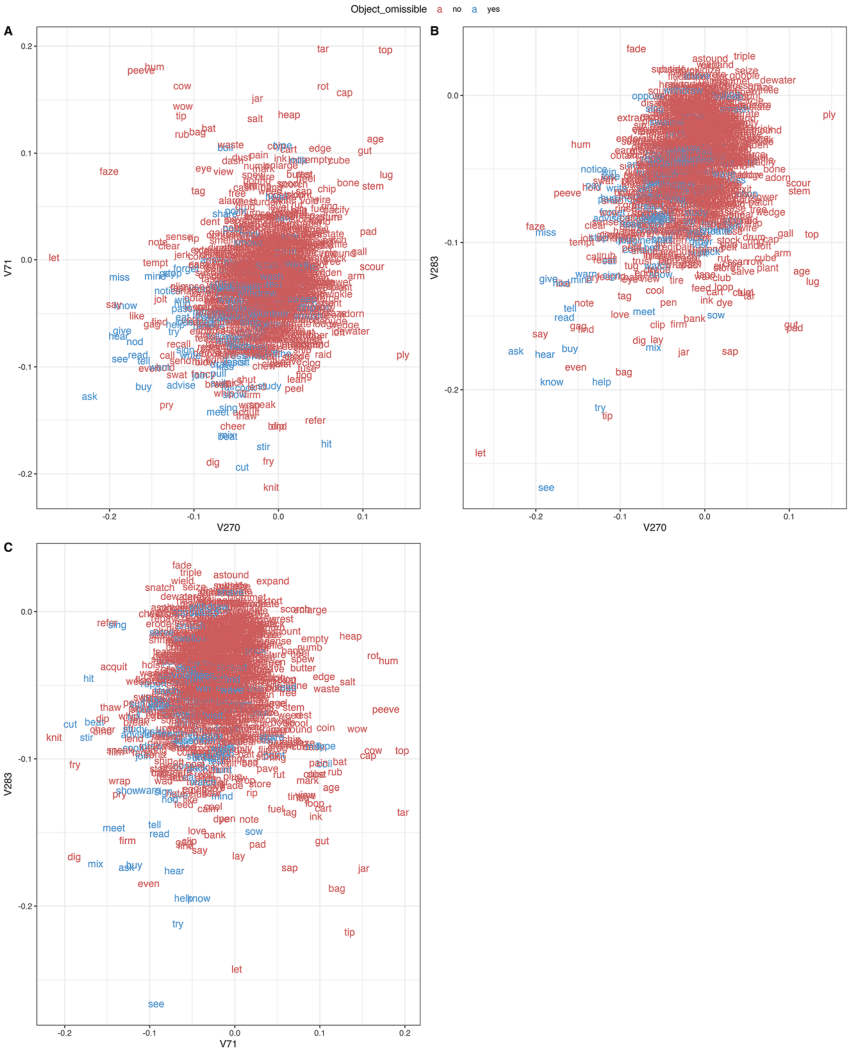

Figure 8: Visualisation of word vectors with strongest negative correlations with discriminant scores.

the scores, the more semantically diffuse the underlying lexical groups become (cf. Panel B). Slightly lower scores move towards general surface and object manipulation (*lay*, *lick*, *tap*, *wield*, *drag*) and interactions with containers (*jam*, *cap*, *jar*, *bag*, *pack*, *milk*), which are more concrete in nature. Verbs with negative loadings on V145 also tend to be conversion-friendly (*jar*, *file*, *book*, *dry*, *empty*) and not particularly

frequent overall. Notably, (17) highlights that depending on the frame involved, verbs may or may not license object omission; for instance, *tell* may realise the MESSAGE in the STATEMENT frame as in (17a), but drop the PHENOMENON if it evokes the BECOMING_AWARE frame. Since word vectors are type-level features, frame polysemy of this kind is collapsed into a single numeric representation, potentially biasing the LDA classification for verbs whose omissibility is strongly frame-dependent.[31]

(17) a. STATEMENT
*I'll **tell** you [a funny story about working class]*$_{\text{MESSAGE}}$ later (S1A-037#11:1:A)
b. BECOMING_AWARE
*How can they **tell** Ø*$_{\text{DNI: PHENOMENON}}$ (S1A-037#181:1:A)

(18) EMPTYING
a. *[…]a single cowman **milks** [up to 100 cows]*$_{\text{SOURCE}}$ *per hour* (W2A-033#15:1)
b. *Labour is needed for **milking** Ø*$_{\text{DNI:SOURCE}}$ *per hour, watering and supervision* (W1A-011#19:1)

Dimension V157 is harder to interpret than V200 or V145 (see Panel C). Higher values pattern similarly to V145, again picking up on substance manipulation, whereas lower ones range from acquisition broadly construed (e.g., *win*, *buy*, *pick*) to force-dynamic, causative senses (*gut*, *murder*, *numb*, *swallow*) to loss or consumption (*smoke*, *exhaust*, *leach*, *shear*, *rob*). Nearly all alternating verbs score positively on V157; in other words, these are events in which some physical or mental content is typically gained rather than lost, albeit with some exceptions (cf. *forget* in (21)). Compared to V200 and V145, V157 has the strongest link to polysemy (r = 0.268).

(19) INGESTION
a. *[…] they tried to convince me that this fish had **swallowed** [the fly]*$_{\text{INGESTIBLES}}$ *before it died* (S1A-055#223:1:B)
b. *Jean **swallowed** Ø*$_{\text{INI:INGESTIBLES}}$ *and nodded* (W2F-015#54:1)

(20) INGEST_SUBSTANCE
a. *They might've **smoked** [a cigarette]*$_{\text{SUBSTANCE}}$ (S1B-004#313:1:A)
b. *And you can't breathe and you can't **smoke** Ø*$_{\text{DNI:SUBSTANCE}}$ *unless you smoke Ø*$_{\text{DNI:SUBSTANCE}}$ *through the nose* (S1B-079#103:1B)

(21) REMEMBERING_TO_DO
a. *And anyway, you **forgot** [to bring the picnic basket]*$_{\text{ACTION}}$*, didn't you?* (W2F-013#137:1)
b. *Maybe he **forgot** Ø*$_{\text{DNI:ACTION}}$ *and that's an excuse* (S1A-093#257:1:B)

31 For a discussion of polysemy in semantic vector space, see Mu et al. (2016).

Turning to the negative correlations in Figure 8, higher values are associated with lower object omissibility. V270 (Panel A) seems to vaguely denote a numerical dimension: verbs such as *top*, *cap*, *age*, and *tip* denote states or processes with continuously scalable properties. Conversely, lower scores often involve subtractive processes (*share*, *miss*, *cut*, *forget*) or those incompatible with a scalar interpretation (*nod*, *ask*, *see*, *tell*). In terms of verb frequency, dispersion, and concreteness, V270 is the exact mirror image of V200.

V71 resembles V270 in capturing verticality and actions performed at a distance (*top*, *cap*, *salt*, *bag*, *eye*, *view*). As scores decrease, so does the implied vertical position of the action (e.g., *water*, *stir*, *hit*). In this respect, *cut* and *dig* aptly extend this logic by denoting movement further down through a surface. These verbs tend to be infrequent and register-specific.

Finally, V283 extends the scalar theme of V270 with addition and subtraction (*fade*, *tripe*, *seize*, *dewater*, *gobble*, *shave*, *withdraw*, *sweep*) (Panel B). Crucially, very few verbs load positively on this dimension, suggesting that lower scores may simply reflect the absence of additive or subtractive semantics.V283 is associated with rare, specialised, abstract, and polysemous verbs with strong selectional preferences. While this profile is not characteristic of most alternating verbs in the sample, it does help account for *withdraw* in particular (see 22).

(22) REMOVING

a. *If he would* ***withdraw*** [*his invading army*]$_{\text{THEME}}$ *from Kuwait and return them to Iraq he would not be attacked* (S2B-030#34:1:A)

b. *But uh I think the authority that the United Nation was given is that he should* ***withdraw*** $\emptyset_{\text{DNI:THEME}}$ *from Kuwait* (S2B-027#74:1:B)

5 Discussion and conclusions

The findings paint a probabilistic profile of object-omitting verbs in which word vectors and distributional measures clearly outrank purely semantic predictors. Most immediately, the results contradict Ruppenhofer's (2004) claim that a verb's ability to license object omission is independent of its frequency of occurrence. The role of dispersion further lends support to the genre-based approach of Ruppenhofer and Michaelis (2010). If word vectors and distributional factors are already controlled for, the effects of purely semantic predictors are negligible. Selectional preference strength (Resnik 1996) has approximately the same explanatory power as frequency or dispersion in the minimally complete Model A, yet they overlap strongly in the deviance they account for. Presumably, these three features may be reflexes of the same latent construct, broadly capturing aspects of usage. No effects

could be confirmed for specificity (Rice 1988), as operationalised via concreteness, number of features, and neighbourhood density, or for polysemy (Ruppenhofer 2004).

The word embeddings likely implicitly encode concreteness, specificity, or polysemy, leaving little unique variance for these variables to contribute independently. This is supported by the fact that the most influential word vectors have low-to-moderate correlations with these features. That semantic factors are nonetheless at play is corroborated by the vector space analysis. There is evidence of natural cluster formation in object omissibility both in the global space (based on 300 dimensions) and with respect to the six most influential vectors. Given the homogeneous, yet ultimately 'basic' nature of the events alternating verbs describe, the routinisation account of Glass (2021) offers a plausible overarching explanation, pending further experimental validation. Alternating verbs tend to cluster towards the stative rather than the force-dynamic pole, with a preference for communicative acts. Nevertheless, the substantive interpretation of the word vector dimensions remains an open question.

The examination of individual corpus instances supports the view that semantic frames represent a key missing factor (Ruppenhofer 2004: 457). A single verb may license object omission under one event schema but not another. Consequently, future work should model object omission patterns as nested within both verbs and frames (see also Buskin 2026: 4–5).

More broadly, the findings suggest that object omissibility, and thus the licensing constraints of null instantiation, cannot be reduced to the purely lexical, discourse-pragmatic, or constructional factors typically invoked in the literature (e.g., Chaves et al. 2025: 6 or Goldberg 2001: 523). Although the joint contributions of these factors remain relevant, the evidence presented here points towards learning effects (cf. Heitmeier et al. 2024: 5) paired with semantic criteria: the rates at which speakers are exposed to these verbs during acquisition (i.e., the learning events), including how evenly these exposure rates are distributed across communicative contexts (i.e., their dispersion), appear to interact with the verbs' locations in semantic vector space. Although the distributional effects strongly support usage-driven explanations of syntactic structure (Bybee 2006; Diessel 2015), there remains the issue of confounding (Baayen et al. 2016). Future work should therefore seek to control such confounds more stringently (e.g., via graphical models).[32]

Prospectively, future quantitative research on argument omission should (i) evaluate the relative importance of semantic frames for predicting argument

32 For example, it is plausible to assume that frequency and dispersion moderate the effects of the lexical semantic predictors. Several exploratory mediation analyses are provided in the supplementary script.

realisation, (ii) assess whether frequency, dispersion, concreteness, specificity, and potentially semantic frames can also account for object realisation (*She lost the game* vs. *She lost* Ø) as well as for (iii) the proportion of realised objects $\hat{\pi}_{\text{obj}}$ across verbs. This would contribute to a more complete understanding of the Understood Object Alternation, encompassing its availability, conditioning factors, and rate of occurrence. Beyond that, it would also be worthwhile to (iv) apply linear discriminative learning to model the production and comprehension of alternating and non-alternating verbs, thereby contributing to the study of lexical processing. Finally, (v) the approach proposed here could also be extended to other lexically constrained syntactic alternations, such as the dative, conative, or causative/inchoative alternation.

Appendix A List of features (examples)

Appendix B Technical details on GAMs

Let $\mathbf{y} = (y_1, \ldots, y_N)^\top$ denote the binary response vector and $\mathbf{X}$ the $N \times p$ design matrix, whose i-th row $\mathbf{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$ contains the covariate values for observation i. The full model equation can be written as

$$\log\left(\frac{P(Y_i = \text{yes} \mid \mathbf{x}_i)}{1 - P(Y_i = \text{yes} \mid \mathbf{x}_i)}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{j=1}^{m} f_j(x_{ij}), \tag{6}$$

where the parametric coefficients $\boldsymbol{\beta}$ are estimated for predictors assumed to enter linearly, and the smooth functions $f_j(\cdot)$ capture non-linear contributions of the m continuous predictors. Each smooth term is represented as a weighted sum of basis functions,

$$f_j(x_{ij}) = \sum_{k=1}^{M_j} \beta_{jk}\, h_{jk}(x_{ij}), \tag{7}$$

where $h_{jk}(\cdot)$ are fixed transformations of the j-th predictor and β_{jk} are estimated coefficients (Hastie et al. 2017: 139–141). Thin-plate regression splines, the default basis in `mgcv`, were used throughout.

To control overfitting, each smooth is subject to a wiggliness penalty weighted by a smoothing parameter $\lambda_j \geq 0$: larger values enforce smoother functions, while smaller values permit greater flexibility. The smoothing parameters λ_j are estimated jointly with the model coefficients via restricted maximum likelihood (REML), which

Table 7: Overview of corpus and psycholinguistic measures for examples (6–22).

Example	Verb	Frequency	Log frequency	Dispersion	Concreteness	$N_{meanings}$	Neighbourhood_density	$N_{features}$	Selectional strength
6	sell	129.00	4.86	0.19	3.35	9.00	96.02	13.00	1.86
7	sign	87.00	4.47	0.27	4.62	20.00	245.84	12.00	–
8	decide	234.00	5.46	0.11	1.92	4.00	198.22	8.00	1.38
9	explain	162.00	5.09	0.19	1.97	3.00	116.09	15.00	1.59
10	remember	371.00	5.92	0.38	2.41	8.00	106.02	20.00	1.17
11	help	281.00	5.64	0.25	2.56	12.00	205.00	19.00	1.20
12	write	456.00	6.12	0.31	4.22	10.00	131.34	39.00	1.11
13	fish	3.00	1.10	3.28	5.00	6.00	84.58	31.00	–
14	cut	180.00	5.19	0.15	4.55	70.00	127.23	26.00	1.88
15	see	2,249.00	7.72	0.10	3.21	25.00	164.72	37.00	0.42
16	hear	500.00	6.21	0.29	3.66	5.00	90.18	35.00	1.24
17	tell	802.00	6.69	0.19	2.90	9.00	140.74	37.00	0.82
18	milk	2.00	0.69	1.48	4.92	7.00	61.36	35.00	–
19	swallow	7.00	1.95	1.33	3.89	11.00	66.24	31.00	–
20	smoke	6.00	1.79	1.56	4.96	10.00	85.53	28.00	6.13
21	forget	218.00	5.38	0.29	2.04	4.00	93.73	10.00	2.00
22	withdraw	80.00	4.38	1.28	3.04	12.00	107.91	–	2.63

tends to yield unbiased estimates of the variance components and is less susceptible to overfitting when predictors are correlated (Wood 2017: 267).

Model fit is summarised by the proportion of explained deviance,

$$\text{Deviance explained} = 1 - \frac{D_{\text{residual}}}{D_{\text{null}}}, \tag{8}$$

where $D = -2 \log \ell$ and D_{null} is the deviance of an intercept-only model.

Relative feature importance is computed using `gam.hp` (Lai et al. 2024) by fitting GAMs across all $2^p - 1$ non-empty subsets of the p predictors and partitioning the explained deviance hierarchically. This isolates each predictor's unique contribution while accounting for concurvity, i.e., the non-linear analogue of multicollinearity, which is particularly relevant here given the strong overlap between frequency, dispersion, and selectional preference strength reported in Section 4.2.

Appendix C Technical details on LDA

In essence, the goal of LDA is to assign an observation $\mathbf{X} = \mathbf{x}$ to a class k. Let $\pi_k = P(Y = k)$ denote the prior probability of class k and $f_k(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} | Y = k)$ the class-specific, multivariate Gaussian density function of the feature vector $\mathbf{x}$. Bayes' theorem yields the posterior probability

$$P(Y = k \mid \mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{x})}. \tag{9}$$

The idea is to find the response class k that maximises this posterior probability:

$$\widehat{k} = \underset{k}{\operatorname{argmax}}\, P(Y = k \mid \mathbf{X} = \mathbf{x}). \tag{10}$$

Proposition C.1. The posterior probability is maximised if and only if the discriminant function $\delta_k(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k$ is maximised as well, yielding the decision rule:

$$\widehat{k} = \arg \max_k \delta_k(\mathbf{x}). \tag{11}$$

Proof. Suppose features follow a multivariate normal distribution of the form $\mathbf{x} \mid Y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_k$ denotes the class-specific mean vector and $\boldsymbol{\Sigma}$ the covariance matrix shared across all K classes. The class-conditional density is

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right). \tag{12}$$

For any two classes k and ℓ, Bayes' theorem yields the log-odds ratio

$$\log \frac{P(Y = k \mid \mathbf{X} = \mathbf{x})}{P(Y = \ell \mid \mathbf{X} = \mathbf{x})} = \log \frac{f_k(\mathbf{x})}{f_\ell(\mathbf{x})} + \log \frac{\pi_k}{\pi_\ell}, \tag{13}$$

since the normalising denominators cancel in the ratio. Substituting the densities, the prefactors $(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}$ cancel because $\boldsymbol{\Sigma}$ is shared across classes, giving

$$\log \frac{f_k(\mathbf{x})}{f_\ell(\mathbf{x})} = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\ell). \tag{14}$$

Expanding each quadratic form and using symmetry of $\boldsymbol{\Sigma}^{-1}$ to combine cross-terms:

$$= -\frac{1}{2}\left(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right) \tag{15}$$

$$+\frac{1}{2}\left(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell\right). \tag{16}$$

The term $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}$ cancels since it does not depend on k or ℓ:

$$= \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \frac{1}{2}\boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell. \tag{17}$$

Adding the log prior ratio and grouping terms by class, one obtains:

$$\begin{aligned} \log \frac{P(Y = k \mid \mathbf{x})}{P(Y = \ell \mid \mathbf{x})} &= \left(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log\ \pi_k\right) \\ &\quad - \left(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell - \frac{1}{2}\boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \log\ \pi_\ell\right) \\ &= \delta_k(\mathbf{x}) - \delta_\ell(\mathbf{x}), \end{aligned} \tag{18}$$

where $\delta_k(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log\ \pi_k$ is the linear discriminant function. Since the log-odds ratio is positive if and only if $\delta_k(\mathbf{x}) > \delta_\ell(\mathbf{x})$, class k has a higher posterior probability than class ℓ if and only if it has a higher discriminant score. As this equivalence holds for every pairwise comparison, the optimal classification rule is

$$\widehat{k} = \arg\ \max_k P(Y = k \mid \mathbf{x}) = \arg\ \max_k\ \delta_k(\mathbf{x}). \tag{19}$$

□

For further details, see Hastie et al. (2017: 106–119).

References

Allerton, David J. 1975. Deletion and proform reduction. *Journal of Linguistics* 11(2). 213–237.

Baayen, Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology* 30(11). 1174–1220.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Nulty Paul, Adam Obeng, Stefan Müller & Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30). 774.

Boas, Hans C. 2020. A roadmap towards determining the universal status of semantic frames. In Renata Enghels, Bart Defrancq & Marlies Jansegers (eds.), *Empirical and methodological challenges*, 21–52. Berlin and Boston: De Gruyter Mouton.

Boas, Hans C., Josef Ruppenhofer & Collin Baker. 2024. FrameNet at 25. *International Journal of Lexicography* 37(3). 263–284.

Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov. 2017. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3). 904–911.

Buchanan, Erin M., K. D. Valentine & Nicholas P. Maxwell. 2019. English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods* 51(4). 1849–1863.

Buskin, Vladimir. 2025. Definite null instantiation in English(es): A usage-based construction grammar approach. *Constructions and Frames* 17(2). 278–312.

Buskin, Vladimir. 2026. Is FrameNet usage-based? Predicting frame element coreness with information theory and gradient boosting. *Language and Cognition* 18. e30.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.

Chanin, David. 2023. Open-source frame semantic parsing. arXiv preprint arXiv:2303.12788.

Chaves, Rui P., Kay Paul & Laura A. Michaelis. 2025. *Unrealized arguments and the grammar of context (Elements in Construction Grammar)*. Cambridge: Cambridge University Press.

Chicco, Davide & Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(6). 1–13. https://doi.org/10.1186/s12864-019-6413-7.

Chicco, Davide & Giuseppe Jurman. 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* 16(4). 1–23.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, Mass. and London: MIT Press.

Diessel, Holger. 2015. Frequency shapes syntactic structure. *Journal of Child Language* 42(2). 278–281.

Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.

Douglas, Benjamin D., Patrick J. Ewell & Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One* 18(3). 1–17.

Fillmore, Charles J. 1986. Pragmatically controlled zero anaphora. *Proceedings of the twelfth annual meeting of the Berkeley linguistics society*, 12, 95–107. Berkeley: Berekely Linguistics Society.

Fillmore, Charles J. 2007. Valency issues in FrameNet. In Thomas Herbst & Karin Götz-Votteler (eds.), *Valency: Theoretical, descriptive and cognitive issues*, 129–160. Berlin and New York: De Gruyter.

Fillmore, Charles J. & Paul Kay. 1995. *Construction grammar*. Stanford: CSLI Publications.

Gao, Chuanji, Svetlana V. Shinkareva & Rutvik H. Desai. 2022. SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research Methods* 55(6). 2853–2884.

Gavruseva, Elena. 2024. Let's drop the object: An experimental study of the implicit object construction in American English. *CogniTextes* 25. https://doi.org/10.4000/12rd2.

Glass, Lelia. 2021. English verbs can omit their objects when they describe routines. *English Language and Linguistics* 26(1). 49–73.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago and London: University of Chicago Press.

Goldberg, Adele E. 2001. Patient arguments of causative verbs can be omitted: The role of information structure in argument distribution. *Language Sciences* 23(4). 503–524.

Goldberg, Adele E. 2005. Argument realization: The role of constructions, lexical semantics and discourse factors. In Jan-Ola Östman & Mirjam Fried (eds.), *Construction grammars: Cognitive grounding and theoretical extensions*, 17–43. Amsterdam & Philadelphia: John Benjamins.

Graf, Ricarda, Marina Zeldovich & Sarah Friedrich. 2024. Comparing linear discriminant analysis and supervised learning algorithms for binary classification – A method comparison study. *Biometrical Journal* 66(1). 1–20.

Greenbaum, Sidney. 1996. *Comparing English worldwide: The international corpus of English*. Oxford: Clarendon Press.

Greenbaum, Sidney & Gerald Nelson. 1996. The international corpus of English (ICE) project. *World Englishes* 15(1). 3–15.

Gries, Stefan Thomas. 2020. Analyzing dispersion. In Magali Paquot & Stefan Thomas Gries (eds.), *A practical handbook of corpus linguistics*, 99–118. Cham: Springer.

Hamilton, William L., Leskovec Jure & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In Katrin Erk & Noah A. Smith (eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics*, 1489–1501. Berlin, Germany: Association for Computational Linguistics.

Hastie, Trevor & Robert Tibshirani. 1991. *Generalized additive models*. London: Chapman & Hall.

Hastie, Trevor, Robert Tibshirani & Jerome H. Friedman. 2017. *The elements of statistical learning: Data mining, inference, and prediction*, 2nd edn. New York, NY: Springer.

Heitmeier, Maria, Yu-Ying Chuang, Seth D. Axen & R. Harald Baayen. 2024. Frequency effects in linear discriminative learning. *Frontiers in Human Neuroscience* 17(1242720). 1–22.

Herbst, Thomas, David Heath, Ian F. Roe & Dieter Götz. 2004. *A valency dictionary of English: A corpus-based analysis of the complementation patterns of English verbs, nouns and adjectives*. Berlin: Mouton de Gruyter.

Heumann, Christian, Michael Schomaker & Shalabh. 2022. *Introduction to statistics and data analysis: With exercises, solutions and applications in R*, 2nd edn. Cham: Springer.

Hoffmann, Thomas. 2011. *Preposition placement in English: A usage-based approach*. Cambridge: Cambridge University Press.

Hoffmann, Thomas. 2022. *Construction grammar: The structure of English*. Cambridge: Cambridge University Press.

Hollis, Geoff & Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin and Review* 23(6). 1744–1756.

Honnibal, Matthew and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–299.

James, Gareth, Daniela Witten, Trevor Hastie & Robert Tibshirani. 2021. *An introduction to statistical learning: With applications in R*. New York: Springer.

Josse, Julie & François Husson. 2016. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software* 70(1). 1–31.

Kabbach, Alexandre, Corentin Ribeyre & Aurélie Herbelot. 2018. Butterfly effects in frame semantic parsing: Impact of data processing on model ranking. In *Proceedings of the 27th international conference on computational linguistics*, 3158–3169. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Krijthe, Jesse H. 2015. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.17. https://github.com/jkrijthe/Rtsne.

Lai, Jiangshan, Jing Tang, Tingyuan Li, Aiying Zhang & Lingfeng Mao. 2024. Evaluating the relative importance of predictors in generalized additive models using the gam.hp R package. *Plant Diversity* 46(4). 542–546.

Lambrecht, Knud & Kevin Lemoine. 2005. Definite null objects in (spoken) French: A construction-grammar account. In Mirjam Fried & Hans C. Boas (eds.), *Grammatical constructions: Back to the roots*, 13–55. Amsterdam: Benjamins.

Lemmens, Maarten. 2006. More on objectless transitives and ergativization patterns in English. *Constructions*. https://doi.org/10.24338/cons-447.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20(1). 1–31.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago & London: The University of Chicago Press.

Lyngfelt, Ben. 2012. Re-thinking FNI: On null instantiation and control in construction grammar. *Constructions and Frames* 4(1). 1–23.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA & London: MIT Press.

Markard, André & Jan-Christoph Klie. 2020. *FrameNet Tools: A Python library to work with FrameNet*. Version v0.1.0. https://doi.org/10.5281/zenodo.3993970.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch & Armand Joulin. 2018. Advances in pre-training distributed word representations. *Proceedings of the international conference on language resources and evaluation*, 52–55. Miyazaki, Japan.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.

Miller, David L. 2025. Bayesian views of generalized additive modelling. *Methods in Ecology and Evolution* 16(3). 446–455.

Mouselimis, Lampros. 2024. fastText: Efficient learning of word representations and sentence classification using R. R package version 1.0.4. https://CRAN.R-project.org/package=fastText.

Mu, Jiaqi, Suma Bhat and Pramod Viswanath. 2016. Geometry of polysemy. arXiv preprint arXiv: 1610.07569.

Olsen, M. B. & Philip Resnik. 1997. Implicit object constructions and the (in)transitivity continuum. *33rd proceedings of the Chicago linguistic society*, 327–336. Chicago.

Peer, Eyal, David Rothschild, Adam Gordon, Zachary Evernden & Ethan Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54. 1643–1662.

Perek, Florent. 2015. *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam: Benjamins.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1). 149–188.

Radford, Andrew. 2005. *Minimalist syntax: Exploring the structure of English*. Cambridge: Cambridge University Press.

Rappaport Hovav, Malka. 2008. Lexicalized meaning and the internal structure of events. In Susan Rothstein (ed.), *Theoretical and crosslinguistic approaches to the semantics of aspect*, 13–42. Amsterdam: John Benjamins.

Rappaport Hovav, Malka & Beth Levin. 1998. Building verb meanings. In Miriam Butt & Wilhelm Geuder (eds.), *The projection of arguments: Lexical and compositional factors*, 97–134. Stanford, CA: CSLI Publications.

Reijnierse, W. Gudrun, Christian Burgers, Marianna Bolognesi & Tina Krennmayr. 2019. How polysemy affects concreteness ratings: The case of metaphor. *Cognitive Science* 43(8). e12779.

Reilly, Megan & Rutvik H. Desai. 2017. Effects of semantic neighborhood density in abstract and concrete words. *Cognition* 169. 46–53.

Resnik, Philip. 1993. *Selection and information: A class-based approach to lexical relationships*. University of Pennsylvania PhD thesis, Philadelphia, PA.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61(1). 127–159.

Rice, Sally. 1988. Unlikely lexical entries. *Annual Meeting of the Berkeley Linguistics Society* 14. 202–212.

Ringgaard, Michael, Rahul Gupta and Fernando C. N. Pereira. 2017. Sling: A framework for frame semantic parsing. arXiv preprint arXiv:1710.07032.

RStudio Team. 2020. *RStudio: Integrated development environment for R. RStudio*. Boston, MA: PBC. http://www.rstudio.com/.

Ruppenhofer, Josef. 2004. *The interaction of valence and information structure*. Berkeley: eScholarship: University of California.

Ruppenhofer, Josef & Laura A. Michaelis. 2010. A constructional account of genre-based argument omissions. *Constructions and Frames* 2(2). 158–184.

Shafaei-Bajestan, Elnaz, Peter Uhrig & R. Harald Baayen. 2022. Making sense of spoken plurals. *The Mental Lexicon* 17(3). 337–367.

Swayamdipta, Swabha, Sam Thomson, Chris Dyer and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold.

Turney, Peter & Pantel Patrick. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.

van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(86). 2579–2605.

Van Rossum, Guido & Fred L. Drake. 2009. *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

Villani, Caterina, Adele Loia & Marianna M. Bolognesi. 2024. The semantic content of concrete, abstract, specific, and generic concepts. *Language and Cognition* 16(4). 867–894.

Wood, Simon N. 2013. On p-values for smooth components of an extended generalized additive model. *Biometrika* 100(1). 221–228.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R.*, 2 nd. Boca Raton: Chapman & Hall/CRC.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Oxford, England: Addison-Wesley Press.