콘텐츠

# Polysemy in Semantics with Contextualized Language Models: Distribution, Boundaries and Interpretation of Polysemous Senses

# Contents

# Chapter 1

# Introduction

## 1.1. Verbal Polysemy in Lexical and Constructional Semantics

Polysemy is a type of lexical ambiguity where a single linguistic form has two or more related senses (Taylor, 2003: 103). Over the past few decades, it has been the focus of numerous studies across multiple disciplines, including linguistics, psychology, neuroscience, and computational linguistics (See Haber and Poesio (2024) for a recent review).

Polysemy is ubiquitous in everyday language, and particularly common with verbs. Verb polysemy plays a central role in theoretical and empirical work on lexical and constructional meaning as it tightly correlates with many properties of verb behavior and uses, e.g., participation in argument structure alternations. The interaction of verbal polysemy with the argument structure alternation can best be illustrated with the verb *break*, a canonical instance of a change-of-state verb that participates in the causative alternation. This alternation is characterized by verbs with causative and noncausative uses, such that the causative use means roughly 'cause to V-intransitive'. Following Levin (2015), we will call a sentence with the transitive form of a causative alternation verb the causative variant, and a sentence with the intransitive form the noncausative variant:

(1) a. Perry broke the fence. (causative)
b. The fence broke. (noncausative)

English *break* is one of the best studied verbs in lexical semantics. It takes on a wide array of senses systematically related to its argument structure. The lemma *break* participates in 59 synsets in WordNet (Fellbaum 1998), a lexical resource that is most widely used as the golden standard for polysemy in Natural Language Processing (NLP) and corpus linguistics. Table 1 provides a preliminary grouping of fine-grained polysemous senses of *break*, with illustrating examples of each sense category, which will be further refined and updated in Section 2.2.

Table 1. Major sense categories of *break*

| Sense categories | Examples |
|---|---|
| Physical breaking | The glass broke from the pressure. |
| Bodily harm | My ankle broke in a skiing accident. |
| Emotional/psychological breakdown | Her heart broke when she heard the news. |
| Violation | Picasso broke his contract with Manach. |
| Decoding | The police finally broke the code. |
| Disclosure | The reporter broke the news early in the morning. |
| Termination | They broke the cycle of violence. |
| Interruption | I broke the tense atmosphere with a joke. |
| Breakthrough | She broke the world record. |
| Change | The medicine helped break the fever. |
| Emergence | The day broke over the quiet village. |

These sense distinctions interact with the causative alternation. Whereas prototypical, physical breaking senses alternate, as shown in (1), some senses have been assumed to be strictly transitive or intransitive (Levin and Rappaport Hovav, 1995; Piñón, 2001; Alexiadou et al., 2006; Schäfer, 2008; Romain, 2022). For example, the violation, breakthrough, and decoding senses have been assumed to occur only in the causative variant, whereas the natural emergence sense is restricted to the noncausative variant. However, as observed by Petersen and Potts (2023) and Lee (2025), there are attested cases of causative uses of the *violation*, *breakthrough*, and *decoding* senses, as illustrated below. These examples indicate that the syntactic distribution of the senses of *break* across the two alternation variants is not a categorical phenomenon. Rather, they suggest that verb senses are deeply intertwined with argument structure realization, reflecting graded links between meaning and syntax.[1]

(2) When <the contract broke> and CSM came out, Pochita was still the heart core inside Denji's body. (Lee 2025: 308)

(3) <The Guinness World Record broke>, our furniture didn't.
(Petersen and Potts, 2023: 492)

(4) Almost sixty years later, Frank Rowlett, a cryptologic pioneer and head of the "Purple" team, remembered that historic day when <the code broke>.
(Petersen and Potts, 2023: 492)

The non-categorical, variable nature of the association between polysemous verb senses and the alternation variants poses a challenge to traditional lexical-semantic accounts of argument alternation, which typically explain properties of verb behavior and uses at the level of verb classes rather than at the level of individual sense variation (e.g., Alexiadou et al. 2006; Levin and Rappaport Hovav 1995; Reinhart 2002, 2016; Rappaport and Hovav 2014, 2020, among others).

Construction-based approaches to argument structure have focused primarily on the semantics of constructions themselves rather than on verbs' potential to alternate. Goldberg's (2002) surface generalization hypothesis epitomizes this view, emphasizing that generalizations are more productively sought across different verbs within the same construction than across different constructions of the same verb. Although constructionist approaches prioritize constructional meaning, they also acknowledge the verb's contribution to meaning composition. As Goldberg (1995: 24) notes, grammar is not entirely top-down; constructions do not simply impose their meanings on verbs. Instead, the central schematic meaning of each construction is linked to the prototypical verb that most frequently occurs with it, while other meanings emerge as peripheral extensions.

Subsequent corpus-based work refined this interaction. Studies such as Stefanowitsch and Gries (2003), Yi (2016), and Yi et al. (2019) demonstrate that verbs cluster with particular constructions according to semantic similarity to a representative or anchor verb. Yet, other researchers (Croft 2003; Gilquin 2013; Perek 2015; Bernolet and Colleman 2016) have shown

[1] In complex sentences which consist of more than two clauses such as (2), we will mark clauses where the alternating verb is used by enclosing them in single arrow brackets ('< >').

that constructional meaning is often distributed across verb classes or even verb senses, not simply radiating from a single prototype. Overall, finer-grained analyses at the verb-class and verb-sense level are now viewed as offering a more realistic picture of speakers' knowledge of verb meaning and behavior. However, neither verb-centered nor construction-centered approaches have provided a fully articulated theoretical model that can account for the graded and continuous links between verb senses and syntactic realization.

Table 1's partial list of *break* senses highlights three aspects of polysemous verb meaning that pose a further challenge to existing semantic theories of polysemy. First, the polysemous senses of *break* are semantically diverse, often stretching across near-opposite meanings. For example, the emerge sense in *The day broke over the quiet village* contrasts sharply with the alleviation *or* weakening sense in *The medicine helped break the fever*. Second, it is difficult to determine with confidence how many distinct and non-overlapping senses *break* actually expresses. For instance, does *break* convey the same violation sense in (5) as in *Picasso broke his contract with Manach* in Table 1, or does it instead carry a breakthrough meaning similar to *She broke the world record*? Third, multiple senses may be simultaneously present in actual usage. For example, *Picasso broke his contract with Manach* in (6) expresses both violation and termination meanings at once.

(5) … "They find me, all right," says <Sam Jethroe, who broke the franchise color line with the Boston Braves in 1950>. "A couple came …"
(COCA News: Atlanta-19970622)

(6) <Picasso broke his contract with Manach> and returned in January, 1902, to Barcelona.
(COCA MAG: USA Today Magazine-1997)

This kind of semantic multiplicity raises a fundamental question: how do such overlapping or co-activated senses arise in context? Meeting this challenge requires a theoretical move beyond purely type-level representations of meaning toward token-level, usage-based models that capture how meanings are dynamically constructed in actual discourse. Despite decades of research in semantics and other disciplines, the phenomenon of verbal polysemy continues to pose unresolved challenges. Traditional frameworks, while offering rich insights into argument realization and sense differentiation, often struggle to explain how multiple related senses emerge and interact within natural usage contexts.

The purpose of this book is to explore how semantic approaches leveraging large neural language models (LLMs)—which provide semantically rich, contextualized representations—open a path toward a context-sensitive, usage-based account of lexical meaning and offer new explanations for long-standing, unresolved challenges in explaining verbal polysemy. The following section discusses these challenges in greater detail, motivating the adoption of contextualized language models as analytical tools for studying polysemy.

## 1.2. Unresolved Challenges in Explaining Verbal Polysemy

### 1.2.1. Polysemy and Argument Structure Realization

Previous studies have attempted to explain the obligatorily transitive nature of some *break* senses by focusing on thematic role requirements. This approach suggests that certain frames

require an agentive subject, which must be overtly expressed, thereby preventing noncausative uses (Levin and Rappaport Hovav 1995; Piñón 2001; Alexiadou et al. 2006; Schäfer 2008). However, this explanation is insufficient because it fails to account for obligatorily causative uses where the subject is not an agent and yet the verb still lacks an intransitive alternative (e.g., *the airbags broke the impact* vs. **the impact broke*). In this sentence, *broke the impact* means to lessen or soften the impact of a crash. Similar usages of *break* referring to a mitigating action that reduces the severity of a sudden, forceful event are discussed by Petersen and Potts (2023: 492): *the cushion broke her fall* vs. **the fall broke*.

There are also strictly noncausative uses like usages referring to natural emergence in (7) and breaking (ceasing to function) from normal wear and tear in (8):

(7) a. The day broke over the quiet village.
b. *The sun/the Earth's rotation broke the day over the quiet village.

(8) a. <My watch broke> after the warranty ran out.
b. *<I broke my watch> after the warranty ran out (does not have the same interpretation as (a)).

(Rappaport Hovav, 2014: 25)

Many accounts of the causative alternation including Levin and Rappaport Hovav (1995) take the causative variant to be basic, deriving the noncausative variant from it via a lexical operation of decausativization. This operation applies to verbs that simply specify their subject is a causer, subsuming agents, natural forces, and instruments. A challenge for such accounts is that verbs like *break* that fit the characterization of the class of alternating verbs have uses that lack a causative counterpart, as in (7) and (8).

Rappaport Hovav and Levin (2012) and Rappaport Hovav (2014, 2020) propose a non-derivational approach to the alternation that diverges from their own earlier analysis, as well as non-derivational analyses embodied in Alexiadou et al. (2006) and others. In the realm of change-of-state verbs, Rappaport Hovav (2014) makes a three-way classification, summarized in Table 2.

Table 2. Rappaport Hovav's Three-way Classification of Change-of-state Verbs

| | |
|---|---|
| Verbs that lexically select an external cause argument (non-alternating in English) | Verbs that specify something about the nature of the involvement of an external cause, such as *murder* and *assassinate* |
| | Verbs that specify nothing about the nature of the causing event, for example, *kill* and *destroy* |
| Verbs lexically associated with an internal argument only (alternating in English) | Verbs that specify the nature of the change of state but nothing specific about the cause of the change of state, such as *break*, *open* and *clear* |

On this non-derivational account, verbs that allow alternation in English are lexically associated solely with their patient. Rappaport Hovav and Levin (2012) and Rappaport Hovav (2014) provide systematic evidence supporting this view. The addition of the cause is non-lexical, driven by the well-known constraint that the cause must be construable as a direct cause

– a cause with immediate control over the eventuality. This constraint contributes to the determination of the semantic type of NPs allowed as the external argument in given contexts. It does not, however, explain why alternating verbs differ with respect to how often they are used as a causative and as a noncausative and which variant is appropriate in a given discourse context. For this, we need a theory that accounts for the (non-)appearance of the causer argument.

Rappaport Hovav (2014) provides a pragmatic account of the (non-)appearance of the causer argument. She begins with the following assumption.

(9) In the description of a change of state, the cause of the change of state is relevant; therefore, since an utterance which specifies the cause of the change of state is more informative than one which expresses just the change of state, it is to be preferred, all things being equal (Rappaport Hovav 2014: 23).

Drawing on Grice's (1975) maxims of conversation, she argues that if the expression of the cause is deemed relevant, it is typically preferred since the causative variant is more informative. The degree of informativeness of the causative variant is influenced by factors such as contextual identifiability and predictability of the causer. For instance, if the causer is recoverable in some way from context, then the sentence with the cause expressed is no longer more informative than the corresponding sentence that expresses just the change of state. For example, the causer may have been mentioned previously in the discourse as in (10). In that case, mentioning the cause seems superfluous and the noncausative may be preferred from economy considerations like Grice's (1975) Maxim of Manner, which dictates avoiding prolixity.

(10) I pushed and pushed on the door, and <it finally opened>.
(McCawley 1978, cited in Rappaport Hovav, 2014: 22)

Sometimes the speaker will leave the agent unmentioned though the causative variant would be more informative. This happens, according to Rappaport Hovav (2014), when the speaker does not know the cause; consequently, the noncausative is required because the speaker cannot truthfully specify the cause. Rappaport Hovav (2014, 2020) shows that this account covers constraints on the alternation which have been taken to be lexical and also those which are more clearly non-lexical in earlier accounts.

Rappaport Hovav (2020) modifies (9) because it is not the case that the causer of a change of state is always relevant. Some states are such that they have a propensity to change in the natural course of events, and this affects the relevance of the cause of the change of state. Rappaport Hovav (2020) states how natural events of change influence the relevance of the cause of the change of state as follows.

(11) For a given state and a given entity there is a default expectation of whether the state (or the degree to which the state holds) will or will not change in the natural course of events, i.e., whether the entity has the *disposition* to undergo a change in state. The cause of a change of state is relevant only if for the given state and the given entity, there is no default expectation of change. (Rappaport Hovav 2020: 242)

Many natural events of change have expected causes recoverable by default. For example, the sky clears, ice melts, and a tree grows from a variety of causal factors which are typically co-occurring, common and predictable. As Rappaport Hovav (2020: 244-245) points out, these causal factors are not good candidates as actual causes (the particular cause deemed to be "the" cause of the change of state). In that case, mentioning the cause is highly costly because the speaker cannot easily identify "the" cause of the change of state that is appropriate as the subject of lexical causatives. For this reason, she is more likely to leave the cause unmentioned to save effort and time on less useful information that is not clearly identifiable as the actual cause. This helps explain why the natural causes exemplified in (7) above occur in the noncausative variant and are predominantly unexpressed in the clause of the change-of-state verb. The obligatorily noncausative nature of usages of *break* illustrated in (8) can be similarly explained in terms of the notion of default cause. When a watch breaks (in the sense of ceasing to function) from normal wear and tear, the normal wear and tear is the default cause, but it is difficult to single out one primary factor.

Drawing on Rappaport Hovav's (2020) account of (non-)appearance of the cause argument, we can further explain why the sentence *the airbags broke the impact in a crash* lacks a natural noncausative counterpart like **the impact broke*. In this case, the entity, the impact, does not have the disposition to change in the natural course of events. Since there is no default expectation of change for an impact, the cause of its mitigation—in this case, the airbags—must be explicitly mentioned to explain the change of state.

The inclusion of contextual constraints in the analysis of the causative alternation has recently gained ground in a series of recent corpus studies by Lee (2023, 2025) and her colleagues. Lee (2023) presents a corpus study that tested the empirical validity of Rappaport Hovav'spragmatic approach on the basis of a systematic analysis of 2,400 instances of causative and noncausative uses of 12 alternating verbs extracted from the automatically parsed British National Corpus (BNC). She revisits the notion of causer identifiability, defining it as the extent to which the event described by the verb and its argument has a clear identifiable causer. Statistical analyses of the corpus data indicate that the causative and noncausative uses of the verbs differ in causer identifiability: verbs that are predominantly used as a causative tend to have a causer with a higher degree of identifiability, while verbs that are more frequently used as a noncausative tend to have a causer with a lower degree of identifiability.

This finding has been corroborated by results of Kim et al.'s (2025) study which tested effects of semantic and contextual factors influencing the degree of informativeness of the causative variant against 3,864 instances of causative and noncausative uses of 135 alternating verbs extracted from the automatically parsed BNC. They provide supporting evidence that causer intentionality and identifiability are significantly associated with the realizations of the investigated verbs and that causer identifiability is the most important predictor for discriminating between the two variants. Extending Rappaport Hovav's (2014, 2020) and Lee's (2023) accounts of contextual constraints on the causative alternation, they interpret their findings as a consequence of general principles of efficient language use (Levshina 2022). In a follow-up corpus study, Lee (2025) analyzed 1,002 instances of *break* and *freeze* extracted from the Corpus of Contemporary American English (COCA; Davis 2008-), focusing on how different types of causers are distributed across the causative alternation variants. The analysis revealed that, among the senses of *break* listed in Table 1, eight senses are either restricted to or strongly preferred in the causative variant (causative-dominant senses), while two are either

restricted to or preferred in the noncausative variant (noncausative-dominant senses). The most significant finding is that causative-dominant senses tend to correlate with contextual properties of the causer that enhance the informativeness of the causative variant, such as high intentionality and identifiability. In contrast, noncausative-dominant senses are closely associated with low intentionality, high predictability, or causer types for which explicit mention is pragmatically costly—all of which reduce the informativeness of the causative variant.

However, the traditional methods relying on labeled data involve manually annotating linguistic features across multiple levels. Currently, there are no sufficiently reliable automated annotation tools available, which means these approaches are inherently subjective and require intensive human labor. Additionally, they face limitations in scalability, generalizability, and in achieving robust triangulation with larger datasets.

With recent advances in deep learning, contextualized language models, which generate context-sensitive embeddings, have become widely available. Embedding refers to the process of converting discrete linguistic units (e.g., words, sentences) into quantitative representations in a high-dimensional, continuous space. This allows text data to be expressed as numerical vectors that LLMs can understand and process.

In contrast to earlier models like Word2Vec (Mikolov et al. 2013), which yield static word representations, modern Transformer-based models—such as GPT (Generative Pre-trained Transformer) (Radford et al. 2018) and BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al. 2019)—produce dynamic embeddings that adjust to context, enabling more precise modeling of polysemous expressions. While unidirectional models like GPT are highly effective in text generation, BERT's bidirectional architecture, which incorporates both preceding and following context, makes it particularly well-suited for capturing subtle distinctions in word meaning—especially relevant for this study.

Despite the impressive performance of these state-of-the-art models across many NLP tasks, a core limitation remains: they function largely as black boxes and do not explicitly reveal linguistic structures or knowledge. To address this, Chapter 2 integrates traditional corpus-based analysis with objective, data-driven methods drawn from BERT and large-scale corpora, aiming to supplement, validate, and extend insights from manually labeled data.

### 1.2.2. Semantic Multiplicity and Boundaries in Polysemy

Polysemous words can assume different interpretations in different contexts; what distinguishes them from homonyms like *match* in (12) is that their interpretations are closely related, and often invoke different aspects of or perspectives on the same concept. Take, for example, the different uses of *school* in (13), each of which intuitively reflects a different facet of the concept *school* rather than entirely unrelated meaning like homonymic alternation in (12).

(12) a. The match fell on the carpet and left a burn mark.
b. The match ended without a winner even after going into overtime.
(Haber and Poesio 2024: 352)

(13) a. The school (building) has a dull brown facede.
b. The school (administration) has prohibited light-up sneakers.

c. The school (sports team) won last year's play-offs.
d. The school (institution) is well respected among researchers.

(Haber and Poesio 2024: 353)

Evidence from psycholinguistic studies supports the distinction between homonymy and polysemy, indicating that polysemes are processed very differently than homonyms (Frazier and Payner 1990; Rodd et al. 2002; Klepousniotou et al. 2008, 2012; Haber and Poesio 2021; Haber 2022). A growing body of recent work also has started to challenge the uniform treatment of polysemous senses postulated by traditional theories such as the Generative Lexicon (Pustejovsky 1995; Asher and Pustejovsky 2006; Asher 2011), and put forward proposals of a more structured mental representation of nominal polysemy (Ortega-Andés and Vincente 2019; Vincente 2019; Ortega-Andés 2021). Using copredication tests, several studies (e.g., Ortega-Andés 2019; Haber and Poesio 2021) have shown that there is a crucial distinction between inherent polysemy and other kinds of conventional polysemy. Copredication refers to the phenomenon in which a single polysemous nominal expression simultaneously participates in multiple predications that correspond to distinct meanings or senses of the word within the same sentence. Representative examples of copredicative sentences are given in (14).

(14) a. The school (building) caught fire and (people) was celebrating 4$^{th}$ of July when the fire started.
b. The books (physical object) are thick and (content) interesting.

(Ortega-Andés 2021: 112)

In (14a), *school* denotes both the building and the people associated with it. In (14b), *books* refers simultaneously to the physical object and the content it conveys. Words of this type—often classified as inherently polysemous—typically allow copredication, since their distinct senses are conceptually integrated and jointly accessible.

In contrast, some conventionally polysemous words, whose senses are only logically related rather than inherently linked, do not permit copredication. For instance, in (15) the noun *door* can denote either the physical object or the aperture, but not both simultaneously.

(15) #They took the door (physical object) off its hinges and walked through it (aperture).

(Haber and Poesio 2021: 2664)

The unacceptability of (15) suggests that the *door* senses are distinct and not jointly activatable within a single predicational frame. This contrast illustrates an important semantic boundary: inherent polysemy involves closely integrated facets of a single conceptual representation, whereas other kinds of conventional polysemy reflect looser, contextually licensed sense relations. Ortega-Andés and Vincente (2019) proposes that the senses involved in copredication form robust activation packages, which allow hearers and readers to access the different senses of a polysemous word in interpretation.

Verbs that denote complex events often display an even more intricate form of semantic multiplicity. As discussed in Section 1.2.1, multiple senses of a verb such as *break* may be simultaneously activated in actual usage, and these overlapping senses can also be causally related. For instance, in (5), repeated here in (16), the violation and breakthrough senses are

causally intertwined, while in (6), repeated below as (17), the violation and termination senses overlap in a causal sequence.

(16) … “They find me, all right,” says <Sam Jethroe, who broke the franchise color line with the Boston Braves in 1950>. “A couple came …”
(COCA News: Atlanta-19970622)

(17) <Picasso broke his contract with Manach> and returned in January, 1902, to Barcelona.
(COCA MAG: USA Today Magazine-1997)

In some instances, multiple senses are simultaneously activated within a single subevent, rather than across sequential phases. These cases present a more complex interaction than those involving clearly distinct subevents, as semantic multiplicity arises within a single event boundary. Consider example (18):

(18) A supermajority of 60 Senators can break a filibuster by invoking a cloture, the cessation on the bill, and forcing a vote.
(http://www.whitehouse.gov/our-government/legislative-branch)

Here, the verb *break* in *break a filibuster* exhibits semantic overlap within both the causing act ($e_1$) and the resulting state ($e_2$):

- In $e_1$, *break* encodes the agentive act of disrupting a resistance mechanism, functioning simultaneously as an interruption and a breach of resistance. The action represents not just a stoppage, but an assertive, rule-based termination of a procedural obstruction.
- In $e_2$, the focus shifts to the resulting state, where the filibuster is no longer in effect and the legislative process resumes. This phase activates senses akin to release**,** termination, and even institutional restoration, as the system regains its procedural flow.

Thus, the verb simultaneously evokes destructive and restorative connotations, demonstrating how overlapping senses can co-occur and interact within a single complex event structure. Such cases highlight that semantic multiplicity in verbal polysemy extends beyond mere alternation to involve fine-grained, dynamically interwoven meanings that challenge categorical sense distinctions.

These intricate overlaps call for models capable of representing meaning construction at the level of contextual instantiation rather than through discrete sense enumeration. From a theoretical standpoint, such phenomena challenge lexical-semantic models that assume clearly delineated sense boundaries. They show that verbal meaning is not a fixed inventory of distinct entries but a dynamic configuration of contextually co-activated components, often linked through causal, aspectual, or perspectival relations within a single event frame.

At the same time, these patterns pose serious challenges for computational semantics and annotation practice. Conventional semantic annotation schemes—such as WordNet- or FrameNet-based tagging—typically rely on the *one-sense-per-token* assumption, which presupposes that a single lexical item in context can be assigned one predominant sense. However, examples like (16)–(18) demonstrate that a single token of *break* may simultaneously

realize multiple overlapping senses, thereby blurring categorical distinctions and lowering inter-annotator agreement.

Furthermore, in distributional and embedding-based approaches, this multiplicity complicates the learning of contextual representations: when different senses are co-activated within the same token, vector spaces tend to collapse or average these semantic dimensions, masking the fine-grained relations that give rise to true contextual meaning.

Addressing these challenges requires an integrated, usage-based approach that bridges theoretical and computational perspectives. Such an approach should model how meanings are *constructed*, *activated*, and *differentiated* in context—capturing graded, token-level variation rather than discrete, type-level categories. In this sense, contextualized language models, which encode semantically rich, context-dependent representations, offer a promising framework for unifying insights from lexical semantics and computational modeling.

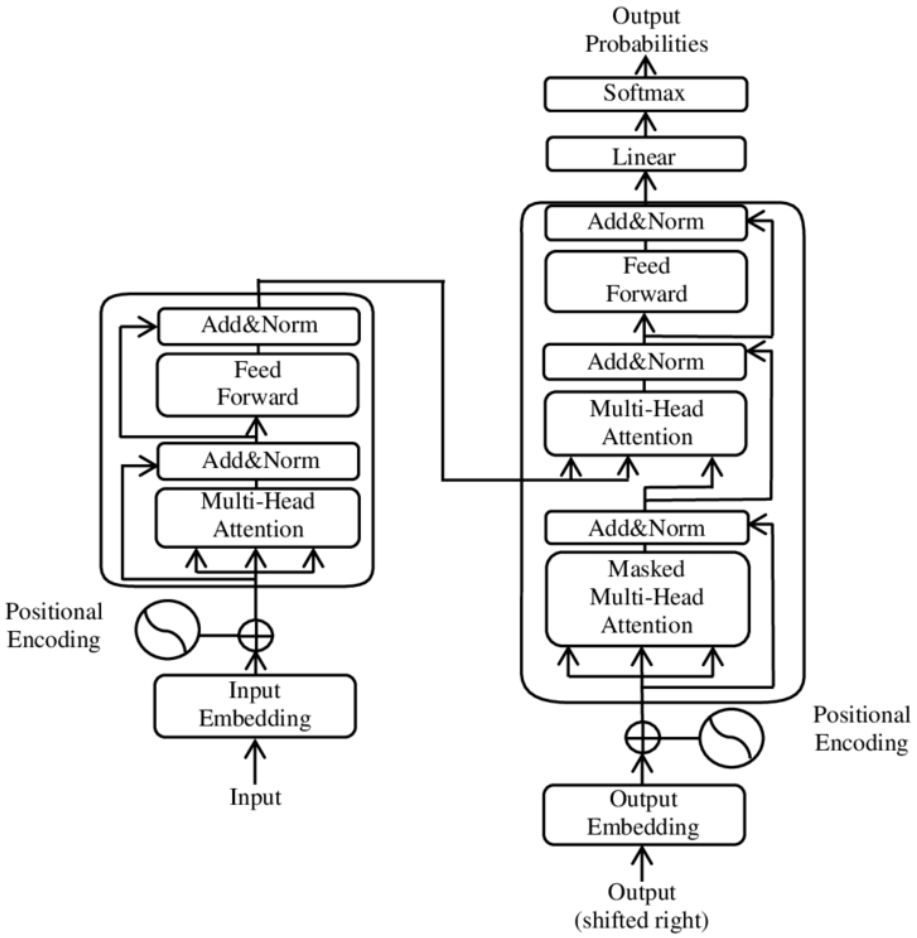
The next section turns to these models, outlining how they can serve as analytical tools for investigating polysemy and contextual meaning construction.

### 1.3. Using Contextualized Language Models to Help Study Polysemy

In this book, we focus on contextualized language models (CLMs) that are built upon the Transformer architecture (Vaswani et al. 2017). The Transformer is a neural network model designed to capture the *contextual relationships* among words in a sequence through a mechanism known as self-attention. Unlike earlier recurrent or convolution-based models, the self-attention mechanism enables the model to compute pairwise relations among all tokens simultaneously, allowing it to integrate both local and global contextual information in a highly efficient way.

Offering a revised architecture that makes it possible to process massive amounts of unsupervised training data in combination with a previously unthinkable number of model parameters, Transformer-based models have dramatically advanced natural language processing. They exhibit a particularly strong capability to model long-range dependencies that are crucial for downstream tasks requiring a deeper, context-sensitive ‘understanding’ of input text. This ability to represent nuanced contextual relations provides the foundation for analyzing polysemous expressions in ways unattainable by earlier distributional or type-based models (Devlin et al. 2019; Haber and Poesio 2021; Petersen and Potts 2023; Lee 2025; Song and Wang 2025). A schematic visualization of the Transformer architecture is presented in Figure 1, illustrating the flow of attention across tokens and the multi-layered structure that enables rich contextual encoding.

Figure 1. Schematic diagram of the Transformer architecture. Figure by Yuening Jia: DOI:10.1088/1742-6596/1314/1/012186, CC BY-SA 3.




At the input embedding stage, the Transformer processes textual data together with positional encodings, which allow the model to recognize the sequential order of words—something not inherently captured by self-attention alone. The encoded input is then passed through a series of multi-head attention mechanisms, each computing the relational weights between a token and all other tokens in the sequence. This operation enables the model to construct a rich representation of contextual dependencies, capturing how each word's meaning shifts in relation to others within the same sentence.

Subsequent add-and-normalization layers stabilize and standardize the attention outputs, while feed-forward neural networks apply non-linear transformations to refine and propagate the encoded information. This entire process—attention, normalization, and feed-forward transformation—is repeated across multiple stacked layers, progressively deepening the contextual abstraction of the input.

At the final stage, a softmax function computes output probabilities, predicting the most likely next word or token based on the learned contextual patterns. In the case of text generation models such as GPT, masked multi-head attention is applied to ensure that predictions for a given position depend only on preceding tokens, thereby preserving the sequential flow of information during language generation.

One of the most influential Transformer-based language models is BERT—Bidirectional Encoder Representations from Transformers (Devlin et al. 2019). Architecturally, BERT

consists of a stack of Transformer encoder modules, each containing multiple self-attention heads. Every layer integrates its outputs through skip connections, layer normalization, and a feed-forward intermediate network, which combines and re-weights information before passing it to the next layer. The BERT BASE model comprises 12 layers with 12 attention heads each, producing hidden states of 768 dimensions and totaling about 110 million parameters. BERT LARGE doubles the depth to 24 layers with 16 attention heads and 1,024-dimensional hidden states, amounting to roughly 340 million parameters.

Although technically *bidirectional*, BERT functions as a non-sequential model: it processes entire text sequences simultaneously rather than word by word. To enable such non-directional pre-training, Devlin et al. introduced two novel objectives—Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)—in place of traditional next-word prediction. In the MLM task (a variant of the *Cloze* test [Taylor 1953]), 15 percent of tokens are replaced with [MASK] symbols; the model must predict these tokens by jointly considering both left and right contexts. NSP, by contrast, trains the model to determine whether one sentence naturally follows another, a task crucial for discourse-level understanding such as question answering.

Following the transfer-learning paradigm popularized by ULM-FiT (Dai and Le 2015; Howard and Ruder 2018; Radford et al. 2018), a pre-trained BERT model can be fine-tuned for specific downstream tasks with comparatively little data—for example, by supplying question–answer or premise–hypothesis pairs. On the GLUE benchmark (Wang et al. 2018), both BERT BASE and BERT LARGE achieved substantial gains, improving average accuracy by 4.5 and 7.0 percentage points over previous state-of-the-art systems.

BERT's main competitor in developing rich, contextualized word representations has been the GPT series introduced by OpenAI (Radford et al. 2018, 2019; Brown et al. 2020). More recently, the family has expanded to include ChatGPT, GPT-4 (OpenAI 2023), and the latest GPT-5, released in 2025. While also grounded in the Transformer framework, GPT (Generative Pre-trained Transformer) models adopt a distinct approach to text processing: they employ a unidirectional, auto-regressive architecture composed of stacked Transformer decoder blocks. Unlike BERT's masked-language modeling objective, GPT models predict each next token sequentially from left to right, allowing them to generate coherent and contextually appropriate text. GPT-2, for instance, contains between 12 and 48 decoder layers with hidden-state dimensions ranging from 768 to 1,600, totaling 1.5 billion parameters—an order of magnitude larger than BERT. Its successor, GPT-3 (Brown et al. 2020), dramatically scaled this design to 175 billion parameters, ushering in a new era of large-scale language modeling. GPT-4 and GPT-5 further refine this paradigm through multimodal and reasoning-enhanced architectures, integrating text, image, and code understanding while maintaining instruction-tuned conversational capabilities characteristic of the ChatGPT interface.

Other Transformer-based architectures have pursued complementary directions. Google AI's T5 (Text-to-Text Transfer Transformer) (Raffel et al. 2020) reformulates all language tasks within a unified text-to-text framework, while the Reformer model (Kitaev, Kaiser and Levskaya 2020) introduces locality-sensitive hashing to efficiently handle context windows extending up to one million tokens. XLNet (Yang et al. 2019) bridges the gap between BERT and GPT by adopting a generalized auto-regressive training method that captures bidirectional dependencies while preserving sequential modeling advantages.

Among the BERT-derived variants, RoBERTa (Robustly Optimized BERT Pre-training Approach; Zhuang et al. 2021) improves training stability through larger mini-batches and

longer training, whereas ALBERT (A Lite BERT; Lan et al. 2019) reduces parameter redundancy via factorized embeddings and weight sharing. BART (Lewis et al. 2020b) combines denoising autoencoding with sequence-to-sequence pre-training, and MARGE (Lewis et al. 2020a) extends this idea to multilingual retrieval and generation.

Finally, Google's Gemini series (2023–2024) represents another major milestone. Built upon a multimodal Transformer architecture, Gemini integrates language, vision, and reasoning capabilities within a unified training framework. The model family (Gemini 1 through 1.5, and Gemini 2 in 2025) demonstrates competitive or superior performance to GPT-4 across diverse benchmarks, emphasizing cross-modal understanding and grounded reasoning—capabilities central to the study of meaning construction and contextual interpretation (Google DeepMind 2023, 2024).

Although contextualized language models have achieved remarkable success across a wide spectrum of downstream NLP tasks, they have also introduced a major epistemological challenge. Owing to their black-box architecture, relatively little is known about *how* such models attain their impressive performance. This opacity not only raises issues of model accountability and explainability, but also, as Rogers et al. (2020) point out, "limits hypothesis-driven improvement of the architecture." The search for greater transparency has given rise to an entire subfield of NLP devoted to probing and interpreting large neural language models (Ethayarajh 2019; Hewitt and Manning 2019; Hewitt and Liang 2019; Lin et al. 2019; Tenney et al. 2019; Petersen and Potts 2023, among others). Among these, two lines of research are especially relevant to the phenomena explored in this book—namely, verbal polysemy and argument-structure realization.

Hewitt and Manning (2019) introduced the structural probe, a diagnostic tool designed to analyze how neural language models internally represent grammatical relations among words. The probe measures how accurately a model such as BERT encodes syntactic distance and hierarchical dependency between tokens. Their findings reveal that syntactic information is most salient in BERT's middle layers, indicating that the model implicitly learns grammatical structure even without explicit supervision. The same probing framework also allows for comparative analyses across architectures and layers, illuminating how different models encode and preserve syntactic information at varying depths.

Building on this line of work, Petersen and Potts (2023) extend the probing paradigm from syntax to lexical-semantic phenomena, focusing particularly on how contextualized Transformer encoders represent verb polysemy and argument-structure alternations. Their study is especially relevant to the present work, as it directly targets the verb *break*, a canonical example of the causative alternation and one of the most extensively analyzed polysemous verbs in both theoretical and computational linguistics. Using a series of semantic and syntactic probes, Petersen and Potts trained lightweight diagnostic classifiers on top of frozen Transformer embeddings to examine whether the models encode distinctions corresponding to *break*'s multiple senses—such as physical destruction, emergence, violation, or termination—and whether these sense representations correlate with argument-structure realization (i.e., causative vs. noncausative use). Their results show that Transformer-based encoder models, including BERT and RoBERTa, exhibit high discriminability among verb senses, suggesting that fine-grained polysemous distinctions are reflected in the distributional geometry of contextual embeddings.

At the same time, their syntactic-probing results demonstrate that the same representations also encode information about argument alternation patterns, capturing whether a verb occurs

in a transitive (causative) or intransitive (noncausative) frame. Together, these findings imply that semantic multiplicity and syntactic realization are jointly encoded within the internal space of contextualized models—an insight that bridges structural and semantic probing traditions and provides an empirical foundation for studying how polysemous verbs are represented and differentiated across contexts.

Building on these insights, the analyses presented in Chapter 3 extend the probing methodology to a broader and more fine-grained investigation of verbal polysemy and argument realization. Specifically, this study combines probing experiments with evaluations of fine-tuned Transformer models, comparing their classification performance and analyzing misclassification patterns across verb senses. Through this dual approach, the chapter aims to uncover the factors that sharpen or blur the boundaries between verb senses, revealing how contextualized representations capture overlapping or transitional meanings. In doing so, it contributes to a deeper understanding of how semantic differentiation emerges within continuous embedding spaces, and how such gradience interacts with syntactic realization in naturally occurring language use.

### 1.4. Structure of the Book

The following chapters outline how this book applies contextualized language models to address the long-standing, unresolved challenges in explaining verbal polysemy discussed in Section 1.2.

Chapter 2 investigates the semantic organization of causative alternation verbs such as *break* and *freeze* by integrating annotated corpus variables with BERT-based distributional analyses. Through principal component and regional distribution analyses, it identifies two major semantic dimensions that systematically structure the causative–noncausative alternation. Building on this foundation, the chapter presents an in-depth case study of the two polysemous change-of-state verbs *break* and *freeze*, which exhibit contrasting constructional preferences. The results demonstrate that sense-level semantic distinctions play a central role in shaping verb-specific constructional patterns in the causative alternation.

Chapter 3 extends this inquiry with probing and fine-tuning experiments on Transformer encoder models. Structural and semantic probes are used to test how internal representations encode verb-sense distinctions and argument-structure frames, while fine-tuned classifiers are evaluated for their predictive accuracy. An analysis of misclassified tokens provides further insight into the factors that sharpen or blur the boundaries between verb senses, highlighting the graded and context-dependent nature of semantic categorization.

Chapter 4 examines how both human participants and generative AI models perceive and evaluate sense boundaries in polysemous verbs, focusing on *break* and *freeze*. Through three complementary experiments—human meaning selection, sense applicability judgment, and usage similarity judgment—the chapter systematically investigates how categorical versus graded sense distinctions emerge and how they are shaped by semantic proximity and overlap. The results of these experiments showed a clear convergent pattern across all tasks: semantic connectivity—manifested as causal, extension-based, or subcategorical relations—consistently predicted gradient sense boundaries, whereas context-dependent or implicature-based relations tended to sustain discrete distinctions. Both human and AI judgments reflected this continuum, differing primarily in degree of variability rather than in underlying representational structure.

Chapter 5 integrates the human and AI results from Chapters 3 and 4 to advance a unified account of the mental and computational representation of polysemous verb meaning, the Generative Activation Package Model (GAPM). The model conceptualizes verb meaning as a network of potential semantic activations dynamically instantiated in context, thereby bridging discrete sense enumeration and continuous semantic representation.

In sum, this study contributes to a deeper understanding of how semantic differentiation emerges within continuous embedding spaces and how such gradience interacts with syntactic realization in natural language use. The final chapter elaborates on the theoretical and methodological implications of the preceding analyses, situating the proposed model within current developments in lexical semantics, cognitive modeling, and large-language-model research. It also outlines future directions for integrating corpus-based, experimental, and computational approaches to the study of meaning.

# Chapter 2

# Verb Uses, Sense Distribution, and the Causative Alternation: Insights from a BERT-based Distributional Semantic Analysis

In this chapter, we discuss the analysis of usage variation and polysemy patterns of change-of-state (COS) verbs participating in the causative alternation, using BERT as a distributional semantic model. The primary aim of this analysis is to uncover the core semantic dimensions underlying these verbs and to explain the divergent preferences in constructional realization shown by polysemous alternating verbs.

The study addresses two central research questions:

1. How does the integration of BERT's contextualized embeddings with subjectively labeled data employed in traditional corpus studies capture the core semantic dimensions and usage variation of alternating COS verbs?
2. How does a BERT-based distributional semantic analysis of polysemy illuminate the fundamental factors driving the constructional preferences of these verbs?

The chapter is organized as follows. Section 2.1 presents a BERT-assisted analysis of the causative and noncausative uses of verbs that show a strong tendency toward the causative alternation. The initial objective here is to validate previous findings from theoretical and corpus-based research regarding the semantic and contextual factors that differentiate causative from noncausative uses. Beyond this, the analysis investigates whether the principal dimensions of BERT-derived semantic spaces can meaningfully capture semantic and contextual variation in verb usage. Section 2.2 presents a BERT-based case study of two polysemous COS verbs, *break* and *freeze*, which exhibit contrasting constructional preferences. By examining the syntactic and semantic distribution of their polysemous senses, this section aims to identify the underlying factors responsible for their divergent preferences in constructional realization. The observed distributional patterns are further explained in terms of pragmatic principles of argument structure realization and language use.

## 2.1. Analyzing Variations in Verb Usage

### 2.1.1. Annotated Dataset

The basis for our investigation is an annotated dataset created by Kim et al. (2025). This dataset consists of 3,864 instances of causative and noncausative uses of 135 alternating COS verbs, extracted from the automatically parsed BNC. For the present analysis, we focus on a subset of the data, namely 2,893 instances of 79 verbs that were identified as *strong alternators* in Lee (2025).

The term strong alternators refers to lexical items that not only exhibit high collostructural strength—the statistical measure of association between lexical items and the two constructions constituting the alternation—but also occur frequently and relatively evenly across the two constructions (Lee 2025: 120). To identify such verbs, Lee (2025) applied the collostructional analysis methodology (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004; Gries

2024), a technique widely used in corpus linguistics to measure the strength of association between lexical items and constructions, to 279 COS verbs participating in the causative alternation.[1] This procedure yielded a set of 137 verbs classified as strong alternators. In this section, we analyze the distribution of instances of 79 strong alternators appearing in Kim et al.'s (2025) BNC dataset by the two causer-type variables listed in Table 1.[2]

Table 1. Annotated causer-type variables

| Variables | Levels | | Examples |
|---|---|---|---|
| Causer Intentionality | Intentional (Intent): Agentive causer | Agent | (1) |
| | | Intentional causing action (Int-act) | (2) |
| | Nonintentional (NIntent): Nonagentive causer (cause) | Animate causer (Cause_anim) | (3a) |
| | | Inanimate causer (Cause_inan) | (3b) |
| | | Causer that can be either animate or inanimate (Cause_anim/inan) | (3c) |
| Causer Identifiability | Specified causer (Spec) | | (4a), (4b) |
| | Recoverable causer (RC) | Previously mentioned causer (RC_mentioned) | (2) |
| | | Hinted causer (RC_hinted) | (5) |
| | | Default causer (RC_default) | (6) |
| | | Other RCs (RC_other): type-inferable causer, omitted subjects of causative uses of finite COS verbs, and subjects of imperative clauses | (7) |
| | Unknown causer (UC) | | (8) |

In what follows, we adopt Kim et al.'s (2025) annotation schemes for the two variables that are based on Rappaport Hovav (2014, 2020), Heidinger and Huyghe (2024), and Lee (2023, 2025). The variable *intentionality* annotates subtypes of causers distinguished by the intention to bring about the event. We use the cover term *causer* for the entity that brings about the change of state in a causative event, and the term *cause* to denote unintentional causers, following the terminology used, for instance, in Heidinger and Huyghe (2024) and in VerbNet (Kipper-Schuler 2005) for semantic role labeling. Following Heidinger and Huyghe (2024), *cause* is further classified into three subtypes according to animacy. The resulting definitions of each subtype are given in (1)–(3), along with illustrative examples.

[1] See Romain (2017) for discussions of other measures of the alternation strength of causative verbs.

[2] The table with the absolute frequencies of causative and noncausative instances for each of the investigated verbs is given in Appendix A. The total number of the causative instances included in the analysis was 1,234, and the total number of the noncausatives was 1,659. The relative frequencies of the instances of the two variants were 42.7 and 57.3 percent, respectively.

(1) Agent: Animate causer that intentionally brings about the event
The protesters broke the window.

(2) Intentional causing action: Causing action that is carried out intentionally
I pushed and pushed on the door, and <it finally opened>.
(McCawley 1978, cited in Rappaport Hovav 2014: 22, (62))

(3) Cause: Inanimate causer or unintentional animate causer that brings about the event denoted by the verb
a. John broke the window when he was playing football (animate causer).
(Levshina 2022: 168)
b. The storm/the rocks broke the window. (inanimate causer)
c. You bought a plastic toy at Christmas from Japan, and <it broke the next day>. (either animate or inanimate) (COCA NEWS: Denver Post-20000227)

In the variable *identifiability*, distinctions are made between specified, recoverable, and unknown causers, following Kim et al. (2025). A specified causer (Spec) is defined as a causer that is realized in the same tensed clause where the change is expressed, and hence is clearly identified as the ultimate cause of the change without requiring reference to the surrounding context. Two examples of a specified causer are given in (4). In (4a), the causer (*Bao*) is realized as the subject of the tensed clause headed by the finite verb *succeeded*, which takes a non-finite clause containing the COS verb *breaking* as its complement. In (4b), the prepositional phrase containing the cause modifies the verb phrase headed by *broke*.

(4) a. I realized that <Bao had succeeded in breaking the endless chain of thought I'd been chasing>, … (COCA FIC: Naamah's blessing-2011)
b. The window broke from the pressure/from the explosion/from Will's banging.
(Heidinger and Huyghe 2024: 191, (23a))

The restriction to causes realized in the same tensed clause where the change is expressed excludes causers mentioned or hinted in the surrounding context from cases of specified causer. Instead, such causers were annotated as recoverable causes, that is, contextually identifiable causers. The following types of causer were considered recoverable causers: (i) previously mentioned causer (RC_mentioned), (ii) hinted causer (RC_hinted), (ii) default causer (RC_default), and (iv) type-inferable causer and others (RC_other). A previously mentioned causer is recoverable because it has been established in the surrounding context, that is, in the preceding or following clause or sentence, as in (2) above.

A cause can also be recoverable when the causer may be hinted at in the surrounding context. An example would be (5). Although the cause of the sudden bead of sweat is not explicitly stated in this sentence, the preceding context implies that it is due to the narrator's fear of the machine gun and the potential for violence.

(5) He probed my pannier with his machine gun.…<A bead of sweat broke from my forehead> and trickled down the side of my face.
(COCA MAG:Bicycling-1995 (Sep))

A different case of recoverability is what Rappaport Hovav (2014, 2020) refers to as default cause, namely causes that are part of what discourse participants know about the way the world works: they are recoverable by default. Examples of a default cause are causes of changes which happen in the normal course of events illustrated in (6).

(6) a. My watch broke after the warranty ran out. (Most likely indicates cessation of functioning due to normal wear and tear) (Rappaport Hovav 2014: 18, (45a))
b. The days lengthened. (Most naturally understood as resulting from the regular astronomical cycle of the seasons)
(Levin and Rappaport Hovav 1995: 105, (58b))

Another type of causer that is recoverable from world knowledge about properties of events is what Kim et al. (2025) call a type-inferable causer. Like a default cause, a type-inferable causer combines characteristics of both recoverable causers and unknown causers, and can thus be seen as lying at the boundary between these two causer types. An example with a type-inferable causer is given in (7). The *when*-clause in this example describes the closure of a hospital effected through the involvement of an agent who can be inferred to be typically responsible for such an event, namely, the hospital's owner. The precise identity of the agent is unimportant and need not be specified. Type-inferable causers were categorized as other types of recoverable causers (RC_other), along with imperative subjects and omitted subjects of finite verbs.

(7) Careful plans have been made for these people so that when <the hospital eventually closes> they will not find themselves on the streets.
(BNC W:non-ac:polit_law_edu, HH3-9982)

Finally, an unknown causer (UC) is defined as a causer that is unidentifiable in context. An example with an unknown causer is given in (8). In this example, the direct cause of the measuring equipment's breaking is not mentioned or clearly identifiable in the context.

(8) When the 12 Japanese transmissions were tested, an engineer reported that <the measuring equipment had broken>. (COCA MAG: Smithsonian-1990)

Deciding whether a causer for a change is recoverable or unknown requires a careful contextual analysis. When further context was needed, we consulted the original text by querying the online version of the BNC hosted at Lancaster University.[3]

[3] We used the CQPweb interface of the BNC to access preceding and subsequent context. CQPweb (Hardie 2012) is an online corpus analysis tool that serves as an interface to the Corpus Workbench software (CWB) and its effective Corpus Query Processor (CQP) search utility (https://cqpweb.lancs.ac.uk).

### 2.1.2. Integrating BERT with a Distributional Semantics Framework

The annotated variables discussed in Section 2.1.1 were processed through distributional semantic analysis. Distributional semantics is based on the Distributional Hypothesis, the assumption that "similarity in meaning results in similarity of linguistic distribution" (Firth 1957; Lenci 2008, 2018), and aims to approximate word sense by inferring the relationships between words from large amounts of corpus data. This usually is done by abstracting words and their contexts to vectors in semantic space and measuring the similarity between vectors of given target expressions.

Models of polysemy have previously been proposed in distributional semantics (see for example Boleda et al. 2013), but for the most part, such models found limited impact and applications. This changed with the emergence of a new generation of contextualized language models.

One of the most influential contextualized language models is BERT (Devlin et al. 2019), a Transformer-based model (Vaswani et al. 2017) jointly trained on masked language modeling and next-sentence prediction. Fundamentally, the BERT architecture is a stack of Transformer encoder modules consisting of multiple so-called self-attention heads. Each layer of self-attention heads is wrapped with a skip connection, and followed by layer normalization and a fully connected intermediate layer to com bine and weigh outputs, turning them into the next layer's inputs. In the BASE model, BERT is made up of 12 layers each consisting of 12 self-attention heads, and creates hidden states of 768 dimensions (for a total of 110 million parameters). BERT Large contains 24 layers, each with 16 attention heads and hidden state representations of size 1,024, boasting a total of 340 million parameters.

Devlin et al. (2019) and Haber and Poesio (2021, 2024) highlight BERT's ability to disambiguate the meanings of polysemous words in context. As demonstrated in the experimental study by Haber and Poesio (2021), BERT can distinguish between polysemy and homonymy and exhibits semantic processing patterns similar to human judgments of meaning similarity and acceptability. However, even the best-performing model in their study, BERT Large, while handling certain types of nominal polysemy in a human-like manner, failed to consistently reproduce human-like similarity patterns for other types of polysemy. This study not only sheds light on the complexity of lexical meaning and the limitations of neural language models, but also suggests that BERT is likely to face even greater difficulty in consistently interpreting and distinguishing the more complex polysemy of verbs compared to that of nouns. Lee (2025) proposes an extended distributional semantics model that integrates verb meaning, syntactic structure, and argument information with neural embeddings, to better capture fine-grained and context-sensitive semantic phenomena in BERT (see also Petersen and Potts 2023, and Song and Wang 2025). This extended model takes as input to BERT sentences annotated with labels such as verb sense and syntactic realization, thereby generating BERT embeddings that jointly encode meaning and its syntactic realization. Such embeddings enable the analysis of how syntactic realization patterns of individual verb senses and distributional differences among senses are reflected in BERT's representational space, thereby enhancing a language model's capacity to replicate linguistic phenomena, as well as its analytical accuracy and interpretability. In the present study, we apply Lee's (2025) BERT-based neuro-symbolic integrated distributional semantics model to the analysis of usage variation of alternating COS verbs and the polysemy of *break* and *freeze,* with the goal of explaining the divergent

preferences in constructional realization shown by (polysemous) alternating verbs. The BERT-based analysis procedure is as follows:

1. Data preprocessing: Read the Excel file and extract sentences, along with sense, realization and causer-type labels, to be used as BERT input.
2. Embedding extraction: Feed each extracted sentence into BERT and obtain the embedding (768 dimensions) for the position where the target word (COS verb token) occurs.
3. Dimensionality reduction and visualization: Apply PCA or t-SNE to reduce the embedding dimensions to two or three, and visualize the resulting data.
4. Analysis and interpretation: Examine distributions, semantic cluster distances, and pathways of sense extensions.

Our analysis is implemented in Google Colaboratory and we used the pretrained BERT Base provided by the HuggingFace transformers library.[4] Data and code to reproduce our results are available at the following GitHub repository: https://github.com/hanjung-25/clmsemantics.[5] For the purpose of analyzing the semantic distribution of COS verbs, we use verb embeddings extracted from layer 12 of the BERT Base model, following evidence that lexical content is well-preserved in higher layers. As Yu and Ettinger (2020) observe, although compositional information tends to peak in middle layers, the final layer still retains strong word-level properties.

### 2.1.3. Results and Discussion

This section presents the results of the analysis of causative and noncausative embeddings of the investigated COS verbs in the BERT semantic space. The goal of the BERT-assisted analysis of verb uses is to demonstrate how BERT embeddings validate and further supplement the findings of the corpus-based studies of variables distinguishing between causative and noncausative uses.

We first report the proportional distributions of causative and noncausative uses with respect to the two causer-related variables, intentionality and identifiability. Figures 1 and 2 present the distribution of intentionality levels and identifiability levels across the two variants, respectively. As shown in Figure 1, causative uses are predominantly realized with intentional causers (888 out of 1,314 cases, ≈68%), whereas noncausative uses are strongly biased toward nonintentional causers (1,234 out of 1,659 cases, ≈74%). This distribution highlights that intentionality functions as a robust semantic discriminator between the two variants, aligning with prior corpus-based observations. Figure 2 further demonstrates that causative uses overwhelmingly involve specified causers (1,202 out of 1,234 cases, ≈97%), while noncausative uses are largely associated with recoverable (1,219 out of 1,657 cases, ≈74%) or unknown causers (293 cases, ≈18%). This sharp contrast underscores the role of identifiability in shaping constructional preferences.

---

[4] https://huggingface.co/models?library=transformers&sort=trending

[5] Code: https://github.com/hanjung-25/clmsemantics/tree/code
Dataset: https://github.com/hanjung-25/clmsemantics/tree/data

Figure 1. Proportional distribution of intentionality levels in the causative and noncausative variants

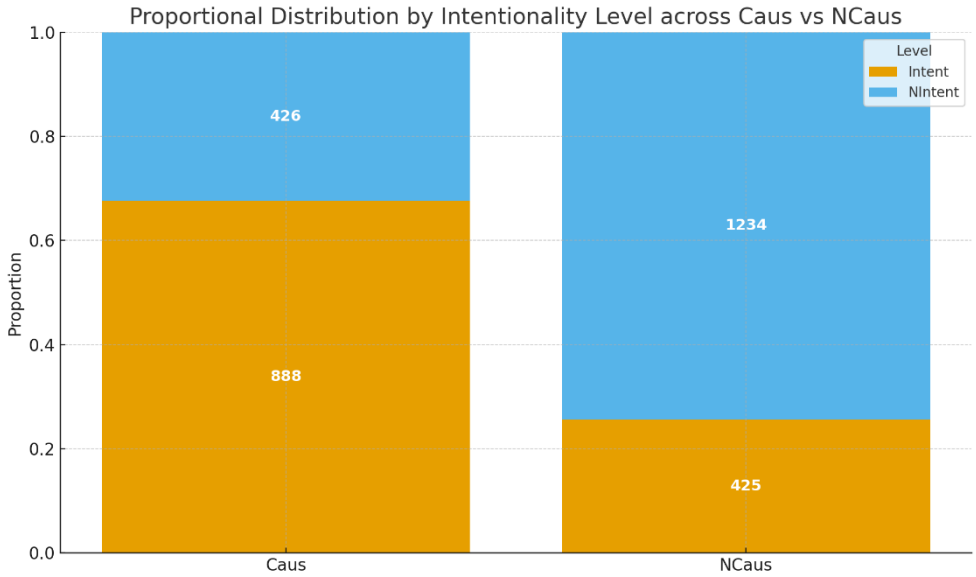
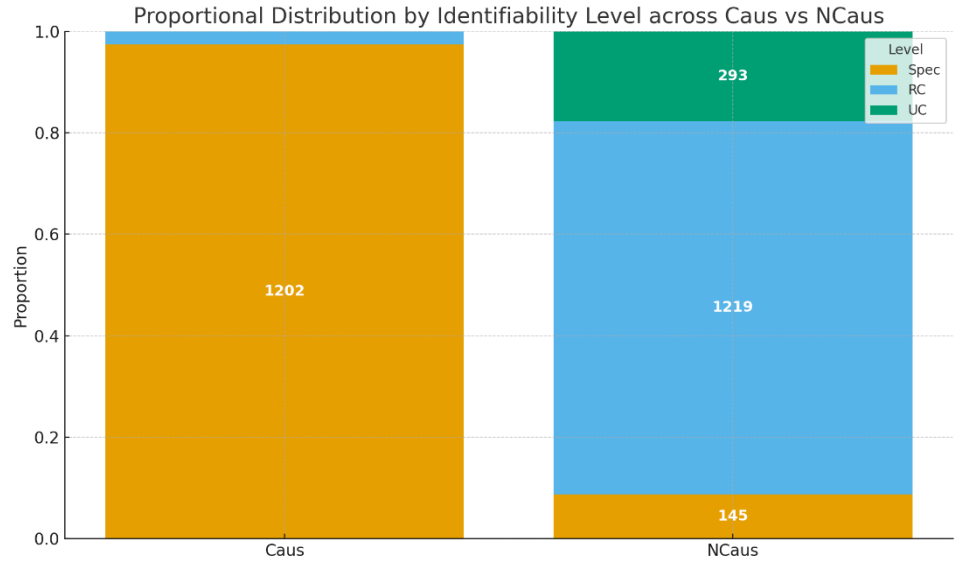



Figure 2. Proportional distribution of identifiability levels in the causative and noncausative variants

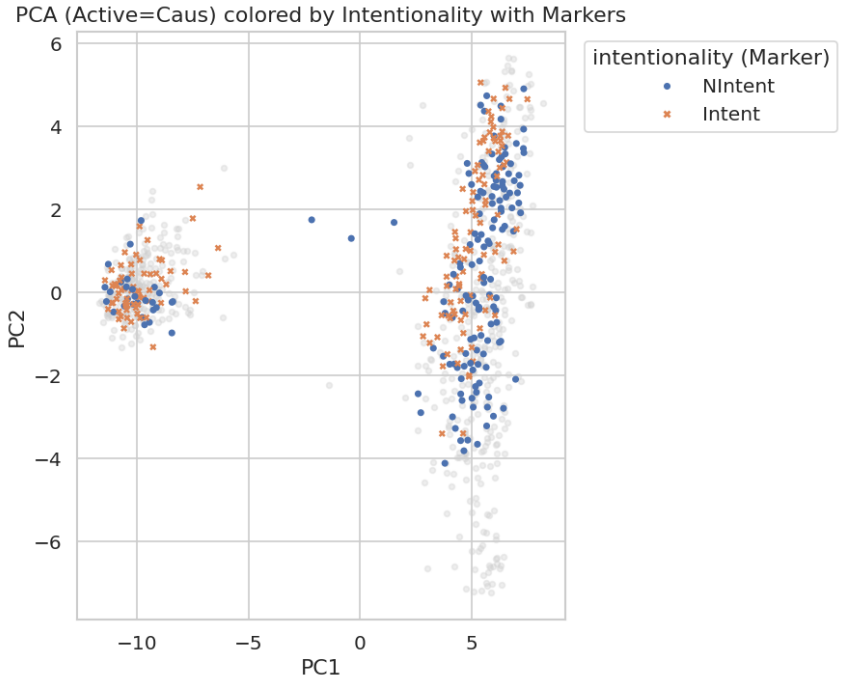



Next, let us examine the visualizations of BERT representations for the 2,893 tokens in our annotated dataset. We provide PCA (Principal Component Analysis) visualizations of BERT embeddings in Figures 2–5. PCA simplifies the complexity of high-dimensional data by projecting them onto a subspace that maximizes variance, thereby preserving key trends and patterns. This dimensionality reduction facilitates more effective visualization and interpretation of the data. The explained variance for the first two dimensions of PCA is 33.85%. The first dimension (PC1), which captures the primary axis of variation in the data, shows an explained variance of 33.85%, while the second dimension (PC2) accounts for a considerably smaller proportion, namely 3.59%.

Figures 3 and 4 present PCA scatter plots of BERT embeddings of causative and noncausative uses, respectively, with data points colored and shaped by the intentionality

variables. The two plots exhibit several shared distributional tendencies. First, the embeddings are concentrated on the left and right regions along PC1, with a noticeably larger proportion clustered in the right-hand region. Second, while the embeddings located on the left cluster tightly around the center of PC2, those on the right are dispersed more widely along the vertical axis, stretching both upward and downward. These distributional tendencies will be revisited in subsequent analyses, where we examine the semantic dimensions captured by PC1 and PC2 along with regional verb distributions and qualitative patterns of verb use.

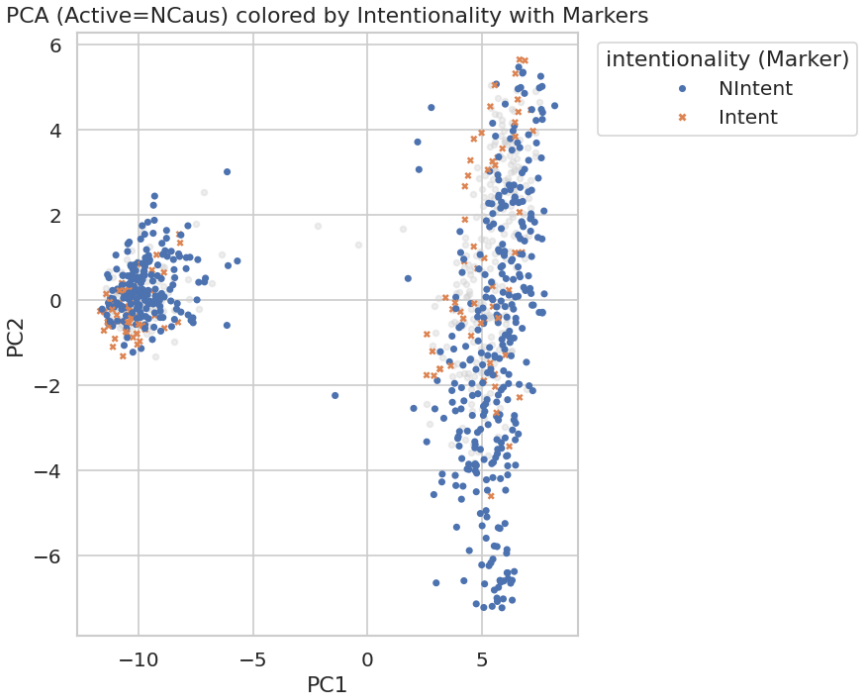
Figure 3. PCA scatter plot of BERT embeddings of causative uses of COS verbs colored by causer intentionality




A comparison of the two plots reveals a clear separation between causative and noncausative uses in terms of causer intentionality. In the causative plot, intentional causer types are relatively predominant in both the left and right clusters. By contrast, in the noncausative plot, nonintentional causer types overwhelmingly dominate across both clusters, with intentional causers appearing only sparsely. This visual impression is statistically supported by the MANOVA results. For the causative uses, intentionality had a significant effect on the distribution of embeddings (Wilks' $\lambda = 0.9018$, $F(2, 325) = 17.70$, $p < .001$). Similarly, in the noncausative uses, intentionality was also found to be a significant predictor of distributional variation (Wilks' $\lambda = 0.9739$, $F(2, 605) = 8.10$, $p < .001$). These results confirm that intentionality levels form significantly distinguishable clusters in the BERT semantic space for both causative and noncausative variants, highlighting the crucial role of causer intentionality in distinguishing between the two syntactic realizations (see also Kim et al. 2025; Lee 2025).

\

Figure 4. PCA scatter plot of BERT embeddings of noncausative uses of COS verbs colored by causer intentionality




Let us now examine the distributions of BERT embeddings of causative and noncausative uses in terms of causer identifiability. A comparison of Figures 5 and 6 reveals that causative and noncausative uses are more distinctly separated with respect to this variable. In the causative plot (Figure 4), specified causers—actual causers of the change of state that are explicitly identified—predominate and minimally overlap with recoverable causer categories. In contrast, noncausative uses are dominated by lower-identifiability types (recoverable and unknown causers), as shown in Figure 5. The clearer separation of specified causers and lower-identifiability types in the two plots visually underscores the strong predictive power of causer identifiability for distinguishing between causative and noncausative uses—an observation consistent with previous findings that this variable is a highly reliable discriminator (Kim et al. 2025; Lee, 2023, 2025). This visual evidence is corroborated by the MANOVA results. For the causative uses, identifiability significantly affected the distribution of embeddings (Wilks' $\lambda$ = 0.9662, $F(2, 325) = 5.68$, $p < .01$). For the noncausative uses, the effect of identifiability was even stronger (Wilks' $\lambda$ = 0.9283, $F(4, 1208) = 11.44$, $p < .001$). These results confirm that identifiability levels form statistically distinguishable clusters in the BERT semantic space for both causative and noncausative variants, further reinforcing the conclusion that causer identifiability is a highly robust factor in differentiating between the two syntactic realizations.

Figure 5. PCA scatter plot of BERT embeddings of causative uses of COS verbs colored by causer identifiability

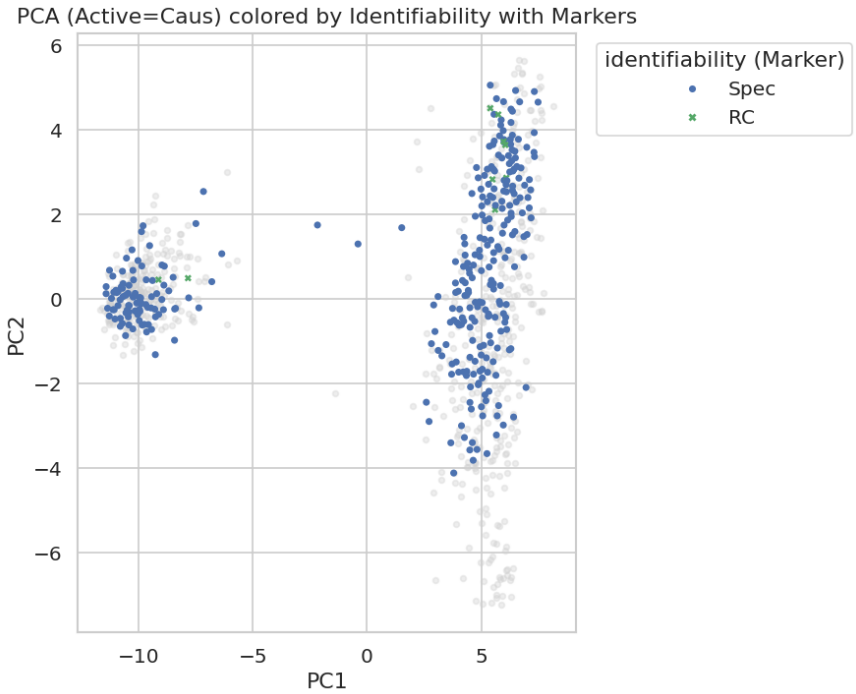
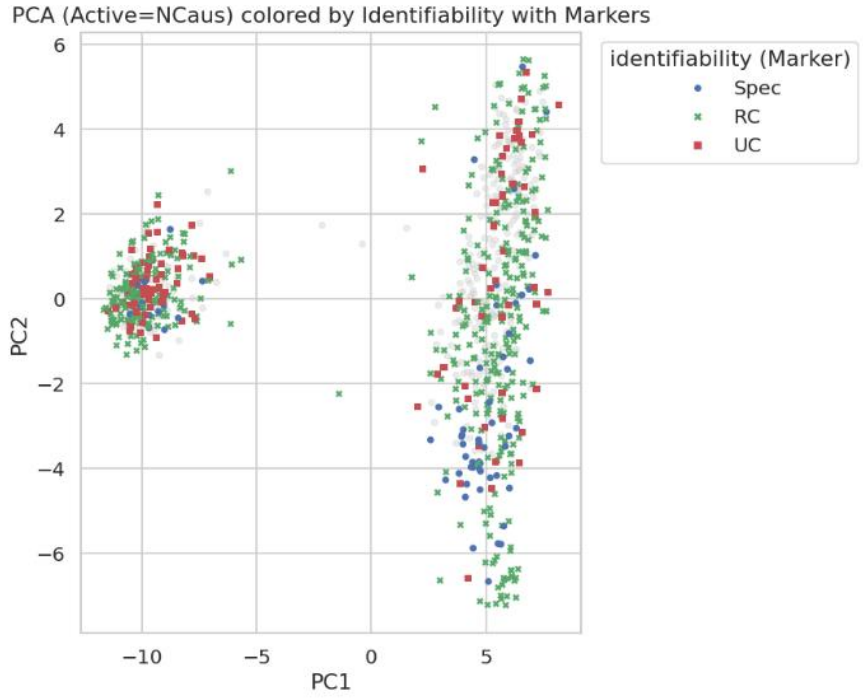



Figure 6. PCA scatter plot of BERT embeddings of noncausative uses of COS verbs colored by causer identifiability




We now turn to an analysis of which verbs and verb uses are distributed across different regions of the PCA space defined by PC1 and PC2. The primary purpose of this regional verb distribution analysis within the BERT-based PCA space of COS verbs is to gain a conceptual understanding of the distributional characteristics of these embeddings and to interpret the key semantic dimensions captured by the two principal components. By examining which verbs

cluster together in different regions of the PCA space, we aim to identify potential semantic groupings or tendencies associated with the verbs based on their contextual usage as represented by the BERT embeddings. This analysis complements the quantitative MANOVA results by providing qualitative insight into the semantic landscape of the verbs and into how linguistic features such as intentionality and identifiability may relate to these spatial distributions.

The regional verb distribution analysis was conducted through the following steps:

1. Data preparation and filtering: The dataframe containing the PCA results (PC1 and PC2 values), along with verb, syntactic realization, intentionality, and identifiability information, was used as the basis for analysis. Data points with missing PC1 or PC2 values were excluded, and the cleaned dataset was then split into two subsets according to syntactic realization: Causative ("Caus") and Noncausative ("NCaus").
2. Defining regional boundaries: For each subset, the PCA space was divided into a 3×3 grid of regions based on the distribution of data points along the PC1 and PC2 axes. The boundaries were defined using the tertiles (1/3 and 2/3 quantiles) of the PC1 and PC2 values within each subset. This created three regions along the PC1 axis (Left, Center, Right) and three regions along the PC2 axis (Bottom, Center, Top).
3. Identifying verbs in each region: Within both subsets, data points belonging to each of the nine regions (3 PC1 × 3 PC2) were identified.
4. Analyzing verb frequency: For each region, the frequency of each unique verb was calculated. "Each unique verb" refers to each distinct verb type found within that specific region's data points.
5. Identifying top verbs: The top 5 most frequent verbs in each region were identified and listed, along with their frequency counts.
6. Generating regional plots: Scatter plots of the PCA space were generated for both causative and non-causative subsets, with the defined PC1 and PC2 region boundaries visually marked to illustrate the spatial division.
7. Conceptual interpretation: Based on verb distributions across regions and the visual patterns observed in the PCA plots, a conceptual interpretation was developed regarding the semantic or contextual dimensions captured by PC1 and PC2. This involved examining which verbs or linguistic features tended to align with higher or lower values on each principal component.

This procedure enabled a detailed exploration of the PCA space, highlighting regions where particular verbs or groups of verbs tend to cluster and providing a foundation for interpreting the semantic distinctions captured by the principal components. The analysis was conducted in Colab using Gemini integrated with the Colab environment.

Tables 2 and 3 summarize the results of the verb distribution analysis across the 3×3 grid of PC1 and PC2 regions for the causative and noncausative subsets, respectively. In each region, the top five verbs are presented in descending order of frequency. For causative constructions (Table 2), the analysis of regional verb distribution reveals distinct patterns. The PC1 dimension appears to differentiate between verbs based on certain semantic or usage characteristics, while PC2 captures another orthogonal aspect.

Table 2. Top verbs in causative PC1/PC2 regions

| PC1 region | PC2 region | Top 5 verbs | Freq. counts |
|---|---|---|---|
| Left | Top | *close* (1), *empty* (1), *freeze* (1), *multiply* (1), *sear* (1) | 6 |
| | Center | *light* (17), *empty* (11), *break* (8), *sink* (4), *multiply* (4) | 62 |
| | Bottom | *break* (10), *light* (7), *empty* (6), *dirty* (3), *close* (2) | 41 |
| Center | Top | *reopen* (5), *sear* (3), *bend* (3), *rekindle* (3), *compress* (2) | 25 |
| | Center | *loose* (5), *alter* (4), *loop* (4), *tame* (3), *chip* (2) | 33 |
| | Bottom | *dim* (9), *sour* (8), *level* (5), *dull* (5), *blunt* (4) | 51 |
| Right | Top | *rekindle* (20), *clog* (12), *multiply* (10), *degrade* (8), *disintegrate* (5) | 79 |
| | Center | *level* (3), *sear* (2), *thaw* (2), *blunt* (1), *chip* (1) | 14 |
| | Bottom | *sour* (6), *dull* (4), *abate* (2), *blunt* (2), *chill* (1) | 17 |

- In the Causative Left–Center region of the PCA space, verbs such as *light*, *empty*, and *break* are prominent, suggesting that these verbs, when used causatively, tend to cluster in this area. The state changes expressed by these verbs are typically associated with surface alterations (*empty, dirty*), boundary changes (*close*), or material and physical transformations (*break, freeze, sear*).
- In the Causative Center–Top region, verbs like *reopen* and *compress* are concentrated, representing physical or surface-level changes that often result from external force or pressure (e.g., *compress* and *bend*).
- Conversely, the Causative Center–Bottom region is dominated by verbs such as *dim*, *sour*, and *level*, which characterize gradual changes in sensory or qualitative properties.
- In the Causative Right–Top region, verbs such as *rekindle*, *clog*, and *multiply* occur most frequently, indicating processes of reactivation, accumulation, or increase that often involve dynamic external interventions.
- Finally, the Causative Right–Center and Right–Bottom regions are populated by verbs like *thaw*, *blunt*, *abate*, and *dull*, which describe shifts in sensory or internal properties, often gradual in nature.

While some verbs appear in multiple adjacent regions, there are clear concentrations of specific verbs in certain areas. This suggests that PC1 largely contrasts physical, material, and surface transformations with qualitative property changes, whereas PC2 differentiates between more immediate versus more gradual changes.

For noncausative constructions (Table 3), the regional analysis of verb distribution in the PCA space also reveals distinctive patterns, which differ from those observed in causative constructions. While the PC1 and PC2 dimensions capture variance in verb meaning and usage, they distribute noncausative verbs in ways that highlight particular semantic tendencies.

Table 3. Top verbs in noncausative PC1/PC2 regions

| PC1 region | PC2 region | Top 5 verbs | Freq. counts |
|---|---|---|---|
| Left | Top | *grow* (11), *crumble* (7), *vary* (7), *shrink* (4), *multiply* (3) | 52 |
| | Center | *shrink* (32), *vary* (24), *grow* (18), *sink* (17), *break* (6) | 130 |
| | Bottom | *shrink* (8), *sink* (5), *empty* (2), *advance* (1), *break* (1) | 21 |
| Center | Top | *reopen* (8), *unfold* (6), *vary* (5), *multiply* (4), *loop* (3) | 52 |
| | Center | *vary* (10), *shrink* (5), *chip* (4), *level* (4), *loop* (4) | 42 |
| | Bottom | *firm* (24), *pale* (22), *abate* (9), *fray* (9), *mature* (8) | 108 |
| Right | Top | *crumble* (23), *sprout* (11), *unfold* (11), *subside* (10), *thaw* (8) | 99 |
| | Center | *shrivel* (8), *crumble* (6), *thaw* (6), *abate* (4), *dissolve* (1) | 30 |
| | Bottom | *pale* (20), *dim* (15), *sour* (11), *abate* (10), *shrivel* (6) | 74 |

- In the Noncausative Left–Top and Center regions, verbs such as *shrink*, *vary*, and *grow* are most frequent, indicating a clustering of verbs that, according to Lee's (2025) collostructional analysis, are most strongly associated with the intransitive construction.
- The Noncausative Left–Bottom region is dominated by *shrink* and *sink*, but also includes verbs like *empty* and *break*, which are characteristic of the causative Left region, as well as *advance*, which denotes an abstract type of state change.
- The Noncausative Center–Top and Center regions contain verbs reflecting boundary changes (*reopen*), surface or external changes (*unfold, shrink, chip*), quantitative expansion (*multiply*), and functional variation (*vary*).
- The Noncausative Center–Bottom region shows a strong concentration of verbs denoting gradual changes in qualitative properties, such as *firm*, *pale*, and *abate*.
- In the Noncausative Right–Top region, verbs like *crumble*, *sprout*, and *unfold* are particularly frequent, pointing to processes of disintegration, emergence, or outward transformation.
- The Noncausative Right–Center and Right–Bottom regions further feature verbs such as *shrivel*, *thaw*, *abate*, and *dim*, which likewise highlight qualitative and sensory changes, often gradual in nature.

Similar to the causative examples, the distinct clustering of verbs across different regions suggests that the PCA dimensions capture meaningful contrasts in the semantic and contextual properties of noncausative verbs. PC1 appears to oppose physical or surface-level transformations to changes in qualitative properties, while PC2 differentiates more immediate from more gradual types of change.

Overall, both causative and noncausative realizations are organized along two principal

dimensions: PC1, contrasting physical/material transformations with qualitative property changes, and PC2, contrasting immediate with gradual changes. Yet the two constructions diverge in their distributional tendencies. In causative uses, verbs denoting surface, boundary, or externally induced material changes (*break, empty, sear, compress*) dominate the Left and Center–Top regions, while verbs encoding qualitative or sensory changes (*dim, sour, dull*) concentrate in the Center–Bottom and Right regions. By contrast, in noncausative uses, verbs strongly associated with the intransitive construction (*shrink, vary, grow*) cluster in the Left–Top and Center regions, while gradual qualitative changes (*firm, pale, abate*) are especially salient in the Center–Bottom and Right regions. Furthermore, verbs of emergence and outward change (*sprout, unfold*) appear prominently in the noncausative Right–Top region but are largely absent from the causative distribution. These complementary patterns underscore how shared semantic dimensions are differentially weighted across the two syntactic realizations.

The regionalized PCA plots in Figures 7–10 provide additional evidence for the distinct clustering patterns of causative and noncausative realizations. In the causative plots (Figures 7 and 9), the Left and Center–Top regions are densely populated with verbs denoting surface, boundary, or externally induced material changes (e.g., *break, empty, sear, compress*), whereas the Center–Bottom and Right regions show concentrations of verbs encoding qualitative or sensory changes (e.g., *dim, sour, dull*). This pattern confirms that causative uses gravitate toward verbs of externally induced, physical transformations. In contrast, the noncausative plots (Figure 8 and 10) highlight a different set of preferences. The Left–Top and Center regions are dominated by verbs such as *shrink, vary,* and *grow*, which Lee (2025) identifies as strongly associated with the intransitive construction. The Center–Bottom and Right regions are characterized by verbs denoting gradual qualitative changes (e.g., *firm, pale, abate*), while the Right–Top region prominently features verbs of emergence and outward change (e.g., *sprout, unfold*), which are largely absent in the causative distribution.

Figure 7. Regionalized PCA scatter plot of BERT embeddings of causative uses of COS verbs colored by causer intentionality

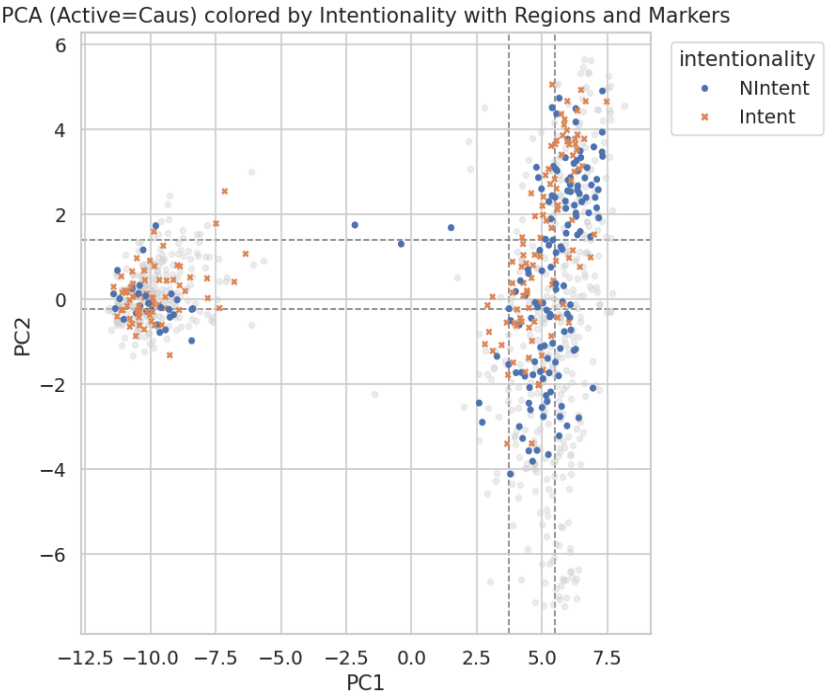


Figure 8. Regionalized PCA scatter plot of BERT embeddings of noncausative uses of COS verbs colored by causer intentionality

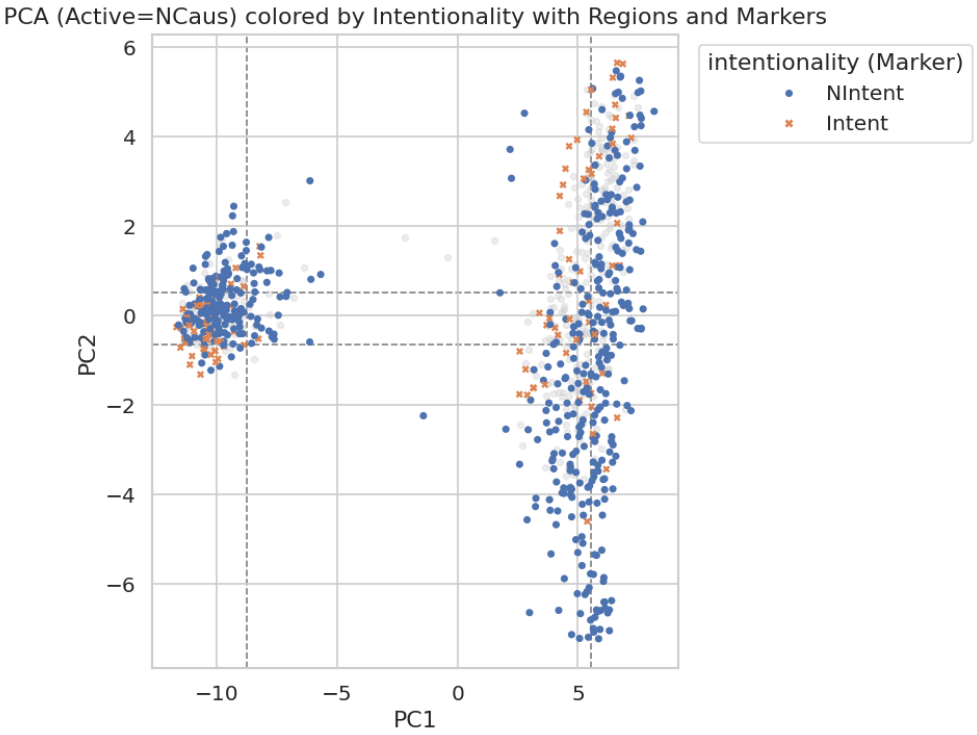



Figure 9. Regionalized PCA scatter plot of BERT embeddings of causative uses of COS verbs colored by causer identifiability

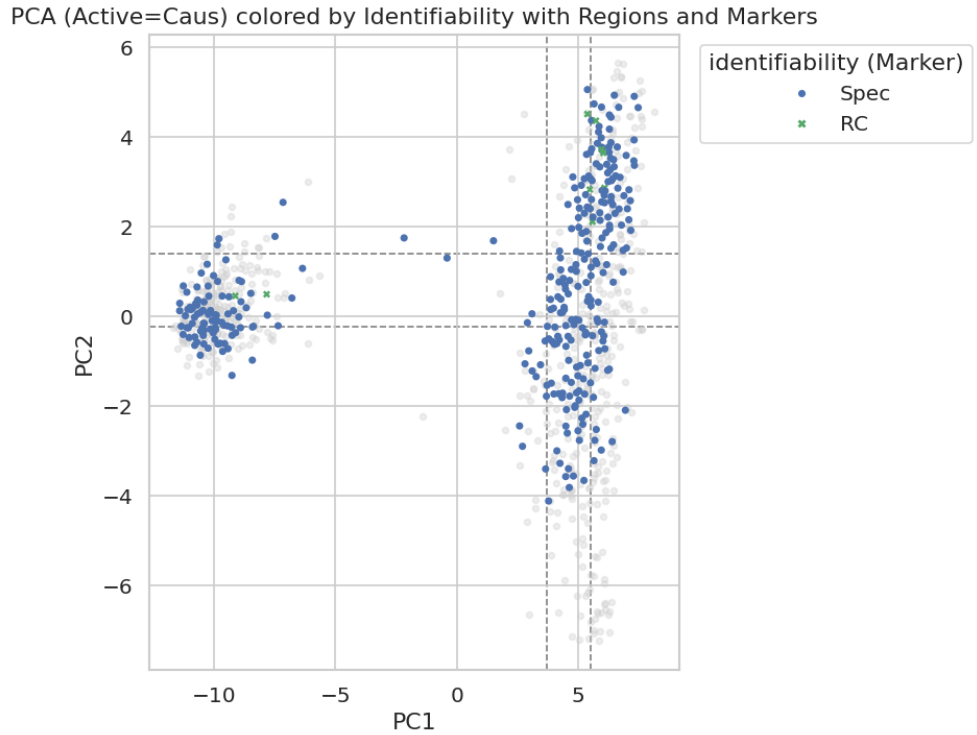
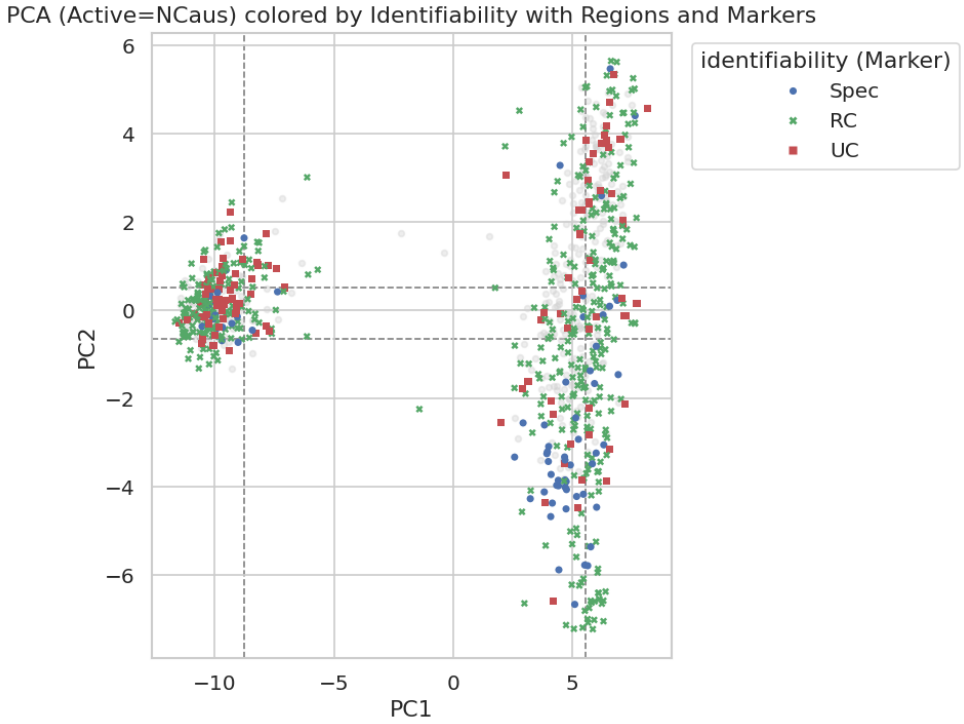


Figure 10. Regionalized PCA scatter plot of BERT embeddings of noncausative uses of COS verbs colored by causer identifiability




The regionalized PCA plots also clarify the two global tendencies noted above. First, although a larger mass of points occurs on the right side of PC1, this region is not driven by physical/material change verbs. Rather, the left cluster houses most verbs denoting concrete, externally induced material or boundary changes (e.g., *break, empty, close, sear*; and, in the noncausative set, *shrink, grow*), whereas the right cluster gathers verbs that encode qualitative/sensory shifts, attenuation or increase, reactivation/accumulation, and emergent or outward change (e.g., *dim, sour, dull, abate; rekindle, clog, multiply; sprout, unfold*). Thus, the greater density on the right reflects the prevalence of these qualitative/emergent patterns rather than physical change per se. Second, the right cluster shows a much wider spread along PC2—consistent with a mixture of more immediate and more gradual trajectories—while the left cluster remains comparatively compact, indicating a narrower range along this dimension. Together, these patterns accord with the regional verb distributions reported in Tables 2–3 and with the interpretation of PC1 as contrasting material/boundary vs. qualitative/emergent change, and PC2 as capturing gradience in the temporal/qualitative development of the event.

Finally, we turn to the question of whether the principal dimensions of BERT-derived semantic spaces are fine-grained enough to capture subtle variation in verb usage. For this analysis, Gemini in Colab was used to extract sentences for specific verbs in the PCA regions from the dataframe that contained embeddings, corpus example sentences, and PCA results. Among the verbs extracted by this procedure, we focus on *empty* as a case study.

As shown in Tables 2 and 3, *empty* is confined to the PC1 Left region in both the causative and noncausative subsets, but its occurrences are distributed across the Top, Center, and Bottom regions of PC2. As noted above, embeddings in the left cluster tend to be tightly grouped around the center of PC2, suggesting a relative homogeneity in how these verbs pattern along the gradual–immediate axis. Because *empty* occupies a relatively narrow range in the

PCA space, it provides a suitable test case for examining how subtle variation in verb usage is manifested across PCA regions.

Examples (9)–(12) illustrate causative uses of *empty*. In (9), the verb describes an external event in which factories discharged waste into the river, an instance of an immediate and concrete change (Causative–Left–Top). In the Left–Center region, similar cases occur, but they more often describe gradual processes or extended changes. For example, (10a) and (10b) involve incremental emptying events—a glass being drained or bladders being emptied at timed intervals—while (11) metaphorically depicts an internal psychological change, "The War had emptied him," unfolding over a period of time (Causative–Left–Center). In contrast, the Left–Bottom region includes examples like (12a–c), which highlight a relative gradualness in the emptying of multiple entities (several glasses, guns, or the diverse contents of an envelope). These causative uses of *empty* exhibit subtle variation along the immediate–gradual axis captured by PC2. At the same time, they share the common feature that the causer of the change is clearly identified, underscoring the role of specified agency in causative realizations of *empty*.

(9) Causative-Left-Top
But in time many factories and many sewers emptied their waste into the Mersey.
(BNC W:non_ac:soc_science, B1H-363)

(10) Causative-Left-Center
a. He emptied his glass and poured himself some more beer.
(BNC W:fict:prose, GV6-177)
b. At 30 and 0 minutes before and at 15, 30, 45, 60, 75, 90, 120, 180, 240, 300, 360, and 420 minutes after the onset of the mean, <the subjects emptied their bladders>, …
(BNC W:ac:medicine, HWT-1166)

(11) Causative-Left- Center
The War had emptied him; stripped him of all illusions.
(BNC W:fict:prose, G04-2798)

(12) Causative-Left-Bottom
a. They both emptied their glasses. (BNC W:fict:prose, H0R-2824)
b. But while it was being chopped down the Collector and <his men had emptied their guns into the hacking sepoys>, and … (BNC W:fict:prose, EFW-1613)
c. … <Elinor emptied the contents of the envelope onto the table>: family snaps, letters, postcards belonging to the Chatwins, the coat of arms …
(BNC W:fict:prose, ACK-2230)

Examples (13)–(16) illustrate noncausative uses of *empty*, with surrounding context provided to identify potential causer types. In the Left–Top region, *empty* describes either temporary psychological states, as in (13) "Her mind emptied," which signals a sudden cessation of current thoughts without an explicitly identified cause (a hinted causer, triggered by the emptiness of the room), or immediate spatial changes, as in (14), where the streets emptied directly as a result of extreme heat (RC–previously mentioned causer).

(13) Noncausative-Left-Top

Nothing remained to show that the headmaster had had his daily being here, perhaps here conducted interviews with backsliding pupils, commended scholars. There was no desk, only a big bed someone had made up with whose sheets she did not know, a n armchair and a table, a cupboard, a window through which the passing trains could be seen, tube trains and Metropolitan trains and the trains that went up to the Chiltern s. She had seen a phone in the big hall they called the vestibule. <Her mind emptied>. She sat on the bed and real life came rushing in to fill the vacuum. She thought of Mi ke starting his holiday today.

(BNC W:fict:prose, EDN-1148)

(14) Noncausative-Left-Top

Because of the heat all work stopped about lunch-time and the city came to a halt. <The streets emptied>, the shops shut, the donkey boys retreated into the shade, and government offices closed. Most people took a siesta.

(BNC W:fict:prose, HTX-1458)

In the Left–Center region, noncausative *empty* tends to describe more gradual processes. In (15a), the room empties incrementally as people leave to watch the Battle (RC–previously mentioned causer), while (15b) describes "gastric emptying," a natural biological process explained as a default cause (the human body's digestive system) without an overtly named causer (RC–default cause).

(15) Noncausative-Left-Center

a. 'Just to make sure th'don't take off. When Battle's over I'll be back for thee.' Jess snatched after her skirt, but he disappeared into the crowd. <Gradually the room emptied> until the only people who remained besides herself, were the landlord and an old woman asleep on a stool by the far wall.

(BNC W:fict:prose, C85-168)

b. Figure 1 shows the mean gastric emptying of the test meal. The liquid phase ( lem onade ) began to empty in a rapid exponential fashion immediately after ingestion. After 0.7 ( 0.09 ) h ( mean ( SEM ) ) , <half of the liquid phase ( T50 ) had emptie d from the stomach>.In all subjects , the digestible phase ( omelette ) of the test m eal emptied more slowly than the liquid phase and the mean T50 was 2.4 ( 0.38 ) h .

(BNC W:ac:medicine, HU4-535)

Similarly, in the Left–Bottom region empty tends to describe more gradual changes as in (16), where *empty* conveys a relative gradualness in psychological change: "her head emptied of Fenna," signaling a mental shift toward ease in social interaction. Here, the change has no clear directly identified causer (UC).

(16) Noncausative-Left-Bottom

And suddenly it could be a story, a funny story that was making Hermione laugh. Sud denly she knew how to do this, this chatting and joking and telling her life, sharing it, warming herself from the fire of affection she had lit in her friend. Distantly she reme mbered that she had never understood the art of it, had to work on it, conscious and str

uggling all the time. <Now, her head emptied of Fenna>, it flowed out smooth and bubbling, almost unnoticed. This was hermion, this was her friend.

(BNC W:fict:prose, A6J-1985)

Across these noncausative uses, *empty* consistently reflects nonintentional causation, with causer identifiability classified as either RC or UC. At the same time, the PCA regions reveal subtle differences along the immediate–gradual axis, demonstrating that the principal dimensions of BERT-derived semantic spaces effectively capture these fine-grained distinctions in verb usage.

In sum, this section has examined the usage variation of strong alternators in the causative alternation by integrating annotated corpus variables with BERT-based distributional semantic analysis. The results demonstrate that causer intentionality and identifiability are robust discriminators between causative and noncausative uses, as evidenced by proportional distributions, MANOVA tests, and PCA visualizations. The analysis identified two principal dimensions of the BERT-derived semantic space—PC1 (physical/material vs. qualitative change) and PC2 (immediate vs. gradual change)—and demonstrated how these dimensions structure verb distributions across constructions. The case study of *empty* further illustrated that BERT embeddings capture subtle usage variation, complementing quantitative findings with qualitative insights. This sets the stage for the in-depth case studies of two polysemous COS verbs, *break* and *freeze*, which exhibit contrasting constructional preferences.

## 2.2. Analyzing Sense Distribution: A Case Study of *Break* and *Freeze*

### 2.2.1. Annotated Dataset

The dataset for this case study consists of a total of 1,060 examples of *break* and *freeze* (620 *break* examples and 440 *freeze* examples). These examples were extracted from the *Corpus of Contemporary American English* (COCA: Davies 2008–) and are based on the annotated dataset created by Lee (2025), which was subsequently updated to include additional examples and sense distinctions.

The data prepared as BERT input were saved in an Excel spreadsheet and manually annotated by the author and two graduate students. Each token was coded for five variables: sense, syntactic realization, causer intentionality, cause identifiability, and theme concreteness. These variables are summarized in Table 4. In this section, we provide an overview of the annotation schemes for the sense and concreteness labels. The two causer-related variables follow the same annotation schemas applied in the analysis of Section 2.1.

Table 4. Annotated variables

| Variables | Levels | |
|---|---|---|
| Sense | 11 *break* senses (Table 5) and 10 *freeze* senses (Table 7) | |
| Realization | Causative (Caus) vs. Noncausative (NCaus) | |
| Causer Intentionality | Intentional (Intent): Agentive causer | Agent |
| | | Intentional causing action (Int-act) |
| | Nonintentional (NIntent): Nonagentive causer (cause) | Animate causer (Cause_anim) |
| | | Inanimate causer (Cause_inan) |
| | | Causer that can be either animate or inanimate (Cause_anim/inan) |

| | | |
|---|---|---|
| Causer Identifiability | Specified causer (Spec) | |
| | Recoverable causer (RC) | Previously mentioned causer (RC_mentioned) |
| | | Hinted causer (RC_hinted) |
| | | Default causer (RC_default) |
| | | Other RCs (RC_other): type-inferable causer, omitted subjects of causative uses of finite COS verbs, and subjects of imperative clauses |
| | Unknown causer (UC) | |
| Theme Concreteness | Concreteness vs. Abstract | |

Annotation reliability was ensured by high inter-rater agreement: Light's kappa values exceeded 0.80 for all annotation dimensions, confirming that the scheme is sufficiently reliable for subsequent analyses.

*BREAK* SENSE ANNOTATION. The lemma *break* is listed in 59 SynSets in WordNet. Many of these fine-grained senses involve combinations of *break* with a range of predicates and particles (e.g., *break down, break out, break in, break even, break to the right*). Because such phrasal expressions often denote meanings outside the domain of state change—which is the primary focus of this study—they were excluded from the present analysis. Instead, we consider only one-word uses of *break*.

Among the 59 WordNet SynSets, the one-word uses of *break* that encode change-of-state meanings were classified into 11 sense categories. This classification builds on insights from previous studies (Kellerman 1978; Jung 2019; Romain 2022; Petersen and Potts 2023; Lee 2025).

1. Physical breaking (destruction, separation, detachment, etc.)
- Sense 2: Become separated into pieces or fragments
- Sense 3: Destroy the integrity of (usually by force)
- Sense 44/45: Become or cause to be punctured/penetrated

2. Bodily harm
- Sense 57: Cause injures or bodily harm to

3. Operational failure and cessation
- Sense 4/17: Stop operating or functioning
- Sense 5: Ruin completely (e.g., render useless)
- Sense 23: Come to an end; cease

4. Escape and rupture
- Sense 7: Move away or escape suddenly
- Sense 8/19: Make a rupture (e.g., break ranks)

5. Disruption of continuity (interruption, termination, and change)
- Sense 1/10: Terminate or stop an activity, flow, or state
- Sense 18/47/48: Pause or interrupt a continued activity, relationship, or flow
- Sense 25/26: Disrupt routine (e.g., giving up and changing a habit, routine, or

monotony)
- Sense 34/36/52/55: Change suddenly (e.g., directions, voice or tone quality)
- Sense 58: Diminish or discontinue abruptly (e.g., fever breaking)

6. Emotional, psychological and social damage
- Sense 59: Weaken or destroy in sprit or body
- Sense 12: Make submissive or docile
- Sense 29: Cause the failure or ruin of

7. Revelation and discovery
- Sense 15/46: Make known to the public
- Sense 51: Find a solution or key

8. Non-compliance or violation
   - Sense 6/13: Violate, go against

9. Breakthrough (overcoming barriers and restrictions)
   - Sense 14: Surpass

10. Emergence
    - Sense 35: Emerge from a surface
    - Sense 16: Come into being
    - Sense 27: Come forth or begin from a state of latency (e.g., storm breaking)

11. Economic, transactional, and miscellaneous senses
    - Sense 38: Exchange for smaller units of money
    - Sense 20: Curl over and fall apart

Among the 11 higher-level categories of *break* senses identified above, our analysis considers only those categories that are represented by at least ten examples in the dataset. Table 6 presents these sense categories, together with illustrating examples and the frequency counts for each category.

Table 6. Sense categories of *break*

| Sense categories | Examples | Freq. counts |
|---|---|---|
| Destruction | The glass broke from the pressure. | 129 |
| Bodily harm | My ankle broke in a skiing accident. | 37 |
| Emotional/psychological/ social breakdown | Her heart broke when she heard the news. | 64 |
| Violation | The company broke the law. | 83 |
| Decoding | The police finally broke the code. | 12 |
| Disclosure | The reporter broke the news early in the morning. | 42 |
| Termination | They broke the cycle of violence. | 56 |
| Interruption | I broke the tense atmosphere with a joke. | 42 |
| Breakthrough | She broke the world record. | 93 |
| Change | The medicine helped break the fever. | 36 |
| Emergence | The day broke over the quiet village. | 26 |

The senses listed in the table represent only a partial inventory; we cannot hope to be

comprehensive (indeed, there may not even be a fixed stock of senses), but the examples provided convey the nature of the attested variation. Following standard annotation practice, each corpus example was assigned a single sense. However, as noted by Petersen and Potts (2023) and Lee (2025), multiple senses may simultaneously be present in a single usage. For instance, in *the dawn broke*, *broke* may be interpreted as WordNet sense 27 (*natural emergence*), but at the same time it denotes a spontaneous change in environmental or temporal states, thus also fitting into the broader category of *disruption of continuity*. Such cases of multiple sense realization will be examined in more detail in Chapter 3.

*FREEZE* SENSE ANNOTATION. WordNet lists ten verbal senses for *freeze*. These senses can be grouped into three higher-level categories: physical/bodily/natural freezing**,** sudden immobility, and cessation. Within the sudden immobility category, this study distinguishes two subtypes: physical immobilization (Sense 2: stop moving or become immobilized) and emotional freezing. WordNet explicitly recognizes one subtype of emotional freezing (Sense 10: become formal, haughty, or unfriendly), but does not list a further meaning that we treat as distinct here: *becoming mentally or emotionally paralyzed due to shock, fear, or confusion*.

1. Physical/bodily/natural freezing
- Sense 1: Change to ice
- Sense 3/6: Be or become very cold
- Sense 7: Change from a liquid to a solid when cold
- Sense 4: Cause to freeze (preserve by solidifying through rapid refrigeration)
- Sense 9: Anesthetize by cold; die or be damaged by exposure to cold

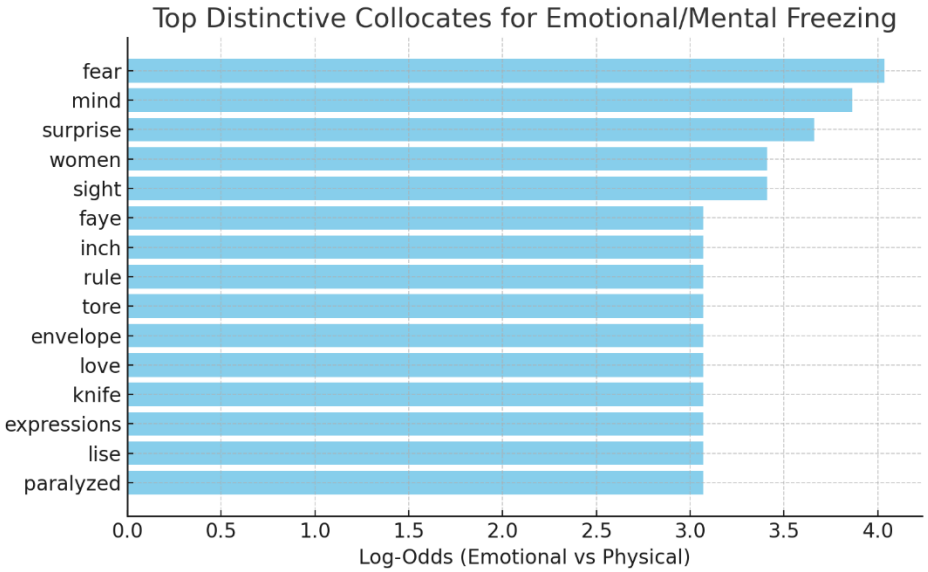
2. Sudden immobility
- Sense 2: Stop moving or become immobilized (physical immobilization)
- Sense 10: Become formal, haughty or unfriendly (emotional freezing 1)
- Becoming mentally or emotionally paralyzed due to shock, fear, or confusion (emotional freezing 2)

3. Cessation
- Sense 5: Stop a process or a habit by imposing a freeze on it
- Sense 8: Block; prohibit the conversion or use of (assets)

The decision to treat this "emotional freezing 2" sense as distinct rests on both empirical and theoretical grounds. Empirically, corpus analysis reveals that emotional freezing exhibits a collocational profile quite different from physical immobilization. The collocational analysis, based on the same COCA-derived dataset used for the sense distribution study of *freeze*, is summarized in Figures 11 and 12. The figures display the top distinctive collocates for the emotional/mental freezing sense and for the physical immobilization sense, with their relative distinctiveness measured by log-odds scores. Positive log-odds values indicate that a collocate is strongly biased toward the emotional freezing sense, whereas negative values signal an association with physical immobilization. In the case of emotional freezing (Figure 11), the top collocates include *fear, mind, surprise, paralyzed, love,* and *expressions*. These words cluster within affective and cognitive semantic fields, pointing to psychological paralysis triggered by emotional states or mental overload. The high log-odds values for *fear* and *mind* (exceeding +4 and +3.8, respectively) show that such collocates are not incidental but systematically and

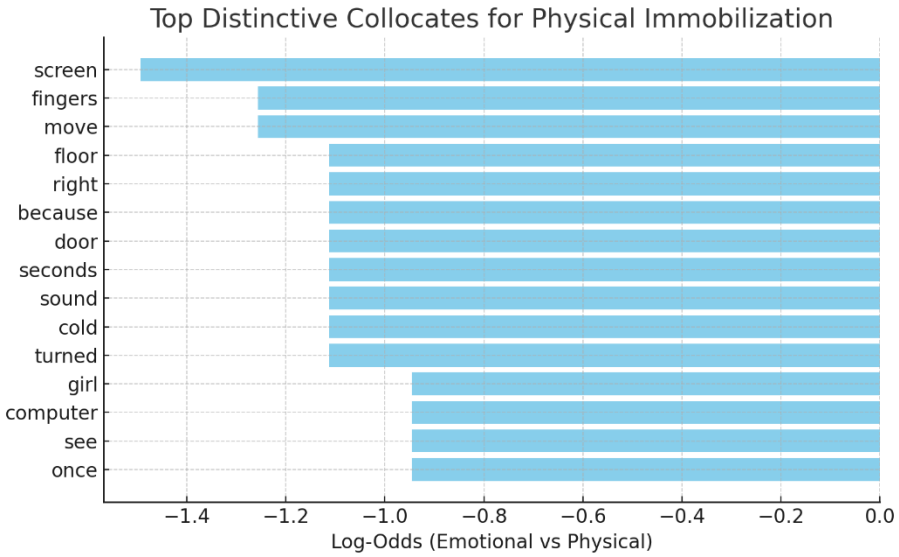
disproportionately linked to emotional freezing contexts.

Figure 11. Top distinctive collocates for emotional/mental freezing sense




By contrast, the physical immobilization profile (Figure 12) highlights collocates such as *screen, fingers, move, door, cold,* and *computer*. These belong to domains of bodily motion, material objects, and device malfunction. The strongly negative log-odds values (e.g., –1.5 for *screen*, –1.3 for *fingers*) indicate that these words are highly characteristic of physical immobilization uses and rarely occur with emotional freezing.

Figure 12. Top distinctive collocates for physical immobilization




The divergence between the two figures demonstrates that the emotional freezing sense is not merely an online extension of the physical immobilization sense but is associated with its own collocational ecology. The log-odds measures provide quantitative confirmation: collocates such as *fear* and *paralyzed* are statistically diagnostic of emotional freezing, while

*screen* and *fingers* are diagnostic of physical immobilization. This clear separation supports the claim that emotional freezing has conventionalized distributional properties, thereby justifying its treatment as a distinct sense of *freeze*.

The theoretical framework of principled polysemy (Tyler and Evans 2001, 2003) further strengthens the decision to treat the "emotional freezing 2" sense as distinct. According to their criteria, a distinct sense is warranted when (i) additional meaning beyond the prototypical sense is encoded, rather than being predictable from context, and (ii) there are context-independent attestations that demonstrate conventionalized use in speakers' long-term semantic memory (Tyler and Evans 2001: 65–70). *Emotional freezing* satisfies both conditions: it introduces the specific notion of affective or cognitive paralysis, and corpus attestations such as *her mind froze* or *my brain froze* demonstrate stable, conventionalized uses that cannot be reduced to the physical immobility sense.

Taken together, both distributional evidence and principled criteria for sense individuation support the analysis of *emotional/mental freezing* as a distinct sense of *freeze*, even though it is not separately enumerated in WordNet and related lexical resources.

Among the categories of *freeze* senses identified, our analysis focuses only on those represented by at least ten examples in the dataset. Table 7 presents these sense categories, along with illustrative examples and their frequency counts.

Table 7. Sense categories of *freeze*

| Sense categories | Examples | Freq. counts |
|---|---|---|
| Physical freezing (freezing of artificial objects) | The storm froze the pipes, causing them burst. | 43 |
| Bodily freezing | My fingers froze in the cold. | 33 |
| Natural freezing (freezing of natural elements) | The lake froze overnight. | 29 |
| Preservation (freezing for preservation) | I froze the meat for preservation. | 57 |
| Physical immobilization | He froze in place when he saw the bear.<br>My hands froze in mid-clap. | 143 |
| Mechanical breakdown / Technical failure | The engine froze due to the cold. | 17 |
| | My computer froze, and I lost my work. | |
| Emotional/mental freezing | Her mind froze completely. | 48 |
| Suspension | The company froze production. | 42 |
| Economical freezing | The bank froze their assets. | 28 |

As discussed above, there is sufficient evidence to treat physical immobilization and emotional/mental freezing as distinct senses, although they may co-occur within a single example (e.g., *Fear froze her in place*). In addition, certain senses such as *suspension* and *economic freezing* exhibit blurred boundaries, making their categorization less clear-cut. Issues concerning sense boundaries will be addressed in detail in Chapter 3.

THEME CONCRETENESS ANNOTATION. Semantic concreteness refers to the extent to which a word or concept denotes a physical, tangible entity that can be perceived through the senses. A concrete theme refers to an object or entity that can be seen, touched, heard, smelled, or tasted, whereas an abstract theme designates ideas, states, qualities, or relationships that lack physical form.

Theme concreteness has been identified as one of the key semantic dimensions structuring the distributional space of polysemous COS verbs in Lee (2025). In Lee's multifactorial analysis, which simultaneously considered causer-related and theme-related factors in the causative alternation, concreteness of the theme emerged as the second most significant factor—following causer-related features—in distinguishing between the two variants. In this section, we validate these findings through a BERT-based distributional semantic analysis.

Following Lee (2025), each theme was annotated as concrete or abstract based on the following operational criteria: lexical meaning (conceptual properties of the referent), psycholinguistic concreteness ratings**,** WordNet semantic fields, and context-based judgment.

1. Lexical meaning (conceptual properties of the referent)

- If the theme noun refers to a material object or physical entity, annotate as concrete.
- If it refers to a conceptual, emotional, or abstract entity, annotate as abstract.

2. Predefined psycholinguistic ratings

- Use Brysbaert el al.(2014)'s psycholinguistic concreteness ratings to classify words numerically:
  - Rating > 3.0 → Concrete
  - Rating ≤ 3.0 → Abstract

3. WordNet semantic fields

- Words linked to 'noun.artifact', 'noun.body', 'noun.animal' → Concrete
- Words linked to 'noun.attribute', 'noun.cognition', 'noun.communication' → Abstract

4. Context-based judgment

- In ambiguous cases (e.g., *account, heart, record*), context is checked to determine whether the theme is used literally (concrete) or figuratively/conceptually (abstract).

Brysbaert et al. (2014)'s psycholinguistic concreteness ratings were used as the primary criterion for annotation, with the other three criteria serving as supplementary measures. In their study, Brysbaert et al. evaluated 60,000 English words and 2,900 two-word expressions by collecting concreteness ratings from over 4,000 U.S.-based participants through an online survey. Participants rated each item on a five-point scale (1 = highly abstract, 5 = highly concrete), resulting in approximately 1.7 million valid judgments. The study produced an extensive database that has since become a widely used resource in psycholinguistic research. Following the authors' recommendations, subsequent studies—including the present one—classify items with a mean rating above 3 as concrete, and those with a rating below 3 as abstract. Applying this threshold consistently ensures clarity, reliability, and replicability in annotation.

For items not included in Brysbaert et al.'s dataset, as well as plural nouns, pronouns, and proper names, classification was made by identifying the referent. In cases where a singular noun was not rated in Brysbaert et al. (2014), theme concreteness was determined by consulting WordNet semantic fields and by examining the contextual meaning of the item. Words associated with WordNet fields such as `noun.artifact`, `noun.body`, and `noun.animal` were annotated as concrete, while those linked to fields like

`noun.attribute`, `noun.cognition`, and `noun.communication` were annotated as abstract. In ambiguous cases—for example, *account*, *heart*, or *record*—the immediate context was examined to determine whether the theme referred to a literal, tangible entity (concrete) or a figurative or conceptual notion (abstract).

### 2.2.2. The Hypothesis and its Predictions

The central hypothesis to be tested in the case study presented in the following section concerns the association between verb senses and causer types. Building on insights from Rappaport Hovav (2014, 2020) and Lee (2025), this hypothesis can be formulated as follows: the sense distribution of polysemous causative alternation verbs correlates with semantic and contextual properties of the causer that affect the degree of informativeness of the causative variant.

From this hypothesis, several predictions follow with respect to differences in syntactic distribution across senses. Causative-dominant senses—those disproportionately realized in the causative variant—are expected to correlate with causer properties that increase the informativeness of the causative construction. By contrast, noncausative-dominant senses—those disproportionately realized in the noncausative variant—are predicted to be associated with causer properties that decrease the informativeness of the causative variant.

In verbs such as *break* and *freeze*, which differ markedly in their constructional preferences, this hypothesis further predicts systematic divergences in how their senses associate with causer types. For *break*, senses that involve intentional or explicitly identified causers are expected to dominate, reinforcing the causative preference of the verb. In contrast, *freeze* is predicted to show a stronger alignment with noncausative-dominant senses involving nonintentional or less clearly identifiable causers, thereby reflecting its bias toward the noncausative variant.

Table 8 schematically summarizes the predicted associations between sense type, causer properties, and constructional preference. As shown, causative-dominant senses are expected to cluster with intentional and explicitly identifiable causers, thereby favoring realization in the causative variant, with *break* serving as a prototypical example. In contrast, noncausative-dominant senses are predicted to align with nonintentional or less identifiable causers, favoring realization in the noncausative variant, as illustrated by *freeze*.

Table 8. Hypothesized association between sense type, causer properties, and variant preference

| Sense type | Associated causer properties | Variant preference | Illustrative verb tendency |
|---|---|---|---|
| Causative-dominant senses | High causer salience (intentional, explicitly identified causer) | More frequent in causative variant | Significantly attracted to causative variant (e.g., *break*) |
| Noncausative-dominant senses | Low causer salience (nonintentional, less clearly identifiable causers) | More frequent in noncausative variant | Significantly attracted to noncausative variant (e.g., *freeze*) |

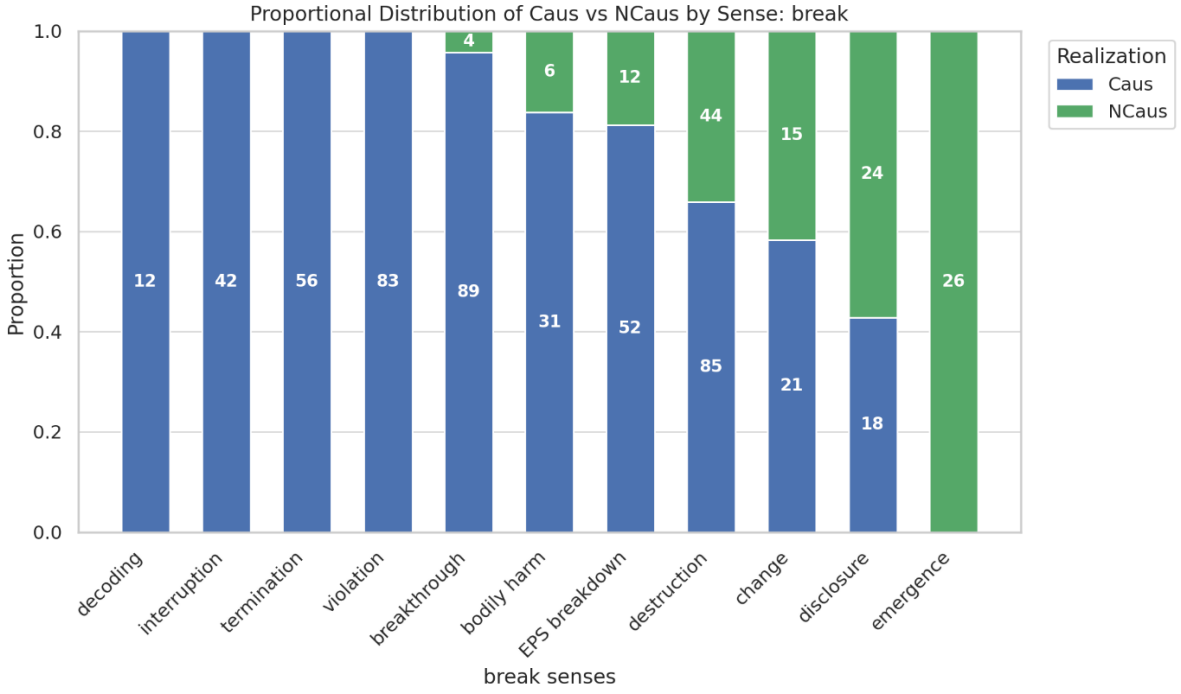
This framework provides the empirical basis for the analyses in Section 2.2.3.

### 2.2.3. The Syntactic and Semantic Distribution of *Break* Senses

This section presents the results of the analysis of the sense distribution of *break*. Beginning with the proportional distribution of causative alternation variants, the analysis shows that among 620 tokens, causative variants account for 489 cases (78.9%), while noncausative variants account for 131 cases (21.1%). This indicates a strong preference for causative realization, consistent with the findings of Romain (2017) and Lee (2025), whose distinctive collostructional analyses showed that *break* is significantly attracted to the causative variant.

Let us now examine the sense-level distributions in the two variants. As Figure 13 shows, out of 11 sense categories, 9 senses exhibit a clear preference for the causative variant, with causative proportions ranging from roughly 70% to near 100%. In contrast, two senses, i.e., disclosure and (natural) emergence, show a strong preference for the noncausative variant, with causative proportions below 30%.

Figure 13. Proportional distribution of the *break* senses in the alternation variants




This figure's sense-level distributions allow us to classify the senses of *break* into causative-dominant and noncausative-dominant categories as shown in Table 9.

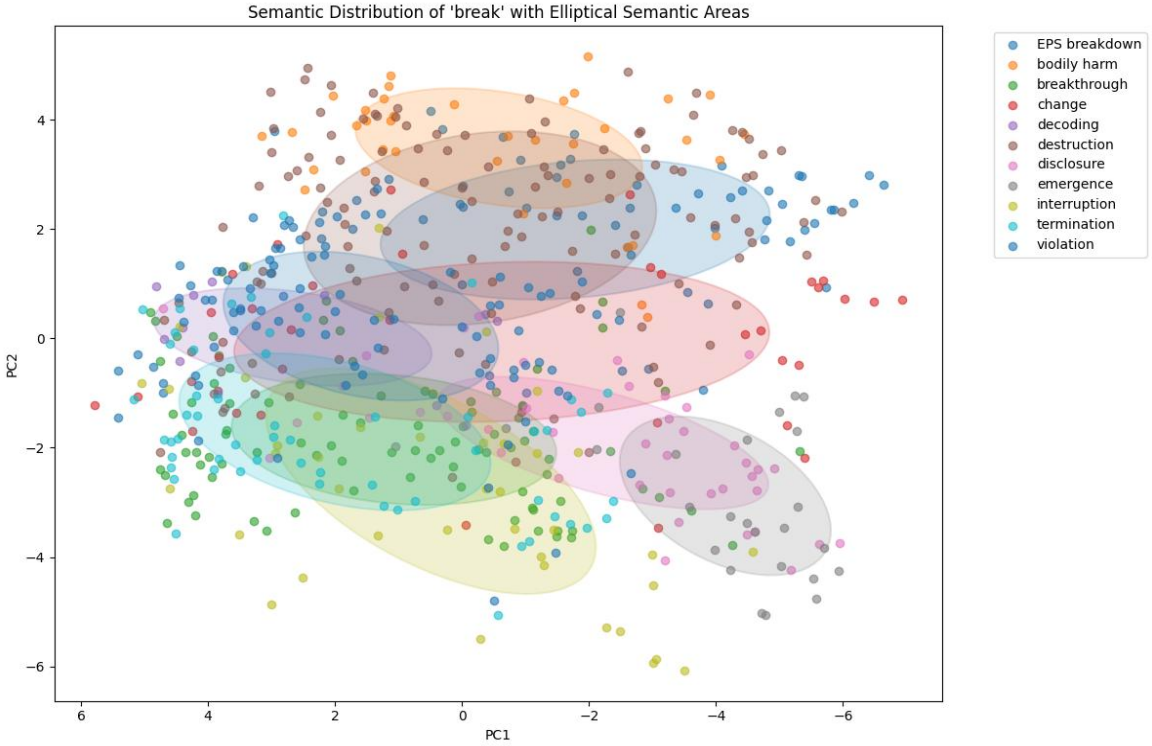
Table 9. Constructional distribution of the *break* senses

| Constructional distribution | | Senses |
|---|---|---|
| Causative-dominant | Strongly skewed | decoding, interruption, termination, violation, breakthrough |
| | Moderately skewed | bodily harm, emotional/psychological/social (EPS) breakdown, destruction, change |
| Noncausative-dominant | Non-alternating | emergence |
| | Alternating | disclosure |

Next, let us examine the visualizations of BERT representations for the 620 tokens in our annotated dataset. Data and code to reproduce our results and figures are available at the GitHub repository (https://github.com/hanjung-25/clmsemantics). We provide PCA visualizations of BERT embeddings in Figures 14-31. The explained variance for the first two dimensions of PCA is less than 15%. This low ratio is likely due to the over-high 768 dimensions of the embeddings, making it difficult to encapsulate the major features of the data within a few dimensions. The first dimension (PC1) of *break*'s PCA scatter plot shows an explained variance of 8.58%, and the second dimension (PC2) 5.35%.

Figure 14 is a PCA scatter plot of *break* sense categories, with distributional areas highlighted by ellipses. In this figure, physical senses (bodily harm and destruction) cluster in the upper-central region, overlapping with the metaphorically extended EPS breakdown sense (emotional/psychological/social breakdown). This overlap suggests early semantic extension from physical to emotional domains. Lower regions of the space in Figure 14 show abstract senses (decoding, violation, change) differentiated along PC1/PC2, with bridging physical/emotional and causative-skewed senses. At the bottom are discontinuity subcategories (breakthrough, termination, interruption), while two noncausative-dominant senses (disclosure and emergence) occupy the far left. Senses with distinct syntactic preferences are clearly separated in the semantic space.

Figure 14. PCA scatter plot of BERT embeddings of *break* colored by sense




Let us now examine the distribution of causative and noncausative embeddings in the BERT semantic space. Figure 15 is a PCA scatter plot of embeddings, colored by syntactic realization (red = causative, blue = noncausative) and shaped by sense category. This plot shows that causative uses (red) occupy much of the embedding space, forming dense clusters, while noncausative uses (blue) are concentrated in smaller, distinct regions. Limited areas of

overlap reflect senses shared across realizations, whereas non-overlapping zones correspond to senses with strong syntactic preferences. From the distribution of causative and noncausative embeddings observed in this figure, we may further infer that PC1 captures differences in causer salience, contrasting higher causer salience on the left with lower causer salience on the right.

Figure 15. PCA scatter plot of BERT embeddings of *break* colored by syntactic realization

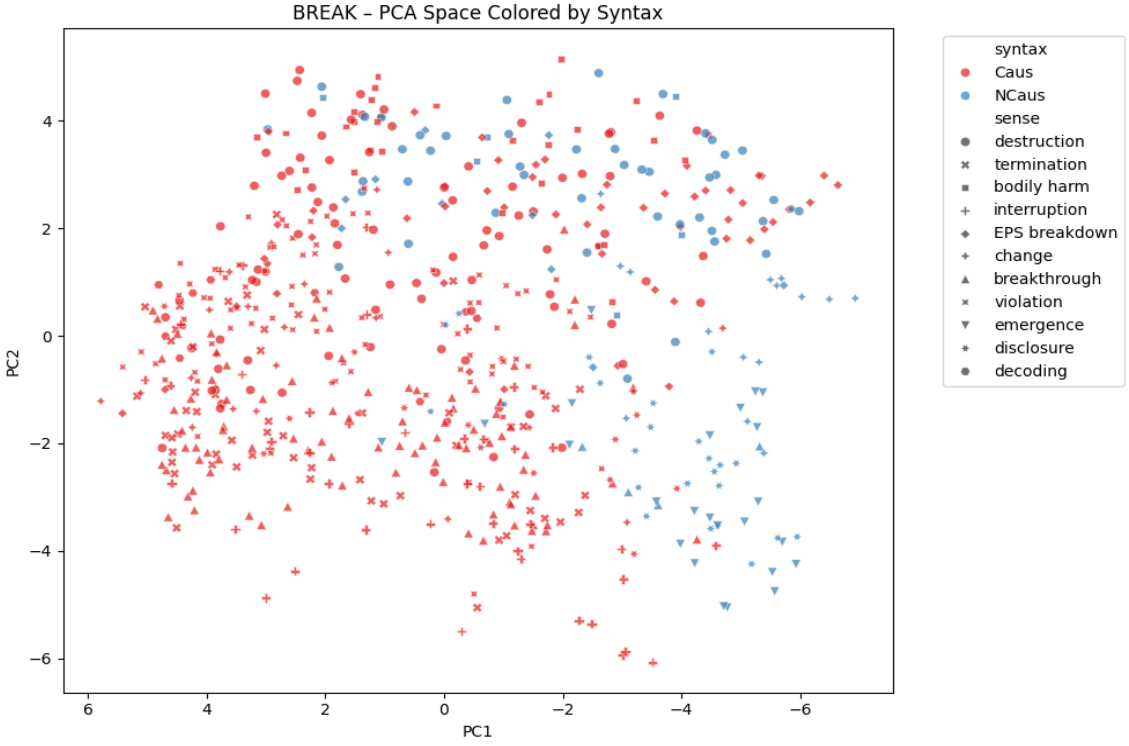
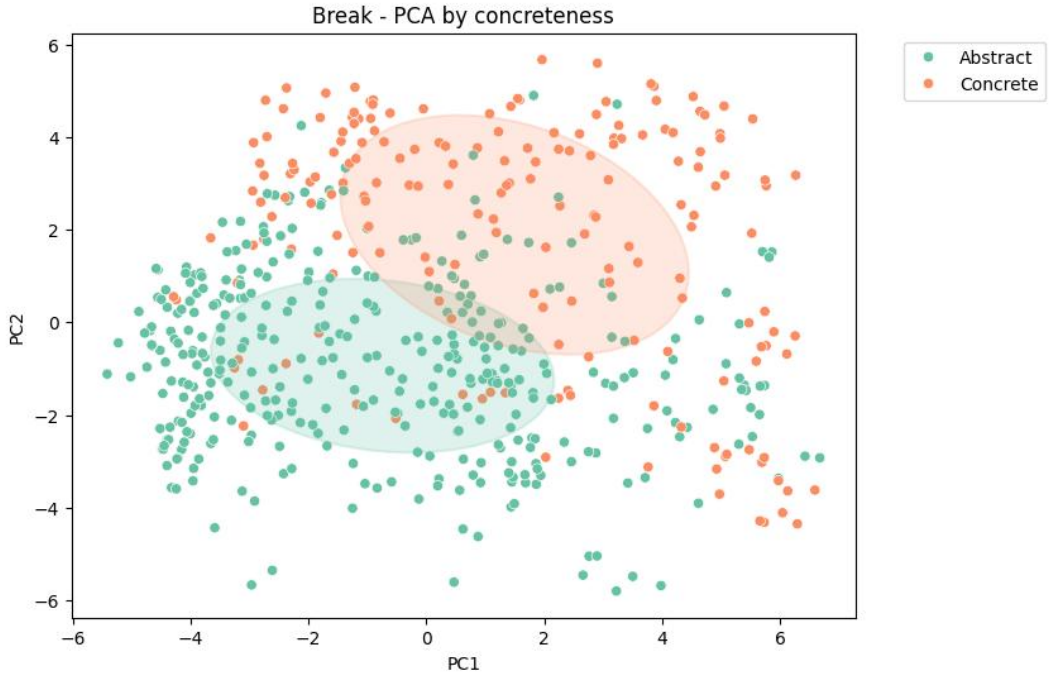



Figure 16 presents the distribution of break embeddings according to the concreteness of the theme argument. In this plot, concrete themes overlap substantially with the embedding distribution of noncausative uses observed in Figure 15, whereas abstract themes show considerable overlap with the embedding positions of causative uses. The clear separation of concrete and abstract themes along PC2 suggests that PC2 captures the semantic dimension of theme concreteness, with higher positions corresponding to concrete themes and lower positions corresponding to abstract ones.

Figure 16. PCA scatter plot of BERT embeddings of *break* colored by theme concreteness




Applying color coding to other variables, the BERT-assisted analysis demonstrates how BERT embeddings validate findings of the corpus-based studies of variables distinguishing between causative and noncausative uses. Figures 17 and 18 are PCA scatter plots of BERT embeddings of causative and noncausative uses, respectively, colored by the intentionality variable. A comparison of these two plots reveals a clear separation between causative and noncausative uses in terms of causer intentionality. In Figure 17 (causative uses), intentional causer types—agent andintentional causing act—are densely clustered in the central and left regions of the semantic space, overlapping only partially with nonintentional types (Cause_inan and Cause_anim). By contrast, Figure 18 (noncausative uses) shows a predominance of nonintentional causer types (Cause_inan, Cause_anim, and ambiguous categories such as Cause_anim/inan and Agent or Cause_anim), which are concentrated in the lower and right regions of the space, with intentional types appearing only sparsely and at the periphery. This distributional contrast visually confirms that causer intentionality plays a crucial role in distinguishing between the two syntactic realizations, as demonstrated by Kim et al. (2025) and Lee (2025).

Figure 17. PCA scatter plot of BERT embeddings of causative uses of *break* colored by causer intentionality

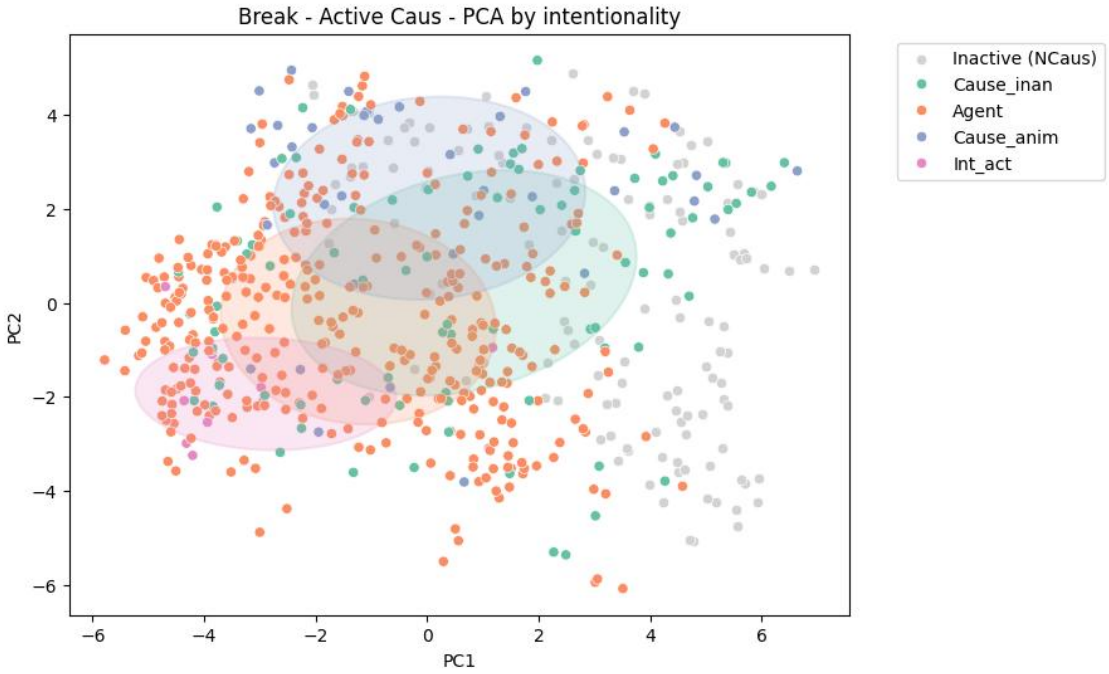
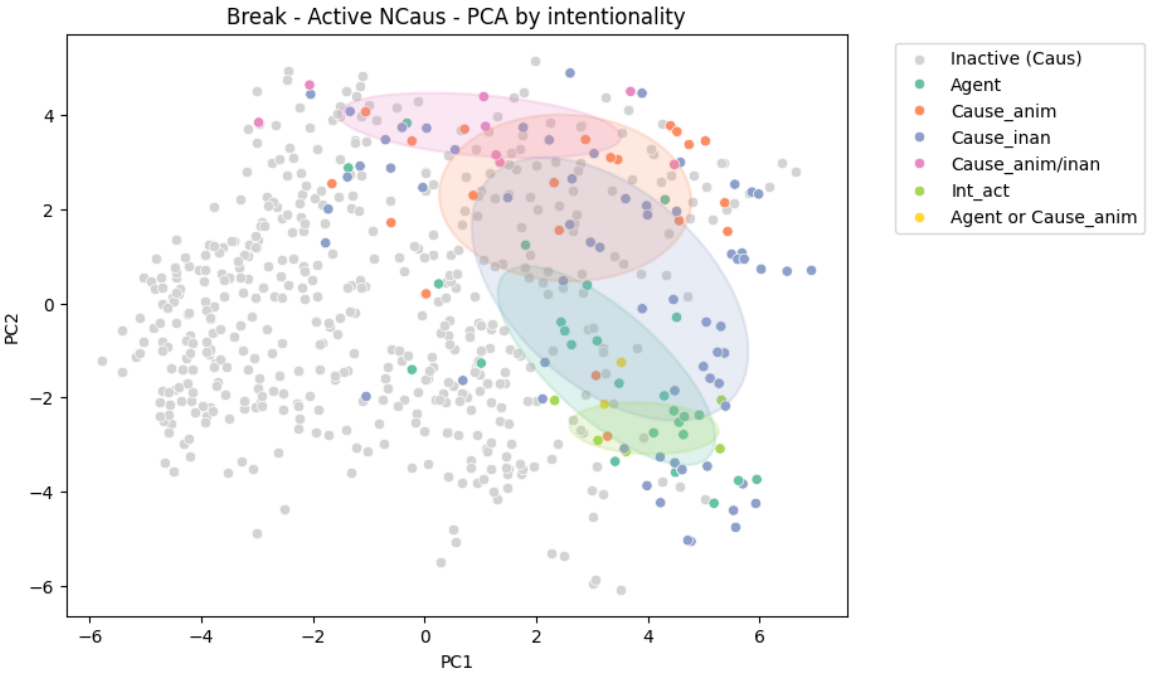



Figure 18. PCA scatter plot of BERT embeddings of noncausative uses of *break* colored by causer intentionality




Let us now examine the distributions of BERT embeddings of causative and noncausative uses in terms of causer identifiability. A comparison of Figures 19 and 20 reveals that causative and noncausative uses are more distinctly separated when the embeddings are color-coded by causer identifiability.

Figure 19. PCA scatter plot of BERT embeddings of causative uses of *break* colored by causer identifiability

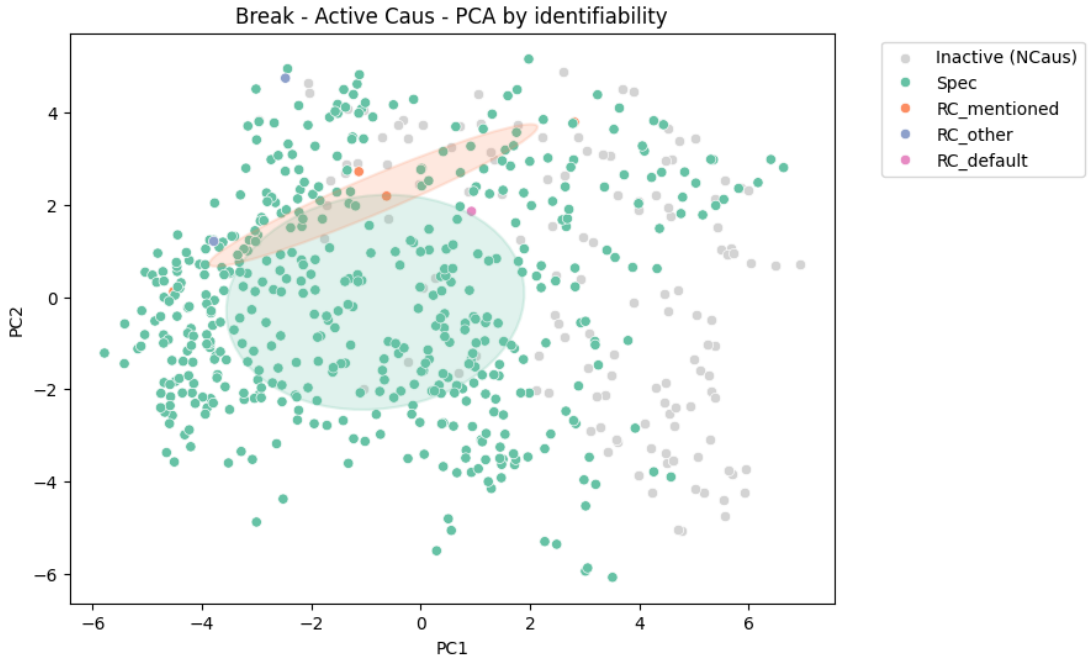
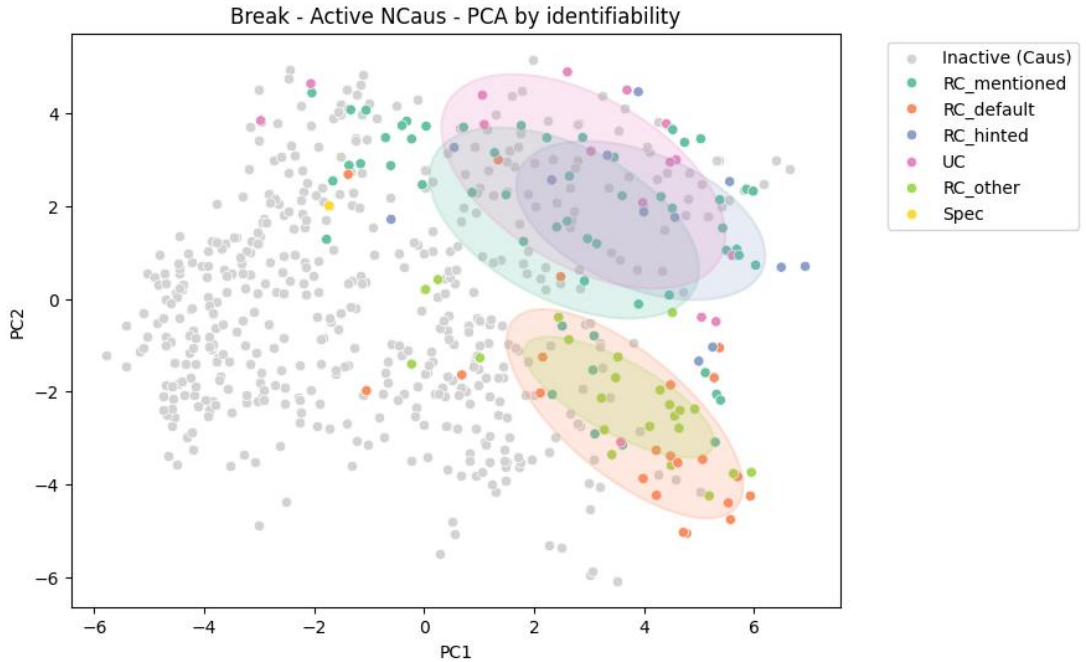



Figure 20. PCA scatter plot of BERT embeddings of noncausative uses of *break* colored by causer identifiability




In the causative plot (Figure 19), specified causers—explicitly identified agents responsible for the change of state—are the dominant type, with only limited overlap with recoverable causers. By contrast, noncausative uses are primarily associated with lower-identifiability categories, namely recoverable and unknown causers. The noncausative plot (Figure 20) further reveals that RC_mentioned, RC_hinted, and UC cluster in the upper region, whereas RC_other and RC_default are concentrated in the lower region. Compared with the causative distribution,

this sharper division between specified causers and lower-identifiability types highlights the strong discriminative role of causer identifiability in distinguishing the two variants—confirming earlier findings that this factor is a highly robust predictor (Kim et al. 2025; Lee 2023, 2025).

A corpus analysis reveals that causative and noncausative uses, even when interpreted in the same sense, show a clear difference in the degree of causer identifiability (for a more detailed discussion, see Lee (2025)). Provided below in (17) and (18) are examples of causative and noncausative uses from one of the causative-dominant senses: physical/material/structural destruction. While the causative example in (17) describes the destruction of a water pipe caused by floating debris in a river, the noncausative example in (8), repeated here as (18), describes the destruction of the measuring equipment that occurred due to an unclear external cause.

(17) … Los Gatos Creek near Coalinga, which was caused by <floating debris that broke a pipeline used by Chevron USA> to transport crude from the Bakersfield area to its Richmond refinery …

(COCA NEWS: SanFranChron-1995)

(18) When the 12 Japanese transmissions were tested, <an engineer reported that the measuring equipment had broken>.

(COCA MAG: Smithsonian-1990)

Just like the examples of physical destruction, the causative use in (19) and the noncausative use in (20) also exhibit a distinct difference in causer identifiability, even though both employ *broke* in the same sense of ‘alleviation' or ‘weakening' (change sense category). The causative example in (19) explicitly identifies the causer as ‘my shoe', while the noncausative example in (20) leaves the cause of the fever's breaking to be inferred as a default cause, representing several simultaneously-acting factors.

(19) I reached out for her and her face hit my shoe. It was actually <my shoe that broke her fall>. (COCA NEWS: Houston Chronicle-19930613)

(20) After a few restless nights and bottle after bottle of tea, the boy's coughs gradually softened and <his fever broke>.

(COCA FIC: The Dalhousie Review-2018)

Figures 21 and 22 present PCA scatter plots of BERT embeddings of causative and noncausative uses, respectively, colored by theme concreteness. In Figure 21 (causative uses), concrete themes substantially overlap with the distribution of destruction and bodily harm senses located in the uppermost region of Figure 14, while abstract themes align with the left-central and lower regions that correspond to causative-dominant senses. In Figure 22 (noncausative uses), concrete themes cluster with destruction, bodily harm, and emotional/psychological/social breakdown senses situated in the upper-left region of Figure 14, as well as with the emergence sense, which is separately distributed in the far lower-left region. Abstract themes, by contrast, overlap with the change sense distributed across the left-central and lower regions, and also intersect with the disclosure sense occupying an intermediate zone

between change and emergence. The fact that abstract themes are markedly more dominant in the causative uses plot corroborates the findings of Lee (2025), whose corpus-based analysis demonstrated that theme concreteness is one of the significant factors distinguishing the two variants.

Figure 21. PCA scatter plot of BERT embeddings of causative uses of *break* colored by theme concreteness

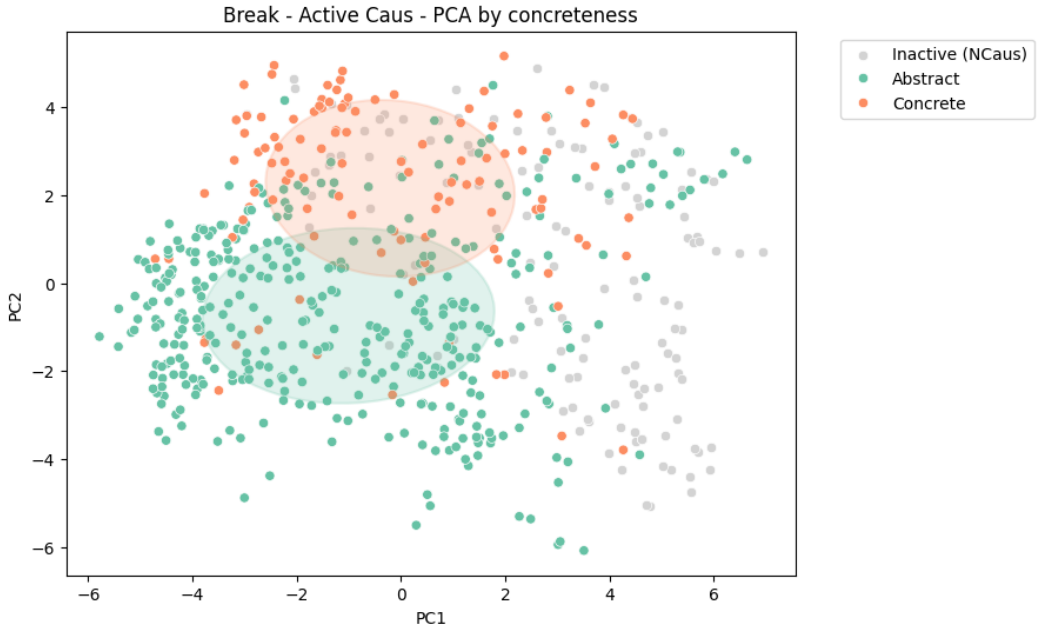
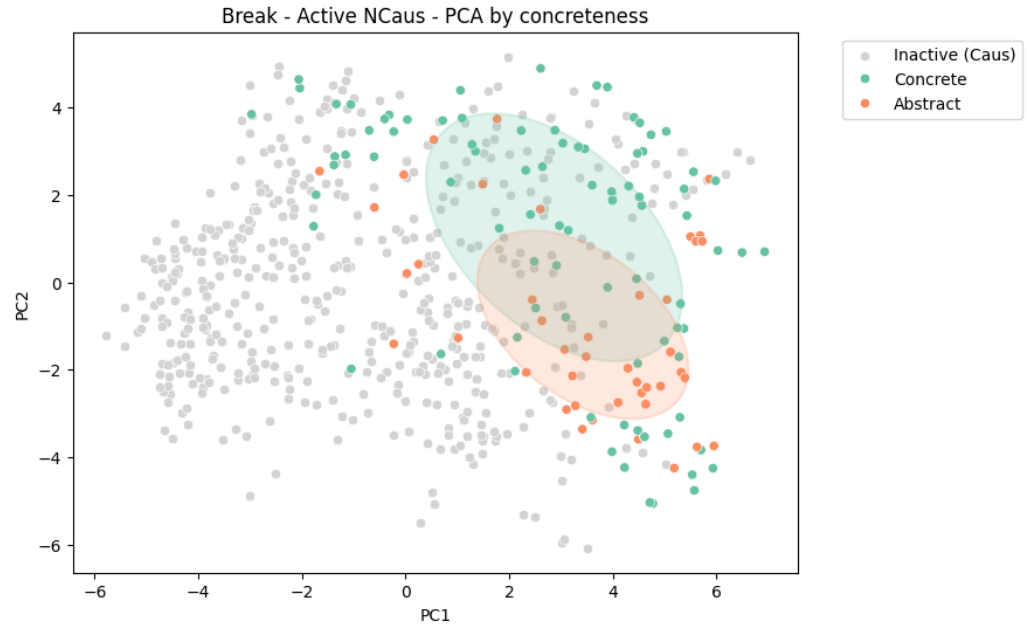



Figure 22. PCA scatter plot of BERT embeddings of noncausative uses of *break* colored by theme concreteness




Overall, the distributions in Figures 17-22 align with the results from the corpus studies, where causer intentionality and identifiability and theme concreteness emerge as the primary explanatory variables of variance.

Let us now turn to the distributions of sense embeddings in terms of causer intentionality and identifiability. Figures 23 and 24 show PCA plots for the five *break* senses most strongly skewed toward the causative variant, colored by intentionality and identifiability, respectively. In Figure 23, agent dominates, with smaller inanimate and inanimate causer clusters, while intentional act is rare. In Figure 24, specified causer leads, followed by smaller RC_mentioned and RC_other clusters. Statistical tests show that for intentionality, ANOVA on PC1 is non-significant ($p$ = 0.1284), but MANOVA on PC1–PC2 is significant ($p$ = 0.0228), indicating multivariate separation. For identifiability, both ANOVA ($p$ = 0.0005) and MANOVA ($p$ = 0.0080) are significant, confirming that identifiability strongly shapes the embedding space.

Figure 23. PCA scatter plot of BERT embeddings of strongly causative-skewed senses of *break* colored by causer intentionality

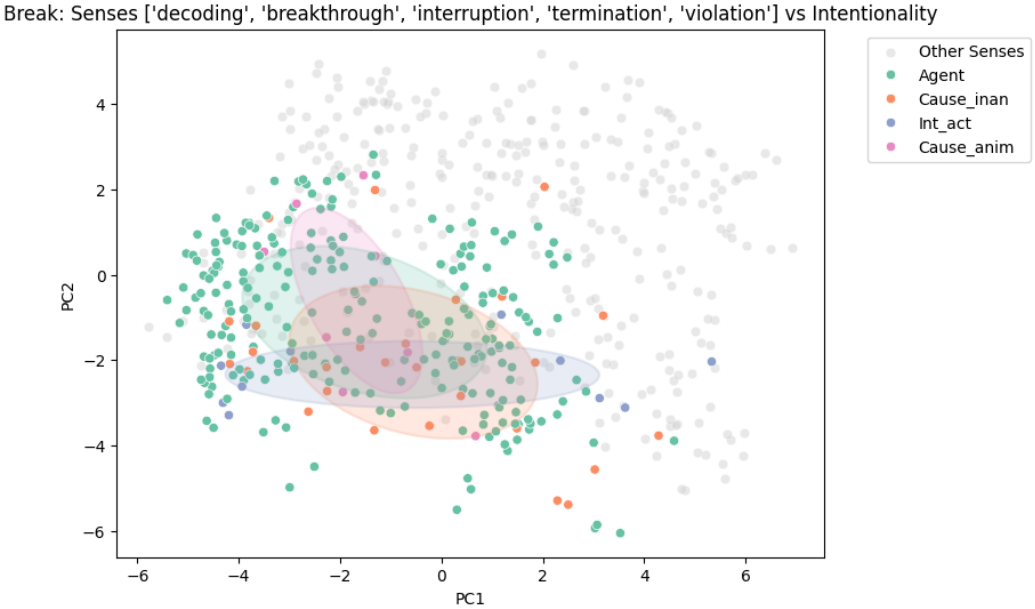
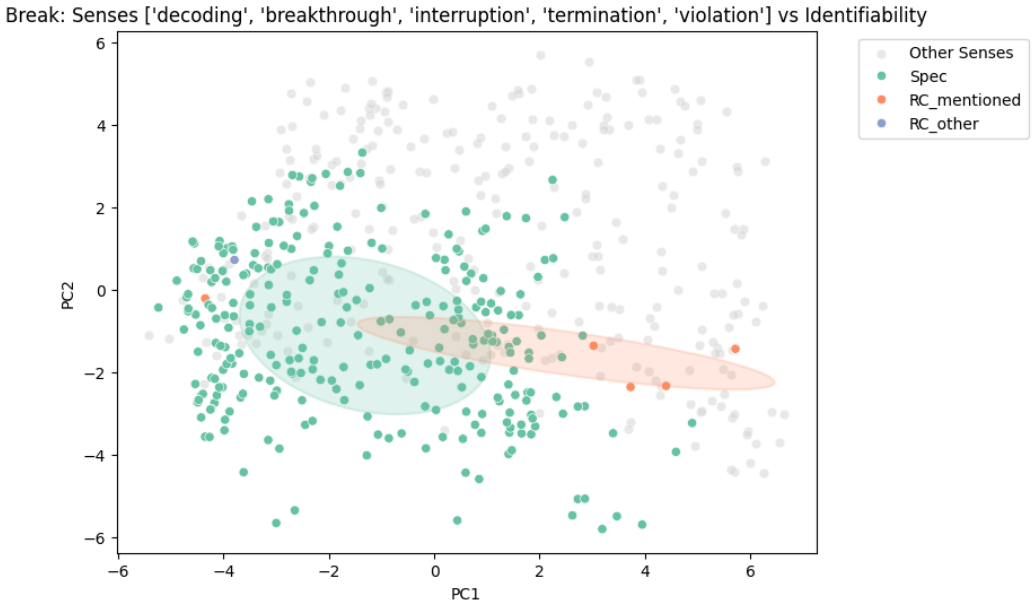



Figure 24. PCA scatter plot of BERT embeddings of strongly causative-skewed senses of *break* colored by causer identifiability



Figures 25 and 26 show PCA visualizations of BERT embeddings for *break* senses with a moderate causative-preference pattern (bodily harm, destruction, EPS breakdown, change), colored by causer intentionality (Figure 25) and identifiability (Figure 26). In Figure 25, intentionality levels form visually distinct clusters, a pattern supported by significant effects in ANOVA ($p$ = 0.0004) and MANOVA ($p$ < 0.0001). In Figure 26, identifiability levels also cluster clearly, with ANOVA and MANOVA both confirming highly significant effects ($p$ < 0.0001).

Figure 25. PCA scatter plot of BERT embeddings of moderately causative-skewed senses of *break* colored by causer intentionality

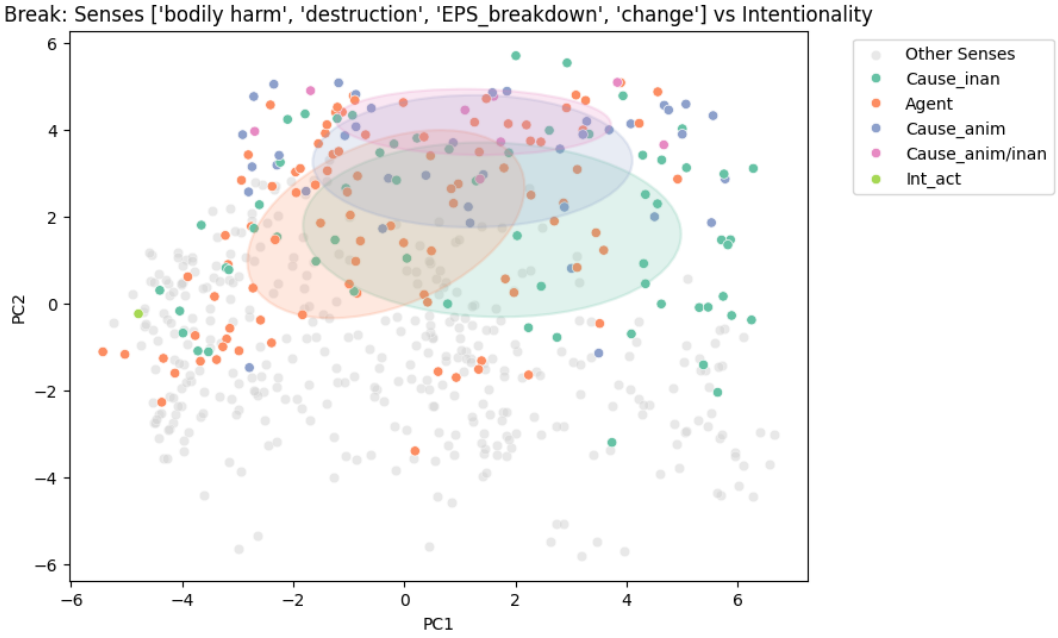
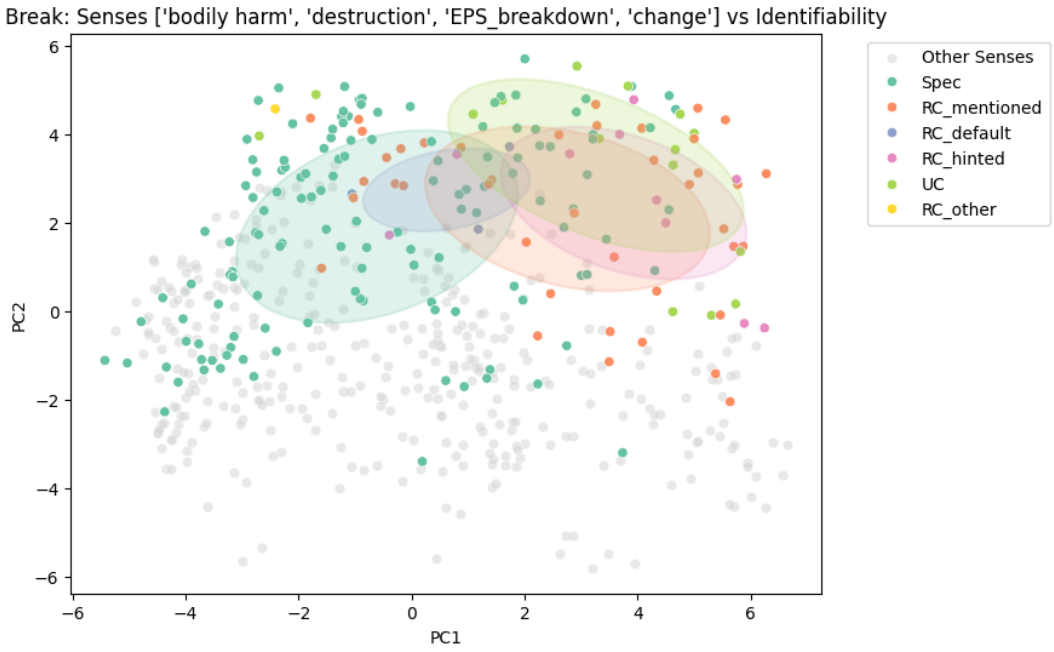


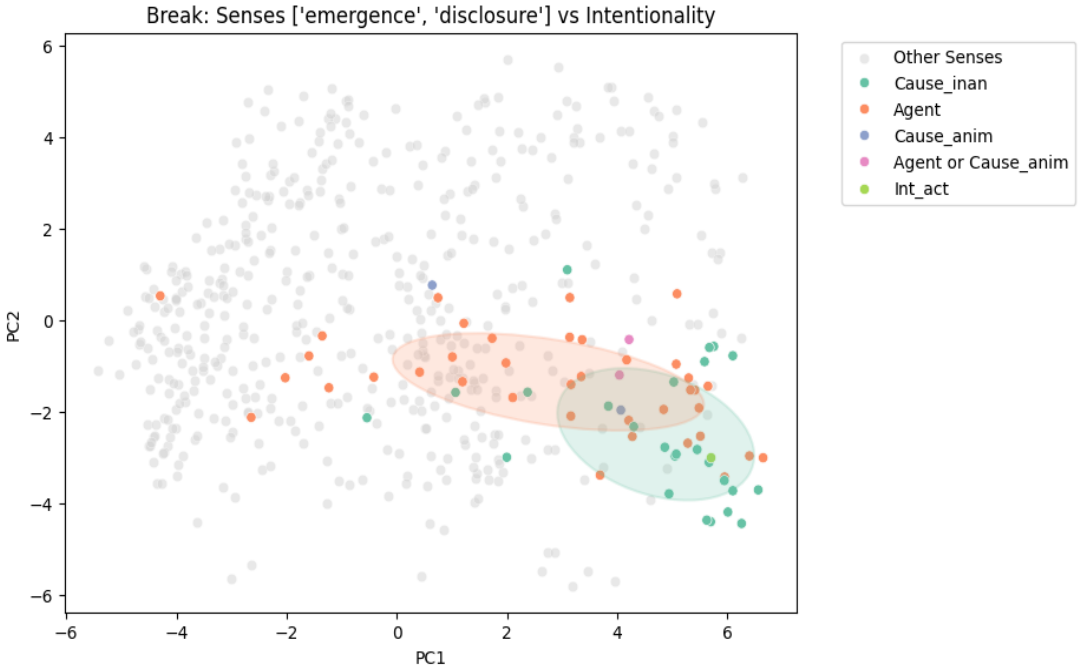

Figure 26. PCA scatter plot of BERT embeddings of moderately causative-skewed senses of *break* colored by causer identifiability



In these plots, it can be observed that high-intentionality/identifiability causer types are predominantly distributed in the right region, where the embeddings of causative uses are concentrated, whereas lower- intentionality/identifiability causer types are clustered in the left region, which is dominated by the embeddings of noncausative uses. These patterns indicate that both causer intentionality and identifiability are strong factors structuring the embedding space for these senses.

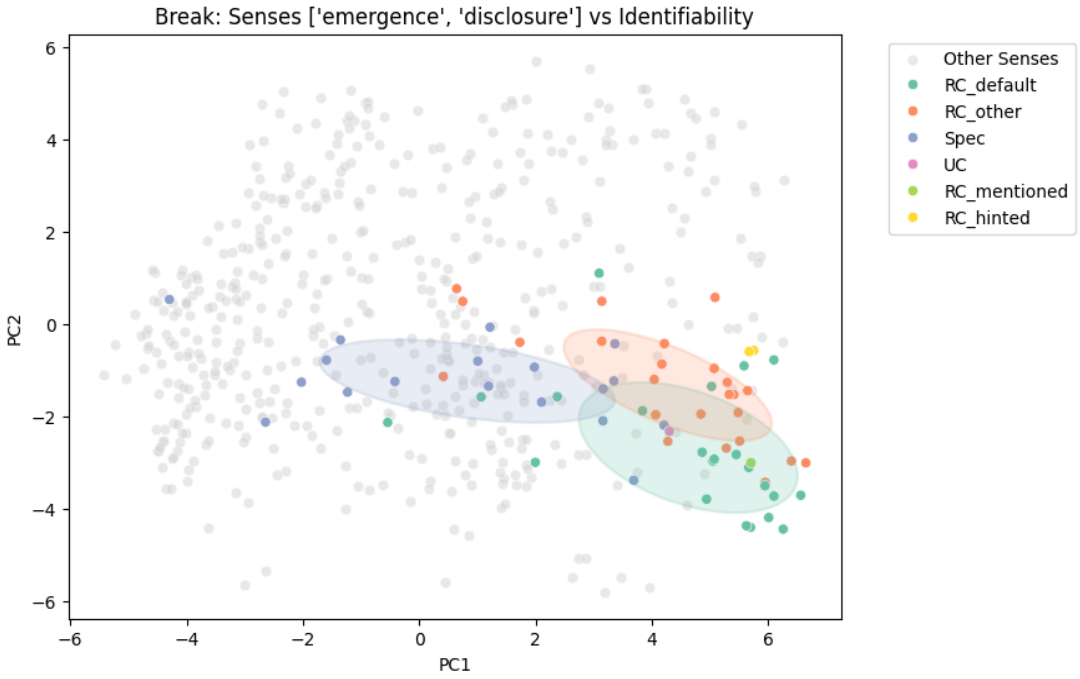
Figures 27 and 28 show PCA plots for the disclosure and emergence senses, which are skewed toward the noncausative variant, colored by intentionality and identifiability. In Figure 27, inanimate causer dominates, with a smaller agent cluster and sparse other categories. In Figure 28, RC_default is most frequent, alongside RC_other and Spec. Statistical tests confirm these patterns: for intentionality, ANOVA ($p$ = 0.0235) and MANOVA ($p$ = 0.0152) indicate significant effects, while for identifiability, both ANOVA and MANOVA ($p$ < 0.0001) show strong effects. These results demonstrate that even in noncausative-dominant senses, causer intentionality and identifiability significantly shape the embedding space.

Figure 27. PCA scatter plot of BERT embeddings of noncausative-dominant senses of *break* colored by causer intentionality




In Figure 27, the cluster location of inanimate causers almost exactly matches that of the (natural) emergence sense in Figure 14, while the location of the agent cluster almost perfectly aligns with the cluster location of the disclosure sense. In Figure 28, the position of RC_default almost perfectly aligns with the cluster location of the emergence sense in Figure 14. Additionally, RC_other (type-inferable causer) is located in a region that almost entirely overlaps with the noncausative embeddings of the disclosure sense, while specified causers are found in nearly the same area as the causative embeddings of the disclosure sense.

Figure 28. PCA scatter plot of BERT embeddings of noncausative-dominant senses of *break* colored by causer identifiability




Examples of the causative and noncausative uses of *break* interpreted in the disclosure sense are provided below in (21) and (22). In the causative example in (21), Currie, who was attempting to contact Lewinsky to report news about her, is explicitly mentioned as the agent. In contrast, in the noncausative example in (22), the subject of the news report is not explicitly mentioned or clearly identified, and is thus inferred to be a person who typically performs the act of reporting (i.e., a reporter).

(21) By either interpretation, Currie testified that <she was trying to reach Lewinsky to break the news> that her name surfaced during Clinton's deposition.

(COCA NEWS: Chicago-1998)

(22) … they considered Woods less of a role model since <news of his extramarital affairs broke in November>.

(COCA NEWS: USA Today-2010)

The distributions of *break* sense embeddings with respect to causer intentionality and identifiability, as shown in Figures 23–28, can be summarized as follows: First, the five senses most strongly skewed toward the causative variant are overwhelmingly associated with higher-intentionality and higher-identifiability causer types. Second, the two non-causative dominant senses are dominated by lower-intentionality and lower-identifiability causer types. Third, the four senses that are only moderately skewed toward the causative variant exhibit a more balanced distribution, where higher- and lower-intentionality/identifiability causer types are clearly separated but relatively evenly represented. This overall pattern aligns well with the predictions articulated in Section 2.2.2: causative-dominant senses correlate with causer

properties that enhance the informativeness of the causative construction, whereas noncausative-dominant senses are tied to causer properties that reduce such informativeness.

Let us now examine the distributions of *break* sense embeddings with respect to theme concreteness. Across the three sets of analyses (Figures 29–31), consistent patterns emerge linking theme concreteness with sense-specific distributions of *break*. Figure 29 presents a PCA scatter plot of BERT embeddings of the causative-dominant senses of *break*, colored by theme concreteness. Although abstract themes are more prevalent in this subset, the separation between abstract and concrete themes is not pronounced. This observation is supported by the statistical tests: the *t*-test for PC1 shows no significant effect of concreteness ($p = 0.4545$), and the MANOVA for both PC1 and PC2 likewise indicates no significant contribution of concreteness (Wilks' $\lambda = 0.9971$, $p = 0.6649$). Thus, while the plot visually suggests a predominance of abstract themes, concreteness does not emerge as a statistically reliable factor distinguishing embeddings in these causative-dominant senses.

Figure 29. PCA scatter plot of BERT embeddings of strongly causative-skewed senses of *break* colored by theme concreteness

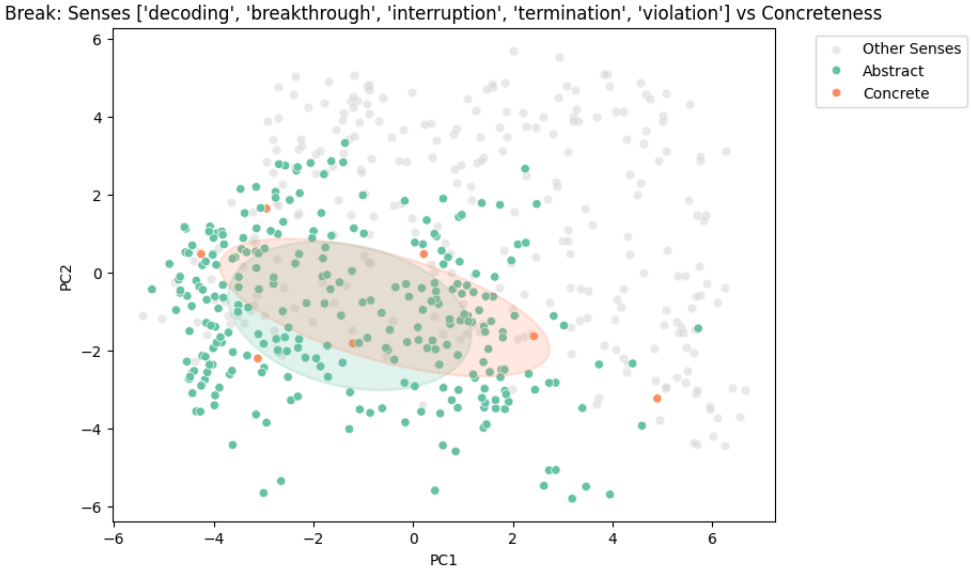
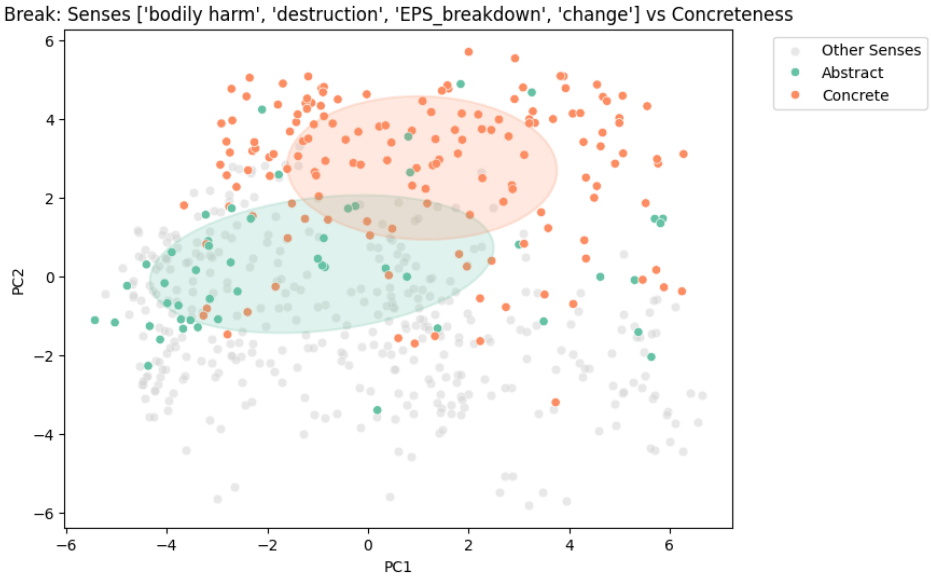


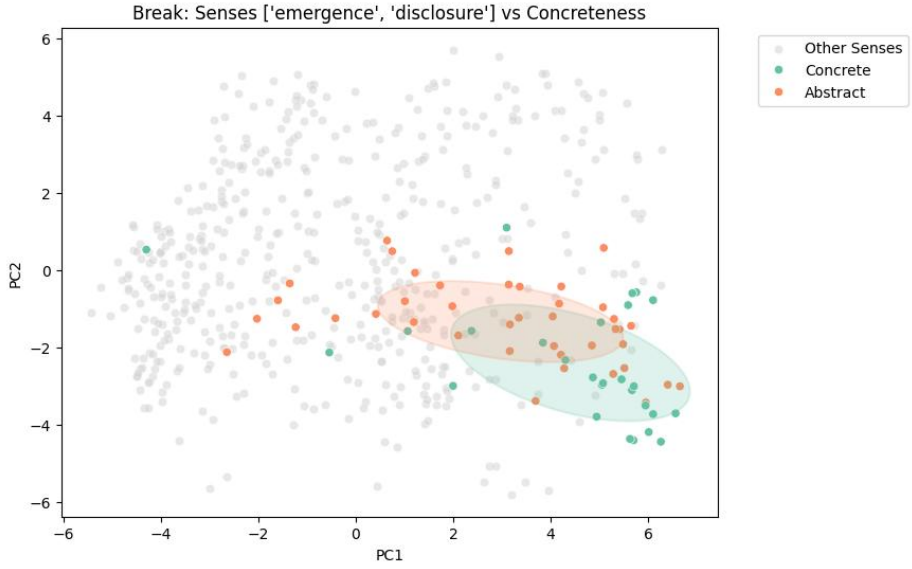

Figure 30 presents the PCA scatter plot of BERT embeddings for the moderately skewed causative-dominant senses of *break* (bodily harm, destruction, EPS breakdown, and change), colored by theme concreteness. Unlike the strongly skewed senses (Figure 29), here the distributions of concrete and abstract themes are more evenly balanced, yet still exhibit a notable separation pattern. Concrete themes (orange) tend to cluster in the upper region of PC2, while abstract themes (green) are more concentrated in the lower and central regions, indicating a systematic contrast along the vertical axis. This impression is statistically supported. A *t*-test for PC1 shows a highly significant effect of concreteness ($p < 0.001$), and MANOVA including both PC1 and PC2 likewise confirms a strong overall effect of concreteness (Wilks' $\lambda = 0.704$, $F = 41.77$, $p < 0.001$). These results demonstrate that for the moderately skewed causative-dominant senses, concreteness is a reliable factor structuring the embedding space, with PC2 in particular capturing the contrast between abstract and concrete themes.

Figure 30. PCA scatter plot of BERT embeddings of moderately causative-skewed senses of *break* colored by theme concreteness




Finally, for the noncausative-dominant senses (emergence and disclosure), the concreteness distinction aligns neatly with the semantic profiles of the senses themselves. As shown in Figure 31, embeddings annotated as concrete cluster in the region associated with the emergence sense in Figure 14, reflecting its link to perceptible, externally observable changes. By contrast, embeddings annotated as abstract are concentrated in the area corresponding to the disclosure sense, consistent with its orientation toward conceptual or informational revelation rather than physical change. Statistical tests again confirm that concreteness significantly differentiates embeddings along the PCA dimensions (*t*-test for PC1: $p = 0.0232$; MANOVA for PC1 and PC2: Wilks' $\lambda = 0.8380$, $p = 0.0032$).

Figure 31. PCA scatter plot of BERT embeddings of noncausative-dominant senses of *break* colored by theme concreteness



Taken together, these results indicate that while causer-related properties remain primary discriminators between causative and noncausative variants, theme concreteness systematically conditions the distribution of verb senses, with abstract themes especially prevalent in strongly causative-skewed senses and causative change contexts and concrete themes salient in noncausative-physical breaking and emergence contexts.

In sum, the analysis of *break* presented in this section confirmed its strong bias toward the causative variant, with nine senses classified as causative-dominant and only two (emergence, disclosure) as noncausative-dominant. PCA visualizations showed that PC1 reflects causer salience and PC2 captures theme concreteness, while distributional patterns aligned with the central hypothesis: causative-dominant senses clustered with intentional and identifiable causers, whereas noncausative-dominant senses were associated with nonintentional or less identifiable causers. Abstract themes were especially prevalent in strongly causative-skewed senses, while concrete themes dominated in noncausative- physical breaking and emergence contexts.

These results not only corroborate predictions about sense–causer associations but also set the stage for the analysis of *freeze*, whose distributional profile stands in sharp contrast to *break* in its stronger affinity with the noncausative variant.

### 2.2.4. The Syntactic and Semantic Distribution of *Freeze* Senses

We now turn to the syntactic and semantic distribution of *freeze* senses. Figure 32 compares the proportional distribution of causative and noncausative realizations across the two verbs examined in this study. As discussed in Section 2.2.3, *break* strongly favors the causative variant, with nearly four times as many causative tokens as noncausative ones. By contrast, *freeze* displays the opposite tendency: noncausative tokens (51.4%) slightly outnumber causative tokens (48.6%), indicating that *freeze* is more evenly balanced between the two variants and, if anything, leans toward the intransitive side of the alternation.

Figure 32. Proportional distribution of causative alternation variants

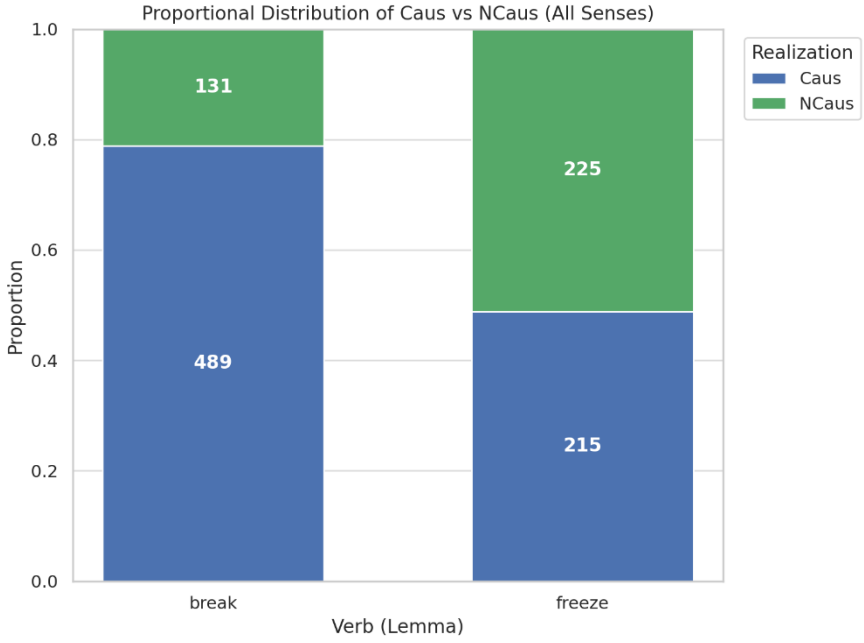
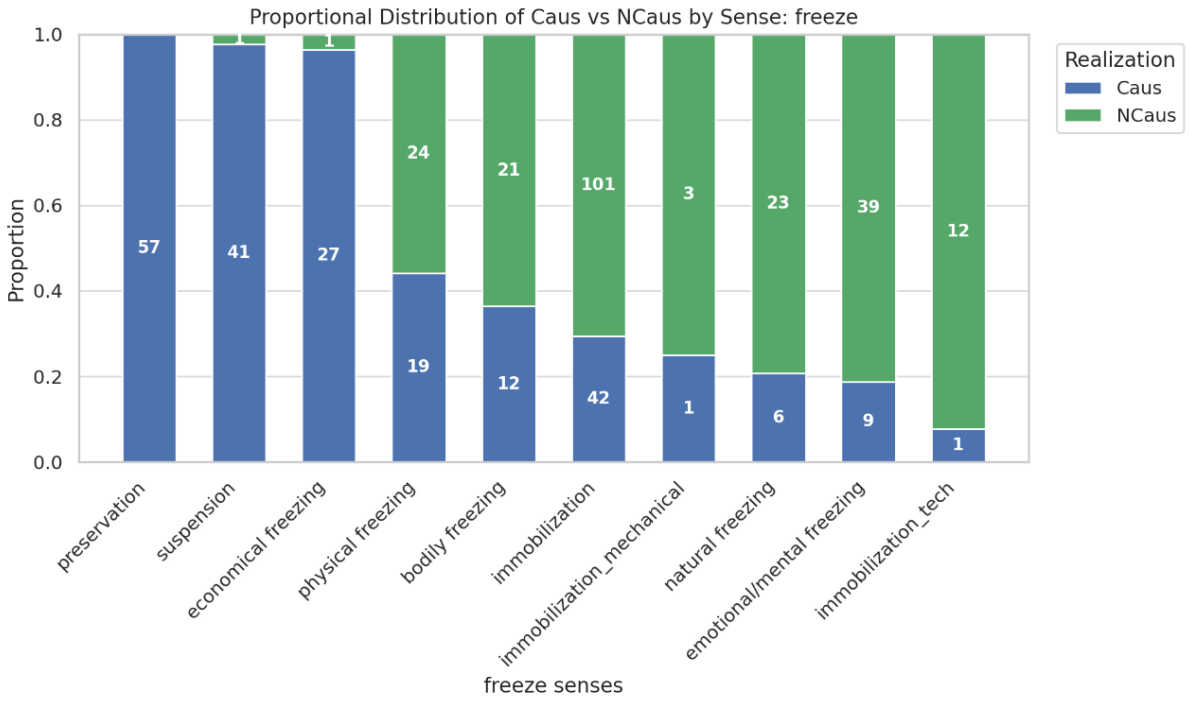


Figure 33 presents the proportional distribution of the individual senses of *freeze* across the two alternation variants. The results show a clear asymmetry: while three senses—preservation, suspension, and economical freezing—are strongly skewed toward the causative variant, the majority of the other senses show a strong preference for the noncausative variant. This pattern confirms that, unlike *break*, which tends to privilege the causative realization across most senses, *freeze* demonstrates a systematic skew toward the noncausative realization, with only a small subset of senses anchoring the causative side of the alternation.

Figure 33. Proportional distribution of the *freeze* senses in the alternation variants



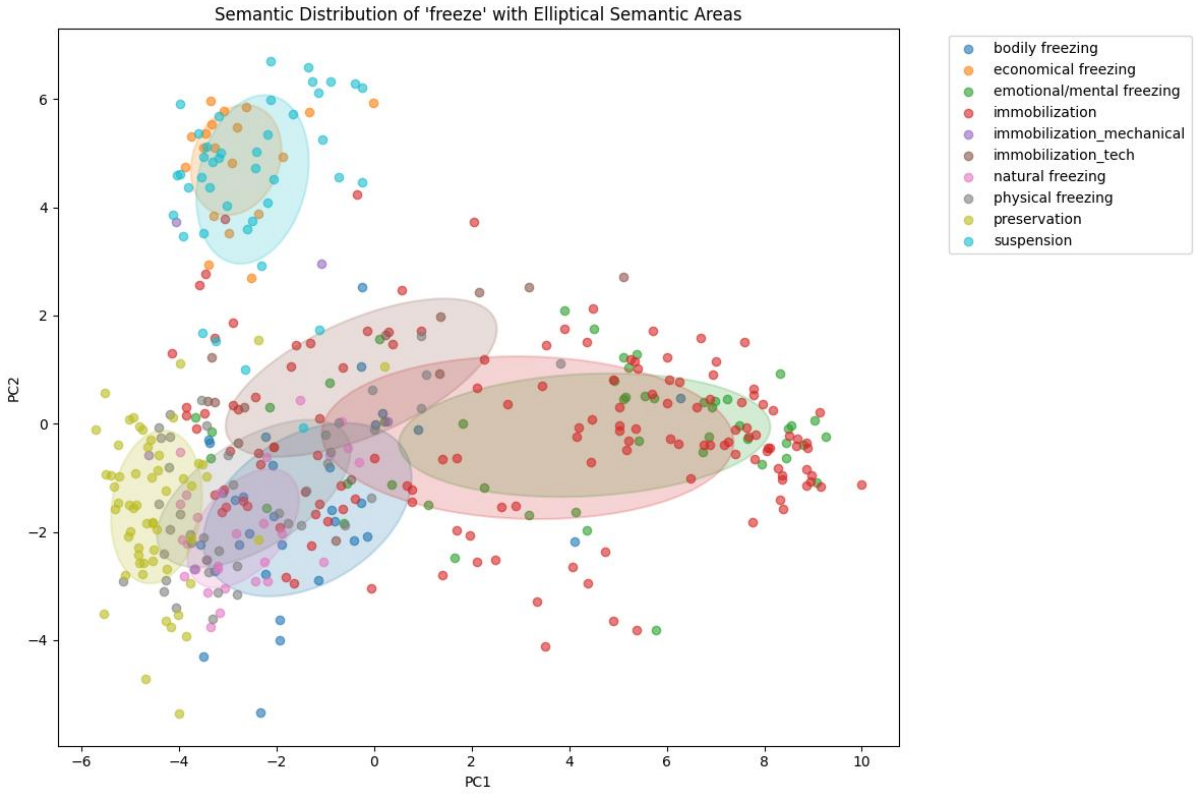

This figure's sense-level distributions enable us to classify the senses of *freeze* into causative-dominant and noncausative-dominant categories, as shown in Table 10.

Table 10. Constructional distribution of the *freeze* senses

| Constructional distribution | Senses |
|---|---|
| Causative-dominant | preservation, suspension, economical freezing |
| Noncausative-dominant | physical freezing, bodily freezing, natural freezing, physical immobilization (immobilization), mechanical breakdown (immobilization_mechanical), technical failure (immobilization_tech), emotional/mental freezing |

Figure 34 displays the semantic distribution of *freeze* embeddings in the PCA space, with points colored by sense categories and ellipses indicating major semantic areas. Similar to the distributional patterns observed for *break* senses, the plot reveals a differentiation along PC1 and PC2 between concrete, physical change-of-state senses and more abstract, derived meanings whose syntactic distributions diverge markedly.

Figure 34. PCA scatter plot of BERT embeddings of *freeze* colored by sense




Along PC1, the lower-left quadrant is dominated by the basic senses of *freeze*. At the far left, the preservation sense is located, followed by natural freezing and then bodily freezing. The bodily freezing cluster overlaps the transition zone toward sudden immobility, *while physical freezing* occupies the central region of the plot, serving as a bridge between the basic freezing senses and the immobilization cluster. On the far right, physical immobilization and emotional/mental immobilization overlap considerably, suggesting the possibility of sense blending. Notably, emotional/mental freezing is concentrated further to the right, highlighting its semantic distinction from physical immobilization.

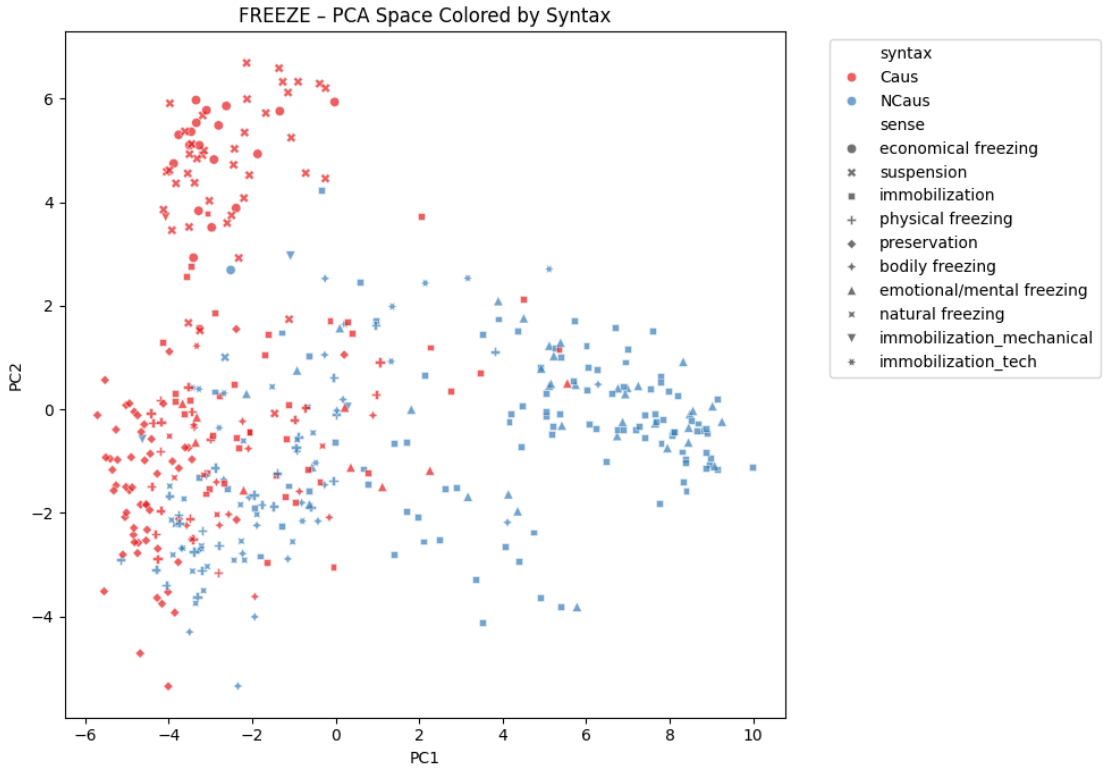
Along PC2, the lowest region of the plot houses the basic freezing senses—preservation, natural freezing, and bodily freezing—with physical freezing positioned just above them. Higher on PC2, though not emphasized with ellipses due to fewer tokens, we find the two specialized immobilization senses (mechanical breakdown and technical failure). At the very top of the distribution, the suspension and economical freezing senses form a distinct cluster. The extensive overlap between these two senses further suggests their close semantic affinity within the broader cessation category.

Notably, the explained variance for the first two principal components is higher for *freeze* than for *break*: PC1 accounts for 19.72% of the variance and PC2 for 6.59%, together explaining 26.31% of the variance—substantially above the <15% accounted for by the first two dimensions in the *break* analysis. This indicates that the semantic space for *freeze* is somewhat more structured and compressible into the lower dimensions of PCA, yielding clearer separations among sense clusters.

Figure 35 presents the PCA scatter plot of *freeze* embeddings, distinguished by syntactic realization (red = causative, blue = noncausative). The causative uses are clustered

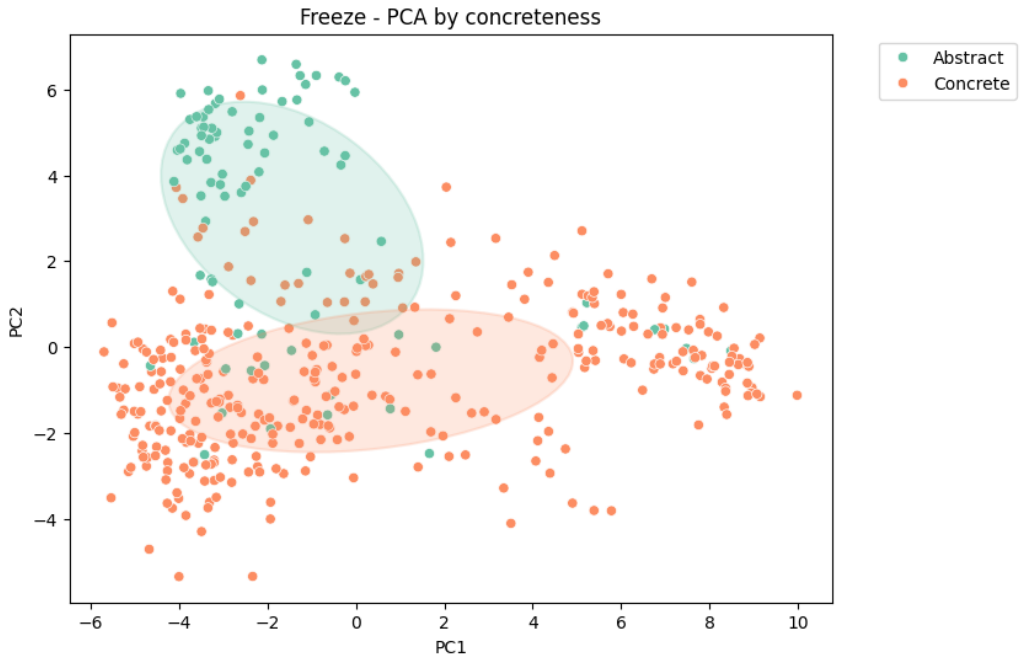
predominantly on the left side of the PCA space, while the noncausative uses spread broadly along the lower regions toward both the left and right. The overlapping area in the lower-left and central zones corresponds to the semantic core of *freeze*, where its basic freezing senses are concentrated.

Figure 35. PCA scatter plot of BERT embeddings of *freeze* colored by syntactic realization




From this distributional pattern, two principal dimensions of the PCA space can be inferred. PC1 appears to capture causer salience, contrasting higher causer salience on the left with lower salience on the right. PC2 reflects the concreteness of state change, ranging from concrete changes at the bottom to more abstract changes at the top. The distribution of concrete and abstract themes in Figure 36 further supports the interpretation of PC2 as encoding concreteness of state change. Concrete themes (orange) cluster predominantly in the lower regions of the plot, while abstract themes (green) are concentrated in the upper regions. This vertical separation mirrors the axis interpretation proposed in Figure 35, where PC2 contrasts concrete physical changes with more abstract or figurative changes. Notably, the concentration of abstract themes in the upper-left quadrant also reflects the strong association of *freeze* with abstract state changes, whereas concrete themes extend across the central and right-hand areas, corresponding to physical or bodily freezing events.

Figure 36. PCA scatter plot of BERT embeddings of *freeze* colored by theme concreteness




Taken together, Figures 34–36 illustrate the structural properties of the shared PCA semantic space for the causative alternation verbs *break* and *freeze*, as well as the ways in which their senses diverge and extend along the principal dimensions. These patterns highlight both commonalities in the organization of change-of-state verbs and verb-specific expansions into abstract or specialized domains. We now turn to an examination of the distribution of BERT embeddings of causative and noncausative uses of *freeze* with respect to causer types and theme concreteness.

Figure 37 shows that agents dominate the causative distribution, extending broadly along the upper and lower sections of the left-hand region. These overlap substantially with the nonintentional causer types (animate and inanimate causers), which are distributed slightly further to the right. A small cluster of intentional causing actions appears in the upper-left quadrant, but their frequency is comparatively low, underscoring the overwhelming prevalence of agentive causation in causative uses of *freeze*. In contrast, Figure 38 highlights that inanimate causers are the most frequent in noncausative uses, spreading widely across both the left and right regions. Other causer types appear only sparsely: agents cluster on the left, animate causers occupy the right, and intentional causing actions are centered in the middle of the plot. This distribution pattern reflects a clear shift from agent-driven causation in causative uses to inanimate-driven, lower-intentionality contexts in noncausative uses.

Figure 37. PCA scatter plot of BERT embeddings of causative uses of *freeze* colored by causer intentionality

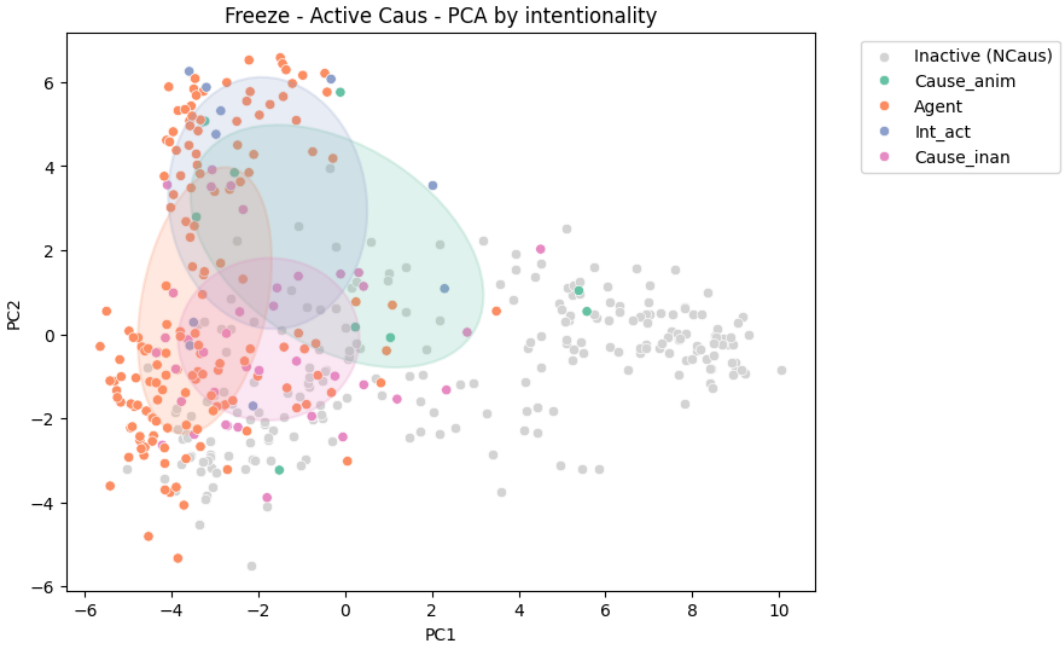
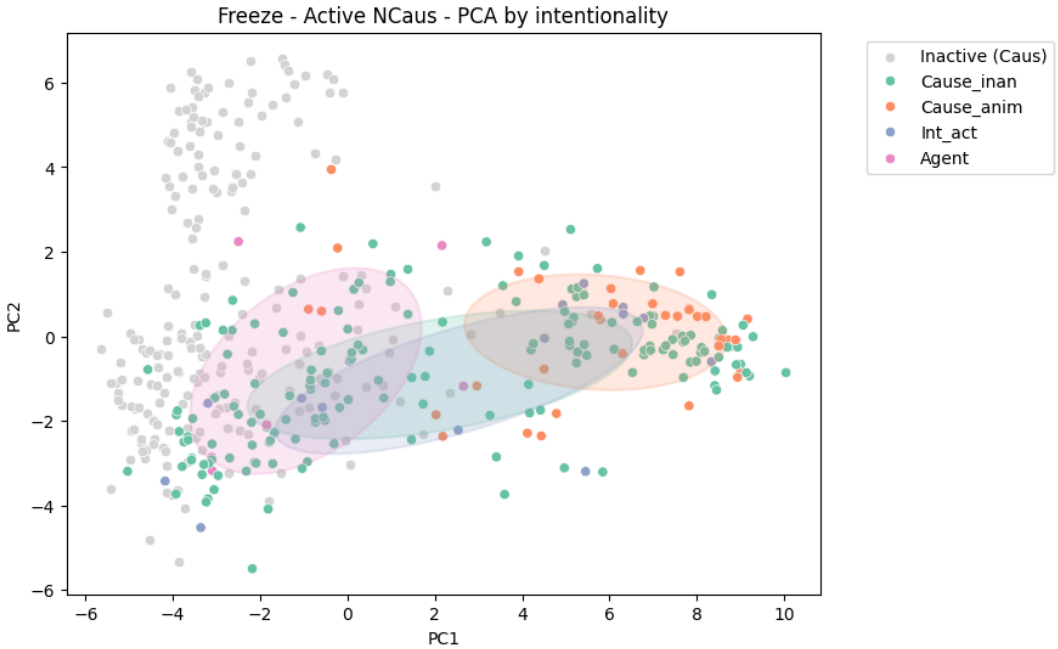


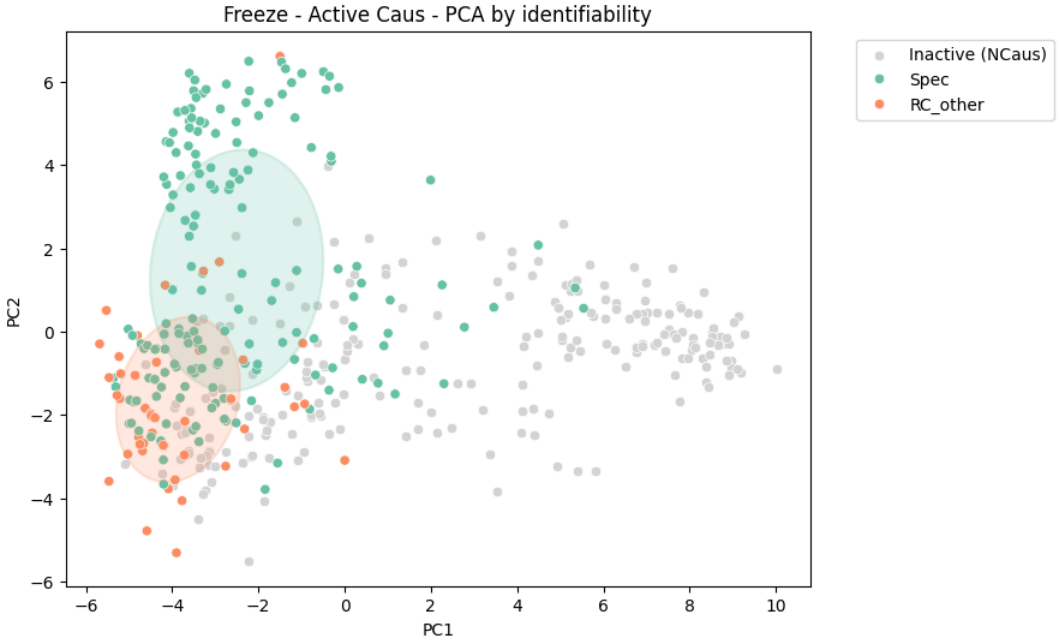

Figure 38. PCA scatter plot of BERT embeddings of noncausative uses of *freeze* colored by causer intentionality




Figures 39–40 show PCA scatter plots of BERT embeddings of *freeze* colored by causer identifiability, separated into causative and noncausative uses. As in the case of *break*, the two variants are more distinctly divided by causer identifiability than by intentionality. In Figure 39 (causative uses), specified causers clearly predominate, clustering across most of the left-hand and central regions of the plot. RC_other points are densely concentrated in the lower-left corner, where preservation senses are distributed. This clustering, however, should not be interpreted as a general property of *freeze*; rather, it reflects characteristics of the dataset,

particularly the frequent occurrence of imperative examples such as (21) in the preservation sense.

Figure 39. PCA scatter plot of BERT embeddings of causative uses of *freeze* colored by causer identifiability




(21) … to pan, and mix well. Cover and freeze until mixture is beginning to freeze at the edges, about 2 hours. Stir, cover, and freeze until …

(COCA MAG, Sunset-2009)

In Figure 40 (noncausative uses) recoverable causer (RC) types dominate overall, with UC (unknown causer) spread across the left and central regions along PC1. RC_other and default cause (RC_default)—often associated with imperative constructions and natural causes, respectively—cluster in the left-hand area. In particular, default cause aligns closely with the distribution of the natural freezing sense, exemplified in (22), where freezing arises from cold temperatures without an explicit causer.

(22) a. In 1947, <when the canal froze>, you could skate from here to Abingdon --- nearly ten miles.

(BNC W:misc, AB4-854)

b. Not very much is known about <how plants freeze>, according to ARS plant physiologist Michael Wisniewski.

(COCA ACAD: Agricultural Research-1997)

Yet unlike in *break*, where specified causers were rare in noncausative uses, *freeze* shows specified causers distributed more broadly across the horizontal axis. In our dataset, emotional/mental freezing and bodily freezing senses frequently license specified causers as adjuncts (e.g., *from*-, *with*-phrases), as exemplified in (23):

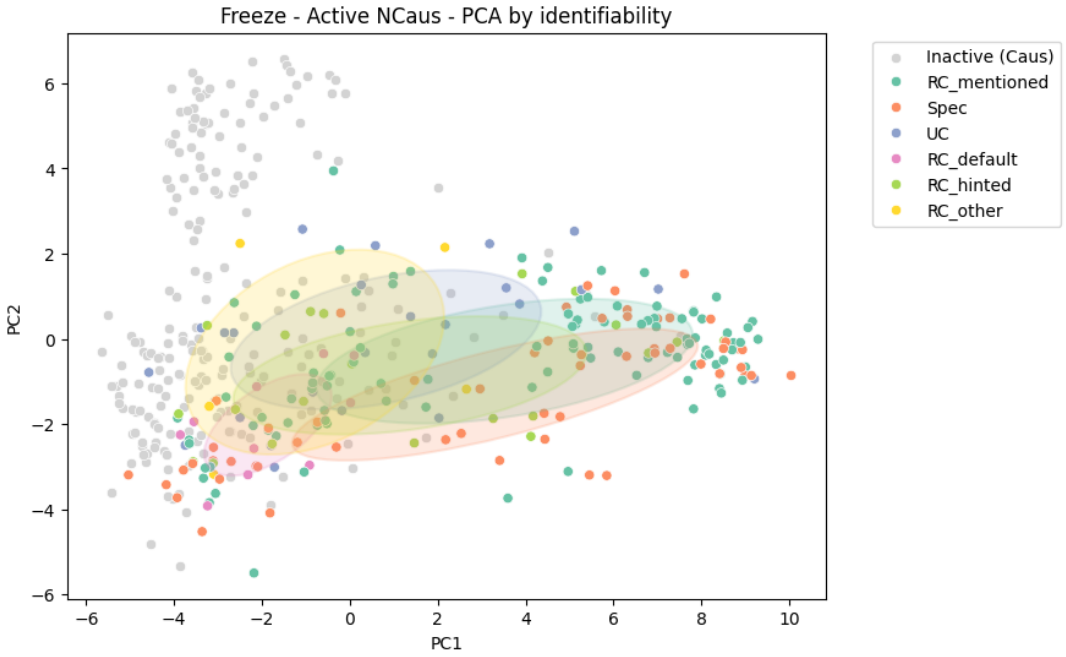
(23) a. If they run positively in the winter, <the exterior of the pile tends to freeze from condensation>.

(COCA ACAD: Canana – United States Law Journal-1996)

b. Dali's brain froze with fear.

(COCA FIC: Analog Science Fiction & Fact-2017)

Figure 40. PCA scatter plot of BERT embeddings of noncausative uses of *freeze* colored by causer identifiability




Figures 41 and 42 compare the PCA scatter plots of BERT embeddings of *freeze* according to theme concreteness, across causative and noncausative uses. In Figure 41 (causative uses), abstract themes are more frequently observed relative to noncausative uses in Figure 42, clustering especially in the upper region of the plot. These abstract embeddings are concentrated in the areas associated with suspension and economic freezing, highlighting that causative uses of *freeze* more readily extend into abstract domains. By contrast, Figure 42 (noncausative uses) shows a strong predominance of concrete themes, distributed widely across the central and right-hand portions of the PCA space. Here, abstract themes occur less frequently and are primarily tied to *emotional/mental freezing* contexts, as illustrated in (23b) and figurative immobilization contexts, as illustrated in (24). This contrast indicates that while both uses support concrete interpretations, causative uses are more prone to abstract extensions (particularly institutional or economic), whereas noncausative uses remain anchored in concrete state changes, with abstract interpretations surfacing mainly in figurative or specific psychological contexts.

(24) … push the tips of her index fingers together and <time would freeze>? Dumbest show ever, but anyway,…

(COCA FIC, Moment-2005)

Figure 41. PCA scatter plot of BERT embeddings of causative uses of *freeze* colored by theme concreteness

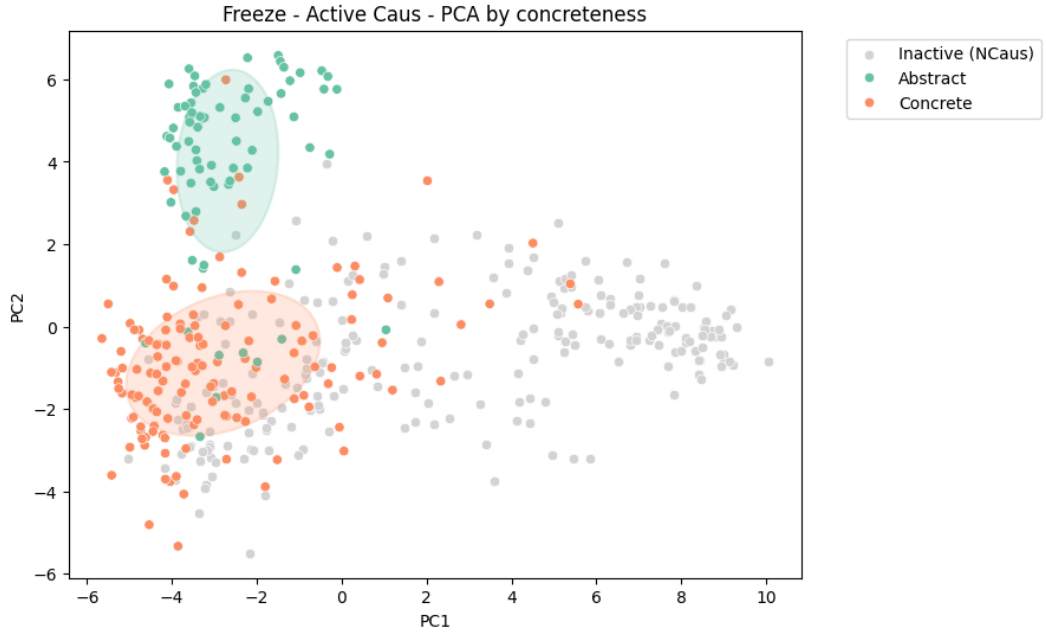
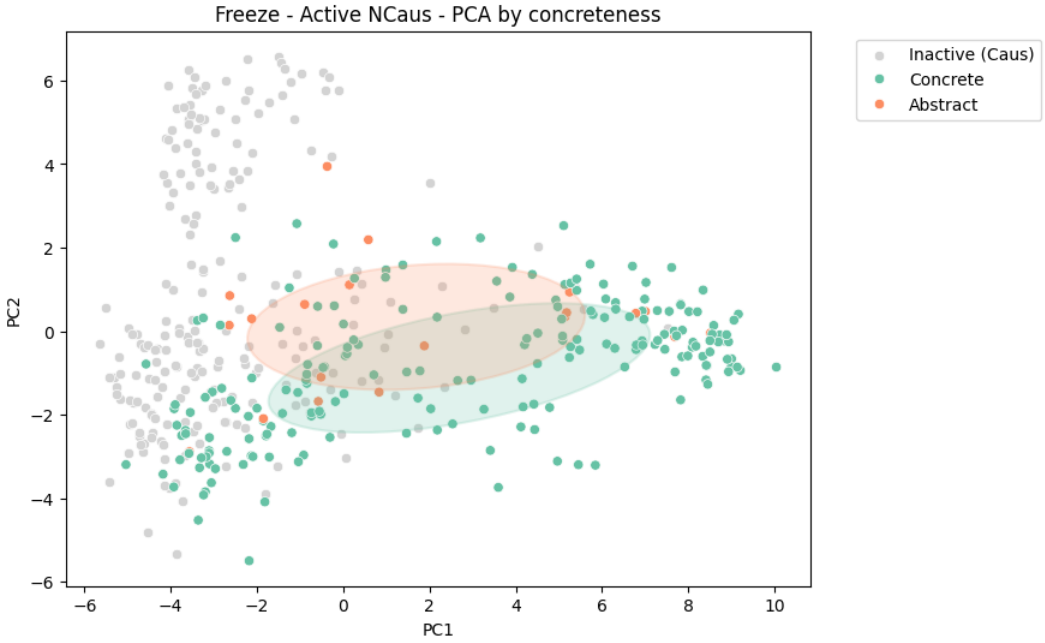



Figure 42. PCA scatter plot of BERT embeddings of noncausative uses of *freeze* colored by theme concreteness




Figures 37–42 demonstrate that, similar to *break*, the causative and noncausative uses of *freeze* are also differentiated by causer-related properties (intentionality and identifiability) and theme concreteness. We now turn to the distributions of *freeze* sense embeddings in terms of these three factors.

Figures 43–44 present the distributions of embeddings for the three causative-dominant senses of *freeze* in terms of causer intentionality and identifiability. In the intentionality plot

(Figure 43), agents are the most frequent causer type, extending across both the upper and lower regions, while animate and inanimate causers cluster more narrowly on the left. This distribution corresponds to the ANOVA and MANOVA results, which indicate a significant effect of intentionality on embedding variation (PC1 ANOVA: $p = 0.00465$; MANOVA: $p < 0.05$). The identifiability plot (Figure 44) shows a clear predominance of specified causers, with recoverable categories (RC_other, RC_mentioned) appearing only at the margins. Statistical tests again confirm that identifiability significantly differentiates embeddings along the PCA dimensions (PC1 ANOVA: $p < 0.0001$; MANOVA: $p < 0.0001$).

Figure 43. PCA scatter plot of BERT embeddings of causative-dominant senses of *freeze* colored by causer intentionality

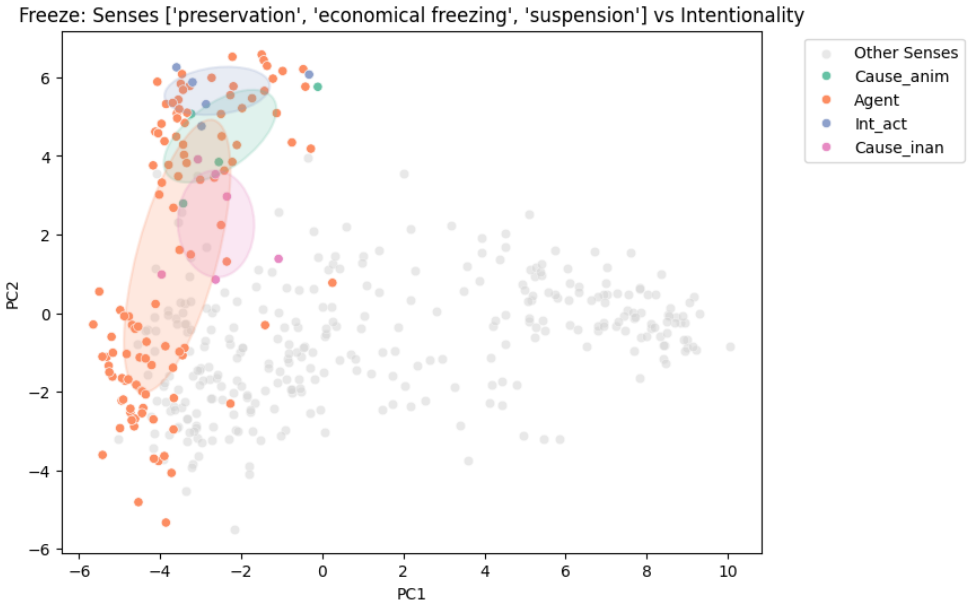
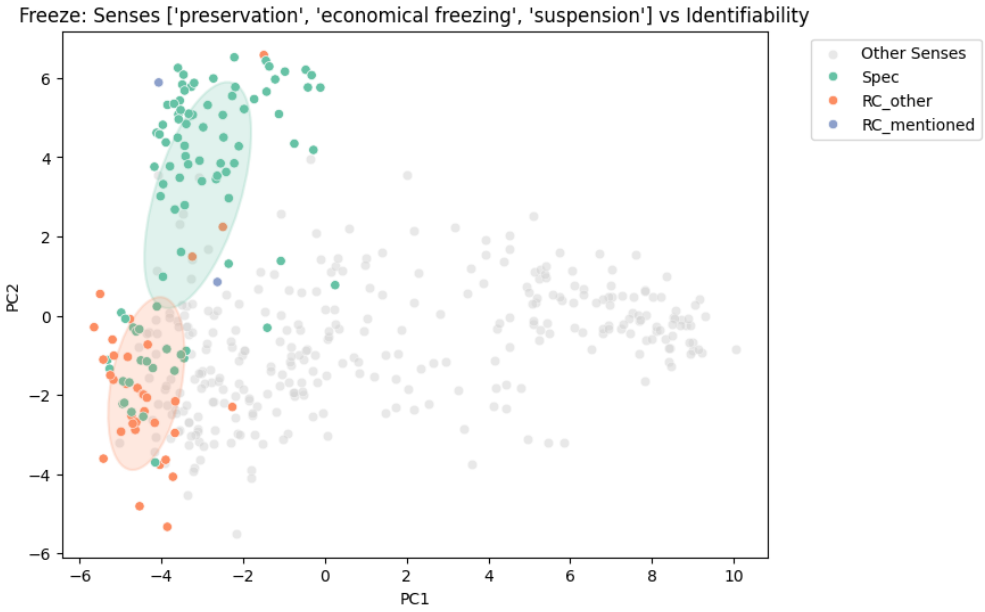



Figure 44. PCA scatter plot of BERT embeddings of causative-dominant senses of *freeze* colored by causer identifiability



These results demonstrate that for causative-dominant senses of *freeze*, both causer intentionality and identifiability exert significant influence on the structuring of the embedding space, consistent with predictions that causative-skewed senses align with causer properties enhancing the informativeness of the causative construction.

For the seven noncausative-dominant senses of *freeze* (bodily freezing, physical freezing, natural freezing, emotional/mental freezing, immobilization, mechanical breakdown, and technical failure), the PCA scatter plots (Figures 45–46) show that causer intentionality and identifiability both structure the embedding distributions. In terms of intentionality (Figure 45), inanimate causers are the most prevalent and spread broadly across the horizontal axis, while animate causers and agents are more confined to the rightward and leftward regions, and intentional causing actions are less frequent and more dispersed. The MANOVA confirms that intentionality exerts a significant effect (Wilks's $\lambda = 0.7705$, $p < .0001$). Identifiability (Figure 46) further distinguishes these senses: recoverable and unknown causers dominate, whereas specified causers occupy more dispersed and peripheral bands. This distributional separation is supported by the statistical test (Wilks's $\lambda = 0.8244$, $p < .0001$). Together, these results indicate that the noncausative-dominant senses of *freeze* are systematically conditioned by both causer intentionality and identifiability, consistent with the hypothesis that lower causer salience aligns with noncausative realizations.

Figure 45. PCA scatter plot of BERT embeddings of noncausative-dominant senses of *freeze* colored by causer intentionality

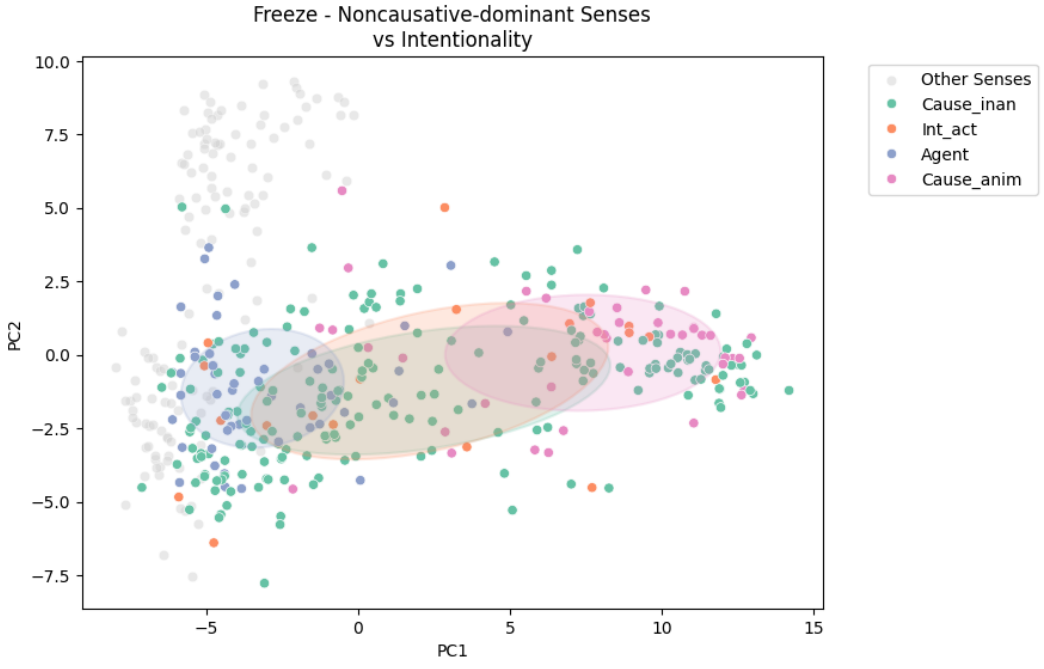
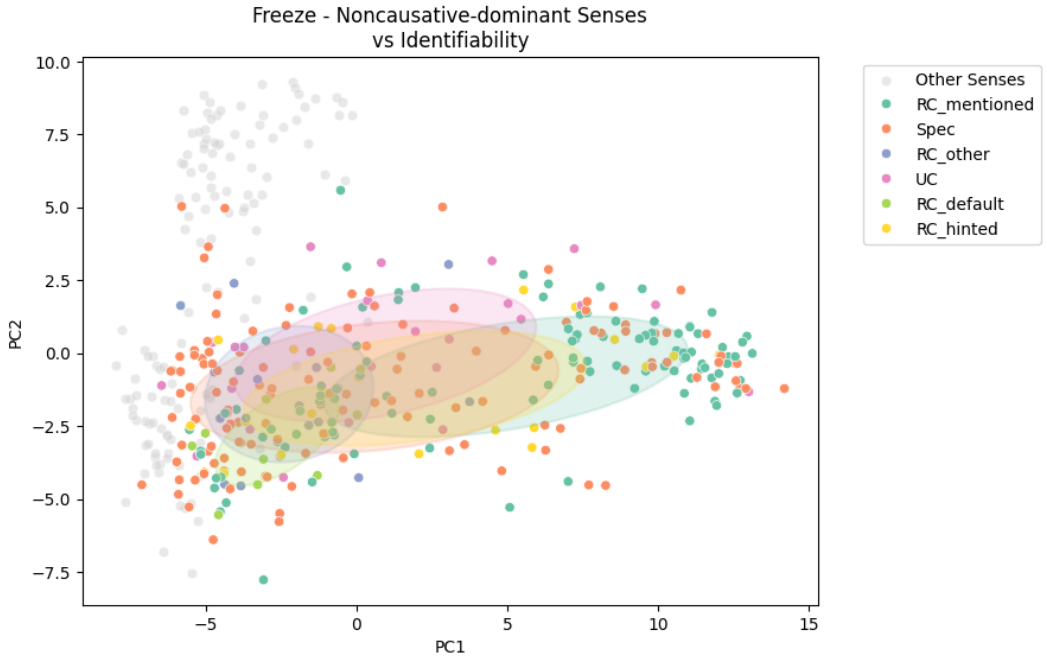


Figure 46. PCA scatter plot of BERT embeddings of noncausative-dominant senses of *freeze* colored by causer identifiability




Causative and noncausative uses of *freeze* that share the same sense exhibit clear differences in terms of causer intentionality and identifiability. To illustrate this, we examine the properties of causers in examples of the physical freezing sense, which is the most evenly distributed across the two alternation variants. Causative uses of *freeze* in this sense typically denote freezing that results from a specific and identifiable causal action—whether intentional or unintentional—or from an abstract triggering event. Example (25) shows a causative use where freezing is caused by an intentional action, namely the deliberate placement of devices in a temperature-controlled chamber.

(25) Environ's job was to freeze the gadgets in a temperature-controlled chamber.
(COCA MAG, PopMech-2009)

By contrast, noncausative uses often specify the cause of freezing in a preceding clause, as in (26), where a cold spell leads to the freezing of pipes. At the same time, noncausative uses frequently include cases where the cause is unknown or left unspecified, as illustrated in (27).

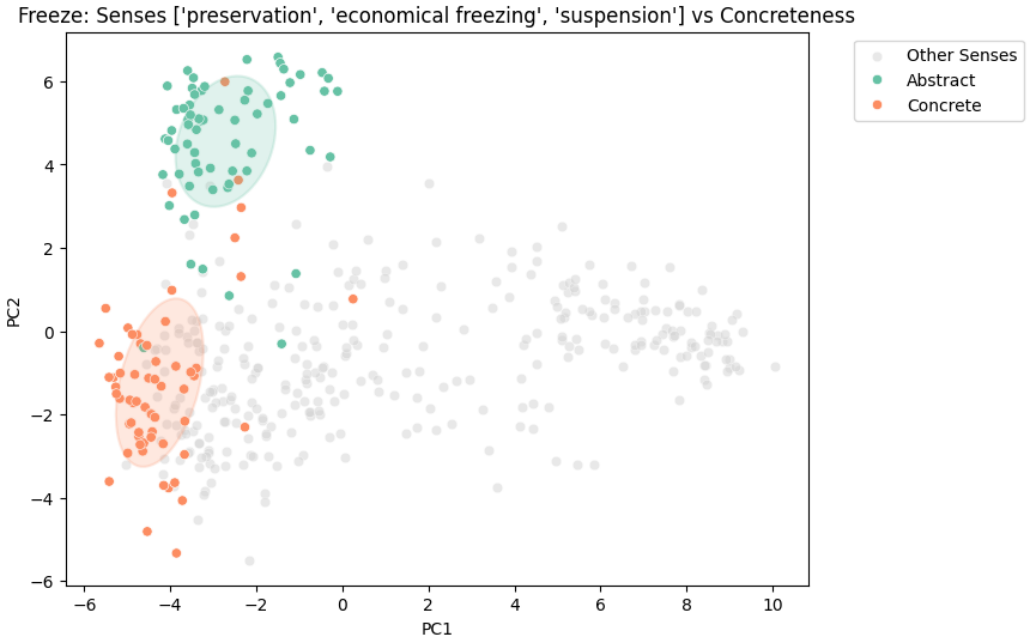
(26) So along comes that cold spell last week and <the pipes froze> and the waterline burst above their bedroom and …
(COCA FIC: SouthwestRev-2015)

(27) … much support from Motorola here in Australia &; <my phone has frozen once for no apparent reason>.
(COCA BLOG: the-gadgeteer.com-2012)

Finally, let us examine the distributions of causative-dominant and noncausative-dominant *freeze* sense embeddings with respect to theme concreteness. As shown in Figure 47, causative-

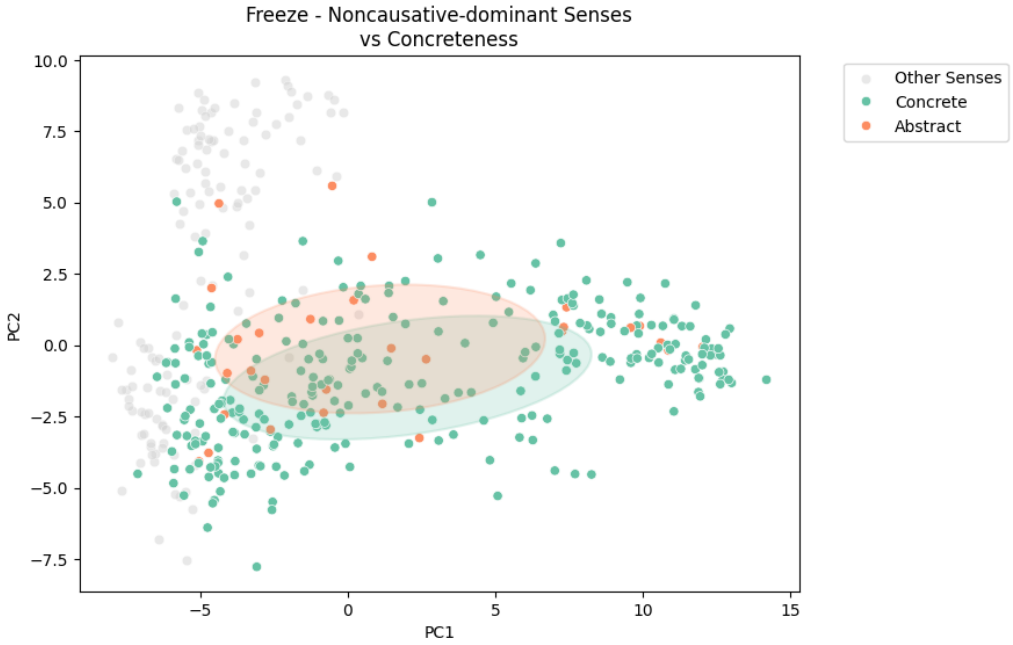
dominant senses (preservation, economical freezing, suspension) display a clear separation between concrete and abstract themes: concrete themes cluster in the lower region of PC2, while abstract themes cluster in the upper region. This distinction is statistically robust, with concreteness emerging as a significant factor (ANOVA for PC1: $p$ = 0.0000; MANOVA for PC1 & PC2: all tests $p$ = 0.0000).

Figure 47. PCA scatter plot of BERT embeddings of causative-dominant senses of *freeze* colored by theme concreteness




By contrast, the noncausative-dominant senses (bodily freezing, physical freezing, natural freezing, emotional/mental freezing, immobilization, mechanical breakdown, technical failure) reveal a markedly different pattern (Figure YY). Here, concrete themes strongly predominate, and abstract themes appear only in limited overlap, often confined to emotional/mental freezing contexts. This distribution is consistent with the statistical results: concreteness is not significant for PC1 ($p$ = 0.4334) but does reach significance in the MANOVA for the combined dimensions ($p$ = 0.0129), indicating a weaker but still detectable effect. These findings show that while concreteness strongly structures the embedding space of causative-dominant senses, it plays a more peripheral role in noncausative-dominant senses, where concrete themes prevail regardless of the embedding dimension.

Figure 48. PCA scatter plot of BERT embeddings of noncausative-dominant senses of *freeze* colored by theme concreteness




Now we turn to a discussion of how the BERT-based analyses of the polysemous sense distributions of *break* and *freeze* presented in this chapter align with the predictions of the hypothesis formulated in Section 2.2.2. The first prediction of the hypothesis concerned the clustering of causative-dominant senses with causer properties that increase the informativeness of the causative variant. The analyses support this prediction: causative-dominant senses are systematically associated with intentional causers or explicitly identifiable causers, making the causative variant highly informative. By contrast, noncausative-dominant senses tend to co-occur with nonintentional or weakly identifiable causers, reducing the informativeness of the causative variant and favoring realization in the noncausative form.

The analyses of *break* and *freeze* confirm the second prediction derived from the central hypothesis in Section 2.2.2. For *break*, which shows a strong causative preference overall (≈79% causative uses), the majority of senses are causative-dominant, clustering with intentional causers and explicitly specified causers. Even moderately skewed senses such as destruction and EPS breakdown align with identifiable causer types. By contrast, only a small minority of senses (disclosure and emergence) are noncausative-dominant, and these systematically correlate with nonintentional or weakly identifiable causers, confirming that break's causative bias is driven by the high informativeness of causer specification

In contrast, *freeze* exhibits the opposite pattern: although a few senses (preservation, suspension, economic freezing) are causative-skewed, the majority of its senses are noncausative-dominant (bodily, natural, physical freezing; emotional/mental freezing; immobilization; mechanical/technical failure). These senses preferentially combine with nonintentional or less identifiable causers, and in the case of emotional/mental and bodily freezing, specified causers appear in adjunct positions that do not enhance causer informativeness and salience in the same way as in *break*. This distribution explains *freeze*'s global bias toward the noncausative variant.

Taken together, the complementary sense distributions of *break* and *freeze* illustrate the predicted systematic divergence: verbs biased toward the causative variant (*break*) are

dominated by senses with intentional and specified causers, while verbs biased toward the noncausative variant (*freeze*) are dominated by senses associated with nonintentional or weakly identifiable causers. These findings corroborate the second prediction of the hypothesis, showing that constructional preferences across verbs emerge from the alignment between sense-level semantics and the informativeness contributed by causer properties.

### 2.3. Conclusions

This chapter has examined the distributional semantics of causative alternation verbs by combining annotated corpus variables with BERT-based analyses, focusing first on general usage patterns and then on verb-specific case studies.

Section 2.1 has examined the usage variation of strong alternators in the causative alternation by integrating annotated corpus variables with BERT-based distributional semantic analysis. The results demonstrate that both causer intentionality and identifiability function as highly robust discriminators between causative and noncausative uses, a finding supported by both proportional distributions and MANOVA tests. The PCA visualizations further revealed two principal semantic dimensions: PC1, contrasting physical/material transformations with qualitative or emergent changes, and PC2, contrasting immediate with gradual changes. Regional verb distribution analyses showed that these dimensions consistently structure the semantic space, while also capturing systematic constructional preferences: causatives cluster around externally induced physical transformations, whereas noncausatives favor gradual, internally motivated, or emergent changes. Finally, the case study of *empty* illustrated that the BERT-derived semantic space can capture subtle variation in verb usage across PCA regions, thereby providing qualitative insights that complement the quantitative analyses.

Building on this foundation, Section 2.2 has provided an in-depth case study of two polysemous COS verbs, *break* and *freeze*, which exhibit contrasting constructional preferences. The BERT-based analyses of sense distributions revealed systematic associations between sense types, causer properties, and constructional biases: causative-dominant senses clustered with intentional or explicitly identifiable causers, while noncausative-dominant senses aligned with nonintentional or weakly identifiable causers. Moreover, the comparison of *break* and *freeze* showed complementary patterns: *break* is dominated by causative-dominant senses that reinforce its causative preference, whereas *freeze* is characterized by noncausative-dominant senses that underpin its bias toward the noncausative variant. Taken together, these results confirm the predictions of the central hypothesis and demonstrate how sense-level semantics systematically shape verb-specific constructional preferences in the causative alternation.

# Chapter 3

# Sense Boundaries in Polysemy: Insights from Experiments with Contextualized Language Models

The BERT-based distributional semantic analysis of polysemous verbs presented in Chapter 2 demonstrated that contextualized language models can be fruitfully employed for polysemy research. However, the PCA visualizations we applied introduce well-known distortions, as they compress high-dimensional semantic spaces into only two dimensions. This limitation highlights the clear need for more robust methodologies that move beyond distributional patterns to capture other, more fine-grained aspects of polysemy.

This chapter shifts the focus to a notoriously problematic issue in theoretical semantic analysis: the problem of sense boundaries. Specifically, we present two experiments designed to investigate the following questions: Which polysemous senses are clearly distinguishable, and which are less distinct and prone to confusion with other senses? What characteristics differentiate the former senses from the latter? And how do the sense classifications and predictions of language models illuminate the still elusive nature of sense distinctions and sense boundaries?

To answer these questions, this chapter examines four different Transformer-based encoder models, BERT-base, BERT-large (Devlin et al. 2019), RoBERTa (Robustly Optimized BERT Pre-training Approach, Liu et al. 2019), and ALBERT (A Lite BERT, Lan et al. 2020). These models recently became the state-of-the art on a variety of NLP downstream tasks. Their strong empirical performance triggered questions concerning the linguistic knowledge they encode in their representations and how it is affected by explicit training or fine-tuning on task-specific, supervised downstream tasks. Section 3.1 discusses an experiment designed to probe for meaning-class and construction-type to evaluate the extent to which the representations of the pre-trained encoder models encode these important linguistic properties. In Section 3.2, we discuss an experiment where we fine-tune them on the meaning classification task to gain new insights about the nature of sense boundaries in polysemy from the observed changes in the models' performance on the task. Based on the meaning classification patterns of the fine-tuned models, Section 3.3 derives the key distributional and semantic factors that contribute to the sharpening and blurring of the boundary between the polysemous senses of *break* and *freeze*. These factors are then used to propose a new unified approach to polysemy that accounts for sense boundaries in polysemy.

## 3.1. Probing Contextualized Language Models

### 3.1.1. Annotated Dataset

The dataset for our probing experiment consists of a total of 1,200 examples from three verbs:

- A subset of 800 instances of causative and noncausative uses of the causative alternation verbs *break* and *freeze*, which were used in the meaning distribution analysis in Chapter 2.

- 400 instances of causative and noncausative uses of the non-causative alternation verb *sprout*.

Our annotation scheme uses the same 11 semantic classes for *break* and 6 semantic classes for *freeze* as applied in Chapter 2. The semantic classes for these two verbs are provided with illustrating examples of each class in Table 1 and Table 2, respectively.

Table 1. Major meaning classes for *break*

| Meaning classes (sense categories) | Examples |
|---|---|
| Physical breaking | The glass broke from the pressure. |
| Bodily harm | My ankle broke in a skiing accident. |
| Emotional/psychological/social breakdown | Her heart broke when she heard the news. |
| Violation | The company broke the law. |
| Decoding | The police finally broke the code. |
| Disclosure | The reporter broke the news early in the morning. |
| Termination | They broke the cycle of violence. |
| Interruption | I broke the tense atmosphere with a joke. |
| Breakthrough | She broke the world record. |
| Change | The medicine helped break the fever. |
| Emergence | The day broke over the quiet village. |

Table 2. Major meaning classes for *freeze*

| Meaning classes (sense categories) | Examples |
|---|---|
| Physical/bodily freezing | The storm froze the pipes, causing them to burst. |
| Freezing for preservation | I froze the meat for preservation. |
| Immobilization | She froze in place when she heard the noise. |
| Economical freezing | The bank froze their assets. |
| Suspension | The company froze the construction project. |
| Emotional/mental freezing | His thoughts froze, and old fears rushed back. |

In Table 2, the three physical senses for *freeze*—physical freezing, bodily freezing, and natural freezing—have been combined into a single class: physical/bodily freezing. Additionally, the three sudden senses—physical/bodily immobility, mechanical breakdown, and technical failure—have been combined into the immobilization class.

In this experiment, the dataset was expanded to include *sprout*, a polysemous verb that, like *break* and *freeze*, participates in transitivity alternation. As Rappaport Hovav (2020) has shown, this verb does not show patterns of argument realization typical of change-of-state verbs, but patterns like verbs of emission, displaying a transitivity alternation known as the source-theme or source-substance alternation. According to Levin (1993: 32-33), this alternation is found with verbs of substance emission, which take two arguments: a source (emitter) and the substance/theme emitted from this source. Unlike the causative alternation, both arguments are expressed in the transitive and intransitive variants of the source-theme alternation, as shown in (1). The substance/theme is expressed in both, and the source is expressed as the subject in the transitive use of the verb and as the object of the preposition *from* in the intransitive use. The intransitive variant of this alternation is analyzed as the unaccusative-emergence frame, and the transitive variant as the source-theme frame.

(1) a. The sun radiates heat.
b. Heat radiates from the sun. (Levin 1993: 32)

In the same way, *sprout* takes two arguments. Examples of the (transitive) emitter-subject variant, analyzed here as the source-theme frame, and the (intransitive) emittee-subject variant, analyzed as the unaccusative-emergence frame, are provided in (2a) and (2b), respectively.

(2) a. The tree has sprouted green leaves.
b. Green leaves sprouted from the tree.

*Sprout* is most frequently used in its biological growth sense. However, as shown in (3), *sprout*'s other meaning—the emergence sense—is also found in both variants of the source-theme alternation.

(3) a. Walk High Street on Friday evening and watch the Oddbins wine shop sprout a queue of bottle-bearing customers in college ties.
(COCA NEWS: Houston Chronicle-19930110)
b. Dozens of sushi restaurants had sprouted all over Warsaw in recent years, and …
(COCA NEWS: Austin American Statesman-20120805)

The transitive uses of *sprout* presented so far are not causative. A piece of evidence for this comes from the fact that these uses do not passivize:

(4) a. *Green leaves have been sprouted by the tree.
b. *A queue of bottle-bearing customers is sprouted by the Oddbins wine shop.

The fact that they do not passivize is not surprising. As pointed out by Rappaport Hovav (2020), transitive uses of verbs of emission do not passivize either:

(5) a. The wound oozed pus.
b. *Pus was oozed by the wound. (Rappaport Hovav 2020: 249)

(6) a. The well gushed oil.
b. *Oil was gushed by the well. (Rappaport Hovav 2020: 249)

Unlike other verbs of substance emission like *bloom, blossom* and *flower*, *sprout* has causative uses that can be interpreted as 'to produce,' 'to create,' or 'to cause to grow/emerge', and, concomitantly, have passive uses as well. As noted by Rappaport Hovav (2020: 250), the causer type can range from agents to natural causes, typical of causative uses, as illustrated in (7) and (8):

(7) a. Various companies have sprouted their own STM products.
(BNC W:pop_lore, ABF-3014)
b. "You know," I say, "you can sprout lentils." Or grow them, if they're so expensive…
(COCA MAG: Popular Science-2014)

c. the warm, rainy weather sprouted the wheat before it could be gathered.
(Rappaport Hovav 2020: 250)

(8) a. The seeds were sprouted by five sprout producers and then sold.
b. These beans were sprouted by Vinitha and they tasted really crunchy!
c. …kumquat blossoms and jasmine. In earlier times, shallot, onion and madder plants were sprouted by the same method. (Rappaport Hovav 2020: 250)

*Sprout* also has another transitive use that is interpreted as a physical/functional change sense, representing a change in external features or a transformation/acquisition. The examples in (9a) and (9b) illustrate this meaning. In example (9a), *sprout* is used to describe a physical change on a person's face. In example (9b), it is used to describe the motors acquiring new functional components (16-valve heads and fuel injection). This represents a functional change or transformation in their physical makeup that enhances performance.

(9) a. Faces can sprout hair and sag with time and circumstance.
(COCA MAG: Popular Science-2014)
b. Both motors have sprouted 16-valve heads and fuel injection.
(BNC W:pop_lore, ACR-2959)

These transitive uses are not causatives, and hence do not passivize, as shown in (10). Because the subject of the sentence acts as the entity undergoing the change, we will analyze the transitive uses of *sprout* that signify a physical/functional change as the unaccusative-change-of-state frame.

(10) a. *Hair and sag can be sprouted by faces with time and circumstance.
b. *16-valve heads and fuel injection have been sprouted by both motors.

The meaning classes and frame types for *sprout* discussed above are summarized in Table 3.

Table 3. Meaning classes and frame types for *sprout*

| Meaning classes | Frame types | Examples |
|---|---|---|
| Creation (produce; create; generate) | Causative (transitive) | (7a) |
| Biological growth (grow; cause to grow) | Causative (transitive) | (7b), (7c) |
| | Source-theme (transitive) | (2a) |
| | Unaccusative-emergence (intransitive) | (2b) |
| Emergence (emerge; appear) | Source-theme (transitive) | (3a) |
| | Unaccusative-emergence (intransitive) | (3b) |
| Physical/functional change (change; transform; acquire; equip) | Unaccusative-change of state (transitive) | (9) |

Summarizing this section, the dataset for our probing experiment consists of a total of 1,200 examples from three polysemous verbs, *break, freeze* and *sprout* (400 examples per verb). These examples are extracted from COCA, with the exception of the causative uses of *sprout*. Since it was difficult to secure a sufficient number of examples from COCA, the causative

sentences for *sprout* were generated using ChatGPT to complete the dataset. As summarized in Table 4, our annotation scheme uses 20 meaning classes and 4 frame types (one causative and three noncausative frames), all of which have substantial representation in the data.

Table 4. Summary of the labeled dataset

| Verb | Primary semantic class | Meaning classes | Frame types |
|---|---|---|---|
| *break* | Verbs of change of state (cos) | 11 classes (Table 1) | Causative (transitive)<br>Unaccusative-cos (intransitive) |
| *freeze* | | 6 classes (Table 2) | |
| *sprout* | Verbs of substance emission | 4 classes (Table 3) | Source-theme (transitive)<br>Causative (transitive)<br>Unaccusative-cos (transitive)<br>Unaccusative-emergence (intransitive) |

We rely primarily on the meaning class distinctions and make secondary use of the frame annotations. For our frame-type probing work, we will use all 1,200 sentences. For the meaning-type work, we will focus on the two change-of-state verbs in probing and fine-tuning to explore whether and how the meaning classification patterns of the language models are related to the complex event structure of these verbs.

### 3.1.2. Probing Methodology and Setup

Our setup largely follows that of previous works (Hewitt and Liang 2019; Hewitt and Manning 2019; Tenny et al. 2019; Mosbach et al. 2020; Petersen and Potts 2023) where a probing classifier is trained on top of the contextualized embeddings extracted from a pre-trained or fine-tuned encoder model. We train logistic regression probing classifiers (L2-regularized classifiers with a cross-entropy loss) and examine four pre-trained encoder models: BERT-base-uncased, BERT-large-uncased, RoBERTa-large, and ALBERT.

For both meaning and frame classification probing tasks, the model performance was evaluated using several metrics: accuracy, F1 score and selectivity score. These metrics are particularly important for a meaning classification task with multiple classes, as they provide a detailed understanding of the model's predictive capabilities. Accuracy represents the ratio of correctly predicted instances to the total instances. The F1 score, which is the weighted average of precision (the ratio of correct positive predictions to the total positive predictions) and recall (the ratio of correct positive predictions to the total actual positives), provides a balance between these two metrics.

Our core metric used to assess the model performance in the probing experiment is the macro F1 score, which assigns equal weight to each class' F1 score regardless of the class size. Following Hewitt and Liang (2019) and Petersen and Potts (2023), we report *selectivity* scores, defined as the difference in macro F1 between the real classification task (e.g., semantic class prediction) and a control task with randomly shuffled labels. We report selectivity scores averaged across 20 random 80%/20% train/test splits.

The experiments are performed on a Google Colaboratory T4 GPU, which provided an accessible environment for training and evaluation. Data and code to reproduce our results and figures are available at the GitHub repository (https://github.com/hanjung-25/clmsemantics/tree/data and https://github.com/hanjung-25/clmsemantics/tree/code)

### 3.1.3. Evaluating Encoded Linguistic Properties

In this section, we discuss the layer-wise probing results along with the classification accuracy by frame type and meaning class. While the performance of the four models is very consistent with each other in terms of layer-wise trends and overall performance, RoBERTa-large generally outperforms other models in the probing tasks. For this reason, we discuss probing results for RoBERTa-large here and report other models in Appendix B.

Table 4 summarize the RoBERTa-large probing results for (a) frame type and (b) meaning class classification. We report accuracy, the macro F1 score for the real (linguistic) task, the macro F1 score for a control task (random assignment of meaningless labels to classes), and selectivity, which is the macro F1 score for the real task minus the macro F1 score for a control task. Performance for both tasks stabilizes around layer 6, after which the classification metrics remain consistent. For the frame classification task, the model achieved an accuracy of over 0.8, a macro F1 score (real task) of over 0.75, and a selectivity score of over 0.63, starting from layer 6. In the meaning classification task (b), the model's performance is slightly lower but still robust. It achieves an accuracy of over 0.74, a macro F1 score (real task) of over 0.68, and a selectivity score of over 0.61, starting from layer 12. These results establish base-lines for our comparison with fine-tuned models to be discussed in Section 3.2.

Table 4. RoBERTa-large probing results

(a) Frame-type probing results

| Layer | Accuracy | F1_Real | F1_Control | Selectivity |
|---|---|---|---|---|
| 1 | 0.667 | 0.595 | 0.099 | 0.495 |
| 6 | 0.829 | 0.777 | 0.087 | 0.690 |
| 12 | 0.820 | 0.752 | 0.082 | 0.669 |
| 18 | 0.812 | 0.755 | 0.116 | 0.638 |
| 24 | 0.804 | 0.752 | 0.110 | 0.641 |

(b) Meaning-class probing results

| Layer | Accuracy | F1_Real | F1_Control | Selectivity |
|---|---|---|---|---|
| 1 | 0.616 | 0.560 | 0.085 | 0.474 |
| 6 | 0.687 | 0.614 | 0.068 | 0.546 |
| 12 | 0.762 | 0.718 | 0.105 | 0.612 |
| 18 | 0.750 | 0.694 | 0.075 | 0.618 |
| 24 | 0.741 | 0.686 | 0.068 | 0.617 |

Figure 1 shows the layer-wise probing results for frame type classification. We can see that performance increases early, starting around layer 6, and reaches a plateau by layer 12. This suggests that syntactic information is encoded in lower-to-mid-layers of contextualized language models, consistent with earlier studies (Lin et al. 2019; Hewitt and Manning 2019; Petersen and Potts 2023).

Figure 1. RoBERTa-large frame-type probing results

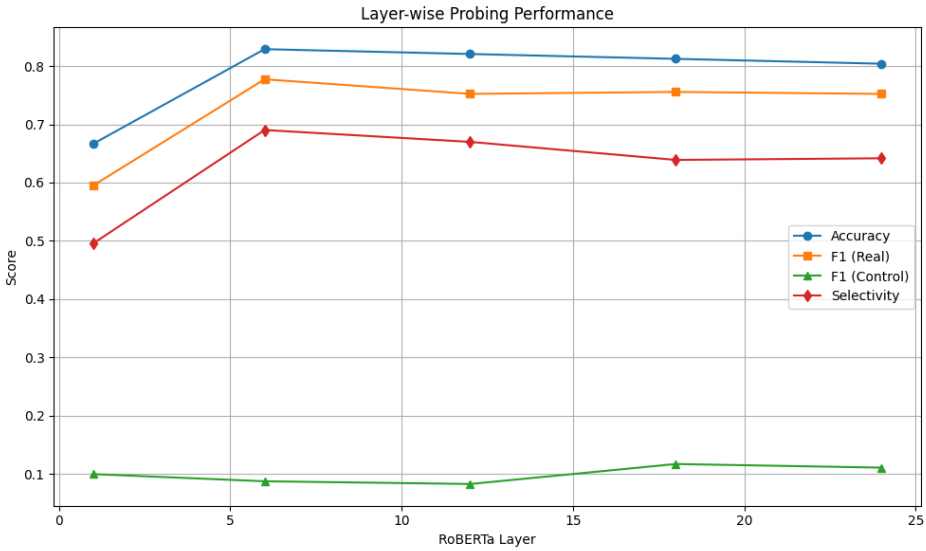
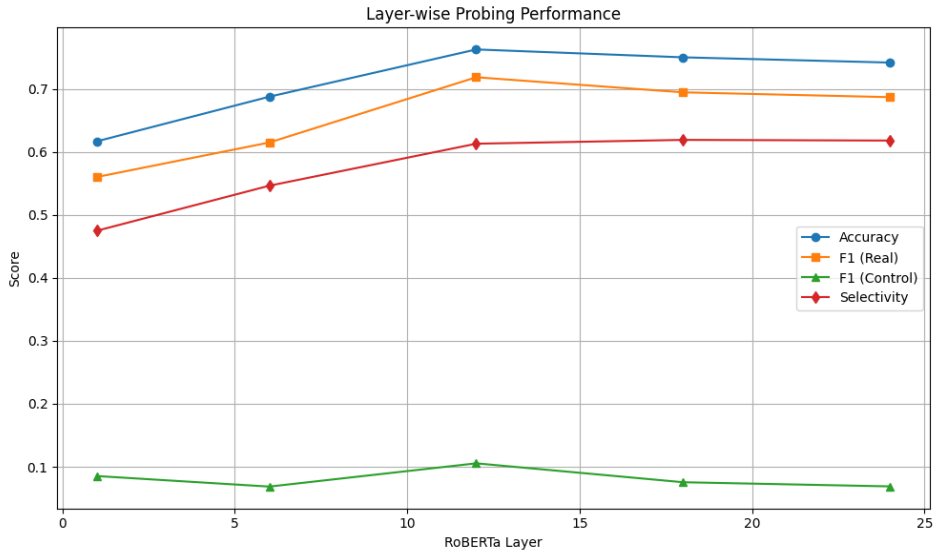


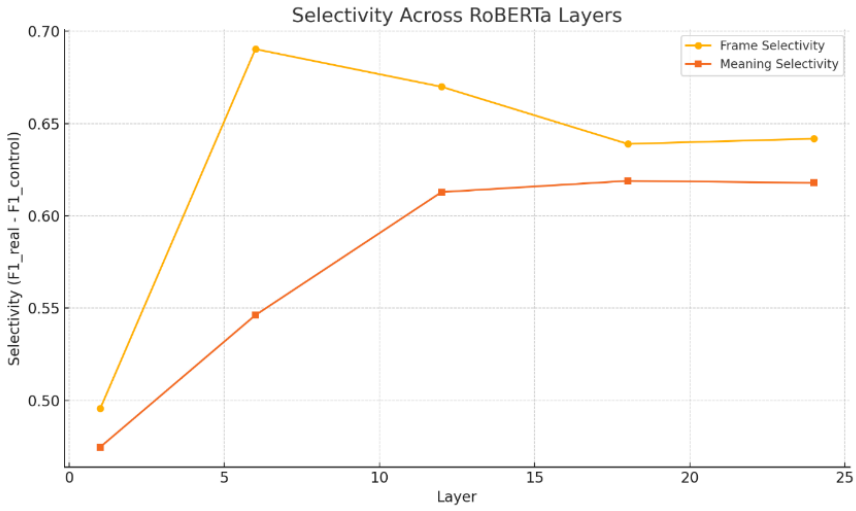

Figure 2 shows that semantic classification improves later, with major performance gains starting at layer 12 and continuing up to layer 24. This supports the idea that deeper semantic distinctions–such as polysemous sense disambiguation–are handled in the deeper transformer layers (Liu et al. 2019; Tenny et al. 2019; Petersen and Potts 2023). Higher overall performance on the frame-type classification task: Again, this is consistent with findings in previous works reporting better performance on syntactic than semantic probing tasks.

Figure 2. RoBERTa-large meaning-class probing results



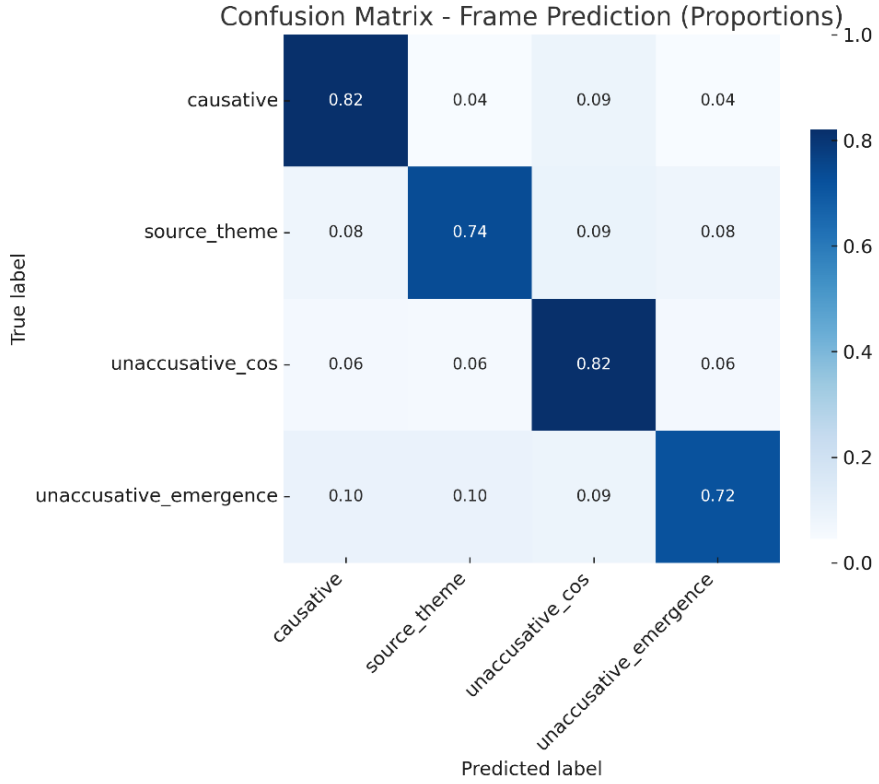

As shown in Figure 3, the layer-wise analysis of RoBERTa reveals a clear increase in selectivity from lower to higher layers, with a marked boost starting at layer 6 (frame classification) and layer 12 (meaning class). This suggests that deeper layers more effectively distinguish real semantic tasks from control tasks, reflecting a growing sensitivity to meaningful linguistic structure.

Figure 3. Selectivity scores across RoBERTa-large layers




Let us now examine the classification accuracy by frame type and meaning class. Figure 4 shows the predictive accuracy by frame type at layer 24 of RoBERTa-large in the form of a confusion matrix, a tabular representation used to evaluate the performance of classification models. It cross-tabulates the predicted categories against the true (gold-standard) categories, allowing researchers to observe not only the overall accuracy but also the distribution of specific errors. In semantic and syntactic prediction tasks, the confusion matrix serves as a crucial tool for identifying which frame types are reliably distinguished by the model and which are frequently confused with others. The diagonal cells represent correct classifications, while off-diagonal cells indicate the relative proportions of misclassified instances.

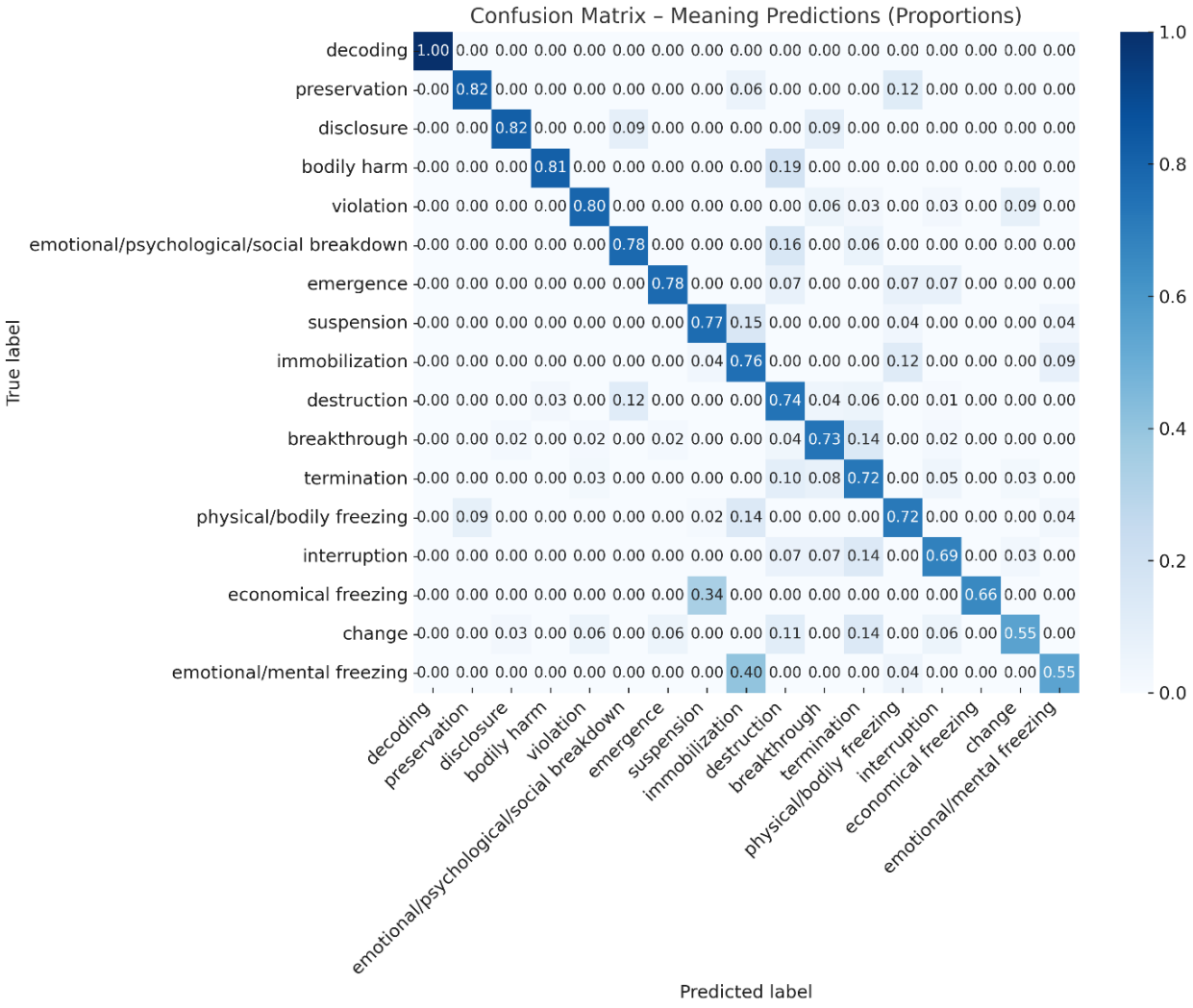
Figure 4. Confusion matrix for frame-type prediction



The results show that the causative and unaccusative-change of state frames, which are shared across all three verbs, are predicted with higher accuracy (≈82%), whereas the source-theme and unaccusative-emergence frames, which are specific to the verb *sprout*, exhibit lower accuracies (≈74% and 72%).

Figure 5 is the confusion matrix for meaning class predictions. It reveals systematic differences across semantic classes. Among the 17 meaning classes of *break* and *freeze*, decoding**,** preservation**,** disclosure, and bodily harm achieve the highest accuracies, indicating that these senses are relatively well distinguished by the model. By contrast, change and emotional/mental freezing show the lowest accuracy, reflecting the difficulty of identifying this sense and its greater overlap with termination and immobilization senses, respectively. The matrix further highlights recurrent misclassification tendencies. For example, meanings related to physical change or impairment (e.g., destruction**,** bodily harm**,** physical/bodily freezing immobilization) are often confused with each other, suggesting that the model struggles to draw sharp boundaries between closely related change-of-state events. Similarly, senses linked to abstract or metaphorical extensions (such as emotional/psychological/social breakdown, termination, interruption, and economical freezing) display substantial cross-confusion, reflecting weaker sense boundaries in figurative or less prototypical contexts.

Figure 5. Confusion matrix for meaning-class prediction



Overall, the results indicate that senses that tend to occur with narrowly restricted theme arguments (e.g., decoding, preservation, disclosure) are captured with higher fidelity, whereas more abstract or metaphorical senses (particularly change and emotional/mental freezing) remain more error-prone. These confusion patterns underscore the uneven granularity of sense distinctions: while some boundaries are robustly recognized, others remain fuzzy and subject to overlap, thereby illuminating the challenges of modeling polysemy in contextualized prediction tasks.

To summarize the above discussion, layer-wise probing of RoBERTa-large reveals early acquisition of syntactic frame patterns by layer 6, while disambiguation of fine-grained verb senses continues to improve through the deepest layers (layer 24), highlighting a structural-to-semantic processing gradient in transformer-based encoder models. Across all layers, RoBERTa-large demonstrates consistently high selectivity, with a marked increase in deeper layers for the semantic task. The sharp divergence between performance on the real and control tasks–especially beyond layer 12–indicates that RoBERTa-large is sensitive to fine-grained semantic distinctions that are not recoverable from surface features alone. These findings reinforce prior claims that transformer models build up a hierarchy of linguistic abstraction, with deeper layers exhibiting greater semantic selectivity and robustness to noise or control manipulations. They also confirm the reliability of probing methods in teasing apart semantic competence from statistical performance. At the same time, lower model performance on the meaning-class classification task and recurrent misclassification tendencies suggest that polysemy is hard to capture and that there is a large room for improvement in overall performance.

## 3.2. Fine-tuning Contextualized Language Models

### 3.2.1. Annotated Dataset

Having established baselines for the probing performance of the pre-trained models, we now turn to the question of how it is affected by fine-tuning. The models are trained using the dataset of corpus examples with meaning class labels. The labeled dataset used for our fine-tuning experiments consists of a total of 1,188 instances for two polysemous verbs, *break* and *freeze*. These examples are an updated version of those used for the sense distribution analysis of the two verbs discussed in Chapter 2, with the addition of *break* instances. The dataset is composed of 788 instances for *break* and 400 instances for *freeze*. The distribution of 788 instances across the 11 meaning classes for *break* is as follows:

- bodily harm: 40
- breakthrough: 92
- change: 45
- decoding: 19
- destruction: 132
- disclosure: 45
- emergence: 32
- emotional/psychological/social breakdown: 78
- interruption: 46
- termination: 71
- violation: 101

The distribution of 400 instances across the 6 meaning classes for *freeze* is as follows:

- economical freezing: 26
- emotional/mental freezing: 57
- immobilization: 173
- physical/bodily freezing: 101
- preservation: 71
- suspension: 59

### 3.2.2. Fine-tuning Methodology and Setup

For fine-tuning, we follow the default setup proposed by Devlin et al. (2019). A single randomly initialized task-specific classification layer is added on top of the pre-trained encoder. As input, the classification layer receives `z = tanh(Wh + b)`, where z is the refined and transformed value of the model's final hidden representation (h), made ready for the classification layer to make its final prediction.[1]

The four models examined in the probing experiment were fine-tuned for the meaning class classification task. All these models were trained using a rigorous cross-validation called a stratified K-fold approach. This approach is a variation of K-fold cross-validation that ensures each fold has approximately the same percentage of samples of each target class as the complete set. It is particularly useful when dealing with imbalanced datasets, where some classes have significantly fewer instances than others. By preserving the class distribution in each fold, it helps to produce more reliable and less biased estimates of model performance compared to standard K-fold, which might create folds with skewed class distributions, especially for minority classes.

We fine-tuned the models with a stratified 5-fold cross-validation strategy, which was appropriate given the size of the dataset. In this process, the dataset was divided into five parts, with four parts used for training and the remaining one for validation. This process was repeated five times, with each part used once as the validation set, following the methodology described in Kici et al. (2021), Shi et al. (2023), and Uçar, (2025). This methodology helps prevent overfitting and provides a comprehensive assessment of the model's general performance. The decision to use 5-fold cross-validation instead of 10-fold was made to balance computational efficiency with maintaining statistical reliability. Training large language models (LLMs) such as BERT and RoBERTa requires considerable computational resources, and the 5-fold approach offered a reliable performance assessment while reducing computational costs. This strategy enabled a robust evaluation of model performance, which is crucial when comparing multiple models. Additionally, training was conducted for 5 epochs for each fold to allow the

[1] `W` and `b` are the randomly initialized weight and bias of the classifier. These are parameters that will be optimized during training to effectively map the model's output to the correct meaning classes. `Wh + b` is a linear transformation that scales and shifts the h vector. This adjusts the dimensionality of the hidden representation to match the requirements of the classification task. `tanh` is a non-linear activation function that squashes the values from the linear transformation into a range between -1 and 1. This helps the model to learn more complex patterns and makes the final output more stable.

models to sufficiently learn from the data without overfitting.

During the training process, we monitored and compared the performance of the model from one fold to another and from one epoch to another, and therefore monitored, recorded, and visualized the performance characteristics of the model. The performance evaluation metrics included accuracy, precision, recall, and F1 score, calculated for each fold and epoch. The average values across all folds were used for the final assessment. In addition to metric calculation, graphical analysis was conducted to plot these metrics across each fold and epoch, facilitating the identification of potential issues like overfitting or underfitting. Training times were also recorded for each model to compare their computational efficiency. Data and code to reproduce our results and figures are available online (https://github.com/hanjung-25/clmsemantics).

### 3.2.3. Evaluating the Fine-tuned Models' Performance

This section discusses the evaluation results for the models. Again, we report results for RoBERTa-large. We see similar results for other models, as reported in Appendix C.

Table 6 presents a RoBERTa-large cross-validation summary. The 5-fold cross-validation results demonstrate the robust performance of the fine-tuned RoBERTa-large model on the sentence classification task. The cumulative average accuracy and macro F1 score across all folds are both 0.87. This represents a significant improvement compared to the probing experiment on RoBERTa-large layer 24, which yielded an accuracy of 0.741 and a macro F1 score of 0.686 on the real task. The consistency in accuracy and F1 scores across the folds, as shown in the summary table, further highlights the stability and generalization capability of the fine-tuned model.

Table 6. RoBERTa-large cross-validation summary

| Fold | Accuracy | F1 |
|---|---|---|
| 1 | 0.844 | 0.849 |
| 2 | 0.844 | 0.830 |
| 3 | 0.882 | 0.880 |
| 4 | 0.864 | 0.882 |
| 5 | 0.894 | 0.900 |

Figures 6 and 7 illustrate the performance of the model across the five cross-validation folds, showing accuracy (Figure 6) and macro F1 score (Figure 7) per fold. Figure 6 indicates that the accuracy is stable in Folds 1 and 2. There is a noticeable increase in accuracy from Fold 2 to Fold 3, followed by a slight dip in Fold 4. The accuracy then rises again to reach its highest point (0.894) in Fold 5.

Figure 6. RoBERTa-large accuracy per fold

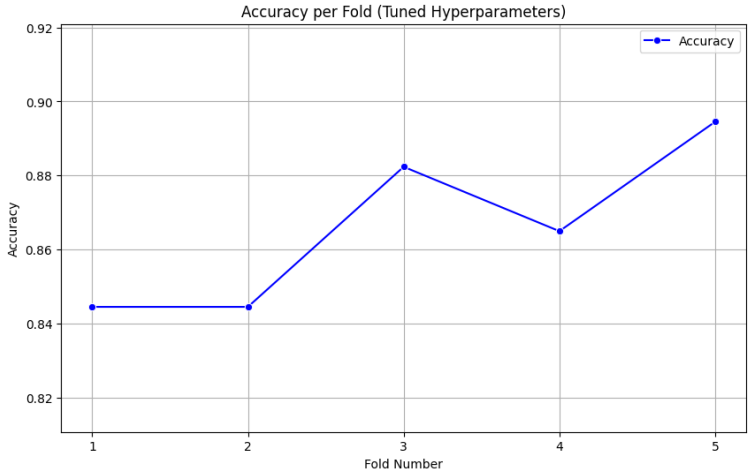
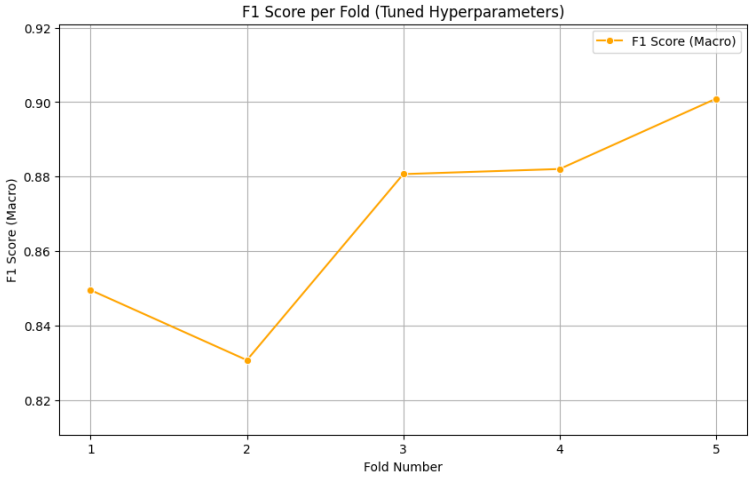



Figure 7 shows a slightly different pattern. The F1 score starts at a similar level in Fold 1 and then drops slightly in Fold 2. Similar to accuracy, there is a significant increase in F1 score from Fold 2 to Fold 3. The F1 score in Fold 4 remains very close to that of Fold 3 and then peaks in Fold 5 (0.900), reaching the highest F1 score among all folds. While both metrics generally show improvement towards later folds and peak in Fold 5, the F1 score exhibits a dip in Fold 2 that is not as pronounced in the accuracy plot. This suggests that the model's performance on minority classes might have been slightly more affected in Fold 2 compared to its overall accuracy.

Figure 7. RoBERTa-large macro F1 score per fold




Let us now look at the plots in Figure 8 that show the average evaluation metrics over epochs (derived from the step-based evaluations): accuracy, F1 score, precision, and recall. We can see that these metrics consistently increase with each epoch. The gain in performance is larger in the earlier epochs and becomes smaller in later epochs, indicating that the model is benefiting from continued training but with diminishing returns. The evaluation loss presented in Figure 9 shows a corresponding downward trend, which is also expected as the model learns to make better predictions. The loss decreases rapidly in the initial steps and then continues to

decrease at a slower rate, indicating that the model is learning effectively over the training duration.

Figure 8. RoBERTa-large average evaluation metrics over epochs

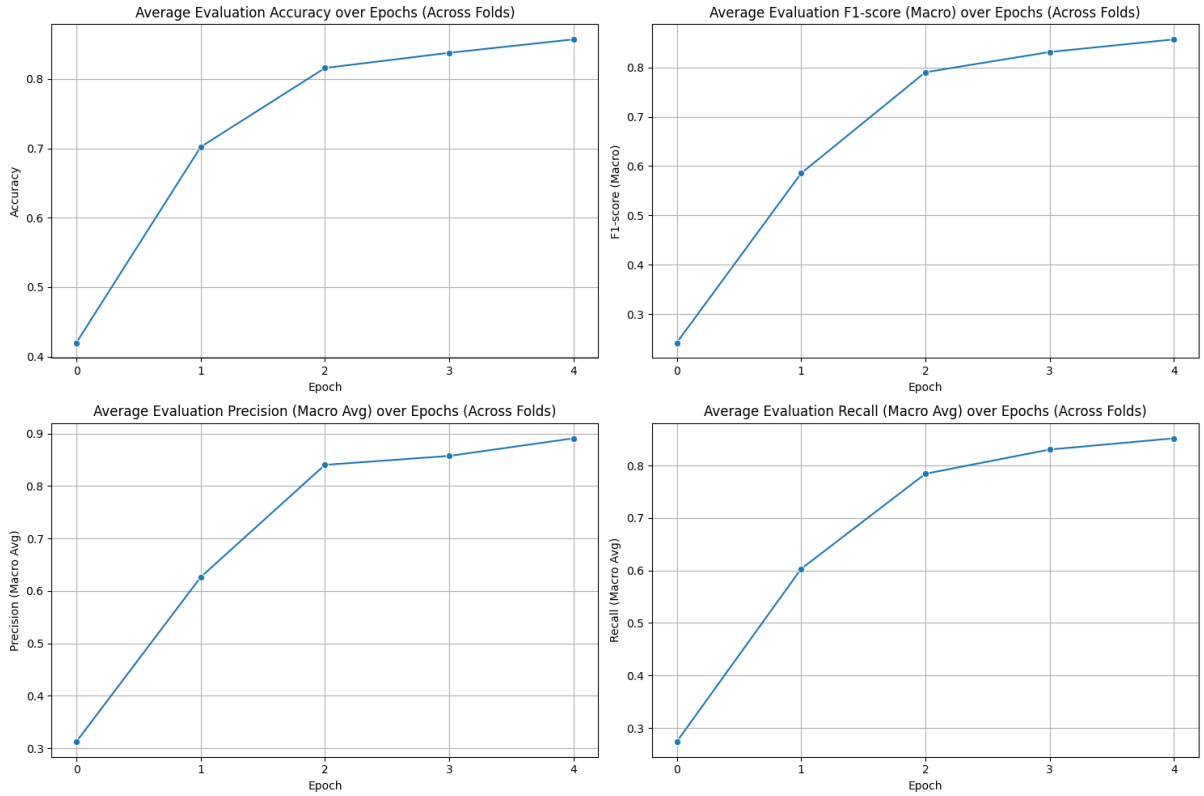
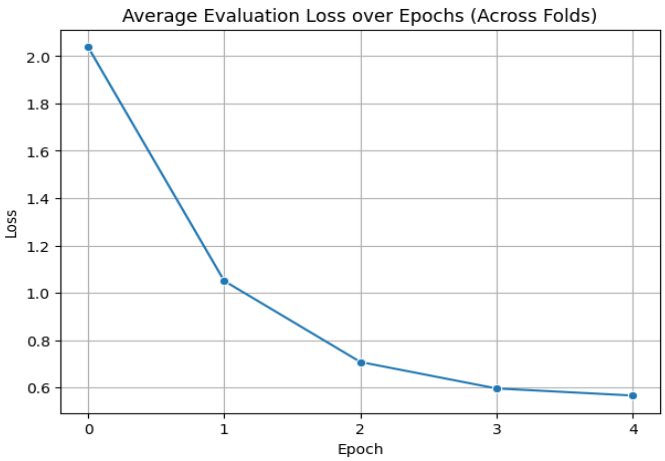



Figure 9. RoBERTa-large average evaluation loss over epochs




The overall upward trend in the aggregated plots in Figure 8 aligns with the fact that our fine-tuned model achieved good average accuracy and F1 scores (around 0.87). The per-fold plots we discussed earlier showed some variability between folds, with Fold 5 achieving the highest scores. The aggregated plots here represent the average learning trajectory across all those folds. The initial lower performance in the early steps/epochs, as seen in the aggregated plots, is a normal part of the training process before the model has fully adapted to the new task. The improvement over steps and epochs demonstrates that the fine-tuning process was

effective in enhancing the RoBERTa model's capability for this specific sentence classification task.

The fact that performance continued to improve up to 5 epochs suggests that 5 epochs was a reasonable training duration for this task and dataset with the chosen hyperparameters. If the curves were still sharply increasing at the end of epoch 5, it might suggest that training for more epochs could yield further improvements, but here they seem to be approaching a plateau. These plots visually confirm that the model is learning and improving throughout the fine-tuning process and provide a more granular view of this learning compared to just looking at the final per-fold scores. They help to confirm that the model is not overfitting significantly within the 5 epochs, as the evaluation metrics are still improving.

We now turn to the fine-tuned model's performance across meaning classes. Table 7 provides a detailed breakdown of the fine-tuned model's performance across all 17 meaning classes. Key findings are presented below the table.

Table 7. RoBERTa-large overall classification results

| Meaning class | Precision | Recall | F1 score | Support |
|---|---|---|---|---|
| Bodily harm | 0.95 | 0.95 | 0.95 | 40 |
| Breakthrough | 0.88 | 0.86 | 0.87 | 92 |
| Change | 0.85 | 0.64 | 0.73 | 45 |
| Decoding | 1.00 | 1.00 | 1.00 | 19 |
| Destruction | 0.85 | 0.86 | 0.86 | 132 |
| Disclosure | 0.96 | 0.96 | 0.96 | 45 |
| Economical freezing | 1.00 | 0.77 | 0.87 | 26 |
| Emergence | 0.91 | 0.91 | 0.91 | 32 |
| Emotional/mental freezing | 0.79 | 0.65 | 0.71 | 57 |
| EPS breakdown | 0.89 | 0.91 | 0.90 | 78 |
| Immobilization | 0.83 | 0.88 | 0.86 | 173 |
| Interruption | 0.82 | 0.80 | 0.81 | 46 |
| Physical/bodily freezing | 0.84 | 0.84 | 0.84 | 101 |
| Preservation | 0.93 | 0.96 | 0.94 | 71 |
| Suspension | 0.84 | 0.90 | 0.87 | 59 |
| Termination | 0.72 | 0.85 | 0.78 | 71 |
| Violation | 0.96 | 0.93 | 0.94 | 101 |
| Accuracy | | | 0.87 | 1188 |
| Macro average | 0.88 | 0.86 | 0.87 | 1188 |
| Weighted average | 0.87 | 0.87 | 0.87 | 1188 |

Key findings:

- Overall performance: Table 7 shows an overall accuracy of 0.87, with macro and weighted average F1 scores also around 0.87. This indicates strong overall performance on the dataset.
- High-performing classes: Several classes show particularly high performance, with precision, recall, and F1 scores close to or at 1.00. Examples include decoding (1.00 F1 score), bodily harm (0.95 F1 score), disclosure (0.96 F1 score), and preservation (0.94 F1 score). This suggests the model is very effective at identifying instances of these classes. As mentioned in Section X, *break* and *freeze* when interpreted with these senses are used with very limited theme arguments that tend not to be shared with other senses.

This characteristic can be said to contribute to the clear boundary distinction of these senses. In our dataset, *break* in the decoding sense is only used with *code* and *encryption*, while most instances of *break* in the disclosure sense are used with *news* and *story*. *Break* in the bodily harm sense typically takes body parts as its theme, and *freeze* in the preservation sense is most frequently used with food (ingredients), and sometimes with other substances/materials intended for preservation, with our dataset's examples showing the same characteristics.

- Classes with room for improvement: Some classes have lower F1 scores compared to the high-performing ones, indicating areas where the model struggles more. Notable examples include 'termination' (0.78 F1 score), 'change' (0.73 F1 score), and 'emotional/mental freezing' (0.71 F1 score).
- Precision vs. recall: For most classes, precision and recall are relatively balanced. However, for classes like change and emotional/mental freezing, the recall is notably lower than precision, suggesting the model is less likely to identify all true instances of these classes (more false negatives). Conversely, for termination, the recall is higher than precision, indicating the model is more likely to predict this class, sometimes incorrectly (more false positives).
- Support: The support column shows the number of instances for each class in the test sets across all folds. There is some class imbalance in the dataset, with immobilization having the most instances (173) and decoding having the fewest (19). While the stratified K-fold helped distribute these across folds, the model's performance on smaller classes can sometimes be more variable.

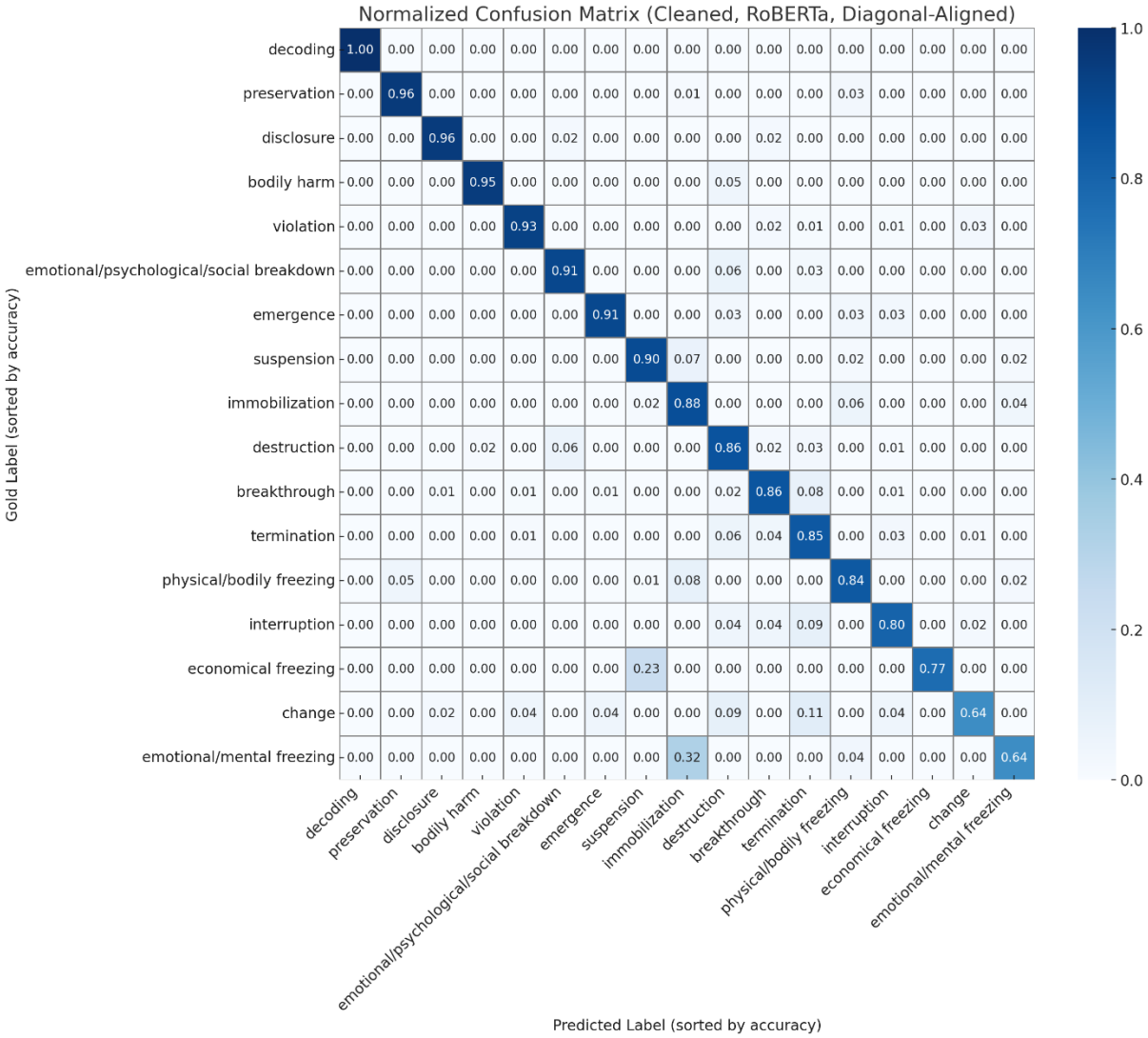
In summary, the overall classification results confirm that the fine-tuned RoBERTa-large model performs well on this multi-class sentence classification task, achieving high overall metrics. It also highlights specific classes where performance is excellent and others where there is still some room for improvement, potentially due to factors like class imbalance or the inherent difficulty of distinguishing certain senses.

Next, we examine the distribution of classification prediction errors shown in the confusion matrix to compare the differences between the meaning classes. Figure 10 presents the confusion matrix of the fine-tuned model for meaning class predictions. Similar to the probing results, systematic variation across semantic classes remains evident. Senses such as decoding, preservation, disclosure, and bodily harm again achieve the highest accuracies, confirming that senses tied to restricted and well-defined theme arguments are most robustly captured. However, compared to probing, fine-tuning substantially improves the recognition of these classes, raising their accuracies close to ceiling levels ($\geq$0.95).

At the lower end, abstract and metaphorical senses (e.g., emotional/mental freezing and change) continue to pose challenges, but their classification has also improved under fine-tuning. For instance, emotional/mental freezing, which showed substantial confusion with immobilization in the probing results, now achieves a markedly higher accuracy ($\approx$0.64), indicating better—but still incomplete—disentanglement.

Cross-class confusions remain most pronounced among semantically neighboring change-of-state events (e.g., physical/bodily freezing and immobilization), though the error rates are significantly reduced after fine-tuning. Similarly, misclassifications across abstract or metaphorical categories (such as interruption vs. termination and suspension vs. economical freezing) are less frequent compared to the probing baseline.

Figure 10. Confusion matrix for meaning-class prediction (fine-tuned RoBERTa-large)




Overall, fine-tuning yields a notable sharpening of class boundaries across the matrix. Whereas the probing model exposed fuzzy overlaps, especially in abstract domains, the fine-tuned model achieves more distinct separations, raising the overall accuracy to 87%. These improvements highlight the role of supervised adaptation in reinforcing sense boundaries that remain opaque under probing alone.

### 3.2.4. Analysis of Misclassified Examples

The recurrent misclassification tendencies observed in the confusion matrices raise the question of what semantic relationships exist between senses that have a high potential for confusion. Petersen and Potts (2023) analyzed the errors made by pre-trained models and found very interesting types of relationships between the predictions of LLM-based probe models and human linguistic annotations. In this section, we will analyze in detail the relationships between the gold and predicted labels observed in 150 misclassified sentences from the fine-tuned

RoBERTa-large model to identify the features that distinguish clearly distinguishable polysemous senses from those that are not.

Our analysis provides new evidence that systematic relationships exist between the gold and predicted labels, as shown in Table 8, and that senses with a high tendency for confusion are closely connected to other senses within this relationship. As noted by Petersen and Potts (2023: 497), most of the misclassification cases are, in fact, not actual "errors" and reflect the systematic relationships that hold between polysemous senses in context. In the following we discuss representative examples of (so-called) errors in detail.

Table 8. Major types of relationship between the gold and predicted labels in misclassified sentences

| Semantic relationship | Description |
|---|---|
| Metaphorical extension | One sense is the basic meaning, while the other is a metaphorical extension of it. |
| Semantic subcategories | Both senses are semantic subcategories of a shared, superordinate meaning. |
| Semantic overlap | Both senses are simultaneously active or are inseparably linked either causally or metaphorically within a specific context. |
| Context-dependent polysemy | Both senses are in principle available and the choice between them depends on the utterance context. |
| Others | - One sense is the primary meaning, and the other is a contextual implication that arises from it.<br>- The predicted meaning is the typical usage.<br>- Model prediction likely reflects a frequent associative link between the two meanings |

Table 9 is a selection of curated examples that involve a common pattern of metaphorical extension. In these examples, the gold meaning is the basic one, and the predicted meaning is a metaphorical extension of it, or vice versa.

Table 9. Metaphorical extension

| | Meaning | |
|---|---|---|
| Sentence | Gold | Predicted |
| 1. Will attending to this now make or <u>break</u> my rep? | destruction (figurative) | EPS breakdown |
| 2. Addison and his brothers continued to <u>break</u> horses the old way. | EPS breakdown | destruction |
| 3. indeed I was at one time fearful my feet would <u>freeze</u> in the thin mockersons which I wore. | physical/bodily freezing | emotional/mental freezing |
| 4. Its rear window <u>froze</u> and shattered as I passed through it. | physical/bodily freezing | immobilization |
| 5. Otherwise your system may <u>freeze</u>. | immobilization | physical/bodily freezing |

| | | | |
|---|---|---|---|
| 6. | power failures and computer malfunctions would bring business to a standstill, ground airlines and freeze automated teller machines. | immobilization | suspension |

In example 1, the gold meaning is the figurative interpretation of the destruction sense, whereas the predicted label is emotional/psychological/social breakdown. This misclassification suggests that the model sometimes links the destruction of abstract social foundations such as reputation or power with the metaphorical "breaking" of a person's emotional or mental state. Both involve a notion of collapse, but one is structural and the other psychological, indicating a metaphorical relationship.

In example 2, the gold label is emotional/psychological/social (EPS) breakdown, but the model predicts destruction. This reversal highlights the reciprocal metaphorical link between psychological collapse and physical destruction: the act of "breaking" horses is interpreted literally as a destructive process, even though in the gold annotation it reflects a metaphorical weakening or breakdown of resistance.

In example 3, the gold sense is physical/bodily freezing, but the predicted label is emotional/mental freezing. Here the model extends the physical loss of bodily movement caused by cold to a figurative inability to act, revealing the metaphorical overlap between physiological immobilization and emotional paralysis.

Across our dataset, the immobilization sense of *freeze* emerges as a recurrent source of misclassification, reflecting its semantic centrality and its frequent metaphorical extensions. The model often confuses immobilization with both literal and figurative freezing events, suggesting that this sense serves as a pivot in the semantic network of *freeze*. In example 4, the gold label is physical/bodily freezing, but the model predicts immobilization. This reflects the close association between literal freezing and the state of being immobilized, with the model prioritizing the static outcome over the physical process.

In example 5, the gold meaning is immobilization, whereas the prediction is physical/bodily freezing. This reversal shows that the model conflates the abstract state of inaction with the concrete physiological freezing event, underlining their metaphorical proximity.

In example 6, the gold sense is immobilization, but the model predicts suspension. Here the metaphorical extension links the inability to move with the halting of processes and systems, showing how immobilization at the bodily level can be conceptually mapped onto the suspension of institutional or technological activities.

Misclassification often arises when both the gold and predicted senses are sibling subcategories under a shared superordinate meaning. In such cases the model's "error" frequently reflects porous boundaries within a higher-level semantic category, rather than a wholesale mismatch. Table 10 is a selection of misclassified examples of semantic subcategories.

Table 10. Semantic subcategories

| Sentence | Meaning | |
|---|---|---|
| | Gold | Predicted |
| 1. At a slight 5.2 pounds and $2,999, it won't break your back or your bank account. | destruction (figurative) | bodily harm |
| 2. I realized that Bao had succeeded in breaking the endless chain of thought I'd been chasing. | interruption | termination |
| 3. Place fragile greens and herbs at top of ice chest, not next to ice (greens may freeze). | physical/bodily freezing | preservation |
| 4. his budget would freeze overall federal domestic discretionary spending for the next five years. | economical freezing | suspension |

In example 1, both labels instantiate the superordinate schema of physical breaking. The gold label targets figurative destruction (damage to resources/reputation), while the model gravitates to bodily harm due to strong lexical cues ("break your back"). Given their shared parent category, the prediction selects a sibling sense within the physical-damage family.

In example 2, both senses are subtypes of disruption of continuity. The gold label treats the event as a pause or break in ongoing thought (interruption), whereas the model reads it as a full stop (termination). The gradient boundary between temporary stoppage and ending explains the model's shift to a closely related sibling.

In example 3, both labels belong under physical freezing. The gold label highlights the freezing event/state, while the model selects preservation, likely guided by the pragmatic goal in the context (keeping produce fresh). Since preservation is often achieved by freezing, the model again prefers a sibling subcategory within the same parent.

In example 4, both are subtypes of cessation. Economical freezing denotes holding spending at a fixed level (no increases), whereas suspension implies halting activity altogether. The model collapses this nuance and selects the stricter cessation. Notably, for *freeze*, we frequently observe economical freezing being misclassified as suspension, indicating a recurrent ambiguity between hold constant and stop in policy/economic contexts.

We also find many cases where multiple senses can be activated and blended in the same example. Table 11 is a selection of curated examples of causal overlap in which the gold and predicted meanings are simultaneously active or are causally linked within a specific context.

Table 11. Causal overlap

| Sentence | Meaning | |
|---|---|---|
| | Gold | Predicted |
| 1. Can both sides exercise restraint or is someone likely to break this deal? | violation | termination |
| 2. Picasso broke his contract with Manach and returned in January, 1902, to Barcelona. | termination | violation |
| 3. The United States may break tradition and dip the US flag to British leaders at the Opening Ceremony today. | change | violation |

| | | |
|---|---|---|
| 4. he'd try to break the strategy of the president and the Democrats. | destruction (figurative) | termination |
| 5. when things get tense someone can break ties. | termination | breakthrough |
| 6. Networking with other teachers can help to break the isolation of the first years of teaching. | breakthrough | termination |
| 7. Indian society is breaking tradition and sanctioning the upward movement of its working-class females. | change | termination |
| 8. Most inexperienced singers break the sound after "stream," rather than connecting and building the sound through "merrily." | change | interruption |

The violation sense of *break* is closely tied to the termination sense in terms of causal relations: violating a rule, agreement, or contract often directly results in its termination. Because of this tight causal link, it is difficult to clearly separate the two categories in actual usage. In example 1, the gold label is violation while the model predicts termination. Far from being a simple error, this illustrates how the two senses are simultaneously active in context: to "break a deal" inherently involves both violating the agreement and terminating its validity. Similarly, in example 2, the gold annotation is termination but the model predicts violation. Here again, the senses overlap causally, since breaking a contract constitutes both the act of ending it and the act of violating its terms. In such cases, the model's prediction cannot easily be considered incorrect, because both senses are conceptually entailed.

Example 3 further shows how violation can overlap with other senses. The gold sense is change, but the prediction is violation. In this case, "breaking tradition" implies both a change in established practice and the violation of a social or cultural norm. The causal overlap between the two senses explains why the model oscillates between them. A similar pattern is found in example 4, where the gold sense is destruction but the prediction is termination. Here the destruction of a political strategy entails its termination, so again the two senses are mutually implicated.

Beyond violation, the subcategories of disruption of continuity—namely breakthrough, change, interruption, and termination—also show frequent causal overlap. Example 5 illustrates this clearly: the gold sense is termination, but the model predicts breakthrough. The contextual meaning of "breaking ties" involves both terminating a relationship and achieving a breakthrough in escaping constraints, showing their causal connection. Likewise, example 6 shows the reverse pattern: the gold sense is breakthrough, while the model predicts termination. The act of breaking isolation is simultaneously a breakthrough and the termination of that isolation.

In example 7, the gold label is change while the model predicts termination. Again, breaking tradition is both an act of change and the termination of a prior state, reflecting their overlap. Finally, in example 8, the gold sense is change and the model predicts interruption. Here the shift in musical flow can be understood as both a change and an interruption in continuity. These cases demonstrate that disruption-of-continuity senses form a network of

causally related subtypes, and that model "errors" often reflect genuine semantic blending rather than misclassification.

Causal overlap does not appear in the misclassification cases of *freeze*, but is instead a phenomenon frequently observed among the polysemous senses of *break*. In contrast, many of the misclassification cases with *freeze* involve gold and predicted senses that are metaphorically connected and overlap through metaphorical extension. Table 12 presents examples where both metaphorical extension and sense overlap are simultaneously at play.

Table 12. Extension-based overlap

| | Meaning | |
|---|---|---|
| Sentence | Gold | Predicted |
| 1. He saw how the boy froze --- how his face grew rigid with fear --- and reached out to hold his arm. | emotional/mental freezing | immobilization |
| 2. They froze on the spot at the sight of that hideous snapshot. | emotional/mental freezing | immobilization |
| 3. Almost immediately, the sound was answered by a low, liquid chuckle that froze Ace's spine. | immobilization | emotional/mental freezing |
| 4. Bahorel's blood ran cold and he froze upon hearing the voice of his enemy. | immobilization | emotional/mental freezing |

The immobilization sense of *freeze* and its extension to emotional/mental freezing frequently co-occur, reflecting their deep interconnection. This overlap arises from fundamental survival mechanisms—most notably the "Fight, Flight, or Freeze" response—in which physical immobility is often accompanied by emotional paralysis under threat. Strong affective shocks typically produce both bodily rigidity and psychological numbness, while physical stillness may in turn reinforce emotional fixation, as observed in trauma responses. The lexical metaphor further extends *freeze* from physical stillness to abstract domains, such as emotions, thereby reinforcing sense overlap.

The examples in Table 12 illustrate this interdependence. In examples 1 and 2, fear triggers both bodily immobility and emotional paralysis, making the distinction between gold (emotional/mental freezing) and predicted (immobilization) labels tenuous. Conversely, in examples 3 and 4, the gold and predicted senses are reversed, yet again, both readings are simultaneously plausible, as fear or shock induces both somatic and affective freezing. These cases demonstrate that the frequent misclassifications are not arbitrary errors but instead reflect the porous boundary between the two senses.

This stands in contrast to *break*, where overlap tends to emerge through causal connections between distinct senses (e.g., violation leading to termination, or destruction leading to interruption). By comparison, *freeze* illustrates a different dynamic: its misclassifications primarily reveal metaphorical extensions from bodily immobility to psychological paralysis. Thus, while *break* exemplifies causal overlap across polysemous senses, *freeze* exemplifies metaphorical overlap grounded in embodied experience.

We also see cases where both the gold meaning and the predicted meaning are in principle available, and the choice between them depends on the utterance context. Examples given in Table 13 illustrate this kind of context-dependent polysemy. In example 1, the interpretation of *break* depends on what bond denotes in context. If it refers to a physical/chemical bond, *break* naturally means destruction. If it instead evokes a social/relational connection, *break* could mean termination. Thus, both readings are contextually plausible, and the model's prediction aligns with one available option. In example 2, the verb *break* with storm can convey two opposite senses: the emergence or onset of a storm, or its relaxation/dissipation. In this context, where escape becomes possible after the storm broke, the gold label change points to relaxation/weakening. The model, however, selects the alternative sense of emergence. Both are licensed by usage patterns, but the context favors dissipation. In example 3, *break* operates in a context-dependent relation between disclosure and breakthrough. The event of a case "breaking" involves both the public disclosure of information and the sense of a significant investigative breakthrough. The two labels overlap causally: disclosure often leads to breakthrough. The model's choice highlights this causal intertwining of senses. In the sports context illustrated in example 4, the gold meaning points to change, i.e., the pitcher altering the motion or trajectory of the ball ("breaking pitch"). The model, however, defaults to the more general and prototypical sense of physical destruction, overlooking the specialized contextual meaning.

Table 13. Context-dependent polysemy

| | Meaning | |
|---|---|---|
| Sentence | Gold | Predicted |
| 1. Wouldn't that behavior break bonds and make some silicon atoms very unhappy? | destruction | termination |
| 2. The storm broke and I was able to escape to the other side of the river and meet up with Foxy and the crew. | change | emergence |
| 3. The murder was widely reported in the newspapers and on television. Not long after the case broke, Paul heard Rebecca's mother say something about Jeffrey's murderer. | disclosure | breakthrough |
| 4. I'm never surprised anymore to see how players break the balls. | change | destruction |

Taken together, the above cases demonstrate that context-dependent polysemy arises when multiple meanings are in principle possible but the utterance context selects the intended sense. Human annotators tend to privilege the contextually appropriate reading, whereas the model often defaults to more general, frequent, or prototypical senses.

Table 14 presents other types of misclassifications, where the relationship between the gold and predicted meanings is not one of straightforward overlap but rather contextual implication, typical usage, or frequent association. Example 1 is a case of contextual implication. The immediate meaning of *breaks* is a voice change, a physical alteration such as cracking or trembling. However, given the preceding context of "the pain of her loss," this physical sign

strongly implies an underlying emotional/psychological breakdown. The model chooses the latter, interpreting the symptom as evidence of the implied state. This is not semantic overlap in itself—voice breaking does not always mean emotional collapse—but in this particular context, the physical change implicates the psychological state.

Example 2 is another case of contextual implication. The gold label captures the physical freezing due to the environment, while the model predicts immobilization, a state entailed by freezing. The former naturally implies the latter, and the model latches onto the inferred state rather than the literal one.

In example 3, the error reflects the model's bias toward the typical usage of *break* as destruction. In agricultural or botanical contexts, however, *break* means "to sprout or emerge." The misclassification thus arises from the model's reliance on the dominant everyday sense rather than the contextually specialized meaning.

In example 4, violation of a rule or restriction is the intended sense, but the model instead predicts change. While the two meanings can sometimes overlap causally (violating a rule leads to change), in this case, such overlap is not at stake. The model's prediction likely reflects a frequent associative link between the two meanings in training data, rather than the contextual sense. Taken together, these examples highlight how misclassification may result not only from genuine semantic overlap but also from contextual implication, default to prototypical senses, or associative biases in the model's learned representations.

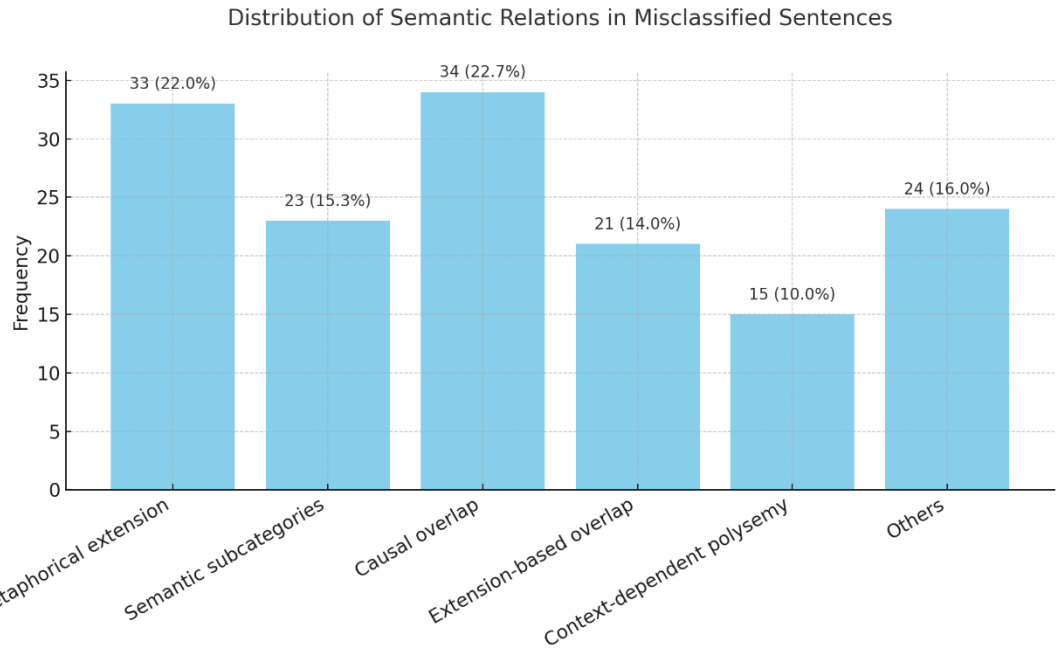
Table 14. Other types of misclassifications

| | Meaning | | Relationship |
|---|---|---|---|
| Sentence | Gold | Predicted | |
| 1. She tried to describe the pain of her loss. She doesn't produce tears, but her voice breaks. | change | EPS breakdown | Contextual implication |
| 2. Some fateful eve a deer had broken through the crust into a crevice of the ice field. It froze there with its head laid down on its forelegs. | physical/bodily freezing | immobilization | Contextual implication |
| 3. This spring, just as the buds broke, we sprinkled ammonium sulfate on top of the ground around the drip line of each plant. | emergence | destruction | Typical usage (predicted meaning) |
| 4. Dieters do not typically break their diet by having just one small biscuit, or one square of chocolate, or one tiny cake. | violation | change | Frequent association (predicted meaning) |

Figure 11 shows the distribution of semantic relation types identified in the 150 misclassified cases. We can observe that different types of semantic relations account for the confusions between gold and predicted meanings. Among them, overlap relations (causal

overlap + extension-based overlap) constitute the largest share, with 36.7% of all cases. This indicates that senses of *break* and *freeze* that are semantically connected through causal chains or extended from one another are particularly prone to confusion. Metaphorical extensions also contribute substantially (22%), reflecting the tendency of both verbs to extend from concrete physical meanings to abstract domains. Semantic subcategories (15.3%) and context-dependent polysemy (10%) highlight situations where finer-grained distinctions or contextual cues are required to determine the intended sense. Finally, a non-negligible portion (16%) falls into miscellaneous others.

Figure 11. Frequency of semantic relations in misclassified examples




Taken together, the results of our analysis suggest that semantic overlap—causal and extension-based—serves as the most significant factor underlying misclassification tendencies.

### 3.3. Explaining Sense Boundaries in Verbal Polysemy

Based on the meaning classification patterns of the fine-tuned models discussed in Section 3.2, we can derive the key distributional and semantic factors that contribute to the sharpening and blurring of the boundary between the polysemous senses of *break* and *freeze*. These factors are summarized as follows:

1. Restricted distribution:
   - Occurrence with restricted and well-defined theme arguments
2. Semantic proximity:
   - Shared superordinate category
   - Restrictedness of shared superordinate category
   - Metaphorical link

3. Semantic overlap:
   - Causal overlap
   - Extension-based overlap

Restricted distribution contributes to making sense boundaries clear, whereas semantic proximity and overlap make the boundaries more indistinct and fuzzy. As discussed in Section 3.2, the highest-performing classes in the fine-tuning experiment—decoding, preservation, disclosure, and bodily harm—share the distributional characteristic of being used with very restricted and distinctly defined theme arguments. In contrast, the senses of *break* and *freeze* that show a higher tendency for confusion have a higher degree of semantic similarity or frequent overlap with other senses. In this section, we show that the degree of confusion observed in these senses can be explained by the difference in the extent to which they are connected to other senses through semantic proximity and overlap. A summary of the core of this explanation is presented in Table 15.

Table 15. Sense-level performance and confusion patterns explained by semantic proximity and overlap

| Model performance | Meaning classes (accuracy) | Degree of connection with other senses through semantic proximity and overlap |
|---|---|---|
| Highest-performing | decoding (1.00)<br>preservation (0.96)<br>disclosure (0.96)<br>bodily harm (0.95) | Low<br>↕<br>High |
| Above-average performing | violation (0.93)<br>EPS breakdown (0.91)<br>emergence (0.91)<br>suspension (0.90)<br>immobilization (0.88) | |
| Below-average performing | destruction (0.86)<br>breakthrough (0.86)<br>termination (0.85)<br>physical/bodily freezing (0.84)<br>interruption (0.80) | |
| Lowest-performing | economical freezing (0.77)<br>change (0.64)<br>emotional/mental freezing (0.64) | |

Table 16 summarizes the cross-class confusions shown in the distribution of classification prediction errors in Figure 10 for the three highest-performing classes, along with the semantic relationship between the gold and predicted meanings. As shown in the table, the observed cases of cross-class confusion cases related to these senses are mainly due to the relationships of semantic subcategories, metaphorical extension and overlap between the gold meaning and the predicted meaning, but the number of such cases is very small. Based on this, we can conclude that the highest-performing meaning classes have the weakest degree of connection with other senses through semantic proximity and overlap, and thus have the most clearly distinguished sense boundaries.

Table 16. Highest-performing classes

| A: Meaning class | B: Confused meaning classes | Semantic relationship between A and B |
|---|---|---|
| preservation | physical/bodily freezing | Semantic subcategories |
| | immobilization | Metaphorical extension |
| disclosure | breakthrough | Causal overlap |
| | EPS breakdown | Other |
| bodily harm | destruction | Semantic subcategories |
| Number of cross-class confusions | | 7 |

Table 17 presents the cross-class confusions shown by the five above-average performing classes, along with the semantic relationship between the gold and predicted meanings. As shown in the table, the number of observed cases of cross-class confusion related to these senses increases to 40. As Romain (2017, 2022) showed through a cluster analysis of theme arguments by sense for *break* and *freeze*, these senses tend to be used with relatively restricted and well-defined theme arguments. Despite this, their higher tendency for confusion compared to the highest-performing classes can be explained by their closer connection to other senses. As discussed in Section 3.2.4, violation is one of the meanings that causally overlap with other *break* senses (see Table 11), and immobilization is metaphorically linked and strongly overlaps with emotional/mental freezing (see Table 12). Furthermore, the emotional/psychological/social breakdown and suspension senses are connected to the concrete and basic meanings of *break* and *freeze*, respectively, through metaphorical extension. Therefore, these senses have a higher degree of connection with other senses through semantic proximity and overlap than the highest-performing classes, which results in more blurred sense boundaries and a higher tendency for confusion.

Table 17. Above-average performing classes

| A: Meaning class | B: Confused meaning classes | Semantic relationship between A and B |
|---|---|---|
| violation | breakthrough, termination | Causal overlap |
| | EPS breakdown | Other |
| EPS breakdown | destruction | Metaphorical extension |
| | termination | Other |
| emergence | destruction | Typical usage |
| | interruption, physical/bodily freezing | Other |
| suspension | immobilization, physical/bodily freezing | Metaphorical extension |
| immobilization | emotional/mental freezing | Extension-based overlap |
| | physical/bodily freezing, suspension | Metaphorical extension |
| Number of cross-class confusions | | 40 |

Table 18 summarizes the performance of the polysemous senses of *break* and *freeze*, highlighting those with below-average accuracy and higher confusion tendencies.

Table 18. Below-average performing classes

| A: Meaning class | B: Confused meaning classes | Semantic relationship between A and B |
|---|---|---|
| destruction | EPS breakdown | Metaphorical extension |
| | bodily harm | Semantic subcategories |
| | breakthrough, termination, interruption | Causal overlap |
| breakthrough | termination, interruption | Semantic subcategories and causal overlap |
| | destruction, emergence, violation, disclosure | Causal overlap |
| termination | destruction, violation | Causal overlap |
| | breakthrough, interruption, change | Semantic subcategories and causal overlap |
| physical/bodily freezing | preservation | Semantic subcategories |
| | immobilization, suspension | Metaphorical extension |
| interruption | breakthrough, termination, change | Semantic subcategories and causal overlap |
| | destruction | Causal overlap |
| Number of cross-class confusions | | 61 |

As shown in the table, the two core meanings of *break* and *freeze*—namely destruction and physical/bodily freezing—exhibit below-average classification performance and high confusion tendencies. This pattern can be explained by three factors.

- Occurrence with a wide range of theme arguments:
  The *destruction* sense of *break* applies to a broad spectrum of breakable entities—physical objects, substances, structures, living beings, even persons—while physical/bodily freezing extends across natural entities, artifacts, organisms, and human subjects that can become rigid from cold. This broad applicability makes these senses less restricted and therefore harder to delimit.
- Sources of metaphorical extension:
  Both senses function as conceptual sources for metaphorical extensions. *Break* extends from physical destruction to social, psychological, and institutional domains (e.g., *break a promise, break down emotionally*), while *freeze* extends from physical immobilization to abstract domains of economic suspension or emotional paralysis. Their role as source domains means that they maintain strong semantic similarity with extended senses, increasing confusion rates.
- Causal overlap for *break* (destruction sense):
  As discussed earlier (see Table 11), the destruction sense of *break* can overlap causally with other meanings, such as termination, change or breakthrough. Destroying an entity often entails ending its function or disrupting continuity, which blurs the boundary between these senses and fuels misclassification.

In addition to these two core meanings, the disruption-of-continuity subcategories—breakthrough, termination, and interruption—also show high tendencies for confusion. As indicated in Table 18, their classification errors are driven by two main factors:

- Semantic subcategory relations**:** these senses represent closely related subtypes under the overarching category of disruption of continuity, such that the difference between a temporary disruption (interruption), a complete end (termination), and a decisive shift (breakthrough) is often context-dependent.
- Causal overlap: in real usage, these events are frequently causally chained—termination can lead to breakthrough, and interruption may culminate in change. These causal links create systematic overlap, which the model reflects in its predictions.

Taken together, the below-average performance of destruction, physical/bodily freezing, and the disruption-of-continuity senses reflects their dual role as central hubs in the polysemy network: they are semantically broad, generative of extensions, and embedded in causal or categorical overlaps with neighboring meanings. As a result, their sense boundaries are inherently more diffuse and thus more prone to confusion in model prediction.

Table 19 highlights three senses that exhibit particularly strong tendencies toward misclassification: economic freezing, change, and emotional/mental freezing. First, economic freezing shows a high rate of confusion with suspension. Both belong to the restricted cessation category, which—unlike broader categories such as physical breaking or freezing—contains only these two subtypes. Their meanings are thus mutually exclusive yet distributionally and contextually very similar. As discussed in Chapter 2, distributional semantic analysis likewise revealed a strong overlap between these two senses, reflecting their near-synonymous behavior in real usage.

Second, change emerges as the sense most extensively overlapped with others. It exhibits causal connections to the largest number of alternative senses and displays semantic overlap with multiple related subcategories. This broad connectivity is further reinforced distributionally, where change frequently co-occurs with and is embedded among its neighboring senses.

Third, consistent with earlier discussion, emotional/mental freezing strongly overlaps with immobilization. Here, too, the distributional patterns reveal significant co-occurrence and embedding, confirming that the figurative extension from bodily immobilization to emotional paralysis creates substantial confusion in model predictions.

Table 19. Lowest-performing classes

| A: Meaning class | B: Confused meaning classes | Semantic relationship between A and B |
|---|---|---|
| economical freezing | suspension | Semantic subcategories |
| change | termination, interruption | Semantic subcategories and causal overlap |
| | disclosure, violation, emergence, destruction | Causal overlap |
| emotional/mental freezing | immobilization | Extension-based overlap |
| | physical/bodily freezing | Metaphorical extension |
| Number of cross-class confusions | | 42 |

These lowest-performing classes illustrate that high confusability is most pronounced when senses are restricted but near-synonymous (economic freezing and suspension), broadly

interconnected through causal overlap (change), or strongly interconnected through extension-based overlap (emotional/mental freezing and immobilization).

To summarize the discussion in this section, we have argued that classification performance differences among the senses of *break* and *freeze* are explained by their degree of semantic proximity and overlap. High-performing senses (e.g., decoding, preservation) have narrow, distinctive distributions and minimal overlap, yielding clear boundaries. Above-average senses show stronger connections to neighboring meanings through causal, metaphorical, or extension-based relations, leading to moderate confusions. Below-average senses, including prototypical meanings like destruction and physical/bodily freezing, function as semantic hubs with broad applicability and strong overlap, which diffuse their boundaries. Finally, the lowest-performing senses (economic freezing, change, emotional/mental freezing) exhibit the greatest confusability because of near-synonymy, extensive causal overlap, or strong extension-based overlap. In short, restricted distribution sharpens sense distinctions, while semantic proximity, connectivity and overlap systematically blur them**.**

### 3.4. Conclusions

This chapter has proposed a framework for investigating polysemous verb sense boundaries through probing and fine-tuning contextualized language models. Sections 3.1 and 3.2 established baseline performance through probing experiments and then evaluated sense classification through fine-tuning. The results showed that, while models achieve relatively high overall accuracy, systematic patterns of misclassification emerge. The error analyses identified key factors underlying sense confusability, including restricted distribution, semantic proximity, connectivity, and overlap.

Section 3.3 advanced a novel explanatory perspective: differences in classification performance across senses of *break* and *freeze* can be systematically explained by their degree of semantic proximity and overlap with other senses**.** High-performing senses were those with narrow distributions and clear boundaries, whereas the lowest-performing senses (economic freezing, change, emotional/mental freezing) were the most deeply embedded in networks of proximity or causal/extension-based overlap. This approach integrates insights from lexical semantics and cognitive linguistics with those from distributional semantics in computational linguistics, using contextualized models as a bridge. Our results support a view of polysemy as a continuum that ranges from near-synonymous relations (e.g., economic freezing and suspension) to homonymy-like relations (e.g., the ‘emergence’ vs. ‘weakening’ readings of *the storm broke*).

These findings raise several important questions for further research:

1. Alignment between human and AI classifications and judgments – To what extent do the classification patterns of contextualized models correspond to human judgments of sense distinctions?
2. Explanatory power for human sense perception – Can the framework developed here account for how humans perceive and demarcate sense boundaries?
3. Mental representation – How are the classifications and judgments of both humans and AI systems integrated into the mental representation of polysemous verbs?

These questions form the basis for Chapter 4, where we directly compare human and model performance on sense distinctions.

# Chapter 4

## Sense Boundaries in Polysemy: Insights from Human and Generative AI Experiments

Chapter 3 demonstrated that variation in model performance can be systematically explained in terms of restricted distribution and the degree of semantic proximity and overlap with other senses. This account predicts that senses more strongly interconnected with others will exhibit fuzzy boundaries and higher rates of confusion. The present chapter extends this perspective by evaluating whether the same explanatory framework can account for both human and AI judgments of sense distinctions. Specifically, we conduct three complementary experiments designed to probe the perception of polysemous sense boundaries.

In Chapter 3, our focus was on contextualized language models, specifically BERT, RoBERTa and ALBERT, which are transformer-based encoder models. These models generate context-sensitive embeddings that allow fine-grained analyses of distributional patterns but are not designed for text generation. By contrast, the present chapter investigates generative AI models, represented by GPT and Gemini. These are transformer-based decoder models (large language models) with powerful generative capabilities, accessible via APIs, and thus particularly suitable for implementing rating and judgment experiments in parallel with human participants. This contrast enables a direct comparison between encoder-based models of contextualization and decoder-based models of generation, highlighting their respective contributions to the study of polysemous sense boundaries.

The overarching research questions addressed in this chapter are as follows:

1. Which polysemous senses are perceived as clearly distinguishable, and which are more prone to confusion or overlap?
2. To what extent do human sense selection and judgments of sense distinctions align with or diverge from the classification patterns of contextualized language models and generative models?
3. Are human and AI judgments of sense boundaries categorical or graded, and how do they reflect semantic proximity and overlap among senses?

This chapter is structured as follows. Section 4.1 reports on the human meaning selection experiment, which directly compares human patterns of sense selection for *break* and *freeze* with the classification patterns observed in contextualized models in Chapter 3. Following the approach of Erk et al. (2009, 2013), we also conduct two rating experiments to further investigate the perception of polysemous sense boundaries. Section 4.2 presents the sense applicability judgment experiment, in which human participants and models provide graded ratings of sense applicability. Section 4.3 introduces the usage similarity rating experiment, designed to assess how humans and models evaluate similarity relations among usage pairs. Both rating experiments explore whether participants make use of a graded scale or persist in making binary decisions even when there is the option for a graded response. The sense applicability judgment experiment tests to what extent the judgments on senses fall into clear-cut boundaries, while the similarity rating experiment allows us to explore perceived similarity of verb usages in context. The findings of these experiments are integrated in Chapter 5 with

the results of Chapter 3 to advance a broader account of the mental representation of polysemous verbs.

### 4.1. Human Meaning Selection Experiment

#### 4.1.1. Methods

The purpose of this study is to verify whether senses that are more strongly interconnected with others also show a high error rate and a strong tendency for confusion in human sense selection, just as they do in language models' sense predictions. For comparison with the models' classification patterns discussed in Chapter 3, the same verbs (*break* and *freeze*) are used as target words in this experiment.

**Materials.** The data for this study consisted of 17 meaning classes (11 for *break* and 6 for *freeze*). For each meaning class, five sentences were created, yielding a total of 85 items. Each item was presented as a pair of a short context sentence and a target sentence. The context sentence was designed to provide disambiguating cues for interpreting the target word. For example:

(1) Context: She was deep in thought, solving a complex problem.
Target: The loud alarm broke her concentration. [interruption sense]

In this example, the context sentence establishes a mental activity that can be disrupted, guiding the interpretation of *broke* in the target sentence as an interruption sense rather than physical destruction.

The next step in the materials design was to construct three-alternative forced-choice options**.** For each sentence, participants were asked to choose one of three alternatives: (1) the Correct label, (2) a Confusable label, or (3) a Non-confusable label.

- The Correct label corresponds to the intended meaning of the target word in the given sentence.
- The Confusable label was selected such that it bears semantic connectivity, proximity, or overlap with the correct label, making it a plausible competitor.
- The Non-confusable label, by contrast, was chosen so that it is semantically less tied to the correct label, with distinct semantic properties of the theme argument and without semantic connectivity or overlap.

Table 1 illustrates five sentences where the correct label is the interruption sense of *break*. In the first example, both the correct and confusable labels (interruption and termination) belong to the subcategories of the disruption-of-continuity domain. In the fourth example, interruption (correct) and destruction (confusable) are related through contextual implication, since interruption of coverage can imply its (metaphorical) destruction. In contrast, the non-confusable labels (decoding, emergence, bodily harm) share no semantic proximity or connectivity with the intended sense and thus function as clear distractors.

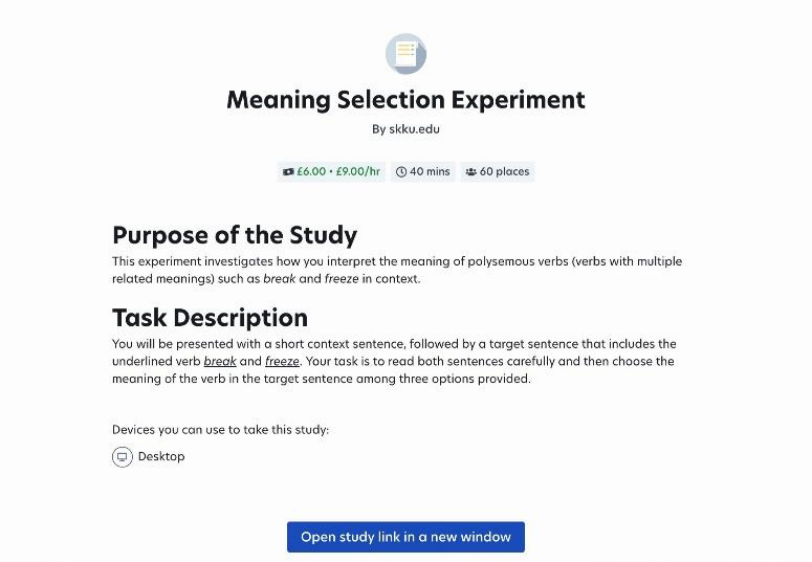
Table 1. Example of label construction (interruption sense)

| Context | Target | Correct label | Confusable label | Non-confusable label |
|---|---|---|---|---|
| We were in the middle of a serious conversation about her future. | A sudden knock on the door broke the flow of our discussion. | interruption | termination | decoding |
| She was deep in thought, solving a complex problem. | The loud alarm broke her concentration. | interruption | termination | emergence |
| The room was silent for a long time. | A phone ringing broke the stillness. | interruption | termination | decoding |
| The live broadcast was covering the press conference. | They broke the coverage to report breaking news. | interruption | destruction | bodily harm |
| He was making great progress on his essay. | A power outage suddenly broke his momentum. | interruption | destruction | bodily harm |

In total, 85 sentence items were constructed in this way and then randomized before being presented to participants. The full experimental dataset is available at the GitHub repository: https://github.com/hanjung-25/clmsemantics/tree/data.

**Participants, task, and procedure.** We collected human responses online through Prolific. Sixty native English speakers (31 male, 29 female) aged between 20 and 67 participated in the study. Participants were instructed to complete a three-alternative forced-choice (3AFC) task. For each of the 85 sentence items, they were presented with a short context sentence and a target sentence containing the verb *break* or *freeze*. Their task was to read both sentences carefully and then select one of three options: (i) the Correct label, corresponding to the intended meaning in context; (ii) a Confusable label, representing a closely related or potentially overlapping meaning; or (iii) a Non-confusable label, representing a sense that was less semantically connected and contextually implausible. A screenshot of the Prolific interface for this task is displayed in Figure 1.

Figure 1. Screenshot of the Prolific interface for the meaning selection (3AFC) task




The instructions emphasized that participants should rely on their intuitive understanding of the sentences and choose the meaning that best fit the target word's use. The full set of participant instructions is provided in Appendix D.

Before beginning the main experiment, participants were provided with brief descriptions and examples of all 17 meaning classes of *break* and *freeze*. They then completed a short comprehension check to ensure familiarity with the sense inventory and task format. All 60 participants successfully passed the comprehension check and completed the task, so their data were included in the final analysis.

The total duration of the experiment was less than 40 minutes. Each participant completed all 85 items and received a payment of £4, resulting in a complete dataset of 60 participants (average expected compensation rate: £6 per hour).

### 4.1.2. Results

**Overall accuracy and response distributions across meaning classes.** A total of 5,100 responses were collected from 60 participants across 85 sentences. As shown in Table 2, 4,293 responses (84.2%) were correct, while 807 (15.8%) were incorrect. Most errors involved the confusable option**:** 711 of 807 incorrect responses (88.1%) selected the confusable label, whereas 96 (11.9%) chose the non-confusable label. The overall human accuracy of 84.2% is slightly lower than the 87% classification accuracy of the RoBERTa-large model reported in Chapter 3.

Table 2. Overall accuracy of human meaning selection responses

| Total responses | Correct responses (accuracy %) | Incorrect responses (error %) | | |
|---|---|---|---|---|
| 5,100 | 4,293 (84.2 %) | 807 (15.8%) | Confusable label | 711 |
| | | | Non-confusable label | 96 |

We now examine the distribution of correct and incorrect responses across meaning classes using confusion matrices. Figure 2 presents the confusion matrix by raw counts, and Figure 3

presents the normalized version showing proportions. The confusion matrices provide a detailed picture of how different sense classes vary in their robustness and susceptibility to confusion, laying the groundwork for subsequent comparisons between human and AI judgments. The matrices reveal clear variation in error tendencies across meaning classes. High-accuracy classes such as bodily harm, destruction, termination, disclosure, and decoding display strong diagonals, indicating that participants consistently selected the correct label with minimal confusion. By contrast, lower-accuracy classes, notably economical freezing, emergence, change and immobilization, exhibit substantial off-diagonal errors. These classes are more frequently confused with semantically related senses—for instance, *change* is often confused with interruption or termination, and immobilization with physical/bodily freezing and emotional/mental freezing.

Figure 2. Confusion matrix for human meaning selection (by count)

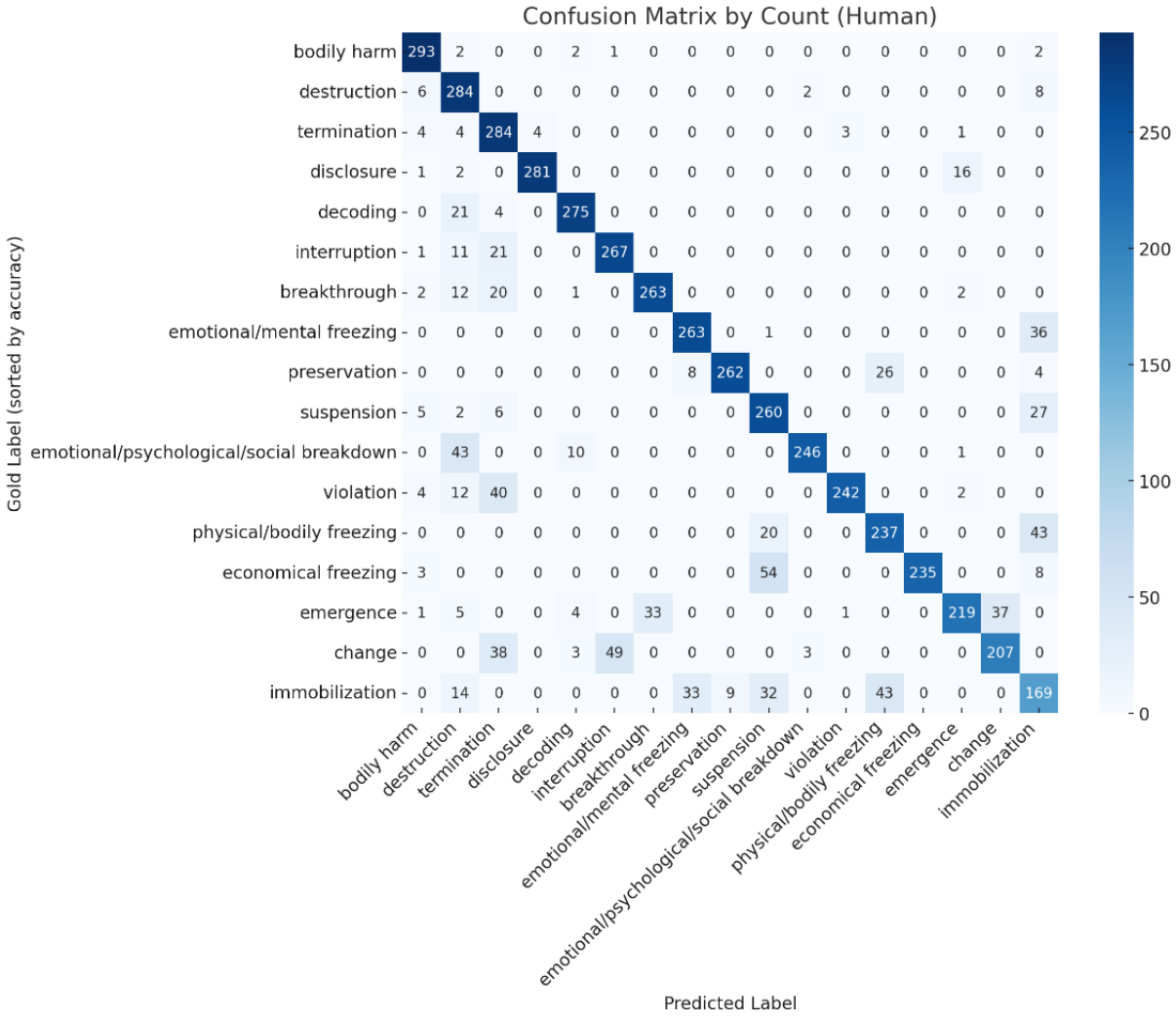
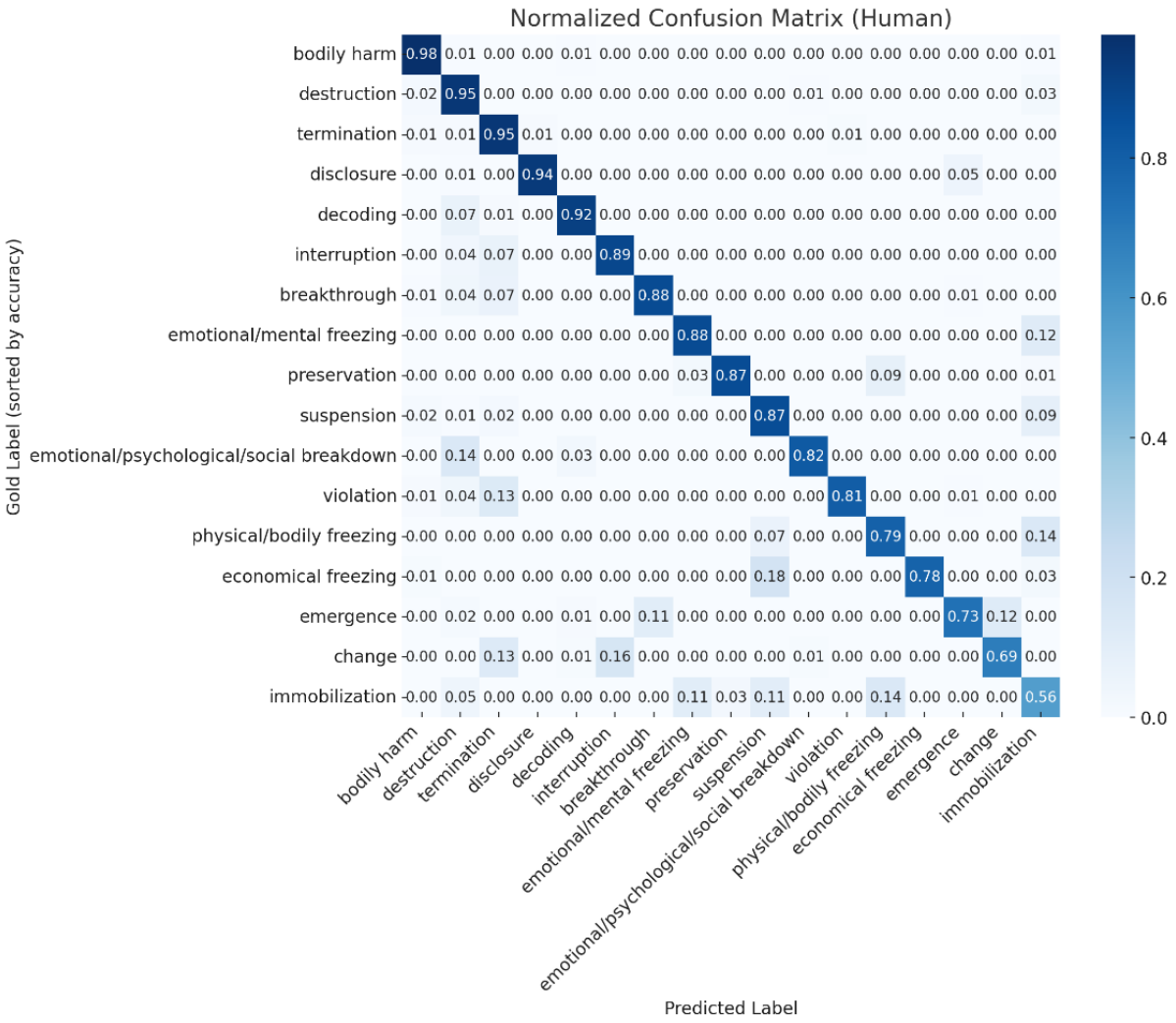


Figure 3. Confusion matrix for human meaning selection (by proportion)




These patterns are further summarized in Table 3, which groups the 17 meaning classes by their accuracy levels. The table shows a clear stratification:

- **Highest-accuracy senses** such as bodily harm, destruction, termination, disclosure, and decoding were identified correctly in over 90% of cases, reflecting robust sense boundaries with little susceptibility to confusion.
- **Above-average senses** including interruption, breakthrough, emotional/mental freezing, preservation, and suspension also performed well, though with somewhat higher rates of confusable errors.
- **Below-average senses** such as emotional/psychological/social breakdown, violation, physical/bodily freezing, economical freezing, and emergence display more frequent misclassifications, consistent with the overlap patterns observed in the confusion matrices.
- Finally, the **lowest-accuracy senses**, change (69%) and immobilization (56%), show the strongest susceptibility to confusion, aligning with their tendency to overlap semantically with multiple other senses.

Table 3. Sense-level accuracy

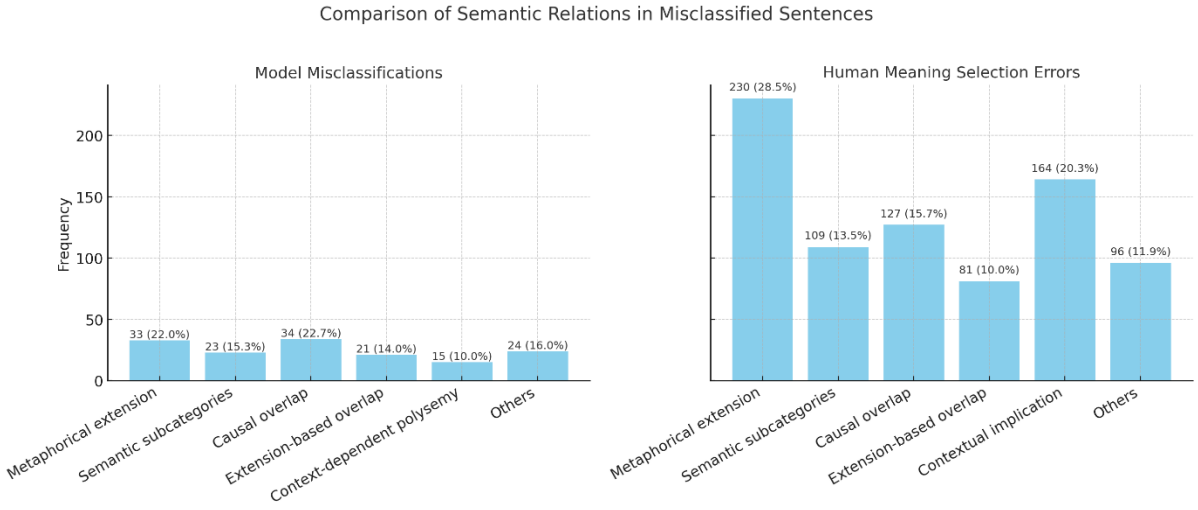
| Selection accuracy | Meaning classes (accuracy%) |
|---|---|
| Highest-accuracy | bodily harm (98)<br>destruction (95)<br>termination (95)<br>disclosure (94)<br>decoding (92) |
| Above-average accuracy | interruption (89)<br>breakthrough (88)<br>emotional/mental freezing (88)<br>preservation (87)<br>suspension (87) |
| Below-average accuracy | emotional/psychological/social breakdown (82)<br>violation (81)<br>physical/bodily freezing (79)<br>economical freezing (78)<br>emergence (73) |
| Lowest- accuracy | change (69)<br>immobilization (56) |

We will now analyze the semantic relationships between the correct and incorrect answers in the 807 incorrect responses. Figure 4 compares the distribution of semantic relations in misclassified sentences produced by the model (left) and in human meaning selection errors (right). Several important patterns emerge. First, both humans and the model show a high proportion of errors involving metaphorical extension and two forms of overlap, confirming that sense boundaries are especially fuzzy where meanings are extended or interconnected. For example, causal overlap and extension-based overlap together account for a substantial portion of model errors, while similar tendencies are observed in human judgments, though at different proportions.

Second, notable differences are observed between the two. Contextual implication appears as a major source of human error, reflecting participants' tendency to infer implicated or pragmatically enriched senses beyond the literal target. By contrast, the model produced only a small number of such cases, mostly subsumed under the "others" category. This indicates that human participants are more sensitive to pragmatic inference and context-driven sense extension, whereas the model relies more heavily on semantic proximity and distributional similarity. Finally, semantic subcategories (e.g., confusion within a shared superordinate domain) play a comparable but somewhat smaller role for both humans and the model, while causal overlap is relatively more prominent in the model's misclassifications.

\

Figure 4. Frequency of semantic relations in human and models' errors




Overall, the comparison highlights both the shared difficulty of handling semantically connected senses and the distinct ways in which humans and generative models exploit context: humans through pragmatic enrichment, and the model through distributional and semantic proximity.

**Analysis of sense-level accuracy in terms of semantic relationships.** Let us now examine in detail the relationships between the correct and incorrect answers (labels) observed in the 807 incorrect responses to determine what factors explain the accuracy differences between the meaning classes shown in Table 3 and the confusion matrices. Table 4 presents the distribution of incorrect answers for the highest-accuracy senses.

Table 4. Highest-accuracy classes

| A:<br>Correct label | B:<br>Incorrect label (selected label) | Semantic relationship<br>between A and B |
|---|---|---|
| bodily harm | destruction | Semantic subcategories |
| | immobilization | Contextual implication |
| | decoding, interruption | Other |
| destruction | bodily harm | Semantic subcategories |
| | EPS breakdown | Metaphorical extension |
| | immobilization | Contextual implication |
| termination | destruction | Contextual implication |
| | bodily harm, disclosure,<br>violation, emergence | Other |
| disclosure | destruction | Contextual implication |
| | emergence | Causal overlap |
| | bodily harm | Other |
| decoding | destruction | Contextual implication |
| | termination | Contextual implication |
| Number of incorrect responses | | 84 |

This set includes bodily harm, disclosure, and decoding, which also ranked among the highest-

performing senses in the RoBERTa-large model results discussed in Chapter 3. By contrast, destruction and termination—which were below-average performing senses in the model's predictions—also appear in this category for the human experiment. This difference can be attributed to the present experimental design. In this study, the five sentences constructed for the *destruction* sense of *break* all involved straightforward physical breaking events (e.g., example (2)), without figurative extensions or overlaps with other senses:

(2) Context: The children were roughhousing in the living room.
Target: They accidentally broke the old wooden chair.

Similarly, the termination items did not include contexts where the sense might overlap with other interpretations, but they do include items that can carry an implicature of figurative destruction, as in example (3):

(3) Context: They had been arguing for weeks, unable to resolve their issues.
Target: She decided to break the relationship once and for all.

For sentences where the correct label was disclosure or decoding, some participants instead selected destruction as the best-fitting meaning. These misclassifications reflect a figurative implicature, as in example (4), where decoding metaphorically entails breaking apart an encoded structure. All of these senses—destruction, termination, disclosure, and decoding—share the overarching conceptual core of disruption of integrity, which underlies these implicature-based connections.

(4) Context: The hacker stared at the encryption for hours, typing lines of code with unwavering focus.
Target: She broke the hidden message behind the ancient script.

Within the highest-accuracy classes, there was only one case of genuine semantic overlap observed. In example (5), disclosure causally overlaps with emergence, as the act of disclosure (cause) by the newspaper directly brought about the story's emergence (effect) into the world. Sixteen responses selected emergence instead of disclosure as the correct label.

(5) Context: The media had been investigating the politician's offshore accounts for months.
Target: A major newspaper finally broke the story this morning.

Taken together, the errors observed in the highest-accuracy senses were predominantly confusable labels linked through contextual implicature, rather than direct semantic overlap. This suggests that, as with the RoBERTa-large model's best-performing senses, these classes are characterized by relatively sharp boundaries and minimal connectivity to other senses, making them the most robust against confusion.

Table 4 presents the distribution of incorrect answers for the above-average accuracy senses. In this set, the total number of incorrect responses rises to 184, with 68 cases of semantic overlap identified.

Table 5. Above-average accuracy classes

| A:<br>Correct label | B:<br>Incorrect label (selected label) | Semantic relationship<br>between A and B |
|---|---|---|
| interruption | termination | Semantic subcategories |
| | destruction | Contextual implication |
| | bodily harm | Other |
| breakthrough | termination | Semantic subcategories and contextual implication |
| | destruction | Extension-based overlap |
| | bodily harm, emergence | Other |
| emotional/mental freezing | immobilization | Extension-based overlap |
| | suspension | Other |
| preservation | physical/bodily freezing | Semantic subcategories |
| | immobilization, emotional/mental freezing | Metaphorical extension |
| suspension | immobilization | Metaphorical extension |
| | destruction, termination | Contextual implication |
| | bodily harm | Other |
| Number of incorrect responses | | 184 |

This group includes preservation and suspension, which were among the higher-performing senses in the RoBERTa-large model results. For the preservation sense, errors were mainly due to participants selecting other physical senses of *freeze*, such as physical/bodily freezing or immobilization, as the most appropriate meanings. For suspension, errors prominently involved immobilization, reflecting the strong semantic connection between these two senses, both grounded in the notion of stopping or stasis.

Interestingly, emotional/mental freezing—which belonged to the lowest-performing senses in the model experiment—appears here as an above-average sense in the human data. Nonetheless, its error patterns remain consistent: participants frequently confused it with immobilization, reflecting their strong semantic overlap.

Among the *break* senses, interruption and breakthrough are included in this set. While these senses were below-average in the model results, their human accuracy falls into the above-average range. The most frequent incorrect response for both senses was termination, consistent with their shared status as subcategories of disruption of continuity. For breakthrough, this relationship can be seen in examples (6) and (7). In both cases, the act of "breaking through" a barrier entails the termination of the previous state:

(6) Context: After months of intense training and preparation,
Target: she finally broke the world record in the 100-meter sprint.

(7) Context: Her success inspired many others in the community.
Target: She broke the long-held prejudices about women in science.

In addition, 12 incorrect responses were found in items where *breakthrough* overlapped with *destruction* (examples (8) and (9)). As discussed in Chapter 3, such cases are difficult to regard as genuine "errors," since the two senses are inherently fused. In (8), the act of

demolishing rubble is simultaneously an act of breakthrough, while in (9), dismantling the metaphorical barrier of gender prejudice is inseparable from the notion of achieving breakthrough.

(8) Context: The rescue team worked tirelessly through the night.
Target: They broke the barrier of rubble and reached the trapped survivors.
(9) Context: For years, she faced rejection in a field dominated by men.
Target: She broke the gender barrier to become the first female engineer at the firm.

Overall, the above-average accuracy senses show greater degrees of semantic proximity and overlap with other senses than the highest-accuracy group. This accounts for their relatively higher rate of incorrect responses compared to the more robust classes described in Table 4.

We now turn to the error patterns found in the below-average accuracy classes. Table 6 presents the distribution of incorrect answers for this group. In total, 321 incorrect responses were observed, including 72 cases of semantic overlap.

Table 6. Below-average accuracy classes

| A:<br>Correct label | B:<br>Incorrect label (selected label) | Semantic relationship<br>between A and B |
|---|---|---|
| EPS breakdown | destruction | Metaphorical extension |
| | decoding, emergence | Other |
| violation | termination | Causal overlap or contextual implication |
| | destruction | Contextual implication |
| | bodily harm | Other |
| physical/bodily freezing | immobilization, suspension | Metaphorical extension |
| economical freezing | suspension | Semantic subcategories |
| | immobilization | Metaphorical extension |
| | bodily harm | Other |
| emergence | change | Causal overlap |
| | breakthrough | Contextual implication |
| Number of incorrect responses | | 321 |

As shown in the table, this group contains three *break* senses—emotional/psychological/social breakdown, violation, and emergence—which had been above-average performing classes in the RoBERTa-large model results. Among these, emotional/psychological/social breakdown showed relatively high confusion with destruction. Of the 54 incorrect responses for this sense, 43 involved selecting destruction as the most appropriate label. This reflects their strong distributional and metaphorical connection, consistent with the pattern observed in the confusion matrix in Figure 2.

For violation, the experimental materials included sentences such as (10), where violation and termination are causally linked. In this case, 35 participants selected the confusable label termination as the correct meaning:

(10) Context: The company was bound by a confidentiality agreement regarding the new technology.
Target: They broke the contract by revealing the design to a competitor.

The emergence sense, which performed strongly in the model results, showed higher error rates in this experiment. This appears to result from the design of test items in which emergence overlapped with change. Examples (11) and (12) illustrate this overlap. In (11), *break* conveys not only the appearance of morning but also the transition from darkness to light, embedding change within the event of *emergence*. In (12), the sun "breaking through the clouds" simultaneously describes the event of emergence and the shift from dark to bright conditions:

(11) Context: The sky remained black all night.
Target: Just before 6 a.m., the day began to break.
(12) Context: The clouds slowly began to part after hours of heavy rain.
Target: Finally, the sun broke through the clouds.

Among the *freeze* senses, both physical/bodily freezing and economical freezing fall into the below-average accuracy group. Both display frequent confusion with immobilization, reflecting their shared grounding in physical stasis. In addition, economical freezing—as in the model predictions—was often confused with suspension, with which it stands in a near-synonym relation.

In sum, the below-average accuracy classes reveal a higher susceptibility to confusion through stronger semantic proximity and overlap, explaining their higher rate of incorrect responses compared to the groups discussed in Tables 4 and 5.

We now examine the error patterns observed in the lowest-accuracy classes, presented in Table 7. This group contains a total of 218 incorrect responses, including 82 cases of semantic overlap. Consistent with the RoBERTa-large results, this set includes the two senses that showed the strongest tendencies toward semantic overlap: change and immobilization. As noted in Chapter 3, the *change* sense of *break* was the most frequently confused with multiple other senses in the model misclassification analysis. In the human meaning selection task, participants frequently chose the confusable labels interruption (49 cases) and termination (38 cases) instead of the correct label. Example (13) illustrates how change and interruption overlap causally:

(13) Context: He had been stuck in a rut for a long time.
Target: He broke his unproductive behavior pattern.

Turning to immobilization, this sense most clearly embodies the core concept of stasis that underlies the semantics of *freeze*. Participants often confused it with other physical senses—particularly physical/bodily freezing, which also serves as a frequent source of metaphorical extension. In addition, strong confusion was observed with emotional/mental freezing, with which immobilization is closely overlapped. These patterns, also evident in the confusion matrices in Figures 2–3, highlight that even in human sense selection, shared conceptual cores and metaphorical extensions play a crucial role in driving errors and overlaps.

Table 7. Lowest-accuracy classes

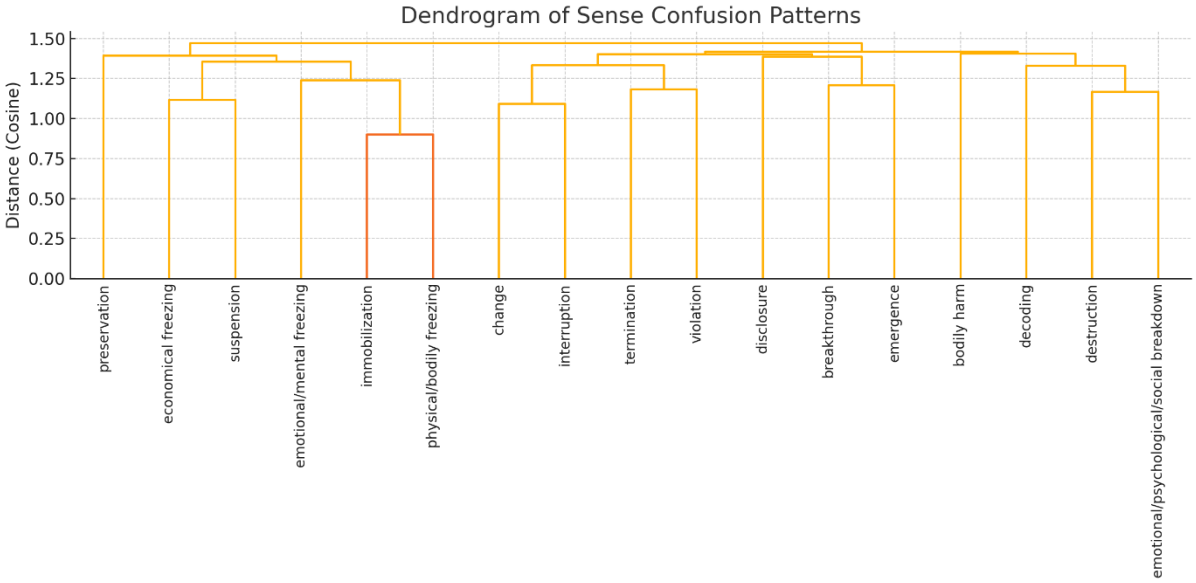
| A:<br>Correct label | B:<br>Incorrect label (selected label) | Semantic relationship<br>between A and B |
|---|---|---|
| change | interruption | Semantic subcategories and causal overlap |
| | termination | Semantic subcategories and contextual implication |
| | EPS breakdown | Other |
| immobilization | emotional/mental freezing | Extension-based overlap |
| | physical/bodily freezing, suspension, preservation | Metaphorical extension |
| | destruction | Other |
| Number of incorrect responses | | 218 |

Overall, the lowest-accuracy classes summarized in Table 7 are characterized by the densest networks of semantic connectivity, where core concepts such as change and stasis readily extend or overlap with multiple other senses, leading to the highest error rates in human responses.

In summary, our analysis of correct and incorrect responses in the human meaning selection task reveals a clear gradient in error patterns. For the highest-accuracy senses, correct and confusable labels are primarily related through implicature. By contrast, as accuracy decreases, errors increasingly involve senses that are connected through stronger semantic proximity and overlap. These findings support our hypothesis that the degree of connectivity to other senses—via semantic proximity and overlap—constitutes a key factor shaping both the model's sense prediction performance and human accuracy in meaning selection.

The hypothesis that senses with higher error rates and stronger confusion tendencies are also characterized by greater semantic connectivity can be tested through hierarchical cluster analysis. This analysis is carried out by computing the cosine distances between the response distributions of meaning classes, as represented in the confusion matrix in Figure 2, and then applying hierarchical clustering to these similarity measures. Figure 5 presents the dendrogram derived from the confusion matrix, showing the hierarchical clustering of meaning classes based on cosine distances between participants' response distributions. The vertical axis represents the cosine distance, which indicates the degree of (dis)similarity between sense distributions: the lower the vertical value at which two classes merge, the more similar their confusion patterns are.

Figure 5. Dendrogram of sense confusion patterns in human meaning selection

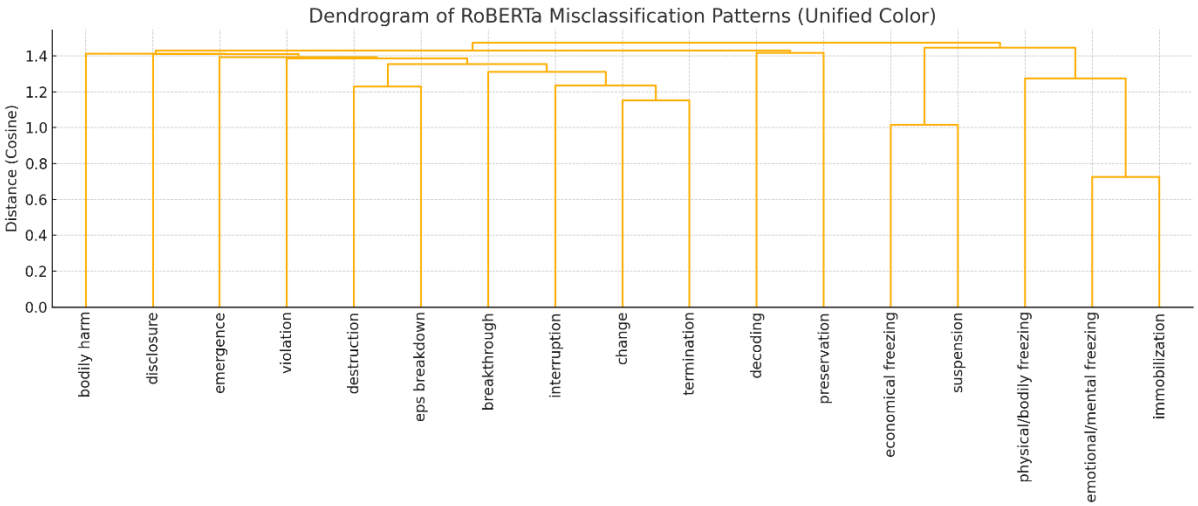



The key interpretation of this dendrogram lies in the relationship between the clustering positions/vertical values and the sense-level accuracies analyzed above. The two pairs that cluster at the lowest vertical values—immobilization–physical/bodily freezing and change–interruption—correspond precisely to the lowest-accuracy senses (change and immobilization), which are grouped with the senses they were most frequently confused with (see Table 7 and Figure 2). Similarly, the four pairs that merge below a distance of 1.25 involve below-average accuracy senses (EPS breakdown, violation, economic freezing, emergence), each of which was identified as among the most frequently misclassified categories (see Table 6 and Figure 2).

By contrast, meaning classes that cluster higher on the dendrogram (e.g., bodily harm, disclosure, decoding) correspond to highest- or above-average accuracy senses, which showed relatively distinct sense boundaries and fewer overlaps. This confirms that the hierarchical clustering of confusion patterns aligns closely with the observed accuracy rankings, reinforcing the conclusion that semantic proximity and overlap drive the distribution of errors in human sense selection.

Figure 6 presents the dendrogram derived from the RoBERTa-large confusion matrix, showing the hierarchical clustering of sense misclassification patterns. As in the human experiment, the vertical axis represents cosine distance, with lower values indicating stronger similarity in confusion tendencies between two meaning classes. The three pairs that cluster at the lowest vertical points—emotional/mental freezing–immobilization, economical freezing–suspension, and change–termination—correspond directly to the lowest-performing senses (emotional/mental freezing, change, and economical freezing), each grouped with the meaning it was most frequently confused with. This finding demonstrates that the weakest-performing senses are precisely those most semantically proximate to their confusable counterparts, confirming that semantic connectivity underlies the systematic error patterns observed in model predictions.

Figure 6. Dendrogram of sense confusion patterns in RoBERTa-large prediction




In sum, the human meaning selection experiment discussed in this section demonstrates that accuracy is systematically shaped by the degree of semantic connectivity among senses. Clear, distinct senses show high accuracy, while senses marked by overlap or metaphorical extension yield the greatest confusion. These results parallel the patterns observed in model predictions, confirming that fuzzy boundaries and dense interconnections are the key drivers of error. Section 4.2 extends this inquiry through sense applicability judgments, probing whether graded boundary effects also emerge in explicit evaluations of sense applicability.

## 4.2. Sense Applicability Judgment Experiment

### 4.2.1. Methods

This section presents the sense applicability judgment experiment. In this task, human participants and two generative AI models—GPT and Gemini—rate the applicability of target senses of *break* and *freeze* on a graded scale. The goal is to examine how applicability scores correlate with the semantic relations among the senses under evaluation.

**Materials.** This experiment used materials from the human meaning selection experiment. For each of the 17 meaning classes of *break* (11 senses) and *freeze* (6 senses), two sentences were selected, yielding a total of 34 items. As in the meaning selection experiment, each item was presented as a pair of a short context sentence and a target sentence, accompanied by three candidate sense labels: (i) the Correct label, (ii) a Similar label, and (iii) a Dissimilar label. The correct label corresponds to the intended meaning of the target word (*break* or *freeze*) in the given sentence. The similar and dissimilar labels were equivalent to the confusable and non-confusable labels used in the meaning selection task.

Across the 34 items, the semantic relation between the correct and similar labels was systematically drawn from six types of relationships that had been identified in the model predictions (Chapter 3) and the human meaning selection experiment (Section 4.1) as frequent sources of confusion: contextual implication, context-dependent polysemy, semantic subcategories, metaphorical extension, extension-based overlap, and causal overlap. Table 8 illustrates two sample sentences where the correct label is the economical freezing sense of

*freeze*. In this case, the correct label and the similar label (suspension) are related by the semantic subcategory relation, since both belong to the broader domain of cessation.

Table 8. Example of label construction (economical freezing sense)

| Context | Target | A: Correct label | B: Similar label | C: Dissimilar label | Semantic relationship between A and B |
|---|---|---|---|---|---|
| The government suspected the funds were linked to illegal activities. | Authorities moved quickly to freeze the company's assets. | economical freezing | suspension | bodily harm | Semantic subcategories |
| After repeated failed login attempts, the system flagged the account. | The bank froze the customer's account as a security measure. | economical freezing | suspension | bodily harm | Semantic subcategories |

In total, 34 sentence items were constructed in this way and then randomized before being presented to participants. The full experimental dataset is available at the GitHub repository: https://github.com/hanjung-25/clmsemantics/tree/data.

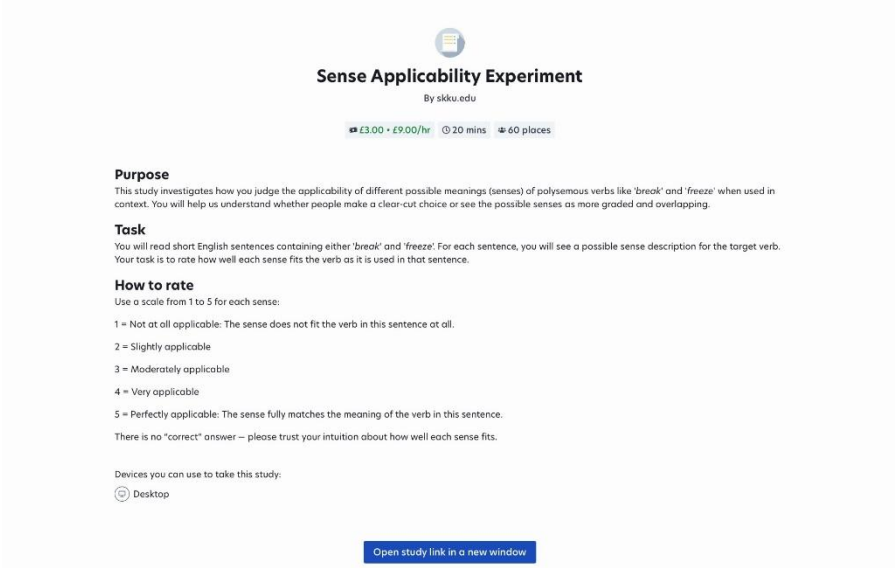
**Participants, task, and procedure.** We collected human judgments online through Prolific. Sixty native English speakers (31 male, 29 female; age range 20–66) participated in the study. None of the participants had taken part in the meaning selection experiment reported in Section 4.1.

Participants were instructed to evaluate the applicability of three candidate senses for each target word (*break* or *freeze*) as it appeared in a short context sentence and a target sentence. For each item, they rated the degree to which each of the three sense labels—the Correct label, Similar label, and Dissimilar label—fit the contextualized use of the verb. Ratings were provided on a 5-point Likert scale:

1 = not at all applicable
2 = slightly applicable
3 = moderately applicable
4 = very applicable
5 = perfectly applicable

This graded scale allowed participants to indicate not only whether a sense applied, but also the extent to which it applied, thereby overcoming the limitations of the forced-choice paradigm used in the previous experiment. A screenshot of the Prolific interface for this task is displayed in Figure 7. The full set of participant instructions is provided in Appendix E.

Figure 7. Screenshot of the Prolific interface for the sense applicability judgment task




As in the meaning selection experiment, participants were provided with brief descriptions and examples of all 17 meaning classes of *break* and *freeze* before beginning the main experiment. They then completed a short comprehension check to ensure familiarity with the sense inventory and task format. All 60 participants successfully passed the comprehension check and completed the task, so their data were included in the final analysis.

The total duration of the experiment was less than 40 minutes. Each participant completed all 34 items and received a payment of £4, resulting in a complete dataset of 60 participants (average expected compensation rate: £6 per hour).

**Model procedure.** For the generative AI models, GPT and Gemini were accessed via their respective APIs. The same 34 items used in the human applicability judgment experiment were submitted as prompts, with the three candidate sense labels (Correct, Similar, Dissimilar) provided for each item. The models were instructed to assign an applicability score on the same 5-point Likert scale (1 = not at all applicable, 5 = perfectly applicable) used in the human experiment.

To account for stochastic variation in generative model outputs, each item was run 10 times per model, and the mean of these ratings was used in the analysis. This procedure ensured comparability across human and model judgments, while also stabilizing the variability inherent in large generative models. All implementation details, including the Colab notebook used to query the models via API and aggregate the results, are available in the accompanying GitHub repository (https://github.com/hanjung-25/clmsemantics/tree/code). The code provides the full experimental pipeline, from API calls to data collection and preprocessing, ensuring replicability of the model-based applicability judgments.

**Hypotheses and predictions.** This experiment specifically addresses the third research question posed at the beginning of this chapter: Are human and AI judgments of sense boundaries categorical or graded, and how do they reflect semantic proximity and overlap among senses? To answer this question, we formulated the following hypotheses:

(14) Hypotheses

a. Hypothesis 1. Judgments involving the Correct label and the Dissimilar label are expected to be largely categorical, whereas judgments for the Similar label will show more graded tendencies**.**

b. Hypothesis 2. The applicability scores of the Correct label and the Similar label, when they are closely connected semantically through semantic proximity and overlap, are expected to show smaller differences or even overlap. By contrast, when the labels are relatively less tied semantically, their scores should display larger differences with little or no overlap.

The predictions of these hypotheses will be tested in Section 4.2.2.

### 4.2.2. Results

A total of 6,120 human responses and 2,040 AI model responses were collected. We begin by examining the distribution of sense applicability scores in order to test the hypotheses. We then apply regression models to evaluate the significance and relative importance of semantic relations in shaping applicability judgments.

**Applicability score distribution across sense types.** Table 9 presents summary statistics of applicability scores across the three sense types (Correct, Similar, Dissimilar). Correct labels received the highest ratings (Mean [M] = 4.62, Median = 5), with a strong concentration at the upper end of the scale (Q1 = Q3 = 5). By contrast, dissimilar labels were rated lowest (M = 1.32, Median = 1), showing a strong floor effect, with the vast majority of judgments clustered at the minimum value. Similar labels occupy an intermediate position (M = 3.14, Median = 3), but with a much wider spread (Standard deviation [SD] = 1.53, IQR = 3). Here, Q1 (first quartile) and Q3 (third quartile) indicate the 25th and 75th percentile values of the distribution, respectively, and the IQR (interquartile range = Q3 – Q1) reflects the middle 50% spread of scores. The fact that the IQR for Similar labels is substantially larger than for Correct or Dissimilar labels highlights the graded and variable nature of judgments for semantically related senses. These results align with Hypothesis 1, confirming that judgments of Correct and Dissimilar labels are predominantly categorical, whereas judgments of Similar labels are graded and more variable, reflecting their semantic proximity to the correct meaning.

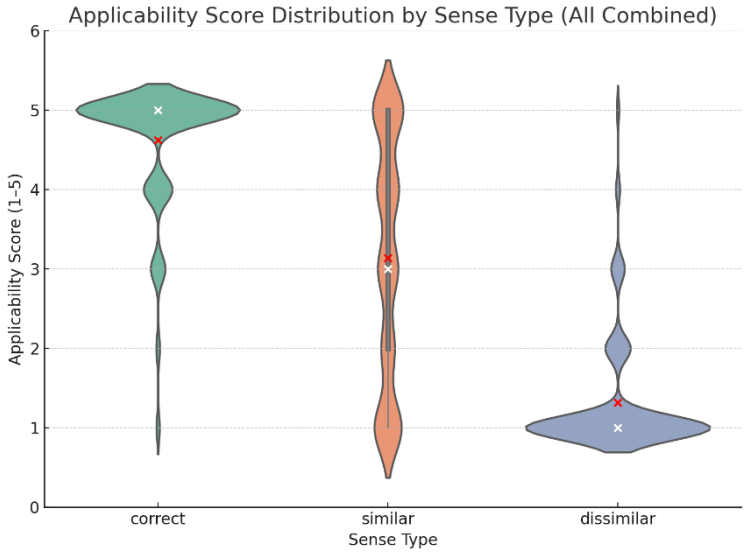
Table 9. Summary statistics of applicability scores across sense types

| Sense type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Correct | 2720 | 4.62 | 5 | 0.81 | 5 | 5 | 1 | 5 | 0 |
| Similar | 2720 | 3.14 | 3 | 1.53 | 2 | 5 | 1 | 5 | 3 |
| Dissimilar | 2720 | 1.32 | 1 | 0.75 | 1 | 1 | 1 | 5 | 0 |

Figure 8 visualizes the distributions of applicability scores across the three sense types using violin plots. The horizontal axis represents the three sense types (Correct, Similar, Dissimilar), while the vertical axis shows the applicability scores on the 1–5 scale. Each violin plot displays the full distribution of scores, with the white dot indicating the mean (M), the red dot indicating the median, the thick dark gray vertical bar showing the interquartile range (IQR,

Q3 – Q1), and the thin gray vertical line (whisker) representing the range within 1.5 × IQR. When no whisker is visible, this reflects the fact that the data points are strongly concentrated on a single value (e.g., a floor or ceiling effect), leaving no variability beyond the IQR.

Figure 8. Applicability score distribution by sense type (all evaluators)




As expected, Correct labels cluster tightly at the top of the scale (Median = 5), forming a ceiling effect with little dispersion. Dissimilar labels, by contrast, are heavily skewed toward the minimum value (Median = 1), showing a clear floor effect and minimal spread. Similar labels occupy an intermediate position (Median = 3) but with the widest distribution, reflected in the largest IQR and extended shape of the violin. This broad spread indicates graded and variable judgments of applicability, consistent with their intermediate semantic status between correct and dissimilar labels. This figure provides strong visual confirmation of the patterns summarized in Table 9: Correct and Dissimilar labels are judged in a predominantly categorical fashion, whereas Similar labels elicit gradient and more variable responses, reflecting their semantic proximity to the intended sense.

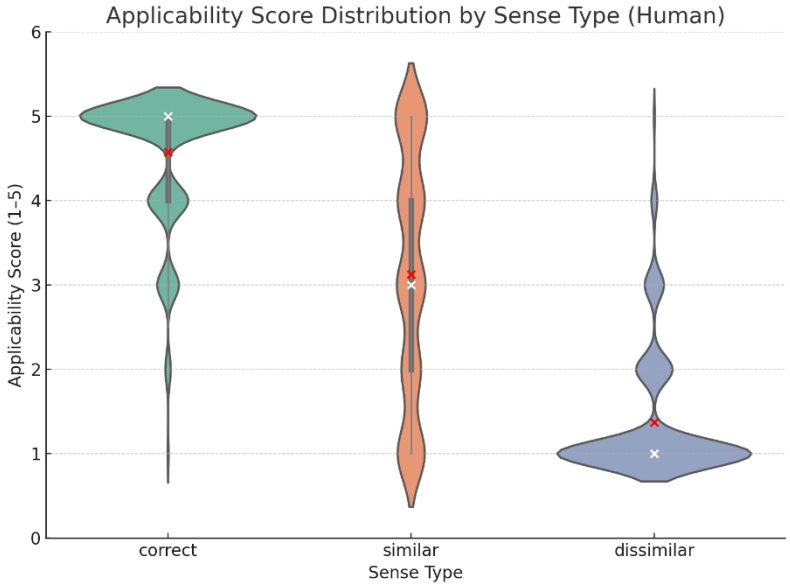
Now we turn to the distribution of applicability scores separately for humans and the two AI models. Table 10 presents the descriptive statistics of human judgments across the three sense types. Correct labels received the highest scores (M = 4.72, SD = 0.78), confirming that participants strongly aligned them with the intended meaning. Similar labels showed a much wider spread (M = 3.13, SD = 1.45), indicating graded judgments and frequent overlap with the Correct label. Dissimilar labels, by contrast, were consistently rated very low (M = 1.37, SD = 0.75), with little or no overlap with the other two categories.

Table 10. Summary statistics of applicability scores across sense types (human)

| Sense type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Correct | 2040 | 4.72 | 5 | 0.78 | 4 | 5 | 1 | 5 | 1 |
| Similar | 2040 | 3.13 | 3 | 1.45 | 2 | 4 | 1 | 5 | 2 |
| Dissimilar | 2040 | 1.37 | 1 | 0.75 | 1 | 1 | 1 | 5 | 0 |

Figure 9 visualizes these results. As seen in the violin plots, Correct senses are tightly concentrated at the upper end of the scale, Similar senses span a broad range from low to high values, and Dissimilar senses cluster narrowly near the bottom. This pattern supports Hypothesis 1: judgments of Correct and Dissimilar labels are categorical and sharply distinguished, whereas judgments of Similar labels are gradient, often overlapping with Correct labels depending on their degree of semantic relatedness.

Figure 9. Applicability score distribution by sense type (human)




Having established the distributional patterns of human judgments, we now turn to the two generative AI models, GPT and Gemini, to examine how their applicability ratings compare. Table 11 summarizes the descriptive statistics for the three sense types (Correct, Similar, and Dissimilar), and Figures 10 and 11 visualize the corresponding score distributions for GPT and Gemini, respectively. As shown in Table 11, both models display a strong categorical separation between the Correct and Dissimilar labels, with the highest mean applicability scores assigned to the Correct label and the lowest to the Dissimilar label. This mirrors the general pattern observed in human judgments and supports the first hypothesis regarding categorical separation.

Table 11. Summary statistics of applicability scores across sense types (AI models)

a. GPT

| Sense type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Correct | 340 | 4.79 | 5 | 0.75 | 5 | 5 | 1 | 5 | 0 |
| Similar | 340 | 3.02 | 3 | 1.44 | 1 | 5 | 1 | 5 | 4 |
| Dissimilar | 340 | 1.14 | 1 | 0.71 | 1 | 1 | 1 | 5 | 0 |

b. Gemini

| Sense type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Correct | 340 | 4.74 | 5 | 0.79 | 5 | 5 | 1 | 5 | 0 |
| Similar | 340 | 3.31 | 4 | 1.46 | 1 | 5 | 1 | 5 | 4 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dissimilar | 340 | 1.17 | 1 | 0.69 | 1 | 1 | 1 | 5 | 0 |

Figure 10 presents GPT's applicability score distributions. The violin plot indicates that GPT consistently assigns high ratings to the Correct label, while ratings for the Dissimilar label cluster near the bottom of the scale. The distribution for the Similar label, however, is more spread out, with considerable overlap with both Correct and Dissimilar scores. This spread suggests that GPT is sensitive to semantic connectivity, producing more gradient judgments when meanings are related, in line with Hypothesis 2.

Figure 10. Applicability score distribution by sense type (GPT)

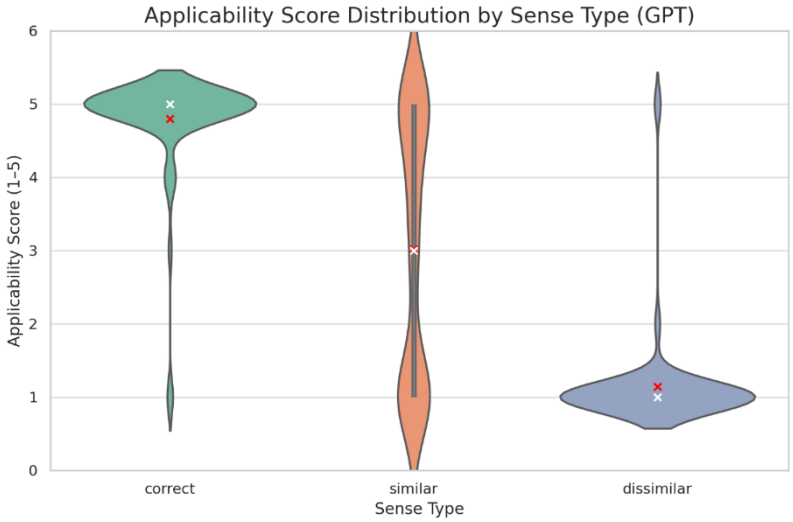
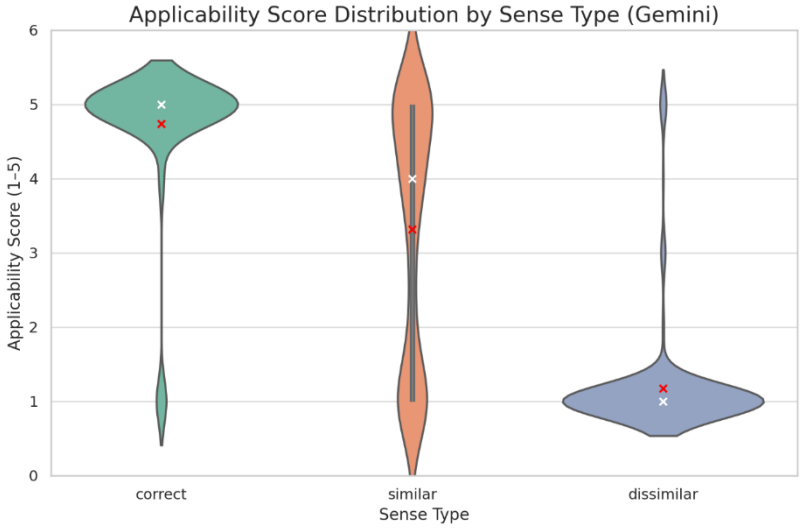



Figure 11 shows Gemini's distributions, which follow the same overall pattern as GPT's but with a somewhat tighter clustering of scores. Gemini assigns the Correct label scores that are nearly categorical at the upper bound, while the Similar label displays a gradient distribution similar to GPT's, though with slightly higher means. This pattern suggests that Gemini is even more conservative in separating Correct from Dissimilar labels, while still capturing graded connectivity in the Similar condition.

Figure 11. Applicability score distribution by sense type (Gemini)



Taken together, these results shown in Table 11 and Figures 10-11 indicate that the two generative AI models exhibit the predicted categorical–gradient pattern across sense types, with minor differences in how sharply they separate the categories. The convergence of human and model results on these patterns highlights the role of semantic connectivity, proximity and overlap in shaping sense applicability judgments.

To further examine whether the observed distributional patterns differed across evaluators, we conducted a mixed-design ANOVA with model type (Human, GPT, Gemini) as the between-subject factor and sense type (Correct, Similar, Dissimilar) as the within-subject factor. Table 12 presents the results of the mixed-design ANOVA on applicability scores. The main effect of model type was not significant, indicating no overall difference in average scores between humans, GPT, and Gemini. By contrast, the main effect of sense type was highly significant ($p < .001$, $\eta^2 = .938$), showing very large differences between correct, similar, and dissimilar labels. No significant interaction was found between model type and sense type.

Table 12. Results of the mixed-design ANOVA on applicability scores

| Effect | df | *F* | *p* | $\eta^2$ (partial) | Results |
|---|---|---|---|---|---|
| Model type (Human vs. GPT vs. Gemini) | (2, 59) | 0.043 | .958 | – | n.s. (not significant) |
| Sense type (Correct vs. Similar vs. Dissimilar) | (2, 118) | 893.56 | <.001 | .938 | *** (highly significant) |
| Model × Sense type | (4, 118) | 0.274 | .895 | – | n.s. |

**Applicability score distribution across relation types.** Now we turn to the relation-specific distributions of applicability scores. Table X in Appendix X presents descriptive statistics of applicability scores broken down by the six relation types identified as sources of sense confusion. Focusing on the data aggregated across all evaluators, the results show that the mean differences between the Correct and Similar labels vary systematically depending on the relation type. For relations characterized by stronger semantic connectivity—such as extension-based overlap and causal overlap—the applicability scores of the Correct and Similar labels are closer to each other, with relatively small mean differences. By contrast, for relations involving looser semantic connections—such as context-dependent polysemy and contextual implication—the Correct label maintains a higher mean score than the Similar label, with a wider gap between the two. The relation type of metaphorical extension falls in between these two extremes, showing moderate differences. These aggregated results are consistent with Hypothesis 2, confirming that the degree of semantic relatedness systematically predicts the extent of differentiation between the applicability scores of Correct and Similar labels.

To further test Hypothesis 2, we quantified the degree of overlap between the applicability score distributions of the Correct and Similar labels across the six relation types. Three complementary indices employed in our analysis are given in (15).

(15) Measures of the degree of overlap between applicability score distributions

a. **Mean difference (|Mean_correct – Mean_similar|):**
Smaller values indicate that the two distributions are closer on average, suggesting that participants evaluated the Correct and Similar senses as more alike.

b. **Median difference (|Median_correct – Median_similar|):**
A small difference between the medians implies that the central tendencies of the

two distributions are aligned, further reflecting the perceived similarity of the two senses.

c. **IQR overlap ratio:**
This index measures how much the interquartile ranges (IQRs) of the two distributions overlap. A value close to 0 indicates virtually no overlap, meaning that the two senses were judged as clearly distinct. A value close to 1 indicates almost complete overlap, suggesting that participants perceived the senses as highly continuous and difficult to separate.

These three measures provide a systematic basis for evaluating the degree of distributional overlap between Correct and Similar labels. Larger overlaps across these indices point to fuzzier or more graded sense boundaries, while smaller overlaps reflect clearer categorical distinctions.

Table 13 summarizes the overlap analysis of applicability scores for human data only. This table compares the distributions of applicability scores for the Correct and Similar labels across six semantic relation types. As noted above, mean difference and median difference show how close the two distributions are in central tendency (smaller values → closer distributions); IQR overlap ratio indicates how much the interquartile ranges overlap (0 = no overlap, 1 = complete overlap).

Table 13. Overlap analysis of applicability scores (human data)

| Relation type | Mean difference | Median difference | IQR overlap ratio |
|---|---|---|---|
| Causal overlap | 0.75 | 1.0 | 0.9 |
| Context-dependent polysemy | 2.64 | 3.5 | 0.1 |
| Contextual implication | 2.54 | 3.0 | 0.1 |
| Extension-based overlap | 1.07 | 1.0 | 0 |
| Metaphorical extension | 1.89 | 3.0 | 0.3 |
| Semantic subcategories | 0.74 | 1.0 | 0.9 |

The human data summarized in Table 13 reveal two distinct patterns. For the three overlapping relations—causal overlap, extension-based overlap, and semantic subcategories—the small mean/median differences and high IQR overlap ratios indicate substantial score overlap, suggesting that participants perceived these senses as gradiently connected**.** By contrast, the three non-overlapping relations—context-dependent polysemy, contextual implication, and metaphorical extension—showed large score differences and negligible IQR overlap, pointing to clear categorical distinctions (see Appendix F, Table 2 for label pairs)**.**

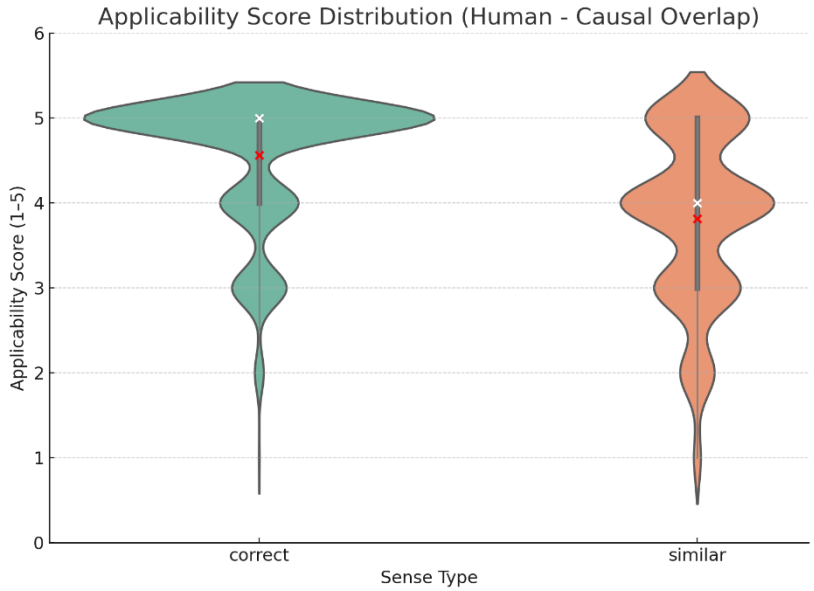
- **Overlapping relations**

Figures 12–14 present the applicability score distributions for the three relation types in which the Correct and Similar labels showed substantial overlap: causal overlap, extension-based overlap, and semantic subcategories**.** These visualizations correspond to the statistical indices reported in Table 13.

For causal overlap (Figure 12), the mean and median differences (0.75 and 1.0) between the Correct and Similar labels are very small, indicating that participants often judged the two labels to be nearly equally applicable. Consistent with this, the IQR overlap ratio approaches

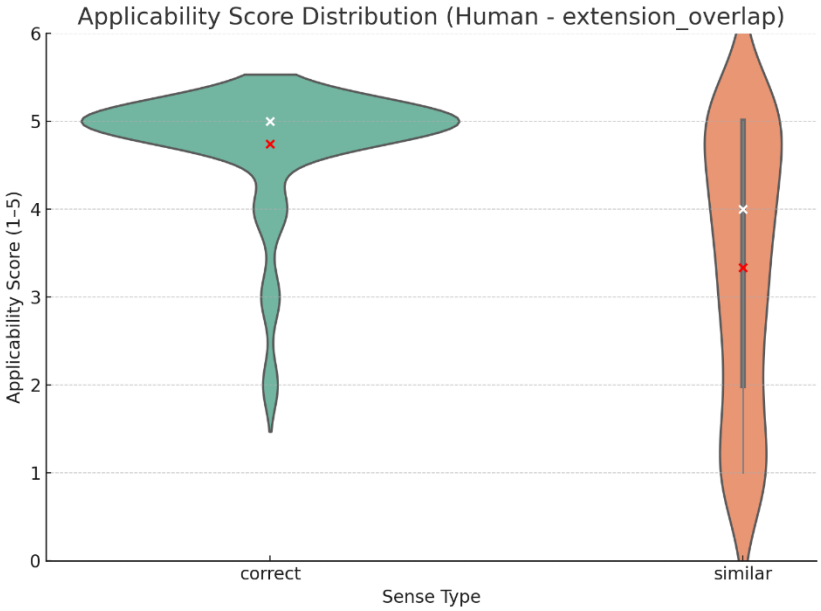
0.9, pointing to almost complete overlap between the two distributions. This suggests that senses in a causal relation are perceived as tightly connected and gradient in boundary.

Figure 12. Applicability score distribution by sense type (Human – Causal overlap)




For extension-based overlap (Figure 13), the mean and median differences (1.07 and 1.0) are again modest, but the IQR does not overlap. This suggests that participants perceived the senses as partially aligned in central tendency but still distinct in boundary definition, owing to divergent distributional spread.

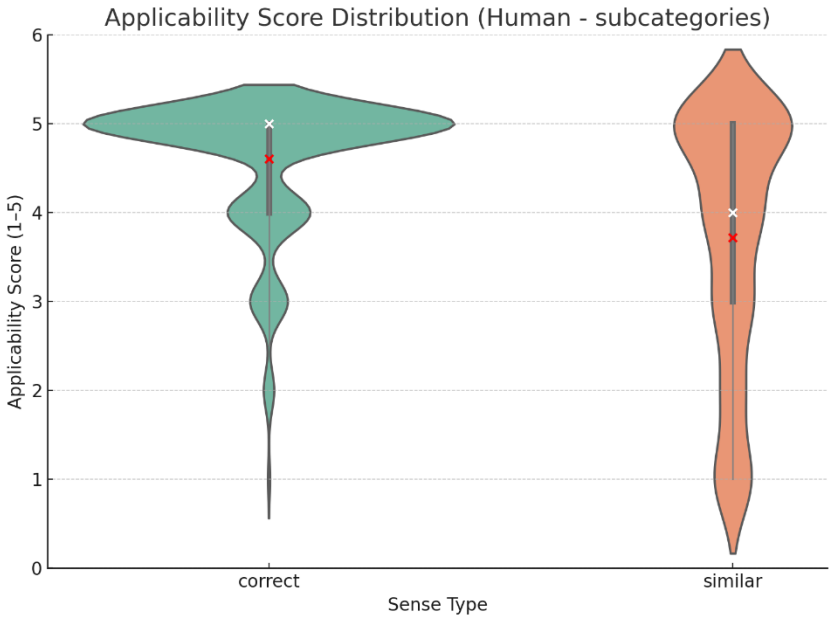
Figure 13. Applicability score distribution by sense type (Human – Extension-based overlap)




For semantic subcategories (Figure 14), the results are similar: the Correct and Similar labels show small mean and median differences (0.74 and 1.0), and their IQRs overlap

substantially (IQR overlap ratio = 0.9). This reflects the fact that when one sense represents a subcategory of another, the two are naturally judged as closely related in applicability.

Figure 14. Applicability score distribution by sense type (Human – Semantic subcategories)




Taken together, the three relation types—causal overlap, extension-based overlap, and semantic subcategories—demonstrate how strong semantic connectivity leads to overlapping distributions of applicability scores, thereby supporting Hypothesis 2 that semantic proximity and overlap yield gradient rather than categorical sense boundaries.

- **Non-overlapping relations**

Figures 15–17 present the applicability score distributions for the three non-overlapping relations: context-dependent polysemy, contextual implication, and extension**.** Again, these visualizations correspond to the statistical indices reported in Table 13.

For context-dependent polysemy (Figure 15), the correct label shows means and medians close to 5 (M = 4.72, Median = 5.0), indicating consistently high applicability scores, while the similar label remains much lower (M = 2.08, Median = 1.5). The mean difference (2.64) and median difference (3.5) are the largest reported in Table 13, and the IQR overlap ratio is only 0.1. These results indicate that the two distributions are sharply separated, suggesting that context-dependent polysemy is perceived by human participants as distinct rather than overlapping senses.

For contextual implication again (Figure 16), the correct label scores are high (M = 4.58, Median = 5.0), while the similar label scores are substantially lower (M = 2.08, Median = 2.0). The mean difference (2.54) and median difference (3.0) are large, and the IQR overlap ratio is only 0.09, pointing to virtually no overlap between the distributions. Visually, the violin plots show a clear separation, confirming that participants regarded implication-based relations as categorically distinct meanings.

Figure 15. Applicability score distribution by sense type (Human – Context-dependent polysemy)

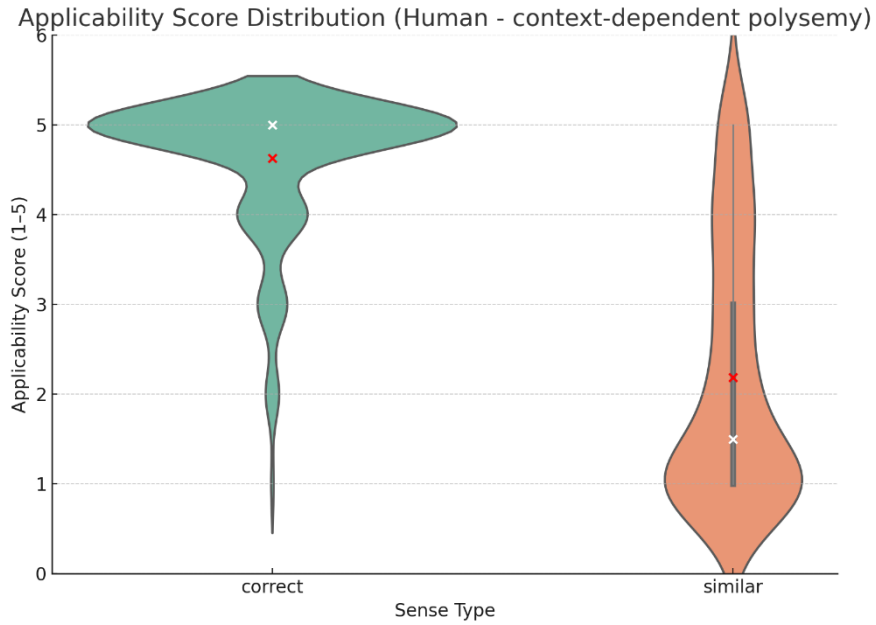
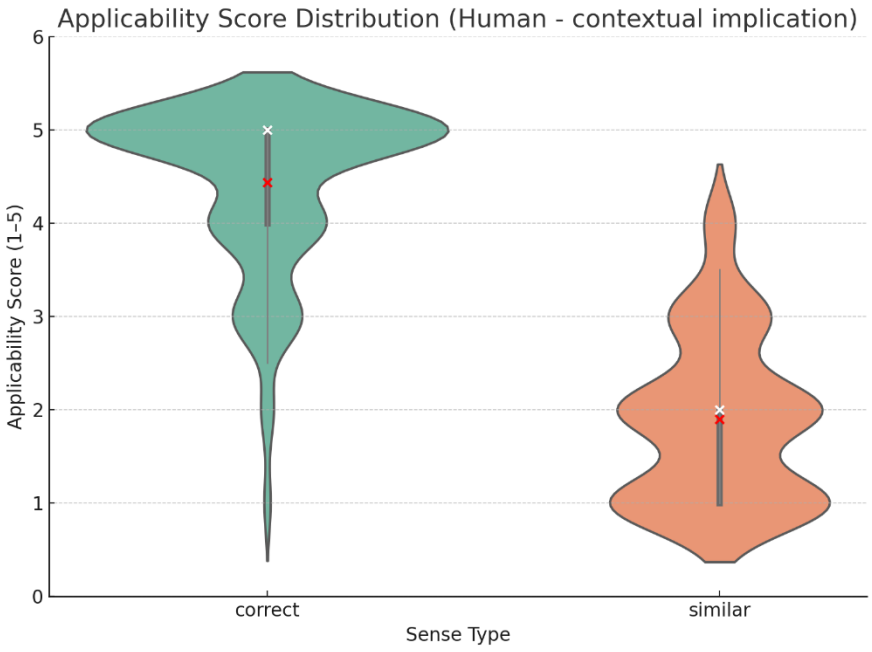



Figure 16. Applicability score distribution by sense type (Human – Contextual implication)




For metaphorical extension (Figure 17), The correct label scores are high (M = 4.54, Median = 5.0), in contrast to the similar label scores (M = 2.65, Median = 2.0). The mean difference (1.89) and median difference (3.0) are moderate—larger than those observed for the three overlapping relations, but smaller than for context-dependent polysemy and contextual implication. The IQRs do not overlap (overlap ratio = 0.30), indicating that the two distributions remain largely separated. The violin plot thus illustrates that metaphorical extension relations were evaluated as relatively distinct senses, with boundaries clearer than in overlapping relations but less sharply separated than in the most categorical cases.

Figure 17. Applicability score distribution by sense type (Human – Metaphorical extension)


Across the three non-overlapping relations—context-dependent polysemy, contextual implication, and metaphorical extension—the results consistently show larger mean and median differences between correct and similar labels, combined with very low IQR overlap ratios. These converging indicators point to minimal distributional overlap, suggesting that human participants perceived these sense relations as more distinct boundaries rather than overlapping meanings. Taken together, the distinct patterns observed in the overlapping and non-overlapping relations support Hypothesis 2, demonstrating that the degree of semantic proximity and overlap systematically predicts the extent of overlap and differentiation between the applicability scores of Correct and Similar labels.

Now we turn to the analysis of applicability score distributions for the two generative AI models and compare them with the human data patterns. Table 14 summarizes the overlap analysis of applicability scores for GPT and Gemini data.

Table 14. Overlap analysis of applicability scores

a. GPT data

| Relation type | Mean difference | Median difference | IQR overlap ratio |
|---|---|---|---|
| Causal overlap | 1.52 | 2.0 | 1.0 |
| Context-dependent polysemy | 3.0 | 4.0 | 0 |
| Contextual implication | 3.22 | 4.0 | 0 |
| Extension-based overlap | 0.1 | 0 | 1.0 |
| Metaphorical extension | 2.9 | 4.0 | 0 |
| Semantic subcategories | 0.56 | 0 | 1.0 |

b. Gemini data

| Relation type | Mean difference | Median difference | IQR overlap ratio |
|---|---|---|---|
| Causal overlap | 1.16 | 1.0 | 1.0 |
| Context-dependent polysemy | 3.43 | 4.0 | 0 |
| Contextual implication | 3.63 | 4.0 | 0 |
| Extension-based overlap | 1.97 | 1.0 | 1.0 |
| Metaphorical extension | 2.9 | 4.0 | 1.0 |
| Semantic subcategories | 0.06 | 0 | 1.0 |

For GPT, the mean and median differences across relation types closely mirror the human data patterns. The smallest differences appear for extension-based overlap and semantic subcategories, followed by causal overlap, while metaphorical extension occupies a middle ground between these relations and the more distinct context-dependent polysemy and contextual implication. Consistent with the human results, the largest mean and median differences are observed for context-dependent polysemy and contextual implication. Importantly, the IQR overlap ratios show overlap only for causal overlap, extension-based overlap, and semantic subcategories—the same set of overlapping relations identified in the human data.

For Gemini, the mean and median differences also align with the human and GPT data: causal overlap, extension-based overlap, and semantic subcategories display the smallest differences, metaphorical extension falls in the middle, and context-dependent polysemy and contextual implication show the largest gaps. However, unlike in the human and GPT data, Gemini assigns an IQR overlap ratio of 1.0 not only to the three overlapping relations but also to metaphorical extension. This indicates that Gemini evaluates metaphorical extension more gradiently, treating it as overlapping rather than clearly separated.

Taken together, these results support Hypothesis 2: the applicability scores of correct and similar labels converge when the senses are semantically proximate (as in causal overlap, extension-based overlap, and semantic subcategories), but diverge sharply when the senses are semantically more distant (as in context-dependent polysemy and contextual implication).

To further examine how the models' applicability judgments compare with human data, we now turn to the sense-relation–specific plots. These plots, presented for each of the six relation types—first the overlapping relations (causal overlap, extension-based overlap, and semantic subcategories), followed by the non-overlapping relations (context-dependent polysemy, contextual implication, and metaphorical extension)—allow us to compare the score distributions in greater detail and to identify both convergent patterns and systematic divergences between humans and models.

- **Overlapping relations**

Figures 18–20 present the applicability score distributions for the three relation types in which the Correct and Similar labels showed substantial overlap: causal overlap, extension-based overlap, and semantic subcategories**.** These visualizations correspond to the statistical indices reported in Table 14. Figure 18 compares the applicability score distributions for GPT and Gemini on the causal overlap relation. Both models reproduce the general human-like pattern: high scores for the correct label and moderately high scores for the similar label, leading to overlap in the middle range. A notable difference is that Gemini's correct-label distribution is

more tightly clustered around the upper bound (5), showing stronger concentration compared to GPT and to the more varied human responses (see Figure 12).

Figure 18. Applicability score distribution by sense type (Models – Causal overlap)

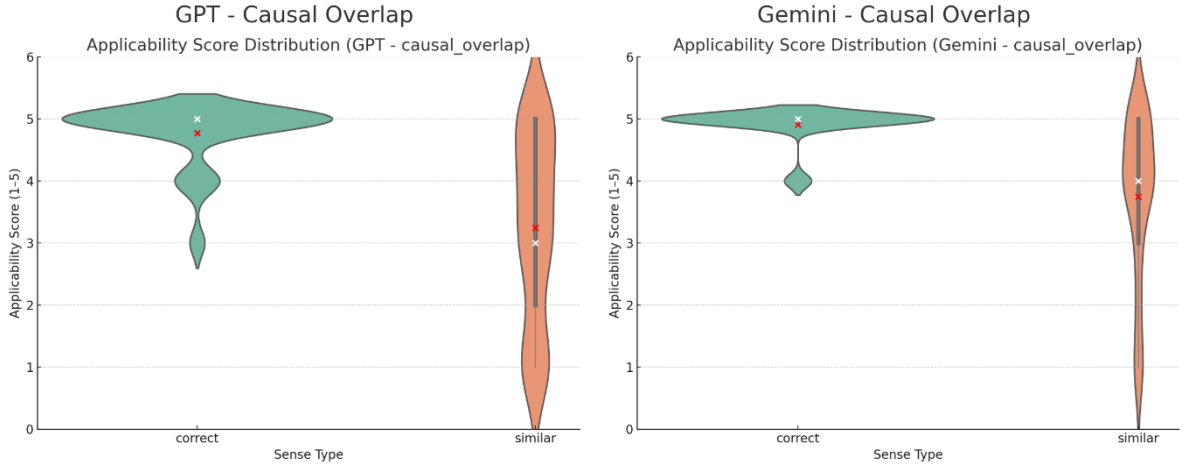



Figure 19 compares the applicability score distributions for GPT and Gemini on the extension-based overlap relation. Both GPT and Gemini show distributions highly similar to those observed in the human data (Figure 13) for the extension-based overlap relation. Correct label scores are concentrated near the upper end of the scale (close to 5), while similar label scores are also high, indicating gradient connectedness between the senses. However, compared to the human data, both models—especially Gemini—exhibit less dispersion, with scores more tightly clustered around the mean and median. This pattern suggests that while the models successfully capture the semantic relatedness between the senses, their judgments are less variable and more uniform than those of human participants.

Figure 19. Applicability score distribution by sense type (Models – Extension-based overlap)

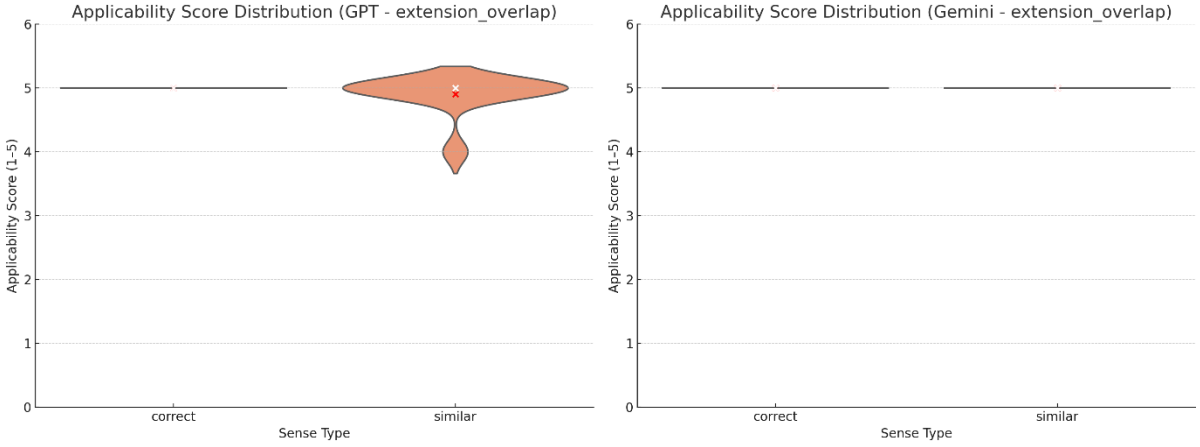
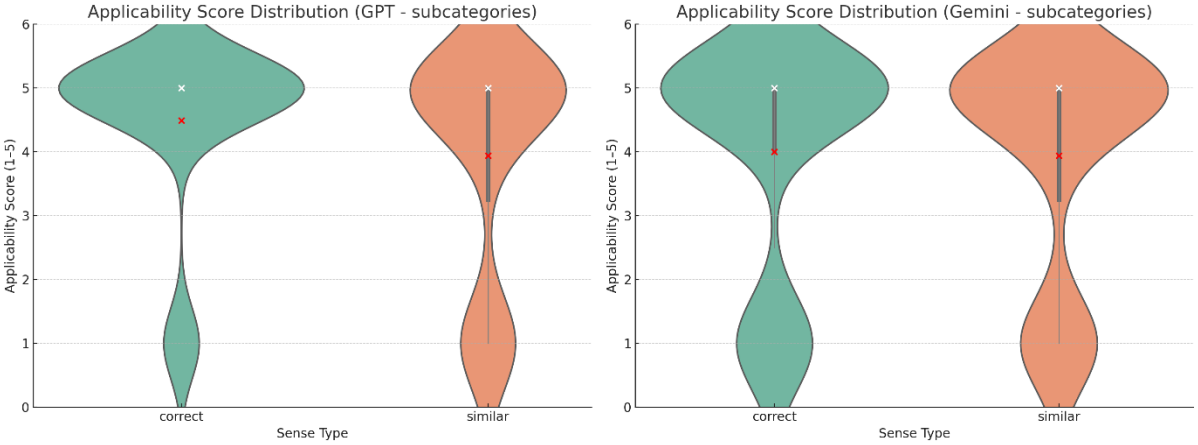



Figure 20 compares the applicability score distributions for the two models on the relation of semantic subcategories. Among the six relations, this relation exhibited the strongest similarity in both human and model judgments. As can be seen, the distributions of the correct and similar labels in this relation showed minimal differences in their central tendencies (means and medians) and substantial overlap in their interquartile ranges (IQRs), indicating that both

GPT and Gemini effectively captured the close semantic relatedness between subcategory senses.

Figure 20. Applicability score distribution by sense type (Models – Semantic subcategories)




- **Non-overlapping relations**

Figure 21 presents the applicability score distributions for GPT and Gemini in the context-dependent polysemy relation. Both models reproduce the human pattern of clear separation between the correct and similar labels. However, compared to human data, the models show a stronger concentration of correct label scores at the maximum value of 5, reflecting a more categorical evaluation tendency. A similar pattern is also observed for contextual implication (Figure 22).

Figure 21. Applicability score distribution by sense type (Models – Context-dependent polysemy)

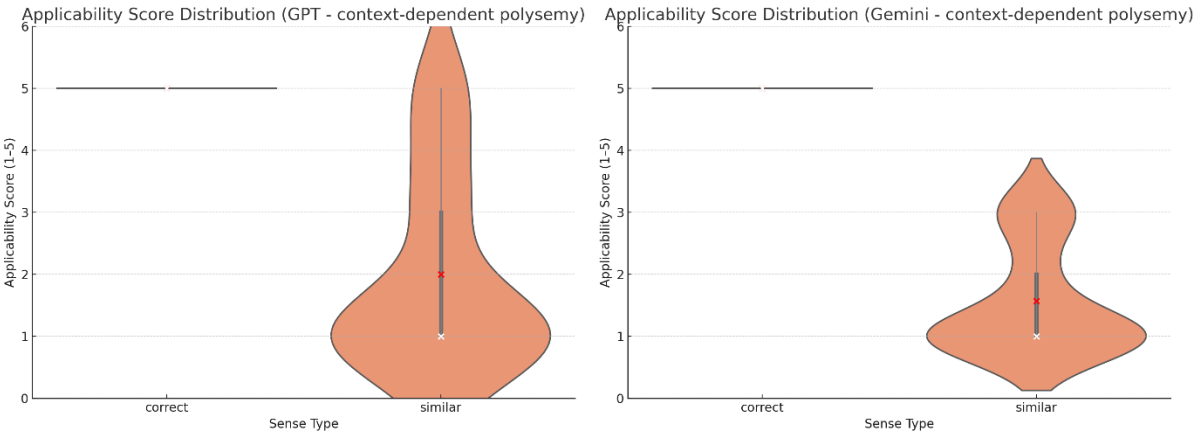
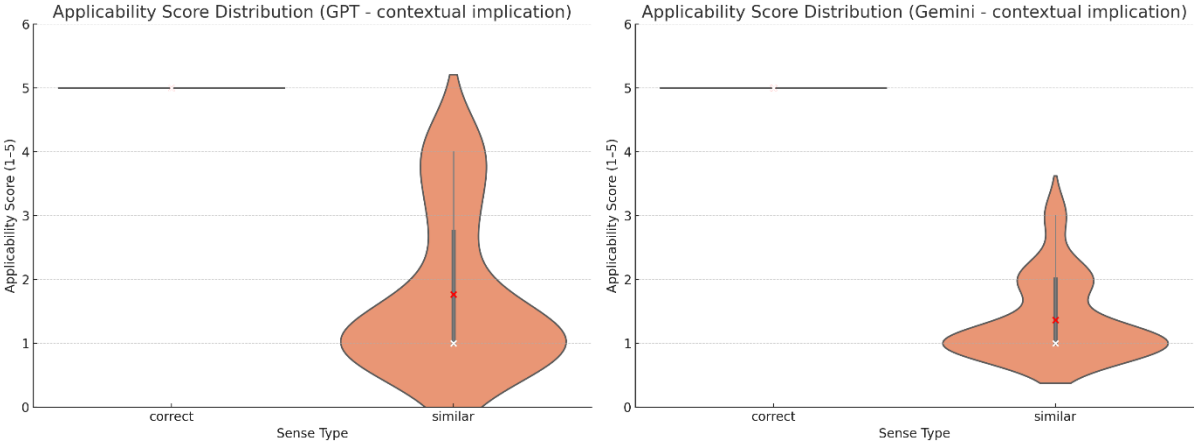
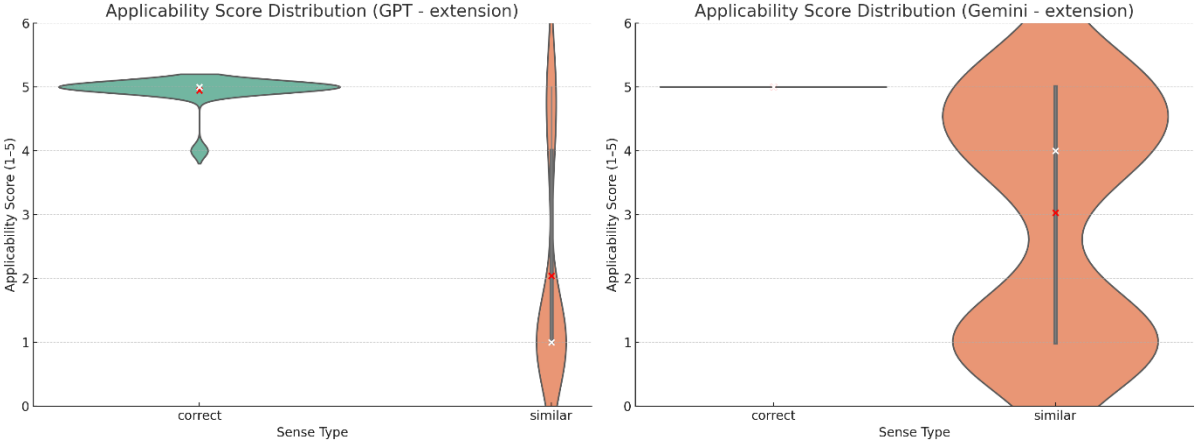


Figure 22. Applicability score distribution by sense type (Models – Contextual implication)




Finally, let us compare the applicability score distributions in the metaphorical extension relation shown in Figure 23. As noted above, this relation exhibited an intermediate pattern across humans and models. In the human data (Figure 17), Correct and Similar labels showed clear separation, with large mean and median differences and negligible IQR overlap, indicating a categorical boundary. GPT produced a distribution that was broadly similar to the human pattern: Correct labels were rated near the maximum, while Similar labels received much lower and more widely dispersed scores. By contrast, Gemini diverged from this pattern: although the correct label scores were still clustered at the top, the Similar label scores were shifted upward, with the median lying closer to that of the Correct label.

Figure 23. Applicability score distribution by sense type (Models – Metaphorical extension)




We have established through distributional and overlap analyses that semantic proximity and overlap play a central role in shaping applicability judgments across humans and generative AI models. Building on this finding, we now turn to regression modeling to examine these effects more systematically.

**Regression analysis of human sense applicability judgments.** To gain a more robust insight into the magnitude of relation type effects and their interaction with sense type, we initially planned to employ mixed-effects regression models with participant as a random effect. The primary analytic goal was to test how the difference in applicability scores between the Correct

and Similar labels varies as a function of semantic relation type. In this specification, sense type and relation type were treated as fixed factors, while participant was included as a random factor to capture between-participant variability.

However, the estimated variance component for the random effect was effectively zero, indicating negligible variability across participants. Consequently, we proceeded with ordinary least squares (OLS) regression, which under these conditions yields equivalent coefficient estimates while simplifying model interpretation. In the OLS models, the dependent variable was the applicability score (1–5 scale). The predictor variables are summarized in Table 15. Specifically, sense type was coded with Similar label as the reference level, and relation type with contextual implication as the reference level. Contextual implication was chosen as the baseline because the human data showed the smallest IQR overlap ratio for this relation, meaning that Correct and Similar label scores hardly overlapped, providing the clearest categorical contrast. An additional interaction term (sense type × relation type) was included as the key predictor of interest. Based on the overlap analysis, the prediction was that for the three overlapping relations—causal overlap, extension-based overlap, and semantic subcategories—the Correct–Similar score difference would shrink relative to the baseline, with Similar label scores becoming closer to those of the Correct label.

Table 15. Specification of the predictor variables

| Variables | Levels |
|---|---|
| Sense type | Similar label (reference level), Correct label |
| Relation type | Contextual implication (reference level), causal overlap, extension-based overlap, semantic subcategories, context-dependent polysemy, metaphorical extension |
| Interaction term | sense type × relation type |

We built the regression models using Python's `statsmodels` package in Google Colab. All analysis scripts and code are available in the accompanying GitHub repository.

Table 16 summarizes the results of the OLS regression model. The model explained approximately 38% of the variance in applicability scores ($R^2 = .385$, Adj. $R^2 = .383$), and the overall fit was highly significant, $F(11, 4068) = 231.7$, $p < .001$. The intercept represents the baseline condition in which the sense type is the Similar label and the relation type is contextual implication, yielding a mean score of 1.90. Relative to this baseline, Correct labels were judged substantially higher (coef. = +2.54, $p < .001$).

Main effects of relation type further showed that applicability scores for causal overlap (+1.91), semantic subcategories (+1.82), and extension-based overlap (extension_overlap) (+1.43) were especially elevated compared to contextual implication, with smaller but still significant increases for (metaphorical) extension (+0.75) and context-dependent polysemy (+0.28).

Most importantly, the interaction terms reveal that the correct–similar score gap significantly decreased under causal overlap (–1.79), semantic subcategories (–1.65), extension (–0.65), and extension overlap (–1.13), indicating that these relations strongly promote score convergence. By contrast, context-dependent polysemy did not differ significantly from the contextual implication baseline ($p = .553$). These results confirm that semantic proximity relations such as causal overlap, subcategories, and extension promote graded polysemy by

reducing the correct–similar distinction, whereas context-dependent polysemy and contextual implication maintain more categorical boundaries.

Table 16. Summary of OLS regression results

| **Predictor** | **Coef.** | **Std. Err.** | **t** | ***p*** | **Interpretation** |
|---|---|---|---|---|---|
| Intercept | 1.90 | 0.08 | 23.90 | <.001 | baseline score |
| Sense type | +2.54 | 0.11 | 22.59 | <.001 | Correct > Similar |
| Relation type | | | | | |
| – Context-dependent polysemy | +0.28 | 0.11 | 2.52 | .012 | Slightly higher scores |
| – Causal overlap | +1.91 | 0.09 | 21.32 | <.001 | Much higher scores |
| – Subcategories | +1.82 | 0.09 | 19.49 | <.001 | Much higher scores |
| – Extension | +0.75 | 0.10 | 7.87 | <.001 | Higher scores |
| – Extension_overlap | +1.43 | 0.13 | 11.40 | <.001 | Higher scores |
| Sense × Relation | | | | | |
| – Correct × Context-dependent polysemy | –0.09 | 0.16 | –0.59 | .553 | n.s. |
| – Correct × Causal overlap | –1.79 | 0.13 | –14.13 | <.001 | Correct-Similar gap shrinks |
| – Correct × Subcategories | –1.65 | 0.13 | –12.51 | <.001 | Correct-Similar gap shrinks |
| – Correct × Extension | –0.65 | 0.13 | –4.83 | <.001 | Correct-Similar gap shrinks |
| – Correct × Extension_overlap | –1.13 | 0.18 | –6.36 | <.001 | Correct-Similar gap shrinks |
| **Model fit:** $R^2$ = .385, Adj. $R^2$ = .383, $F(11, 4068) = 231.7$, $p < .001$ | | | | | |

Figure 24 visualizes the interaction between sense type and relation type in predicting applicability scores. The red line (Correct labels) remains consistently high across relation types, whereas the blue line (Similar labels) shows substantial variation depending on the relation. Crucially, the correct–similar gap narrows for causal overlap, semantic subcategories, and extension-based overlap, in line with our prediction that these overlapping relations yield more graded applicability judgments. By contrast, the gap remains large for contextual implication, supporting its role as a baseline category with clearly discrete boundaries.

Figure 24. Interaction plot of sense type and relation type

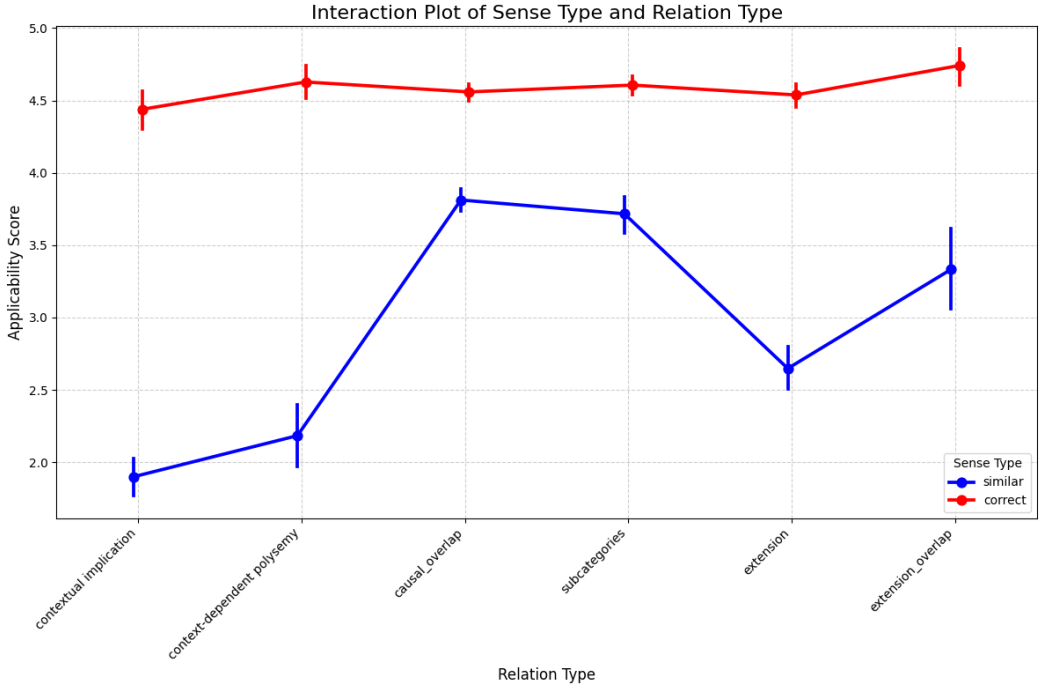
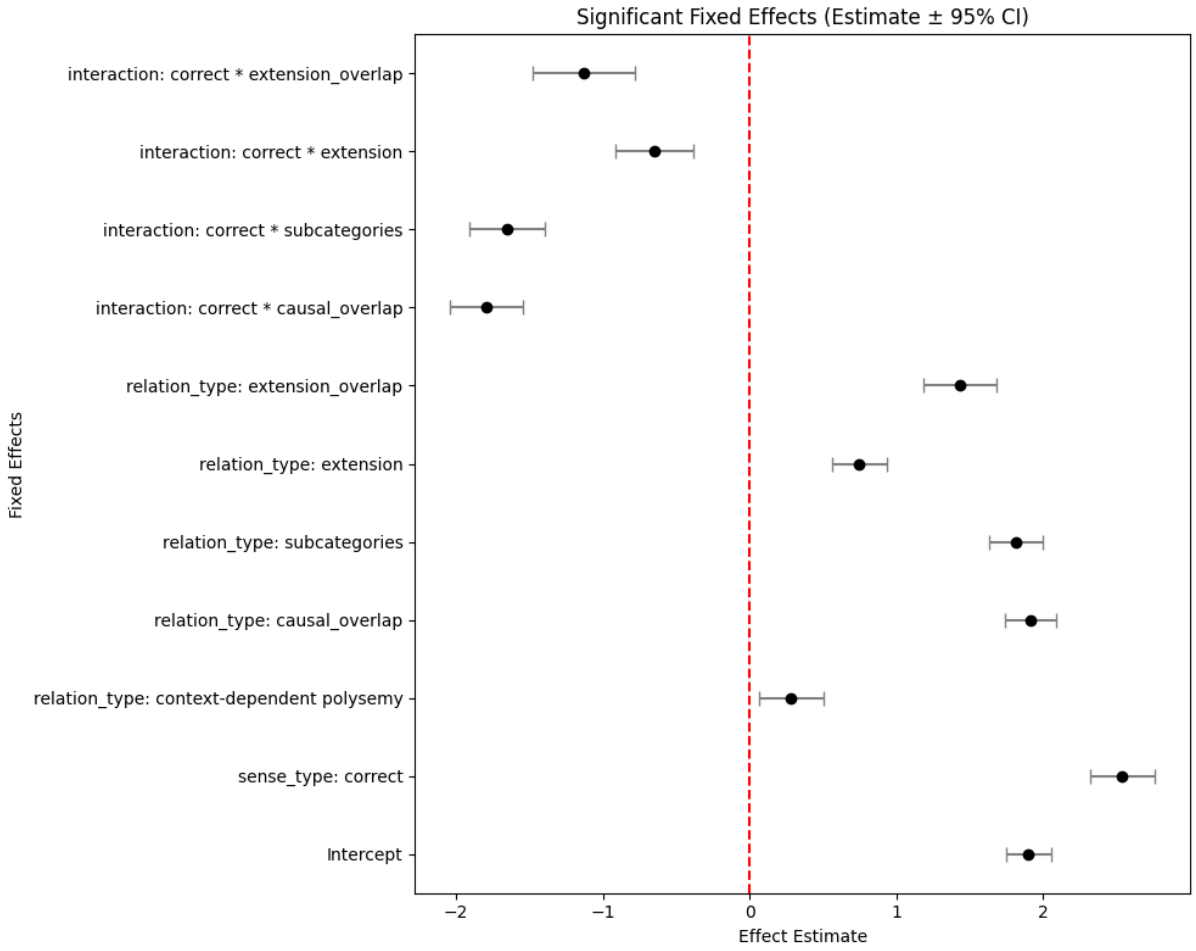



Figure 25 presents the fixed-effect estimates from the OLS regression with 95% confidence intervals. The vertical dotted line at zero represents the reference baseline (similar label × contextual implication). Predictors plotted to the right of the dotted line indicate a positive effect**,** i.e., an increase in applicability scores relative to the baseline. Predictors plotted to the left of the dotted line indicate a negative effect, i.e., a reduction relative to the baseline. As shown, relation types such as causal overlap, subcategories, and extension-based overlap exhibit strong positive main effects, but their interactions with sense type are negative and significant, meaning that the correct–similar gap is reduced under these overlapping relations compared to contextual implication. By contrast, context-dependent polysemy lies close to the zero line and shows no significant interaction effect, behaving more like the discrete baseline condition.

Figure 25. Predictors of the OLS regression model, with 95% confidence intervals




Taken together, the regression analyses corroborate the distributional and overlap-based findings: gradedness in sense applicability judgments emerges systematically in relations characterized by strong semantic proximity and overlap, such as causal overlap, semantic subcategories, and extension-based overlap. Conversely, relations such as contextual implication and context-dependent polysemy maintain large and stable Correct–Similar distinctions, reflecting categorical sense boundaries. These results provide a fine-grained confirmation of Hypothesis 2, showing that semantic proximity and overlap are reliable predictors of when applicability judgments become graded rather than discrete.

To summarize the discussion in this section, the results of distributional analyses consistently show that across humans and generative AI models, semantic proximity and overlap shape whether sense boundaries are judged as categorical or graded**.** Correct vs. Dissimilar labels were treated categorically, whereas Similar labels showed graded variability depending on their semantic relation. Overlap analyses demonstrated that causal overlap, extension-based overlap, and semantic subcategories reduced the Correct–Similar gap, while contextual implication and context-dependent polysemy preserved categorical separation, with metaphorical extension occupying an intermediate position. Regression modeling of human data confirmed these patterns: overlapping relations promoted graded judgments, non-overlapping relations sustained categorical distinctions. Both GPT and Gemini broadly mirrored human responses, though with different degrees of convergence.

### 4.3. Usage Similarity Judgment Experiment

#### 4.3.1. Methods

This section presents the usage similarity judgment experiment. In this experiment, human participants and two generative AI models—GPT and Gemini—rate the similarity of the two uses of the common target verb (*break* and *freeze*) in meaning on a graded scale. The goal is to examine how gradedness in usage similarity judgments emerges systematically in relations characterized by strong semantic proximity and overlap.

**Materials.** A total of 30 sentence pairs were constructed for this experiment. These pairs were divided into two groups.

1. Non-confusable group (10 pairs): Each pair combined two senses that, in the previous meaning classification and selection experiments, showed low confusion tendencies or clearly separated applicability score distributions. One pair was selected per sense combination, yielding 10 pairs in total (5 with *break*, 5 with *freeze*).
2. Confusable group (20 pairs): Each pair combined two senses that, in the earlier experiments, showed high confusion tendencies or strongly overlapping applicability score distributions. For each sense combination, two pairs were selected, yielding 20 pairs in total.

The sentence groups, sentence pair IDs, and the corresponding sense combinations and semantic relations are presented in Table 17**.**

Table 17. Construction of sentence pairs for the usage similarity experiment

| Pair group type | Pair ID | N | Pair labels (sense combinations) | Relation type |
|---|---|---|---|---|
| Non-confusable *break* | B1~B5 | 5 | See Appendix x | Dissimilar |
| Non-confusable *freeze* | F1~F5 | 5 | See Appendix x | Dissimilar |
| Confusable | C1 | 2 | immobilization vs. emotional/mental freezing | Extension-based overlap (extension_overlap) |
| | C2 | 2 | destruction (literal) vs. destruction (figurative) | Different readings of the same sense (same sense_diff reading) |
| | C3 | 2 | economical freezing vs. suspension | Semantic subcategories (subcategories) |
| | C4 | 2 | interruption vs. termination | Semantic subcategories (subcategories) |
| | C5 | 2 | breakthrough vs. destruction | Extension-based overlap (extension_overlap) |
| | C6 | 2 | violation vs. destruction (figurative) | Metaphorical extension (extension) |

| | | | | |
|---|---|---|---|---|
| | C7 | 2 | breakthrough vs. termination | Causal overlap and semantic subcategories (overlap_subcategories) |
| | C8 | 2 | violation vs. termination | Causal overlap (overlap) |
| | C9 | 2 | change vs. interruption | Causal overlap and semantic subcategories (overlap_subcategories) |
| | C10 | 2 | disclosure vs. breakthrough | Context-dependent polysemy and causal overlap (context-dependent_overlap) |

The sentence pairs in the Confusable group involve relations of semantic overlap, semantic subcategories, and metaphorical extension**.** Among these, semantic overlap is most prominent: in such pairs, the meaning of one sentence overlaps with that of the other. The experiment included sentence pairs exemplifying four subtypes of semantic overlap.

Extension-based overlap is illustrated in (16) and (17). In (16), immobilization overlaps with emotional/mental freezing, with the physical freezing of the body extending metaphorically to emotional paralysis. In (17), a destructive act is extended metaphorically to a breakthrough: the act of breaking the defense and advancing is not separate from destruction but condensed into one meaning. In this sense, the destruction of the defensive line is itself the breakthrough event; sentence (17A) overlaps with the meaning of (17B).

(16) C1: Extension-based overlap
A. She froze in place as the shadow moved across the wall. [immobilization]
B. I froze, unable to speak as the truth sank in. [emotional/mental freezing]

(17) C5: Extension-based overlap
A. The striker broke the press and charged toward the goal. [breakthrough]
B. The rioters broke the windows during the protest. [destruction]

Causal overlap is exemplified in (18). Here, termination (18B) is causally entailed by violation (18A): breaking the rules eventually leads to the ending of an agreement.

(18) C8: Causal overlap
A. The athlete broke the anti-doping rules before the finals. [violation]
B. The team broke the agreement after several unmet conditions. [termination]

Examples (19) and (20) combine causal overlap and semantic subcategories. In (19), breakthrough (19A) entails termination (19B) while both belong to the broader category of disruption of continuity. Similarly, in (20), change (20A) overlaps with interruption (20B), and the two again represent subcategories of disruption of continuity.

(19) C7: Causal overlap and semantic subcategories
A. The joint workshop finally broke the isolation and sparked genuine dialogue. [breakthrough]

B. She broke the connection and walked away without looking back. [termination]

(20) C9: Causal overlap and semantic subcategories

A. She broke the monotony with a new hobby on the weekend. [change]

B. A joke from one member broke the uneasy atmosphere. [interruption]

(21) illustrates a case where the two sentences exhibit both context-dependent polysemy and causal overlap. The phrase *break the case* may be interpreted either as disclosure of hidden information or as a breakthrough that marks a turning point in solving the case. In (21B), the police investigation causes the case to emerge and simultaneously initiates its resolution, so that the breakthrough meaning overlaps with the disclosure meaning in (21A).

(21) C10: Context-dependent polysemy and causal overlap

A. The investigation broke the story of widespread election fraud. [disclosure]

B. The police broke the case wide open after years of stalled progress. [breakthrough]

In addition to overlap relations, the Confusable group also included sentence pairs that exemplify semantic subcategories without overlap. These involve two senses that belong to the same higher-level category but do not overlap in meaning. For example, in (22), economical freezing and suspension are both subtypes of cessation, while in (23), interruption and termination are both subtypes of disruption of continuity.

(22) C3: Semantic subcategories

A. The government froze all assets tied to the company. [economical freezing]

B. The board voted to freeze the project until further notice. [suspension]

(23) C4: Semantic subcategories

A. He broke the silence with an awkward laugh. [interruption]

B. They broke their partnership after years of disagreements. [termination]

The experiment further incorporated sentence pairs illustrating metaphorical extension. Example (24) shows how the violation sense of *break* in (24A) extends metaphorically to a figurative destruction in (24B), where the abstract entity “the foundations of democracy” is construed as something that can be broken.

(24) C6: Metaphorical extension

A. He was suspended for breaking the university’s honor code. [violation]

B. The new legislation threatens to break the foundations of democracy. [figurative Destruction]

Finally, beyond the six relation types already described, the Confusable group also included pairs contrasting literal and figurative readings of the same sense. As shown in (25), both sentences instantiate the destruction sense of *break*, but (25A) expresses it literally while (25B) applies it figuratively to the loss of political power.

(25) C2: Different readings of the same sense

A. The protesters broke the glass doors of the city hall. (literal destruction)

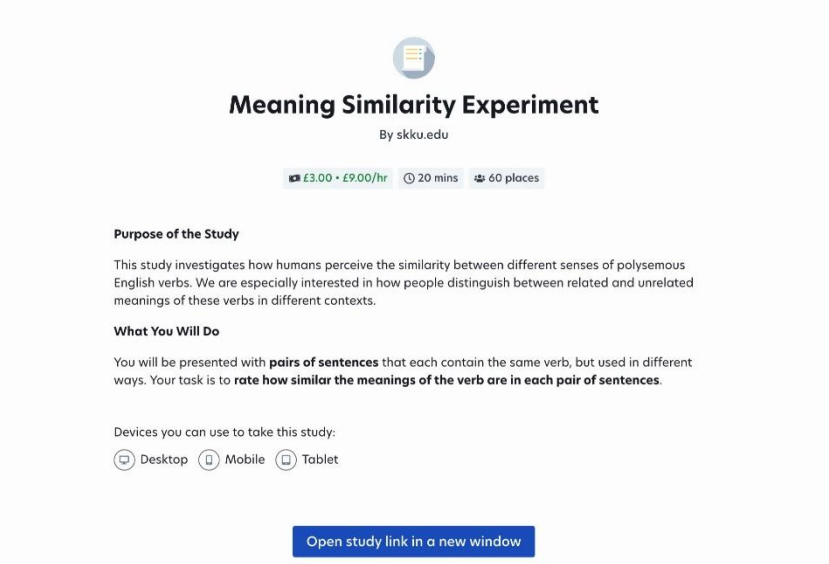
B. The uprising finally broke the dictator's power. (figurative destruction)

A total of 30 sentence pairs (60 individual sentences) were constructed in this way and then randomized before being presented to participants. The full experimental dataset is available at the accompanying GitHub repository: https://github.com/hanjung-25/clmsemantics/tree/data.

**Participants, task, and procedure.** Sixty native speakers of English (30 male, 30 female; age range 21–57) were recruited via Prolific. None of them had participated in the previous experiments. Participants were presented with 30 pairs of sentences containing the same target verb (*break* or *freeze*) in two different usages.

Their task was to judge the similarity of meaning between the two uses of the verb. To do so, they rated each pair on a 5-point scale, where 1 = *completely different meanings*, 2 = *mostly different meanings*, 3 = *somewhat similar meanings*, 4 = *mostly similar meanings*, and 5 = *exactly the same meaning*. The instructions emphasized that participants should focus only on the meaning of the target verb, ignoring other aspects such as tense, subject, or object, and that there were no right or wrong answers. Judgments were expected to reflect participants' intuitive understanding of verb meaning. A screenshot of the Prolific interface for this task is displayed in Figure 26. The full set of participant instructions is provided in Appendix G.

Figure 26. Screenshot of the Prolific interface for the usage similarity judgment task




The total duration of the experiment was less than 25 minutes. Each participant completed all 30 sentence pairs and received a payment of £3**.** In total, data were collected from 60 participants**.** Three participants showed a response pattern in which they assigned the same score to all items and were therefore excluded. The analyses thus included the data from the remaining 57 participants.

**Model procedure.** For the model-based usage–similarity judgments, GPT and Gemini were accessed via their official APIs from a Colab environment. The same 30 sentence pairs used in the human experiment were presented to each model with a standardized prompt that (i) shows

the two sentences sharing the same target verb (*break* or *freeze*), (ii) asks for a meaning–similarity rating on a 5-point scale (1 = *completely different, …,* 5 = *exactly the same),* and (iii) returns a numeric score only.

To account for stochastic variation, each pair was evaluated 10 times per model (independent API calls). We then averaged the 10 scores to obtain a stable similarity estimate for each pair × model. The full Colab notebook (API calls, batching, parsing, and aggregation) is available in the accompanying GitHub repository (same repo as the human materials), ensuring exact reproducibility of prompts and scoring.

**Hypothesis and predictions.** This experiment was designed to test the hypothesis that usage similarity scores systematically reflect the degree of semantic connectivity between senses**.** Specifically, sentence pairs drawn from meaning relations characterized by strong semantic proximity and overlap (e.g., causal overlap, extension-based overlap, semantic subcategories) are expected to receive higher similarity ratings. By contrast, sentence pairs involving relations with weaker or more distant semantic connections are predicted to receive lower similarity ratings. These predictions will be tested in Section 4.3.2.

### 4.3.2. Results

A total of 1,710 human responses and 1,000 AI model responses were collected. We begin by examining the distribution of usage similarity scores in order to test the hypothesis. We then apply regression analyses to assess the significance of semantic relations in shaping similarity judgments.

**Similarity score distribution across group types**. Table 18 presents descriptive statistics of similarity scores for the Non-confusable and Confusable groups, collapsed across all evaluators. The Non-confusable pairs yielded lower similarity scores overall (M = 1.84, Median = 2.0, SD = 0.97, IQR = 1.0), with scores tightly clustered at the lower end of the scale. By contrast, the Confusable pairs received substantially higher ratings (M = 3.21, Median = 3.0, SD = 1.25, IQR = 2.0), indicating broader variability and generally stronger perceived similarity. This pattern confirms the predicted divergence between pairs involving semantically distinct senses and those involving semantically proximate or overlapping senses.

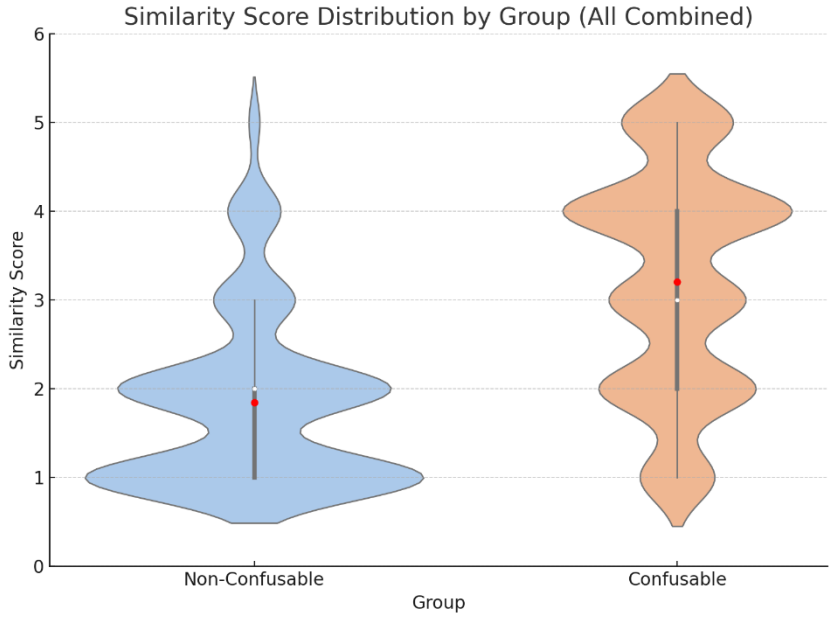
Table 18. Summary statistics of similarity scores across sense types

| Group type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Non-confusable | 770 | 1.84 | 2.0 | 0.97 | 1.0 | 2.0 | 1 | 5 | 1.0 |
| Confusable | 1940 | 3.21 | 3.0 | 1.25 | 2.0 | 4.0 | 1 | 5 | 2.0 |

Figure 27 visualizes these distributions using violin plots. The horizontal axis represents the two sentence-pair groups (Non-confusable vs. Confusable), and the vertical axis represents similarity scores (1–5). Within each plot, the white dot marks the mean, the red dot marks the median, and the thick vertical line indicates the interquartile range (IQR), with whiskers extending to 1.5 × IQR. The Non-confusable group shows a compressed distribution concentrated at lower scores, whereas the Confusable group exhibits a broader and higher distribution. This visualization reinforces the statistical summary, highlighting the systematic elevation of similarity judgments for Confusable pairs. These results provide initial support for

the hypothesis that semantic proximity and overlap systematically increase similarity judgments, distinguishing Confusable pairs from Non-confusable ones.

Figure 27. Similarity score distribution by group type (all evaluators)




Now let us turn to the distribution of similarity scores separately for humans and the two AI models. Tables 19 and 20 summarize the descriptive statistics of similarity scores for the Non-confusable and Confusable groups across the three evaluator types (humans, GPT, and Gemini). To visualize these distributions, Figures 28 and 29 present violin plots of human and model ratings, respectively.

Table 19. Summary statistics of similarity scores across sense types (human)

| Group type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Non-confusable | 570 | 1.87 | 1.0 | 1.08 | 1.0 | 2.0 | 1 | 5 | 1.0 |
| Confusable | 1140 | 3.23 | 4.0 | 1.44 | 2.0 | 5.0 | 1 | 5 | 3.0 |

Table 20. Summary statistics of similarity scores across sense types (AI models)

a. GPT

| Group type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Non-confusable | 100 | 1.77 | 2.0 | 0.42 | 2.0 | 2.0 | 1 | 2 | 0 |
| Confusable | 400 | 2.67 | 3.0 | 0.74 | 2.0 | 3.0 | 1 | 4 | 1.0 |

b. Gemini

| Group type | N | Mean | Median | SD | Q1 (25%) | Q3 (75%) | Min | Max | IQR |
|---|---|---|---|---|---|---|---|---|---|
| Non-confusable | 100 | 1.75 | 2.0 | 0.59 | 2.0 | 1.0 | 1 | 3 | 0 |
| Confusable | 400 | 3.54 | 4.0 | 0.83 | 2.0 | 3.0 | 1 | 4 | 1.0 |

Figure 28 shows that human judgments produced distributions comparable to those observed when all evaluators were combined: Non-confusable pairs tended to receive low similarity ratings (median ≈ 2), while Confusable pairs were rated substantially higher (median ≈ 4). This pattern reinforces the hypothesis that semantic proximity and overlap systematically elevate perceived similarity.

Figure 28. Similarity score distribution by group type (human)

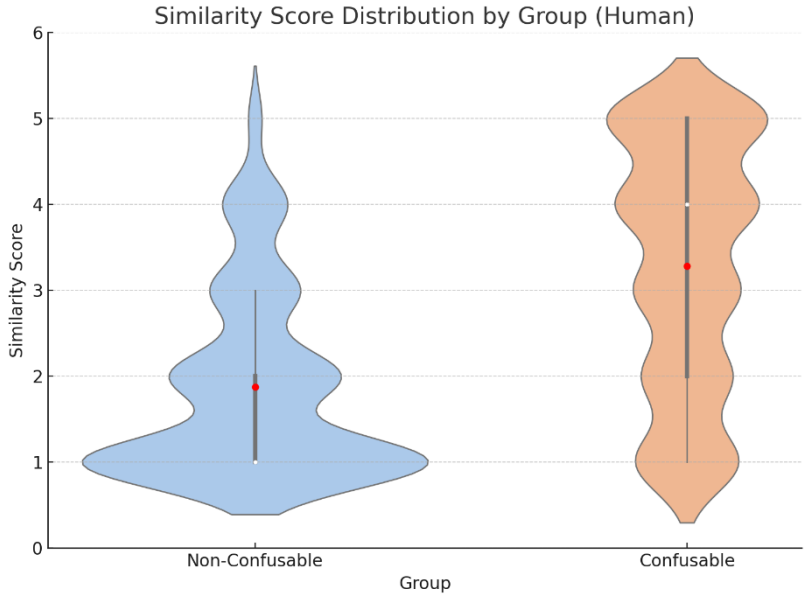
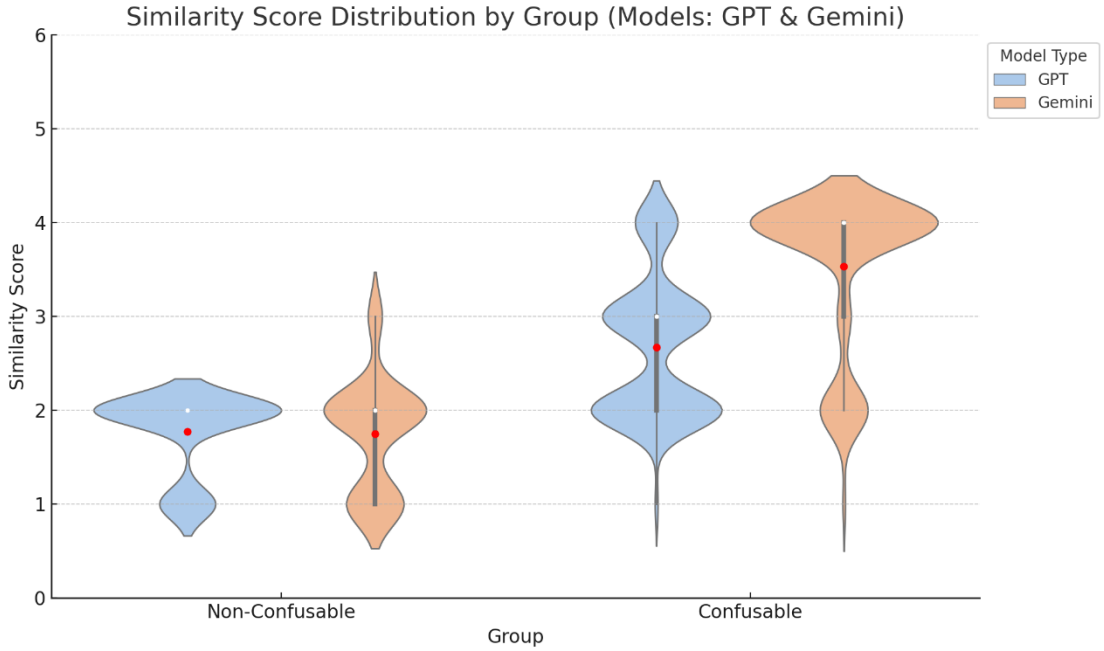



Figure 29 displays the model ratings. Both GPT and Gemini replicated the same overall pattern as humans—higher scores for Confusable than for Non-confusable pairs—but with notable differences in scale and spread. For GPT, the Non-confusable group is compressed at the low end (M = 1.77, Med = 2.0), while the Confusable group shows a broader spread and moderately higher mean scores (M = 2.67, Med = 3.0). This indicates GPT tends to assign more restricted and lower scores overall, but still differentiates the two groups. For Gemini, the pattern is more similar to humans: Confusable pairs are rated clearly higher (M = 3.54, Med = 4.0), while Non-confusable pairs remain low (M = 1.75, Med = 2.0). Compared with GPT, Gemini produces higher overall scores and a closer alignment with human judgments.

Figure 29. Similarity score distribution by group type (GPT and Gemini)




To statistically validate whether the similarity score distributions for the Confusable and Non-confusable groups exhibit consistent patterns across the three evaluator types—humans, GPT, and Gemini—we conducted independent-samples *t*-tests. Table 21 presents the results of independent samples *t*-tests comparing similarity scores between the Confusable and Non-Confusable groups for humans, GPT, and Gemini. Across all three evaluation sources, the Confusable pairs yielded significantly higher similarity scores than the Non-Confusable pairs. For humans, the difference was highly significant, $t(1708) = -22.56$, $p < .001$, with a large effect size (Cohen's $d = 1.06$). GPT also showed a significant difference, $t(998) = -16.05$, $p < .001$, $d = 1.31$, again indicating a large effect. Gemini exhibited the strongest contrast, $t(998) = -24.67$, $p < .001$, with a very large effect size ($d = 2.26$).

Table 21. Independent-samples *t*-test results for similarity scores: Confusable vs. Non-confusable pairs

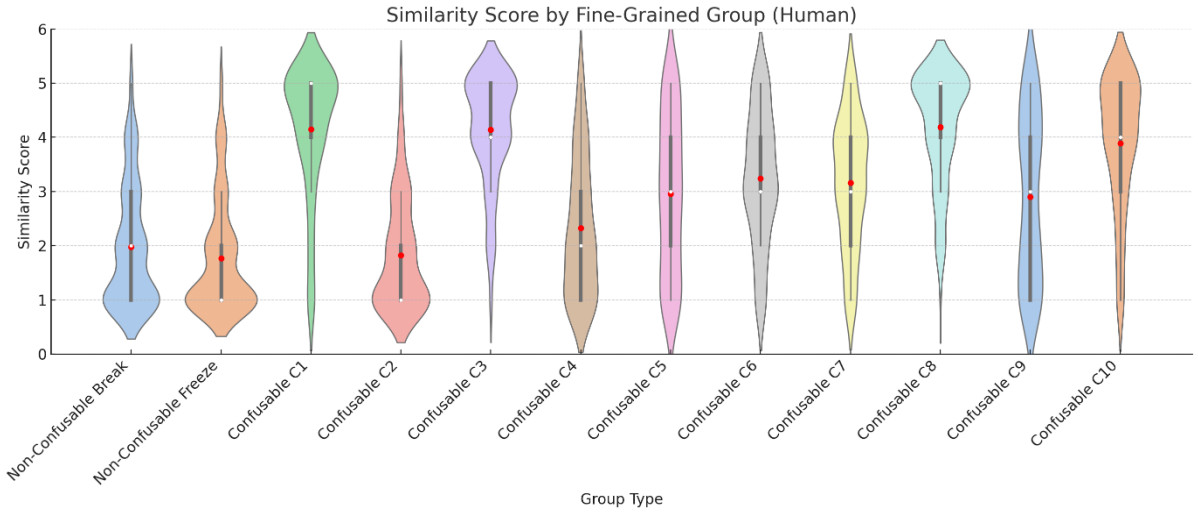
| Evaluator type | *t*-statistic (*t*) | *p*-value | Cohen's *d* | Effect size interpretation |
|---|---|---|---|---|
| Human | -22.56 | $8.07 \times 10^{-37}$ | 1.06 | large effect |
| GPT | -16.05 | $4.53 \times 10^{-41}$ | 1.31 | large effect |
| Gemini | -24.67 | $1.15 \times 10^{-53}$ | 2.26 | very large effect |

Taken together, these results demonstrate a robust and consistent pattern: for all subjects, Confusable pairs were judged substantially more similar in meaning than Non-confusable pairs, with the magnitude of the difference especially pronounced for Gemini.

Now we turn to the similarity score distributions across the fine-grained group types for both human and model data. Figure 30 visualizes the human data, showing the distributions of similarity scores for the 12 sentence-pair groups identified by sentence pair IDs in Table 17. The two Non-confusable groups cluster at the lower end of the scale, clearly separated from the higher scores observed in the Confusable groups. Within the Confusable set, meaningful variation emerges: four groups—C1, C3, C8, and C10—exceed the human Confusable group mean similarity score of 3.23 and show concentrated distributions at the upper end. These

include the semantic overlap relations (C1, C8, C10) and the near-synonym subcategory relation (C3). By contrast, C6 (metaphorical extension), along with C7 and C9 (both semantic overlap), yield scores close to the Confusable-group mean. Notably, C4, representing the interruption–termination subcategory relation for *break*, falls below the average, while C2—capturing the literal–figurative readings of the destruction sense—displays a distribution more akin to the Non-confusable groups and registers the lowest mean among the Confusable pairs.

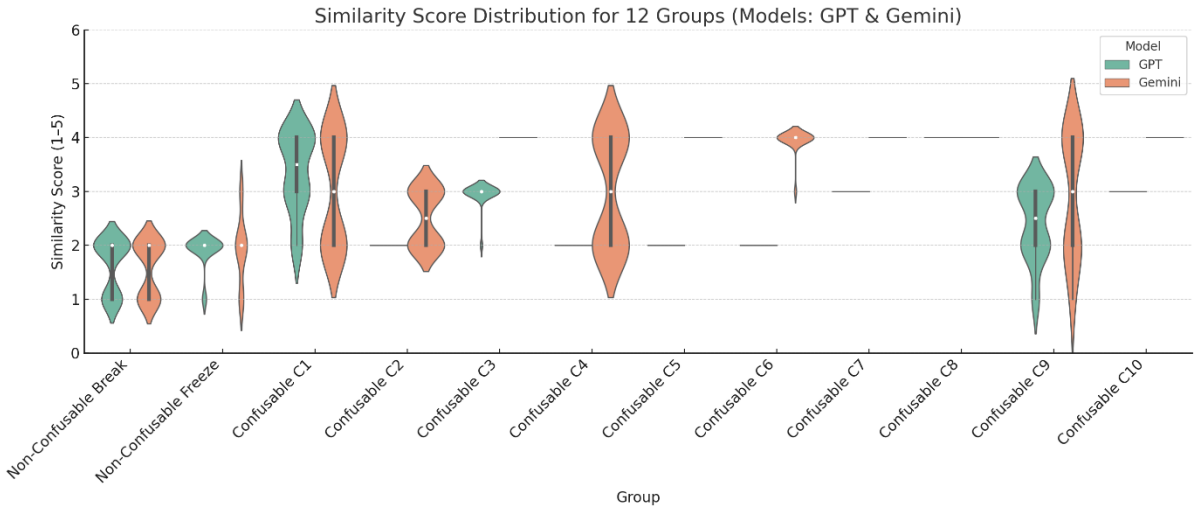
Figure 30. Similarity score distribution by fine-grained group type (human)




Turning to the model data, Figure 31 presents the similarity score distributions for GPT and Gemini across the 12 fine-grained sentence-pair groups. As in the human data, the two Non-confusable groups are clearly separated from the Confusable groups, showing consistently lower similarity scores. We can see that for GPT, the Confusable groups with mean similarity scores above the GPT-wide Confusable mean (2.67) include C1, C3, C7, C8, and C10. Groups below this mean are C2, C4, C5, and C6, with C2 and C4 standing out as the lowest-performing groups, consistent with the human data. C9 hovers around the GPT mean, showing no strong divergence. For Gemini, six groups exceed the Gemini-wide Confusable mean (3.54): C3, C5, C6, C7, C8, and C10. Notably, C3 (a near-synonym subcategory relation) and C8 and C10 (both semantic overlap relations) align with the high-scoring Confusable groups in the human data. Below the Gemini mean fall C1, C2, C4, and C9, with C2 and C4 again emerging as the lowest Confusable groups—a pattern consistently observed across all three evaluators (Human, GPT, Gemini).

Another striking feature of the model data is the stronger tendency toward score concentration at discrete points, compared to the smoother distributions in human judgments. This effect is especially pronounced in 7 Confusable groups for GPT and 5 Confusable groups for Gemini, underscoring a model-specific bias toward categorical scoring.

Figure 31. Similarity score distribution by fine-grained group type (models)




Another striking feature of the model data is the stronger tendency toward score concentration at discrete points, compared to the smoother distributions in human judgments. This effect is especially pronounced in 7 Confusable groups for GPT and 5 Confusable groups for Gemini, underscoring a model-specific bias toward categorical scoring.

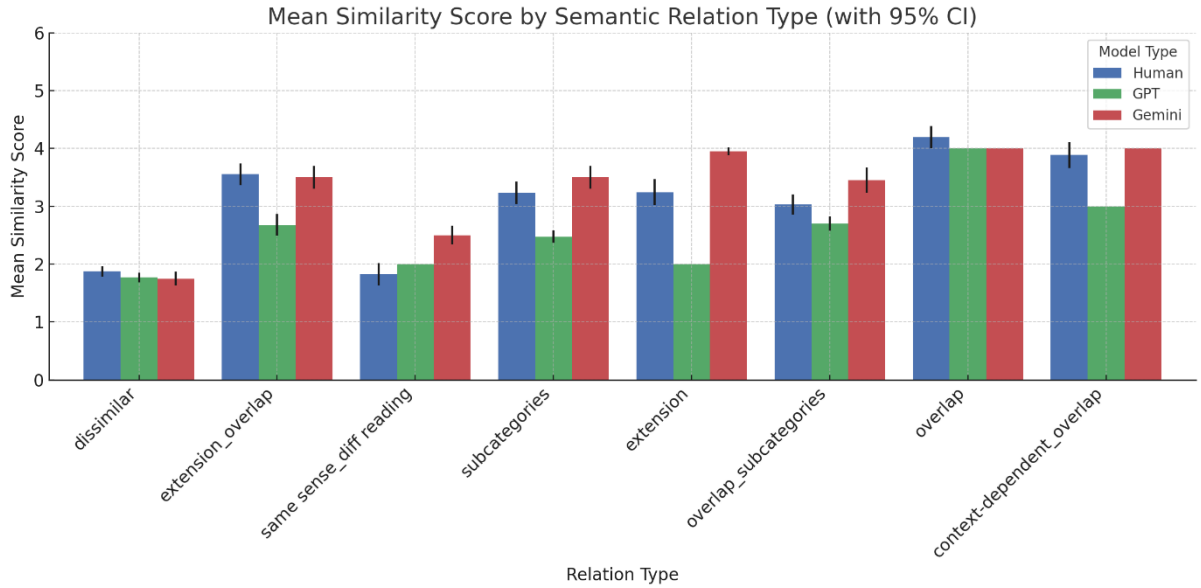
A comparison of Figures 30 and 31 reveals both convergence and divergence between human and model similarity judgments. Across all three evaluators, the two Non-confusable groups remain clustered at the low end of the scale, while the Confusable groups consistently show elevated similarity scores, confirming the strong role of semantic proximity and overlap in shaping judgments. At the same time, differences emerge in the fine-grained profiles. Human judgments display more graded distributions, with smooth variability across the ten Confusable groups, whereas the models exhibit sharper clustering around discrete score values, indicating a stronger tendency toward categorical scoring. GPT captures the relative ranking of several groups (e.g., C2 and C4 as the lowest, C8 and C10 as higher-scoring overlap relations), but its overall means remain lower than those of humans. Gemini, in contrast, assigns uniformly higher scores and more closely replicates the human pattern in high-scoring groups such as C3, C8, and C10, though it also exaggerates the separation between groups. Together, these results suggest that while both models are sensitive to semantic proximity, Gemini aligns more closely with human intuitions in the fine-grained distribution of Confusable relations, whereas GPT provides a flatter and more conservative approximation.

To verify differences in similarity judgments across semantic relations, we compared mean similarity scores across relation types and evaluation subjects. Table 22 presents the mean similarity scores by relation type for humans, GPT, and Gemini, while Figure 32 visualizes these results with 95% confidence intervals. The patterns are clear: with the exception of the *different readings of the same sense* relation, all semantic relations yield substantially higher similarity scores compared to the dissimilar relation (the baseline corresponding to the non-confusable group). This finding provides direct evidence that both humans and models judge verb uses linked by semantic proximity and overlap as more similar than those merely sharing the same sense label.

Table 22. Mean similarity scores by relation type and evaluation subject

| Relation type | Mean similarity score | | |
|---|---|---|---|
| | Human | GPT | Gemini |
| dissimilar | 1.87 | 1.77 | 1.75 |
| Extension-based overlap (extension_overlap) | 3.55 | 2.68 | 3.50 |
| Different readings of the same sense (same sense_diff reading) | 1.82 | 2.00 | 2.50 |
| Semantic subcategories (subcategories) | 3.23 | 2.48 | 3.50 |
| Metaphorical extension (extension) | 3.25 | 2.00 | 3.95 |
| Causal overlap and semantic subcategories (overlap_subcategories) | 3.03 | 2.70 | 3.45 |
| Causal overlap (overlap) | 4.19 | 4.00 | 4.00 |
| Context-dependent polysemy and causal overlap (context-dependent_overlap) | 3.89 | 3.00 | 4.00 |

Figure 32. Mean similarity scores by relation type and evaluation subject




Figures 33 and 34 visualize the distribution of similarity scores across semantic relation types for human participants and the two AI models. In both plots, the six relations characterized by semantic proximity and overlap are clearly distinguished from the dissimilar relation and the relation involving different readings of the same sense, yielding markedly higher similarity scores. This pattern provides explicit evidence that semantic proximity and overlap drive elevated similarity judgments across evaluation subjects.

At the same time, important differences emerge between humans and models. Human judgments display more graded distributions, with smooth variability across the scale, reflecting flexible and continuous assessments of similarity. By contrast, the models—especially GPT—exhibit sharper clustering around discrete score values, revealing a stronger tendency toward categorical scoring.

Figure 33. Similarity score distribution by relation type (human)

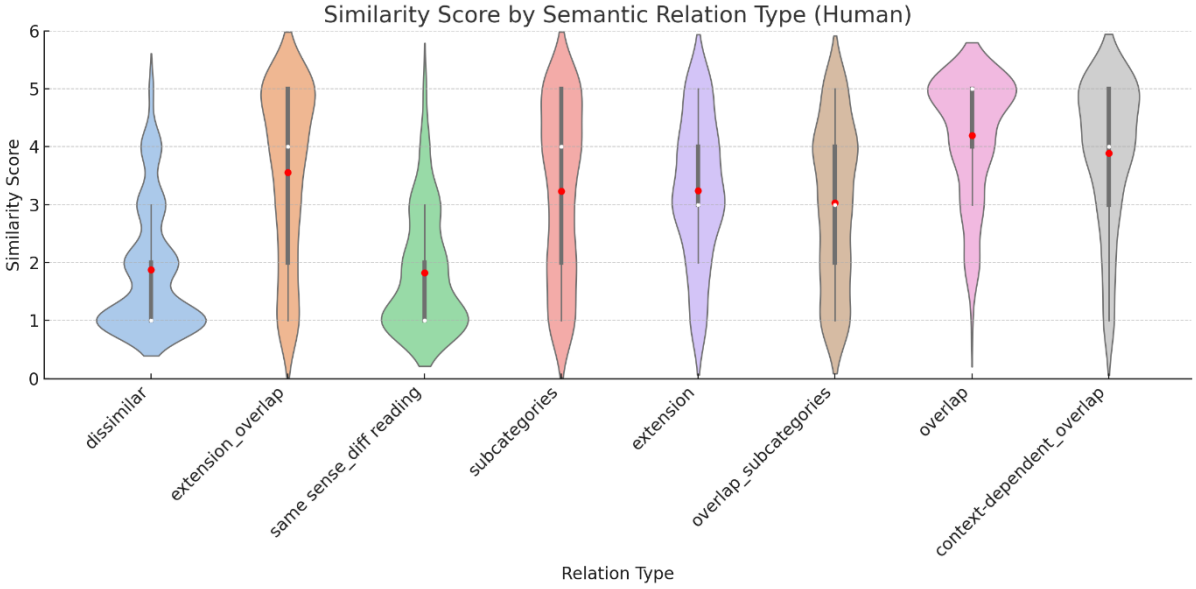
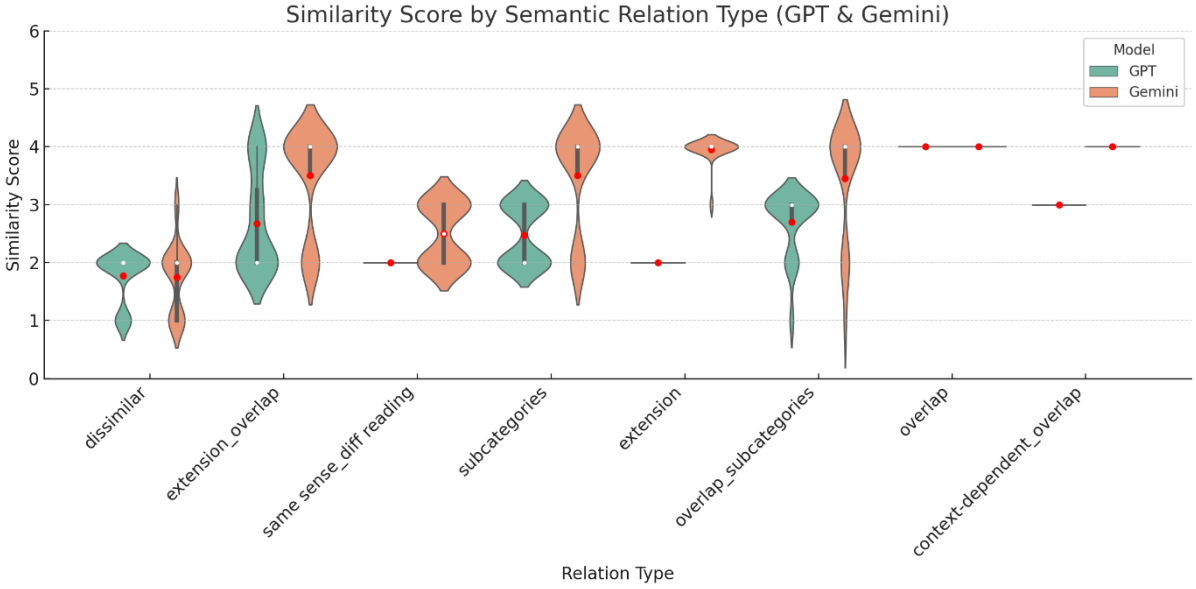


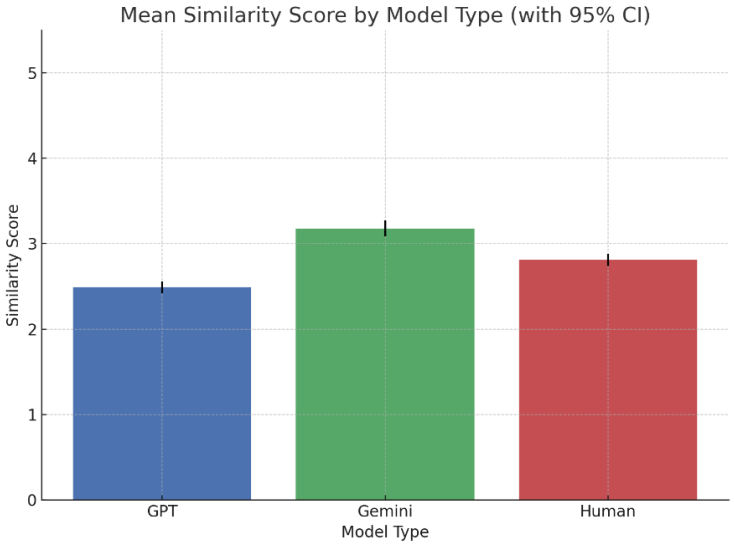

Figure 34. Similarity score distribution by relation type (models)




To further examine differences across the three evaluation subjects, we conducted a one-way ANOVA on the mean similarity scores (Figure 35). The ANOVA revealed a significant main effect of subject type, $F(2, 2707) = 34.65$, $p < .001$, confirming that Human, GPT, and Gemini differ systematically in their similarity judgments. As shown in Figure 35, Gemini

assigned the highest average similarity scores, humans fell in between, and GPT produced the lowest ratings.

Figure 35. Mean similarity score by evaluation subject




To pinpoint which contrasts drive this overall effect, we conducted Tukey's HSD post-hoc tests (Table 23). All pairwise comparisons were significant ($p < .001$): GPT's scores were significantly lower than both Human and Gemini, and Gemini's scores were significantly higher than Human's. These findings indicate that while all three evaluators clearly distinguished Confusable from Non-confusable pairs, their overall similarity levels diverge. In particular, Gemini tends to inflate similarity relative to human judgments, whereas GPT adopts a more conservative stance, assigning lower scores across the board.

Table 23. Tukey HSD pairwise comparisons of mean similarity scores by evaluation subject

| Comparison | Mean difference | 95% CI (Lower, Upper) | *p*-value | Cohen's *d* | Effect size interpretation |
|---|---|---|---|---|---|
| GPT – Gemini | +0.69 | [0.49, 0.88] | < .001 | –0.74 | medium-large |
| GPT – Human | +0.32 | [0.16, 0.48] | < .001 | –0.24 | small |
| Gemini – Human | –0.37 | [–0.52, –0.21] | < .001 | +0.26 | small |

The distributional analyses confirmed the predicted pattern: pairs connected by stronger semantic proximity and overlap consistently received higher similarity ratings than those linked by weaker or more distant relations. To more rigorously test the significance and relative strength of these relational effects, we now turn to regression modeling of the human data.

**Regression analysis of human sense applicability judgments.** As our design and data structure are well suited to mixed-effects regression models, we adopted a specification that included relation type and verb as fixed effects and participant as a random effect. The key predictor variable in this analysis is relation type, specified in Table 24.[1] The relations labeled

[1] Interaction terms were not included in the model, as they are not relevant to testing the hypotheses of the present experiment.

as (i) causal overlap, (ii) causal overlap and semantic subcategories, and (iii) context-dependent polysemy and causal overlap were all collapsed into the category of causal overlap. Based on the distributional analyses of similarity scores, we test the following predictions:

1. **Semantic proximity and overlap**: Relations characterized by semantic proximity and overlap are expected to yield significantly higher similarity scores compared to the dissimilar baseline, and these increases should be larger than those observed for relations where the two uses share the same sense label.
2. **Verb effect**: Since the "same sense" relation type occurs only in the experimental materials for the verb *break*, the similarity scores for *break* are expected to be lower overall compared to *freeze*, which serves as the reference category.

Table 24. Specification of the predictor variables

| Variables | Levels |
|---|---|
| Relation type | dissimilar (reference level), causal overlap, extension-based overlap, semantic subcategories, metaphorical extension, different readings of the same sense |
| Verb | *freeze* (reference level), *break* |

We built the regression models using Python's `statsmodels` package in Google Colab. All analysis scripts and code are available in the accompanying GitHub repository.

The regression analysis reported here differs from that of the sense applicability judgment data in Section 4.2.2. Whereas the earlier analysis converged to ordinary least squares (OLS) estimates due to negligible random-effect variance, the present analysis was executed as a mixed-effects regression model that retained participant as a random effect. In this case, standard model fit indices such as $R^2$ or AIC are not reported, as they are not directly provided by the mixed-effects implementation. Nonetheless, the model converged stably, and the fixed-effect estimates are summarized in Table 25. These coefficient estimates form the basis for the interpretation of how semantic relation types influence similarity judgments.

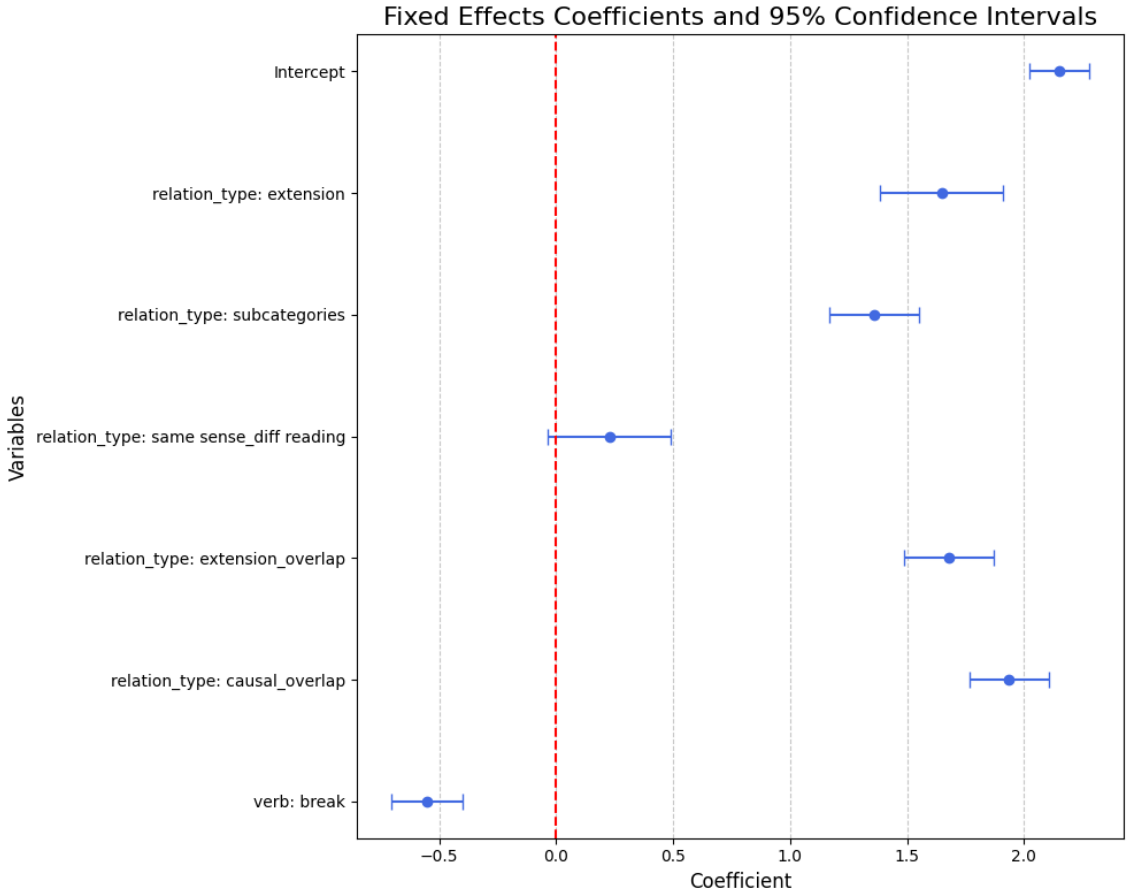
Table 25. Regression results for human usage similarity judgments

| **Predictor** | **Coef.** | **Std. Err.** | **t** | ***p*** | **Interpretation** |
|---|---|---|---|---|---|
| Intercept | 0 | – | – | 1.000 | baseline score |
| Relation type | | | | | |
| – Extension | +1.65 | 0.12 | 13.47 | <.001 | Stronger similarity than baseline |
| – Semantic subcategories | +1.36 | 0.09 | 15.21 | <.001 | Stronger similarity than baseline |
| – Same sense_different readings | +0.23 | 0.12 | 1.85 | .064 | n.s. |
| – Extension-based overlap | +1.68 | 0.09 | 18.79 | <.001 | Stronger similarity than baseline |
| – Causal overlap | +1.94 | 0.08 | 24.21 | <.001 | Strongest effect |
| Verb | | | | | |
| – break | –0.55 | 0.07 | –7.75 | <.001 | Less similar judgments |

Model fit indicates that the predictors explain systematic variance in similarity ratings. The coefficients show strong effects of relation type, consistent with our predictions. Compared to the dissimilar baseline, all semantic proximity and overlap relations—causal overlap (β = 1.94, $p < .001$), extension-based overlap (β = 1.68, $p < .001$), extension (β = 1.65, $p < .001$), and semantic subcategories (β = 1.36, $p < .001$)—significantly increased similarity ratings. These results directly support the hypothesis that stronger semantic connectivity elevates perceived similarity. By contrast, the same-sense different readings relation (β = 0.23, $p = .064$) did not differ significantly from the dissimilar baseline, suggesting that literal–figurative alternations within the same sense are not judged as strongly similar when compared to relations based on overlap or subcategorization. Finally, the coefficient for the verb factor shows a negative effect for *break* relative to *freeze* (β = –0.55, $p < .001$), aligning with our prediction that *break* would receive lower overall similarity ratings due to the inclusion of "same sense" pairs in its items.

Figure 36 visualizes the fixed-effects coefficients with 95% confidence intervals. The plot clearly shows that all semantic relation types characterized by semantic proximity and overlap (extension, subcategories, extension-based overlap, causal overlap) had strong positive effects on similarity scores relative to the dissimilar baseline, while same-sense different readings showed only a weak increase. In contrast, the coefficient for verb: *break* was negative, confirming the lower similarity ratings for *break* compared to *freeze*. Importantly, the confidence intervals for the significant predictors do not cross zero, underscoring the robustness of these effects.

Figure 36. Predictors of the regression model, with 95% confidence intervals



Taken together, the regression analyses of human data show that semantic relation types characterized by semantic proximity and overlap reliably raise similarity scores relative to the baseline, whereas weakly connected or more distant relations preserve low similarity, thereby corroborating the central hypothesis.

In sum, the usage similarity judgment experiment demonstrated that semantic proximity and overlap strongly elevate similarity ratings. Non-confusable pairs were consistently rated low, while Confusable pairs showed higher scores, with causal overlap, extension-based overlap, metaphorical extension and semantic subcategories producing the strong effects, whereas literal–figurative readings received ratings similar to Non-confusable pairs. Human judgments showed graded distributions, whereas GPT and Gemini tended toward categorical clustering, though both broadly paralleled human patterns. Statistical tests confirmed the robustness of these differences, and regression analyses of human data verified that semantic proximity and overlap relations reliably increased similarity scores relative to the baseline. Overall, the findings support the hypothesis that gradedness in similarity judgments systematically reflects the degree of semantic connectivity among verb senses.

### 4.4. Conclusions

This chapter has examined how both human participants and generative AI models perceive and evaluate sense boundaries in polysemous verbs, focusing on *break* and *freeze*. Through three complementary experiments—human meaning selection (4.1), sense applicability judgment (4.2), and usage similarity judgment (4.3)—we systematically investigated how categorical versus graded sense distinctions emerge, and how they are shaped by semantic proximity and overlap.

The human meaning selection task demonstrated that sense confusion patterns are not random but highly systematic. Senses with strong semantic connectivity—such as *change*, *immobilization*, and *economic freezing*—exhibited high confusion rates and clustered closely with semantically related senses in hierarchical analyses. By contrast, clearly delimited senses like *bodily harm* or *termination* yielded minimal errors. These findings confirm that semantic overlap and proximity are key predictors of sense confusability in human judgments, paralleling the model prediction patterns observed in Chapter 3.

The sense applicability judgments experiment further revealed that the gradedness of judgments varies systematically with the type of semantic relation between senses. Correct and dissimilar labels were evaluated categorically, whereas similar labels showed graded, overlapping distributions—especially in relations involving causal or extension-based overlap and semantic subcategories. Regression analyses confirmed that these overlapping relations significantly reduce the Correct–Similar score gap, producing gradient judgments of sense applicability. Conversely, context-dependent polysemy and contextual implication yielded large, categorical distinctions, while metaphorical extension occupied an intermediate position. Both GPT and Gemini broadly mirrored these human patterns, indicating that generative models are sensitive to the same semantic structures that guide human sense evaluation.

The usage similarity judgments experiment extended the analysis to sentence-level meaning comparisons. Human and model ratings jointly showed that semantic proximity among senses produces higher similarity scores and greater overlap in distribution, particularly for causally or extensionally related senses. Literal–figurative readings of the same sense yielded intermediate similarity, aligning with subcategorical relations rather than with distinct

sense boundaries. Mixed-effects regression further demonstrated that similarity ratings were systematically predicted by relation type, confirming the graded continuum between categorical and overlapping senses.

Across all three experiments, a convergent pattern emerged: semantic connectivity—manifested as causal, extension-based, or subcategorical relations—consistently predicts gradient sense boundaries, whereas context-dependent or implicature-based relations sustain discrete distinctions. Human and AI judgments alike reflected this continuum, differing primarily in variability rather than in underlying structure. Together, these results provide robust empirical support for a graded, continuum-based model of polysemy, in which sense boundaries are shaped by structured patterns of semantic proximity and overlap rather than by discrete categorical divisions. Generative AI models, despite differences in architecture and variability, align closely with human evaluative patterns, suggesting that large language models capture key aspects of human-like semantic organization.

These findings set the stage for Chapter 5, which integrates the human and AI results from Chapters 3 and 4 to advance a unified account of the mental and computational representation of polysemous verb meaning.

# Chapter 5

# Interpretation and Representation of Polysemous Verb Senses: A Generative Activation Package Model

The results of the experiments presented in Chapters 3 and 4 provide robust empirical support for a graded, continuum-based model of polysemy, where the boundaries between senses are not cleanly segmented but are instead shaped by structured patterns of semantic proximity and partial overlap. Rather than discrete categories, the senses of polysemous verbs appear to be organized in continuous spaces where neighboring meanings bleed into one another depending on usage context and conceptual association.

These findings invite a fundamental reconsideration of how lexical meaning should be understood and represented. They encourage linguists to embrace a more fluid, usage-based foundation for semantic theory—one that can account for the nuanced and often overlapping nature of word senses as they are realized in actual linguistic contexts.

This chapter explores how best to incorporate the insights gleaned from experimental data into formal linguistic theories of word meaning. In doing so, it aims to bridge the gap between empirical findings on polysemous sense gradience and theoretical models that seek to represent those findings in structured ways.

Section 5.1 surveys major theoretical accounts of how the meanings of polysemous words—especially their underspecified or general meanings—are interpreted and represented. These approaches fall into three broad categories: core meaning approaches, thin semantic theories, and rich semantic models. As a representative of rich semantic modeling, Pustejovsky's Generative Lexicon (GL) theory (1995, 2006) offers a powerful framework for explaining the internal structure of lexical entries and the dynamic generation of senses in context. Section 5.2 investigates how GL accounts for sense selection, particularly in verbs like *break* that show overlapping or layered readings in natural usage. We will show that while GL successfully captures many cases of polysemy via its generative mechanisms and pragmatic enrichment, it also faces limitations in capturing (i) sense dominance or salience within overlapping readings and (ii) non-agentive, naturally occurring changes.

To address these limitations and better account for the empirical phenomena observed in the experimental data, Section 5.3 proposes a hybrid semantic model that integrates key insights from both GL theory and the more recent Activation Package Model (APM) (Ortega-Andrés and Vicente, 2019; Ortega-Andrés, 2021). This new model—termed the Generative Activation Package Model (GAPM)—offers a more flexible and empirically grounded framework for representing polysemous verb senses. It will be shown that this hybrid framework best accommodates the gradedness, sense overlap, and processing asymmetries revealed in human and language model behavior.

## 5.1. Underspecified Meaning and Theoretical Approaches to Polysemy

As discussed in previous chapters, verbs such as *break* and *freeze* take on a wide array of senses that interact with argument alternations and syntactic frames. A key question for lexical semantic theories is whether there is a unifying meaning underlying this diverse array of

senses—or, if not a single meaning, then perhaps a small set of primitive semantic dimensions that give rise to distinct sense clusters.

This position is advanced, for example, by Kellerman (1978:65), who argues that "[t]he various meanings of BREAK [...] can all be subsumed under a 'deep' meaning, '(cause) not to continue in existing state', which links even the most disparate meanings of BREAK" (see also Spalek (2012) for a similar argument regarding *romper* 'break' in Spanish). Another influential approach, especially in the tradition of lexical semantics developed by Jackendoff and also by Levin and Rappaport Hovav, posits that a few primitive semantic features combine to produce a structured combinatorial space of meanings. These dimensions are assumed to correlate with argument structure realization and other distributional properties.

Yet despite these efforts, the precise nature of the unifying meaning—if any—underlying the specific senses of a polysemous verb remains unclear. Nevertheless, one important point of consensus across many semantic theories is that lexical representations of core meaning are highly underspecified and subject to contextual modulation. This view, often called contextual modulation, is a central tenet shared across many otherwise divergent approaches to lexical meaning.

This idea can be traced back to Dowty (1976, 1979), who argued that aspectual analysis must at least include the entire verb phrase to accurately capture meaning shifts (Petersen and Potts 2023: 490). Borer (2005a, 2005b, 2013) takes this further, suggesting that open-class lexical items are "tantamount to raw material, 'stuff' which is poured into the structural mould to be assigned grammatical properties" (2005: 108). According to this perspective, lexical items are highly abstract and unvalued representations, fleshed out through the syntactic and conceptual environment in which they appear. Although we may identify a stable inventory of lexical items, each item carries almost unlimited potential to realize different meanings in different contexts.

A similar view is advocated by Pustejovsky's Generative Lexicon theory (1991, 1995), which models the lexicon as an open-ended system capable of generating new senses in novel contexts by applying structured procedures to base lexical entries (Pustejovsky 2006). This view is also consistent with Clark's (1997) critique of the so-called "Dogma of Sense Selection", which holds that speakers and listeners pick a fixed sense from an enumerable list. Instead, Clark argues that lexical meaning is fluid, context-sensitive, and constrained primarily by the communicative needs of discourse participants (see also Clark and Clark 1979; Searle 1980).

Across all these frameworks, we see the same general conclusion: lexical items are abstract, underspecified entities, whose meanings are realized dynamically and variably across contexts. Building on this idea, semantic theories can be further divided based on what kind of information is assumed to reside in the underspecified core. That is, different theories propose different types of minimal representations as the foundation for sense generation. In this regard, it is useful to distinguish between two major families of semantic theories:

- Thin semantic theories, which posit minimal, often schematic core content, and
- Rich semantic theories, which encode more structured bundles of conceptual and inferential information.

One of the most influential frameworks representing the thin semantic theory approach is the lexical semantic theory developed by Levin and Rappaport Hovav. Their theory posits that

verb meanings consist of two components: an event structure (or event schema) and a root (Rappaport Hovav and Levin 1998; Levin 2009; Levin and Rappaport Hovav 2019). The event structure is a schematic component shared by verbs belonging to the same semantic class, while the root encodes the idiosyncratic part of the verb's meaning. Together, these two components form a lexical semantic template that represents the full meaning of a verb. In other words, verb meanings are assumed to be bipartite, involving:

- one of a small set of event types or schemas (possibly defined by primitive predicates),
- and one of an open-ended set of roots representing the verb's core lexicalized meaning, understood as the meaning components that are entailed across all uses of the verb, regardless of contextual variation.

This framework can account for a wide range of verbal meaning variation, particularly when distinguishing between simple and complex event structures. For example, verbs like *jog*, *run*, or *creak,* which denote relatively simple activities, can be represented using a basic activity schema, as in (1):

(1) [ x ACT$_{<MANNER>}$ ]

In contrast, change-of-state verbs, which describe more complex events involving causation and result, are represented using complex event schemas, such as (2):

(2) [[ x ACT ] CAUSE [ BECOME [ y <*RES(ULT)-STATE*>] ] ]

Here are a few representative examples of this template applied to state-change verbs:

- *break*: [[ x ACT ] CAUSE [ BECOME [ y <*BROKEN*> ] ]]
- *freeze*: [[ x ACT ] CAUSE [ BECOME [ y <*FROZEN*> ] ]]
- *dry*: [[ x ACT ] CAUSE [ BECOME [ y <*DRY*> ] ]]
- *empty*: [[ x ACT ] CAUSE [ BECOME [ y <*EMPTY*> ] ]]

In each case, the root meaning (e.g., *MANNER*, *BROKEN*, *FROZEN*) is integrated into the event schema either as a modifier of the predicate (in simple schemas like (1)) or as the result state (in complex schemas like (2)). The expressions within angled brackets <...> indicate the content of the root, functioning as the lexicalized contribution to the overall event representation.

This bipartite structure provides a way of encoding verb meanings in a minimally specified format, which can then be expanded or interpreted in context. This bipartite structure provides a way of encoding verb meanings in a minimally specified format, which can then be expanded or interpreted in context. Moreover, as discussed in detail by Levin (2009), the notion of verb classes derived from these meaning components offers a powerful theoretical foundation for understanding the typological generalizations about verb syntax and argument alternations.

However, within such thin semantic theories, the specific senses of a verb are not explicitly represented in the assumed meaning structure, and no principled account is provided of how a specific sense is interpreted in context. Lexical semantic theories that aim to capture this

dimension, therefore, need to propose a richer account of meaning, one that can model the structured overlap and contextual modulation of senses rather than treating them as unanalyzed outputs of a sparse lexical entry.

This need for a more detailed account of verb meaning has motivated the development of rich semantic theories, which seek to explicitly represent the structured content of verb senses and their contextual variability. These theories go beyond schematic event templates and minimal roots to offer internally articulated representations of lexical meaning, enabling them to account for both the multiplicity of polysemous senses and their interaction with syntactic and pragmatic patterns. Among the most influential of such models is the Generative Lexicon theory (Pustejovsky 1995), which proposes a generative system for modeling the dynamic interaction between core lexical content and contextual interpretation. More recently, the Activation Package Model (Ortega-Andrés and Vicente 2019; Vincente 2019; Ortega-Andrés 2021) has offered an alternative representational framework grounded in cognitive and discourse-based considerations. In the following sections, we review these rich semantic models and argue that their integration—guided by the empirical findings of this study—offers a promising path toward a more comprehensive account of polysemous verb meaning.

## 5.2. Polysemous Sense Interaction in the Generative Lexicon: Explanatory Power and Limitations

Within the landscape of rich semantic theories, the Generative Lexicon (GL) framework (Pustejovsky 1995, 2006) offers a detailed and structured account of lexical meaning. Unlike thin semantic theories that rely on minimal event templates and unspecified roots, GL posits that lexical items carry internally articulated semantic content, structured across four distinct levels: argument structure**,** event structure**,** qualia structure, and inheritance structure. For the purposes of this chapter, we focus on the first three components, which are most directly involved in capturing polysemous sense interaction.

- Argument structure defines the number and type of arguments a predicate requires, and how they are syntactically realized.
- Event structure represents the internal temporal and causal composition of the eventuality denoted by the verb, including subevents and their relations (e.g., processes, transitions, or states).
- Qualia structure is perhaps the most distinctive contribution of GL: it encodes the intrinsic semantic roles that characterize how an object or event is understood and used. This structure serves as the interface between event structure and argument structure, enabling dynamic composition and the flexible interpretation of lexical items in context.

GL's qualia structure decomposes lexical meaning into four distinct semantic roles that collectively capture why**,** what, and how an object or event is what it is. These are:

- CONSTITUTIVE ROLE: specifies the internal constitution or material of an object (e.g., what something is made of, its parts or components).
- FORMAL ROLE: distinguishes the object within a larger domain, often used to define its essential properties or category.

- TELIC ROLE: describes the purpose or intended function of the object or event (e.g., what it is used for).
- AGENTIVE ROLE: encodes the origin or the process by which the object came into being (e.g., how it was created or caused).

These roles are directly inspired by Aristotle's four causes in his theory of modes of explanation (or generative factors):

- CONSTITUTIVE ↔ Material Cause (what something is made of),
- FORMAL ↔ Formal Cause (what makes a thing the kind of thing it is),
- TELIC ↔ Final Cause (the end or purpose for which a thing exists),
- AGENTIVE ↔ Efficient Cause (the source or agent that brings something into being).

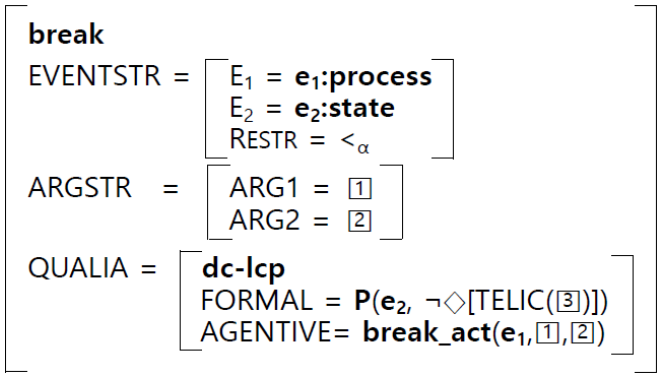
This philosophical grounding provides a rich framework for capturing both the structural and intentional dimensions of meaning within lexical items. In GL, these qualia roles serve as the informational core that interfaces with other lexical structures—such as argument structure and event structure—and also drive the application of generative mechanisms like co-composition to dynamically generate context-sensitive meanings.

A classic example is *bake*, which has a formal qualia role of "changing the state of x." When it combines with a noun like *potato,* the resulting phrase *bake a potato* is interpreted as "changing the state of the potato." However, when *bake* appears with a noun like *cake,* whose agentive qualia is "baking," the meaning shifts to "creating a cake." This interpretation arises through co-composition, a process by which the structures of the verb and the qualia structure of the noun interact to generate a new, context-specific sense. In this case, the noun's agentive role feeds into the compositional process, yielding a derived sense not directly encoded in the lexical entry for *bake* alone (Pustejovsky 1995: 123–125).

In what follows, we explore how GL's co-composition mechanism applies to the polysemous verb *break,* especially in sentences where multiple senses appear to be simultaneously activated or interactively composed. We then turn to the limitations of the GL approach in capturing such dynamic sense interactions.

To begin our discussion, let us consider the lexical representation for *break* shown in (3). In GL theory, the meaning of causative verbs is typically analyzed as involving an initial act or process ($e_1$) followed by a resulting state ($e_2$). These two subevents are reflected in the event structure, with $e_1$ corresponding to an agentive act and $e_2$ to a resulting state. This bifurcated event structure maps directly onto the agentive and formal qualia roles, respectively. In addition to specifying the subeventual composition, the event structure also leaves unspecified the headedness, that is, which subevent serves as the semantic "head" of the overall complex event (Pustejovsky 1995: 72). In the case of *break,* the structure in (3) is headless, i.e., it does not designate either subevent as the main semantic focus by default. This headless representation is what Pustejovsky refers to as the default causative paradigm(dc-lcp)—a simple type of causative relation that is highly productive across syntactic alternations (Pustejovsky 1995: 187).

(3)




Pustejovsky (1995) argues that it is precisely this semantic underspecification of headedness that underlies the causative alternation of verbs like *break*, *sink*, and *open*. That is, depending on contextual and syntactic factors, the event head may be interpreted as either $e_1$ or $e_2$, which in turn affects the mapping between semantic roles and syntactic arguments. This relationship between event prominence and argument realization is formalized by the following mapping principles (simplified here from Pustejovsky (1995: 191)):

(4) When $e_1$ is the head:
   a. $Q_i$: $R(e_1^*, x, y) \rightarrow$ x:SUBJ, y:OBJ
   b. $Q_j$: $P(e_2) \rightarrow$ shadowed

(5) When $e_2$ is the head:
   a. $Q_i$: $R(e_1, x, y) \rightarrow$ shadowed
   b. $Q_j$: $P(e_2^*) \rightarrow$ y:SUBJ

Thus, the same underlying semantic structure in (3) can give rise to either transitive causative structures (e.g., *He broke the window*) or intransitive noncausative ones (e.g., *The window broke*)—depending on which subevent is pragmatically or grammatically highlighted.

The qualia structure in (3) plays a crucial role in determining the interpretation of *break* in context. The formal role encodes the resulting state ($e_2$) of the theme argument (y), specifically representing the inability to fulfill its telic role (i.e., its intended function). For example, when a *window* is broken, it is no longer able to provide transparency or insulation—its functional purpose is impaired. The formal role thus encodes this functional dependency, where the meaning of the verb depends on the argument's inherent purpose (Pustejovsky 1995: 221–224).

Through the process of co-composition, the telic role of the theme argument is dynamically incorporated into the formal role of the verb's qualia structure. This interaction allows the verb to inherit or adapt aspects of the argument's internal semantics, thereby producing an interpretation specific to the nature of that argument. In the case of *break*, the physical sense of "causing something to become unusable or non-functional" arises precisely through the interaction between the verb's semantically underspecified representation and the argument's functional properties contributed via co-composition. In this way, semantic underspecification and co-composition together provide an effective explanation for how the canonical physical breaking sense of *break* is derived and contextually anchored.

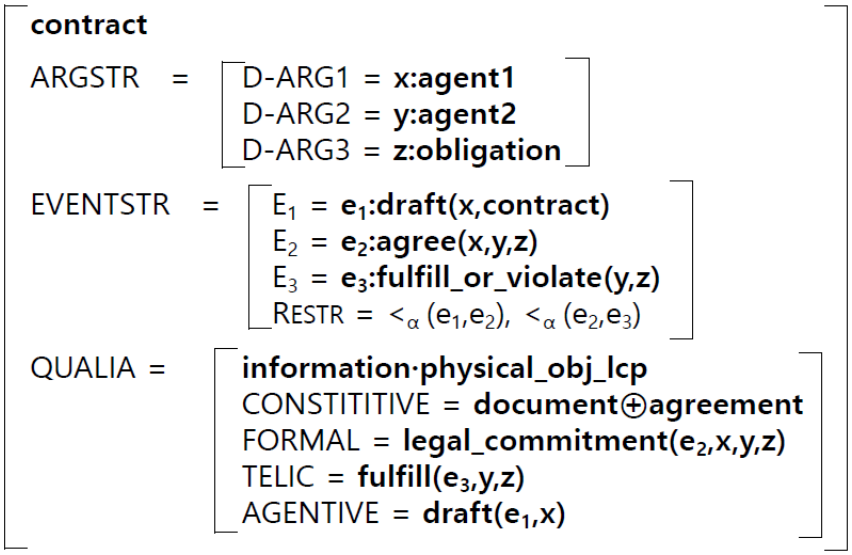
To further explore how the GL framework accounts for sense overlap, we now turn to cases where *break* applies to complex nominal objects, such as *contract*, in which multiple senses

are causally co-activated in context. Consider the example in (6), where the verb *break* simultaneously evokes the meanings of violating and terminating a contractual relationship:

(6) <Picasso broke his contract with Manach> and returned in January, 1902, to Barcelona.
(COCA MAG: USA Today Magazine-1997)

To interpret this sentence within the GL framework, we begin by examining the semantic structure of the noun *contract*, as illustrated in (7):

(7)




In this representation, *contract* is analyzed as a dot-type object (information·physical_obj_lcp), reflecting its dual ontological status as both a document and an agreement. This dual type allows the noun to flexibly participate in both concrete and abstract contexts, enabling the derivation of multiple senses depending on the verb it combines with.

Moreover, *contract* exhibits predicative force—it denotes an event that involves three entities (two agents and one obligation) and unfolds over multiple subevents. These include:

- $e_1$: the act of drafting the contract (typically by agent1),
- $e_2$: the agreement reached between agent$_1$ and agent2 over the obligation (z),
- $e_3$: the fulfillment or violation of that obligation, primarily involving agent2.

These subevents are causally linked via the RESTR relations in the event structure.

Notice that all three entities (x = agent$_1$, y = agent$_2$, z = obligation) are treated as default arguments in the argument structure, meaning they participate in the logical expressions of the qualia roles but are not necessarily overtly realized syntactically. This allows for a flexible interface between semantic and syntactic representation.

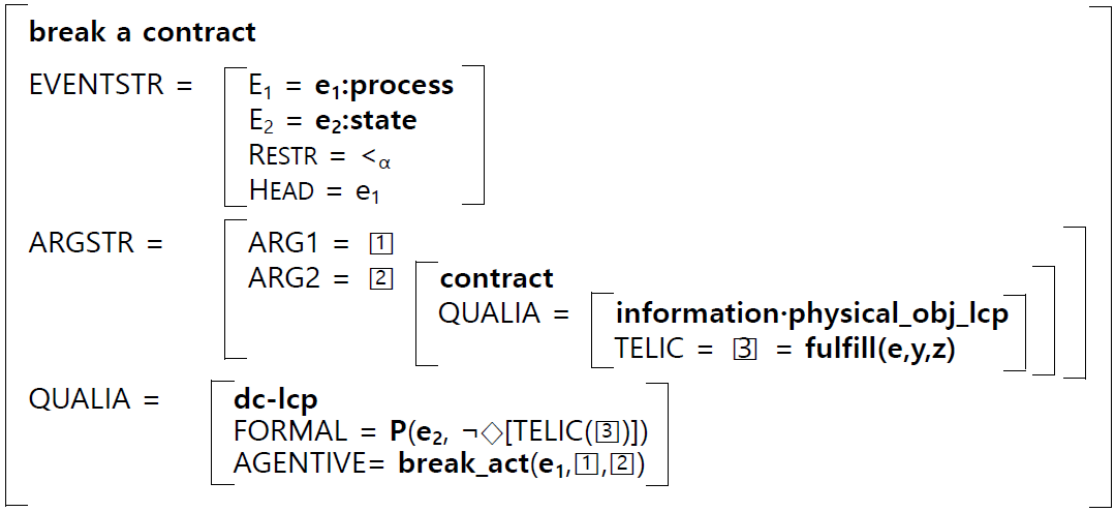
Finally, each qualia role contributes to the rich lexical semantics of *contract*:

- CONSTITUTIVE: document⊕agreement captures the physical and informational components;
- FORMAL: legal_commitment($e_2$,x,y,z) encodes the essential nature of the contract as a binding promise;

- TELIC: fulfill($e_3$,y,z) points to the intended purpose—fulfilling obligations;
- AGENTIVE: draft($e_1$,x) identifies how the object comes into being.

This rich structure enables GL's co-composition mechanism to derive the polysemous interpretation of *break* in (6). As illustrated in (8), the representation of *break a contract* inherits the qualia structure of the noun *contract*, specifically its telic role: `fulfill(e,y,z)`, where y is the agent responsible for fulfilling the obligation z. This role is semantically underspecified until contextual information and co-composition resolve its interpretation.

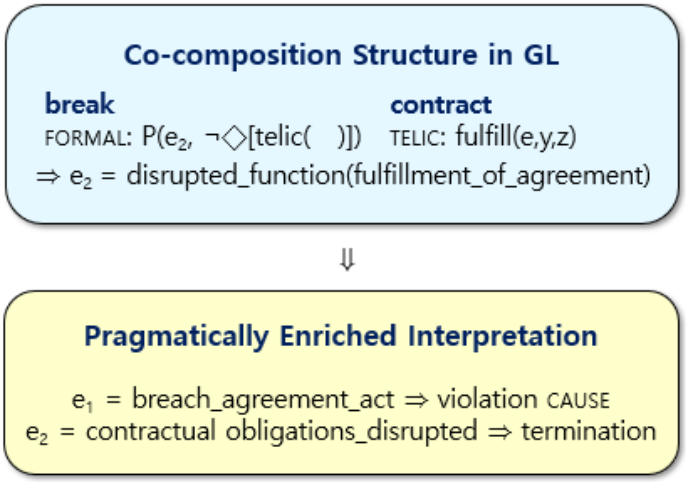
(8)




In this configuration, the formal role of the verb *break*—P($e_2$, ¬◊TELIC(3))—encodes the resulting state as the inability to fulfill the telic role of the object. Through co-composition, the telic role of *contract* (i.e., the fulfillment of obligations) is integrated into the formal role of *break*, yielding an interpretation where the broken state involves disrupted fulfillment.

This semantic configuration, together with contextual cues, enables the inference of a two-phase interpretation: $e_1$, as the agentive act of breaching the contract, and $e_2$, as the resultant state in which the contractual obligations can no longer be maintained. Importantly, such an analysis does not reflect a generative output of GL per se, but rather a pragmatically enriched interpretation grounded in the verb's formal qualia. The causality between $e_1$ and $e_2$ is thus not structurally specified within the lexicon but contextually inferred through discourse knowledge and pragmatic reasoning:

(9) From GL co-composition to pragmatic enrichment: Inference of sense overlap in *break a contract*




As illustrated in (9), the resulting state encoded in the formal qualia of *break*—"disrupted fulfillment of the agreement"—provides the semantic basis for inferring two contextually related subevents: $e_1$, `breach_agreement_act`($e_1$, 1), corresponding to the violation sense, and $e_2$, `obligations_disrupted`($e_2$, 2), corresponding to the termination sense. These are not compositionally generated within the lexicon but inferred through contextual reasoning that operates on the formal role's underspecified meaning and the discourse situation.

This analysis is consistent with Pustejovsky's own observation that "polysemy is not a single, monolithic phenomenon. Rather, it is the result of both compositional operations in the semantics, such as coercion and co-composition, and of contextual effects, such as the structure of rhetorical relations in discourse and pragmatic constraints on co-reference" (Pustejovsky 1995: 236). He further emphasizes that "what is necessary is for research to tackle the difficult question of how other components in the natural language interpretation process interact with the lexicon to disambiguate and fully determine the semantics of words in context."

Thus, while the qualia structure and GL's generative mechanisms provide the formal conditions for interpreting *break a contract*, the emergence of polysemous senses such as violation and termination in this context and the causal linkage between these senses ultimately depend on how lexical underspecification is resolved through inferential mechanisms external to the lexicon.

We now turn to cases where multiple senses are simultaneously activated within a single subevent, or where the resulting state itself is semantically layered, to examine how GL can represent such interpretations. These constructions present a more complex interaction than those with sequentially distinct subevents, as semantic multiplicity arises within a single event boundary. Consider example (10):

(10) A supermajority of 60 Senators can break a filibuster by invoking a cloture, the cessation on the bill, and forcing a vote.
(http://www.whitehouse.gov/our-government/legislative-branch)

Here, the verb *break* in *break a filibuster* exhibits semantic overlap within $e_1$ and $e_2$:

- In $e_1$, *break* encodes the agentive act of disrupting a resistance mechanism, functioning simultaneously as an interruption and a breach of resistance. The action represents not just a stoppage, but an assertive, rule-based termination of a procedural obstruction.
- In $e_2$, the focus shifts to the resulting state, where the filibuster is no longer in effect and the legislative process resumes. This phase activates senses akin to release**,** termination, and even institutional restoration, as the system regains its procedural flow.

Thus, the verb simultaneously evokes destructive and restorative connotations. A similar dynamic is seen in (11):

(11) The status quo can not continue. <We must break the gridlock> and move forward.
(COCA SPOK: PBS_Newshour-19930402)

In this case:

- $e_1$ conveys more than resolution—it encodes a forcible breakthrough, a metaphorical destruction of an entrenched blockage, aligning with figurative destruction and constructive rupture. This blends the coercive force of *break* with a vision of progress or transition**.**
- $e_2$ denotes the elimination of gridlock-induced paralysis, with the system entering a new, normalized phase. Thus, the result is both terminative and restitutive, echoing an institutional reset**.**

To investigate how the sense overlap in (10) arises from a dynamic interplay between GL's underspecified structures and discourse-pragmatic enrichment, we first propose the semantic representation of the noun *filibuster*, as summarized in (12). In this structure, *filibuster* is analyzed as a complex nominal with eventive and informational properties, encoded as a dot object type: `process·information_obj_lcp`. This allows the noun to flexibly participate in both propositional (informational) and procedural (eventive) contexts.

(12)

**filibuster**

ARGSTR =
- D-ARG1 = **x:agent**
- D-ARG2 = **y:proposition**
- D-ARG3 = **z:institutional procedure**

EVENTSTR = E = **$e_1$:prolong_or_block(x,y,z)**

QUALIA =
- **process·information_obj_lcp**
- CONSTITITIVE = **institutional_speech_act⊕parliamentary_procedure**
- FORMAL = **resistence_to_progress($e_1$)**
- TELIC = **delay_or_block_vote($e_1$,y,z)**
- AGENTIVE = **prolonged_speech($e_1$,x)**

The argument structure reflects three default arguments involved in the event:

- x: the speech agent (e.g., a Senator),
- y: the targeted legislative proposition (e.g., a bill),
- z: the parliamentary or institutional procedure being exploited or manipulated.

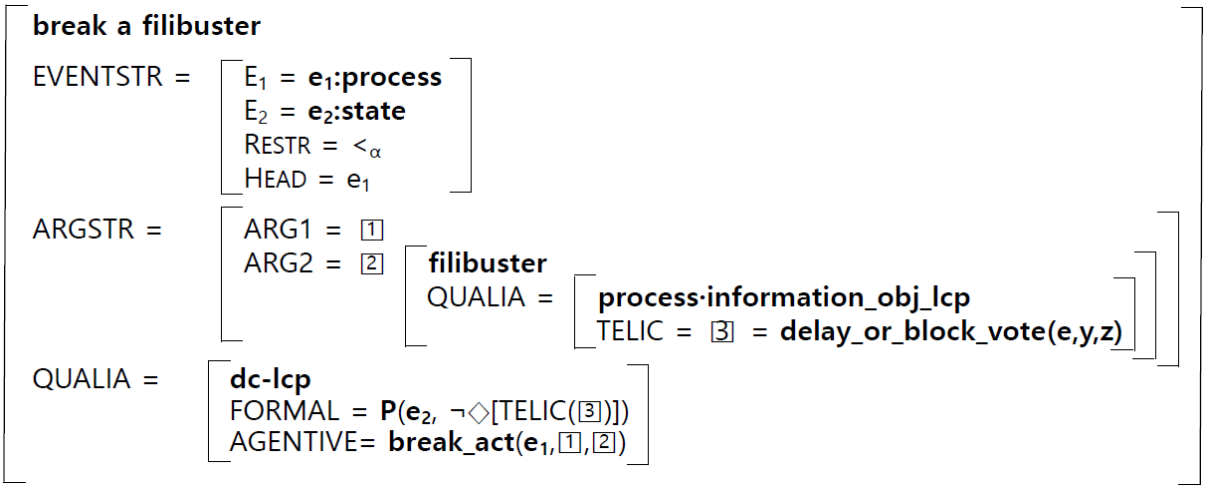
The event structure consists of a single event $e_1$, expressed as `prolong_or_block(x,y,z)`, which links the agent's speech act with the procedural effects on legislative progress.

Finally, each qualia role that contributes to the rich lexical semantics of *filibuster* is as follows:

- CONSTITUTIVE: The *filibuster* is constituted by a combination of institutional_speech_act and parliamentary_procedure. This captures its dual nature as both a speech event and a procedural tactic.
- FORMAL: Its essential nature is the resistance to legislative progress, structurally tied to $e_1$, the act of obstruction.
- TELIC: The intended function is to delay or block a vote on a given proposition (y) within an institutional setting (z).
- AGENTIVE: It is realized through prolonged speech acts, performed by an agent (x) who invokes this tactic to influence legislative outcomes.

The structure in (12), when considered alongside the event and qualia structure of *break* in (3), forms the foundation for semantic underspecification, allowing the verb to support flexible interpretations shaped by predicate–argument composition and discourse context. As previously discussed, *break* introduces a formal qualia that negates the telic role of its theme argument. Through co-composition, the telic role of *filibuster*—namely, delaying or blocking a vote—is projected into the verb's formal role. As illustrated in the representation of *break a filibuster* in (13), this results in a unified interpretation of the broken state $e_2$ as disruption of the filibuster's telic function:

(13)

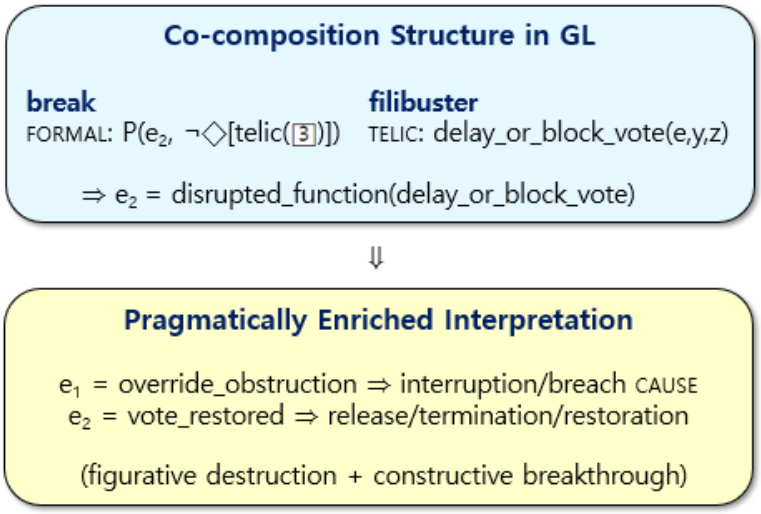


While this co-composition captures the core disrupted-function reading, it does not fully derive the rich, context-sensitive meanings found in (10). As diagrammed in (14), the contextually enriched interpretation pragmatically infers two distinct subevents:

- **$e_1$**: override_obstruction → interpreted as interruption or breach (of resistance)
- **$e_2$**: vote_restored → interpreted as release, termination, or procedural restoration

These two events reflect a figurative mapping of destruction (breaking the obstruction) and constructive breakthrough (restoring democratic procedure), revealing how one lexical item activates multiple senses simultaneously.

(14) From GL co-composition to pragmatic enrichment: Inference of sense overlap in *break a filibuster*




A similar pattern of sense overlap observed in (10) can also be identified in (11). In this case, the telic role of *gridlock* can be taken to denote institutional paralysis or procedural deadlock, while the verb *break* semantically encodes the disruption or removal of such an impeding function. The GL-based co-composition process yields a formal representation in which the resultant state ($e_2$) is interpreted as the functional disruption of the gridlock's telic purpose—namely, its ability to block institutional progress. However, the lexical representation alone does not specify the internal structure of this interpretation. It is through discourse-pragmatic enrichment that a more nuanced interpretation arises: $e_1$ conveys not only resolution but also a figurative breakthrough, connoting a constructive rupture or the forceful dismantling of entrenched obstruction. This blends the coercive destruction sense with a forward-looking, constructive transition. Meanwhile, $e_2$ is interpreted as the resumption of systemic functionality, involving both termination of the impediment and restitution of normal institutional operation. As in the *break a contract* and *break a filibuster* examples, this instance illustrates how underspecified lexical structures in GL interact dynamically with contextual cues to yield a multilayered, polysemous interpretation.

While the GL provides a robust framework for representing underspecified semantic structures and for modeling co-compositional interactions between verbs and nominal arguments, the actual prominence of senses in polysemous expressions—i.e., which subevent

or interpretation is foregrounded in discourse—often exceeds what can be formally encoded within GL representations. To illustrate this, consider the VPs *break a filibuster* (10) and *break the gridlock* (11), both of which involve a dual-event interpretation arising from a co-compositional structure in GL: an agentive event ($e_1$) of breaking and a resultant state ($e_2$) in which the disrupted object no longer fulfills its telic role. As discussed above, these forms yield enriched interpretations where the act of "breaking" blends figurative destruction with constructive breakthrough.

Despite the structural symmetry in their GL event templates, the interpretive emphasis diverges in each case:

- In (10), the discourse focus lies on $e_2$, the termination and resolution of procedural obstruction. The telic function of *filibuster*—to delay or block a vote—is pragmatically inferred to be neutralized, leading to institutional progress (e.g., vote restored).
- In contrast, (11) centers more strongly on $e_1$, the forceful act of disruption, where *break* metaphorically evokes destruction of the entrenched "gridlock." The interpretation foregrounds urgency and agency, with $e_2$ (system reset) emerging as a secondary consequence.

This asymmetry reflects a discourse-driven interpretive hierarchy, one not lexically specified but constructed in context. Such variation in interpretive weight highlights a fundamental limitation of GL: while the framework captures what kinds of meanings are compositionally licensed, it does not fully account for which meanings are *foregrounded* or *profiled* in context.

The Generative Lexicon (GL) framework faces an even more significant limitation when it comes to capturing non-agentive, naturally occurring changes, as illustrated in the following examples:

(15) a. I rid myself of fear. Between my ninth and tenth flying lessons, <the fever broke>. I don' t know how or why.

(COCA MAG: Popular Mechanics-2015)

b. <The storm broke> and I was able to escape to the other side of the river and meet up with Foxy and the crew.

(COCA NEWS: Minneapolis Star Tribune-20180804)

(16) The day broke over the quiet village.

In (15a–b), *break* expresses a process of attenuation or resolution—the fever or storm subsides or loses its intensity—while in (16), *break* conveys a dual sense of natural emergence/revelation and temporal shift/termination: the onset of daylight marking both the appearance of light and the end of darkness. These two semantic components are not sequentially distinct subevents, but rather co-activated within the same resultative dimension ($e_2$), describing the transformation from one state to another as both emergence and termination.

Crucially, such $e_2$-level sense layering cannot be derived through GL's internal generative mechanisms. In the absence of a functional telos in the internal argument (*fever*, *storm*, *day*), there is no telic role available to license co-compositional inference. As a result, the semantic relation between *break* and its argument cannot be compositionally built within the qualia structure. Instead, the interpretation must be constructed through contextual and conceptual

inference based on general world knowledge about natural processes and perceptual change. As Pustejovsky (1995) himself notes, certain entities are characterized by only three features or roles in the qualia structure, lacking a telic role altogether. This omission prevents the formal mechanism of co-composition from operating, forcing interpretation to rely on extra-lexical enrichment—the reader's inference that the event simultaneously involves the emergence of light and the termination of night. In other words, while GL can represent the abstract change-of-state structure of *break*, it cannot generate the multi-aspectual, experiential layering of meanings that arise in natural phenomena.

In this respect, the examples in (15)–(16) mark a boundary of GL's explanatory scope. The framework successfully models functionally motivated causation in events like *break a contract* or *break a filibuster*, where the argument's telic role provides the semantic input for compositional integration. However, when the argument denotes a natural kind concept without functional structure, the model cannot capture how the lexicon, context and cognition jointly construct emergent and transitional meanings.

In sum, this section has demonstrated how *break*'s polysemy can be analyzed within the Generative Lexicon (GL) framework using semantic underspecification, co-composition, and pragmatic enrichment. We showed that contextually enriched interpretations—such as the dual reading of *break a contract* (violation + termination) and *break the filibuster/gridlock* (coercive rupture + systemic restoration)—can be derived when the internal argument contributes a telic role. However, two key limitations emerged. First, GL does not account for sense dominance or salience within overlapping readings; all interpretations remain formally equal. Second, GL's mechanisms fail when the internal argument lacks a telic role, as with natural kind nouns (e.g., *fever*, *storm*, *day*), rendering co-composition inapplicable.

These limitations point to the need for a more general framework for modeling polysemy, such as the Generative + Activation Package Model (GAPM), which integrates compositional structures with context-sensitive activation and interpretive prioritization grounded in discourse and cognition.

### 5.3. Toward a Hybrid Model: The Generative Activation Package Model (GAPM)

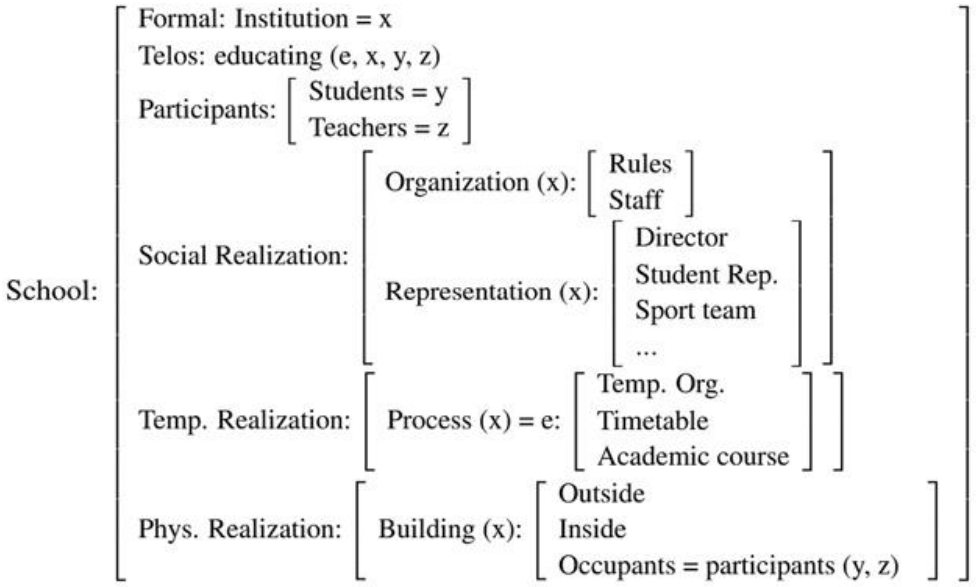
The Activation Package Model (APM), developed by Ortega-Andrés and Vicente (2019) and Ortega-Andrés (2021), presents a cognitively grounded and context-sensitive alternative to static sense enumeration models of polysemy. Rather than representing lexical items as bundles of discrete and fixed senses, APM conceptualizes a word's meaning as a rich, structured network of knowledge that becomes selectively activated in context. This approach enables a dynamic and flexible interpretation mechanism, allowing multiple meaning components (or aspects) to be contextually mobilized depending on discourse cues, conceptual salience, and pragmatic needs. Crucially, the information represented in an activation package is not confined to lexical content, in contrast to more traditional frameworks such as the Generative Lexicon (GL). Instead, it encompasses broader forms of encyclopedic, social, and experiential knowledge, making it less restricted in both its representational format and its inferential potential.

A central concept in APM is that of realization, sometimes referred to as aspects. Realizations capture the different ways in which a concept can be instantiated or foregrounded in context. Each realization reflects a distinct perspective on the same conceptual core, and they may align with different types of polysemous senses depending on which aspects are

activated. The structure in (17) presents a detailed activation package for the noun *school*. This representation illustrates how the concept is decomposed into various realization dimensions, including:

- Formal/Telos: the institution's essential purpose—educating (e.g., *The school educates children.*)
- Participants: agents involved, such as students and teachers (e.g., *The school protested the decision.*)
- Social Realization: institutional structures and roles (e.g., *The school appointed a new director.*)
- Temporal Realization: activities and scheduling (e.g., *The school begins at 9 a.m.*)
- Physical Realization: the physical building and its occupants (e.g., *The school caught fire.*)

(17) Knowledge structure of *school* (Ortega-Andrés and Vincente 2019: 5)




These realizations are not mutually exclusive; they coexist within a shared conceptual structure and can be co-activated in discourse. For instance, in (18), both the physical realization (building) and the participant realization (people) are simultaneously activated. This dual activation enables a nuanced and coherent interpretation without requiring discrete sense switching or coercion.

(18) The school [building] caught fire and [people] was celebrating 4$^{th}$ of July when the fire started. (Ortega-Andrés 2021: 128)

In the APM framework, realizations are not themselves word senses, but structured subcomponents or aspects of a broader concept. Specific senses emerge as contextually shaped instantiations that result from the differential activation of these realizations. The model thus rejects the view of polysemy as a finite list of discrete senses, replacing it with a multi-aspectual

representation from which senses are constructed dynamically in discourse. This view explains how polysemous expressions often resist sharp sense boundaries and exhibit hybrid or overlapping interpretations. It also accounts for the co-presence of multiple senses, as observed in (18), and supports seamless sense transitions in discourse.

While APM provides a compelling account of conceptually grounded, context-sensitive polysemy—especially for nouns like *school*—it also presents several limitations when considered as a general framework for modeling polysemy across syntactic categories. Most notably, the model has yet to be applied to verbs, and no published case studies currently demonstrate its capacity to explain verb polysemy and its interaction with argument alternations, such as those observed in causative alternation. This makes APM less equipped than the GL to account for how different verb senses correlate with systematic patterns of syntactic realization and semantic roles.

Moreover, the rich realization structure proposed in APM—while valuable for modeling noun-based concepts—assumes that all realization dimensions are activated simultaneously upon lexical encounter. This assumption, illustrated in the full realization structure of *school* in (17), lacks strong empirical grounding. Although the model is supported by experimental evidence showing that specific senses are stored in long-term memory and can be retrieved contextually (Ortega-Andrés 2021:126–129), this does not entail that all realizations are active at once, nor that such wholesale activation is cognitively economical or consistent with processing models.

This assumption of full activation introduces theoretical challenges for modeling pragmatic economy, processing cost, and default salience, particularly when multiple realizations are not contextually relevant. For instance, in cases where only the physical building sense of *school* is invoked (e.g., *The school caught fire*), it is unclear what functional role the social or temporal realization structures play during interpretation.

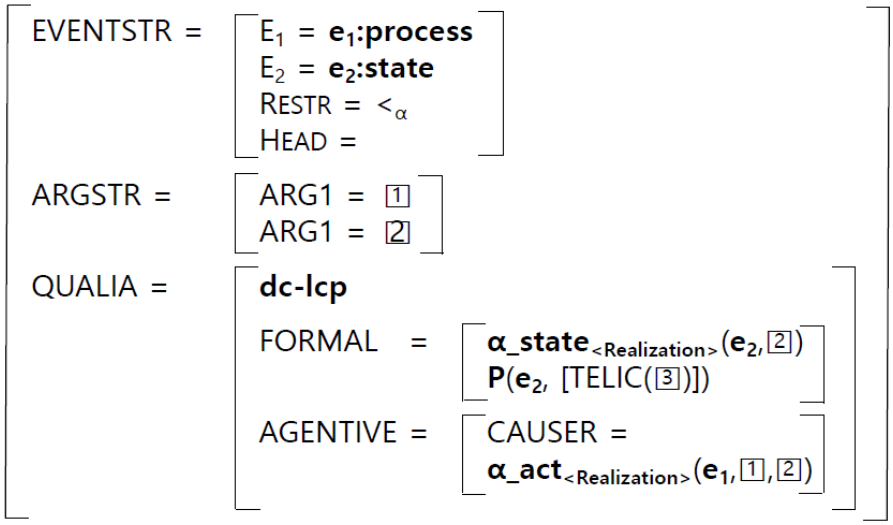
Furthermore, APM currently lacks a formalized mechanism for representing semantic co-composition, which are crucial for explaining sense construction in verbs that undergo systematic alternations or metaphorical extensions. As a result, APM's strength in representing encyclopedic richness comes at the cost of reduced formal compositional transparency, particularly in cases where sense disambiguation is driven by predicate-argument interaction and contextual constraints.

These limitations suggest the need for a hybrid framework that retains APM's flexible, knowledge-rich representations but integrates the generative mechanisms of sense derivation and the mapping principles characteristic of GL. In what follows, we propose such a model: the Generative Activation Package Model (GAPM). This framework models lexical meaning as a dynamically activated set of semantic features, which interact with both co-compositional constraints and discourse-level cues. Unlike APM, which emphasizes the activation of multiple realization dimensions (e.g., physical, temporal, social/institutional), GAPM explicitly incorporates event structural templates and argument realization to better handle verb polysemy, sense overlap, and alternation. Thus, GAPM aims to bridge the strengths of both GL and APM while addressing their respective limitations, offering a more cognitively plausible and computationally tractable account of context-driven sense selection.

To account for the sense variability of change-of-state (COS) verbs that participate in causative alternation, we begin by proposing an extended lexical representation structure, as illustrated in (19), which applies across a wide class of COS verbs (e.g., *break, freeze, open,*

*sink*). This structure builds on the three-tiered GL representation—EVENTSTR**,** ARGSTR, and QUALIA—but integrates two critical innovations: realization types and causer salience.

(19)




What distinguishes this structure from standard GL representations is the introduction of Realization values as part of both the FORMAL and AGENTIVE roles in the qualia structure. Realizations (or *aspects*, in the sense of APM) define the modalities or conceptual dimensions through which an event is instantiated, interpreted, and semantically elaborated. For COS verbs, such realizations include physical, temporal, social/institutional, and emotional/psychological aspects. These realization types are not simply realizers of abstract objects (as in institution in APM), but rather serve as modulators that shape the interpretation of a change event depending on the affected entity and the discourse context. Thus, in (19), the formal quale (state realization) may be instantiated as a temporal shift (*e.g., the day broke*), physical resolution (*e.g., the fever broke*), or structural rupture (*e.g., the window broke*), depending on the conceptual profile of the affected argument and the contextually salient realization.

The second key innovation concerns the integration of causer salience into the AGENTIVE quale and its role in determining event headedness. Traditional GL analyses of causative alternation have offered limited accounts of why certain verbs show preferences for transitive (causative) or intransitive (noncausative) uses in context. GAPM builds on recent work (see Chapter 2) by operationalizing causer salience—based on intentionality, identifiability, and contextual relevance—as a non-lexical but structurally relevant factor.

This salience influences the selection of event head and hence the argument structure mapping, as revised below:

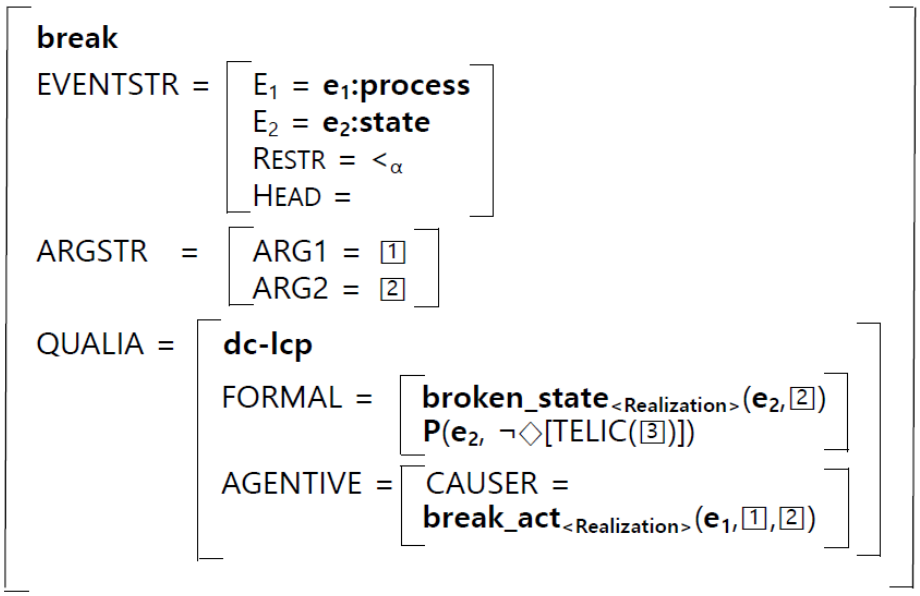
(20) High causer salience → left-headed event ($e_1$ is the head):
  a. $Q_i$: *R*($e_1$*, x, y) → x:SUBJ, y:OBJ
  b. $Q_j$: *P*($e_2$) → shadowed

(21) Low causer salience → right-headed event ($e_2$ is the head):
  a. $Q_i$: *R*($e_1$, x, y) → shadowed
  b. $Q_j$: *P*($e_2$*) → y:SUBJ

These mapping principles capture a core generalization in causative alternation**:** causative uses are associated with intentional and explicitly identified causers, while noncausative uses align with nonintentional, less clearly identifiable causers (Rappaport Hovav 2014, 2020; Kim et al. 2025; Lee 2023, 2025).

Let us now apply this enriched model to *break* and demonstrate how GAPM accounts for both its sense variability and alternation patterns in context. The lexical semantic representation of *break* is shown in (22), which draws on the basic event-qualia-argument structure of GL while extending it with two key innovations: realization types and causer type features.

(22)




The qualia structure in (22) includes:

- A FORMAL role referring to the resulting state, parameterized as `broken_state`$_{\text{<Realization>}}(e_2, \boxed{2})$, where the realization value remains underspecified.
- An AGENTIVE role that introduces the causer (either explicit or implicit) through `break_act`$_{\text{<Realization>}}(e_1, \boxed{1}, \boxed{2})$, capturing the variability in causative structure.

This representation enables a flexible activation of realization types and event construals based on the contextual salience of the causer and the discourse cues. Unlike the realizations proposed in APM for nominal concepts like *school*, the realizations here apply to verbs of change-of-state and are tied to event instantiation rather than object structure. Table 1 summarizes seven empirically attested realization types of *break*, each corresponding to a distinct event construal:

Table 1. Realization types of *break*

| Realization type | Event construal |
|---|---|
| Physical realization | A physical object undergoes rupture or disintegration. |
| Emotional/psychological realization | A psychological state (e.g., tension, focus) is disrupted or collapses. |
| Social/institutional realization | A norm, rule, or collective behavior is violated or dissolved. |
| Temporal realization | The flow or continuity of time or a process is interrupted or terminated. |
| Constructive-navigation realization | An obstacle is surpassed, a constraint breached —yielding forward movement (e.g., "break the impasse"). |
| Revelation realization | Something previously hidden (e.g., information, light, truth) becomes manifest. |
| Attenuation/Resolution realization | Pressure, tension, or conflict dissipates or resolves, often naturally. |

In the Generative Activation Package Model (GAPM), lexical meaning is constructed and refined through two distinct but interdependent interpretive phases:

1. **Phase 1: Co-composition and activation**
   At this stage, the compositional mechanisms of the grammar initiate the activation of relevant semantic features and realization types, based on the lexical structure of the verb and the conceptual nature of its arguments. Underspecified templates (e.g., event headedness and realization values) begin to instantiate, guided by constraints from the verb's qualia roles, causer salience, and event structure.
2. **Phase 2: Pragmatic enrichment and sense selection**
   The activated structures are subsequently refined through pragmatic inference, which incorporates discourse cues, world knowledge, and stereotypical expectations to resolve underspecification and select the most contextually appropriate interpretation. This stage explains how non-literal, metaphorical, or blended readings arise, particularly in verbs with high polysemy like *break*.

Let us illustrate this interpretive architecture through the noncausative use of *break* in (23):

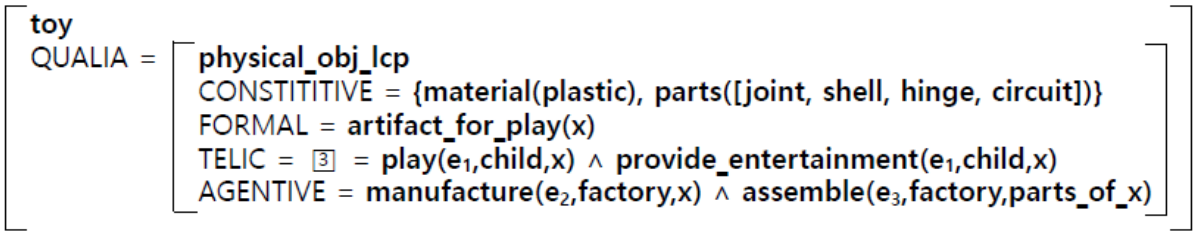
(23) You bought a plastic toy at Christmas from Japan, and <it broke the next day>.
(COCA NEWS: Denver Post-20000227)

This example represents a typical non-agentive change-of-state event, where the subject (*the toy*) undergoes a physical rupture without an identifiable external causer. GAPM explains this interpretation through two interdependent phases: 1. co-composition and realization activation, and 2. pragmatic enrichment and sense selection**.**

**Phase 1: Co-composition and activation**

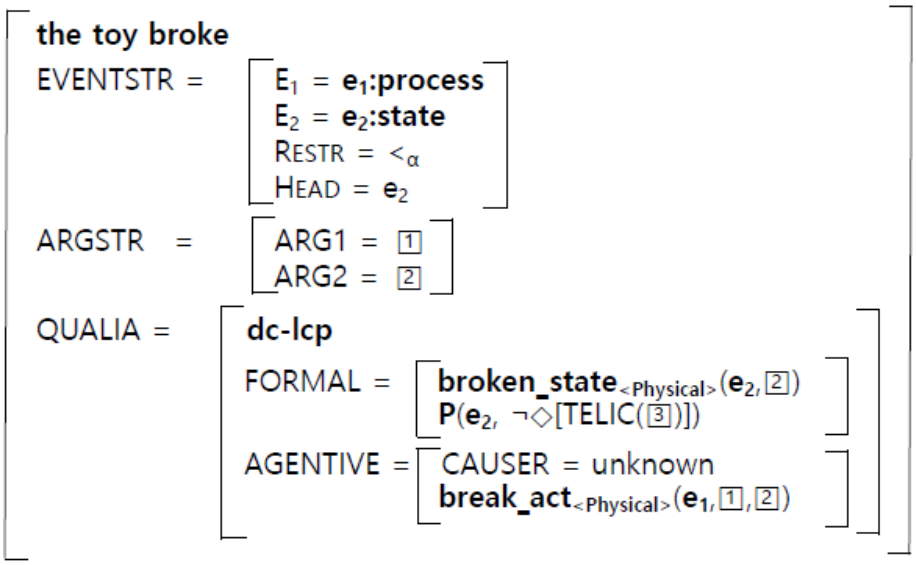
In (23), the subject *it* refers to a toy, whose qualia structure is given in (24):

(24)




The verb *break* is associated with the underspecified representation shown in (25): 수정

(25)




During co-composition, the verb's formal role introduces a predication that negates the telic role of its theme argument**.** This mechanism captures the core semantic relation between the verb and its argument—namely, that *breaking* results in the loss or impossibility of performing the function encoded in the argument's telic role**.**

When *toy* combines with *break*, the following semantic unification occurs:

1. The telic role of *toy* (play / provide entertainment) is imported into *break*'s formal role.
2. Through co-composition, *break*'s FORMAL predication (¬◊TELIC) applies to *toy*'s telic value, yielding the interpretation:
   "the toy entered a state in which its intended function (play) can no longer be realized."
3. Simultaneously, *toy*'s KIND = `physical_obj` and CONSTITUTIVE = `material(plastic)` instantiate the <Physical> realization for both subevents: `broken_state`$_{<Physical>}$($e_2$,2) and `break_act`$_{<Physical>}$($e_1$,1,2)
4. Because the causer is *unidentifiable* in context, the event defaults to a right-headed structure following mapping principle (21).:
   - $e_1$ (process) → *shadowed*
   - $e_2$ (state) → *head*
   - *toy* → subject of $e_2$

This configuration foregrounds the resulting state ($e_2$) as the central predicate of the clause and licenses the noncausative realization of *break*.

**Phase 2: Pragmatic enrichment and sense selection**

The co-compositional structure established above encodes a general result state—"the toy can no longer fulfill its intended function." Pragmatic reasoning then enriches this abstract functional failure into a more specific physical rupture interpretation, guided by the following cues:

1. Constitutive information from *toy* (plastic material, movable parts) makes a *physical* cause of malfunction the most plausible instantiation.
2. World knowledge about the fragility of plastic toys supports the inference that the toy's *physical integrity* failed rather than its design or entertainment function.
3. Discourse context (*it broke the next day*) implies a spontaneous, non-agentive event consistent with wear, impact, *or* material weakness.

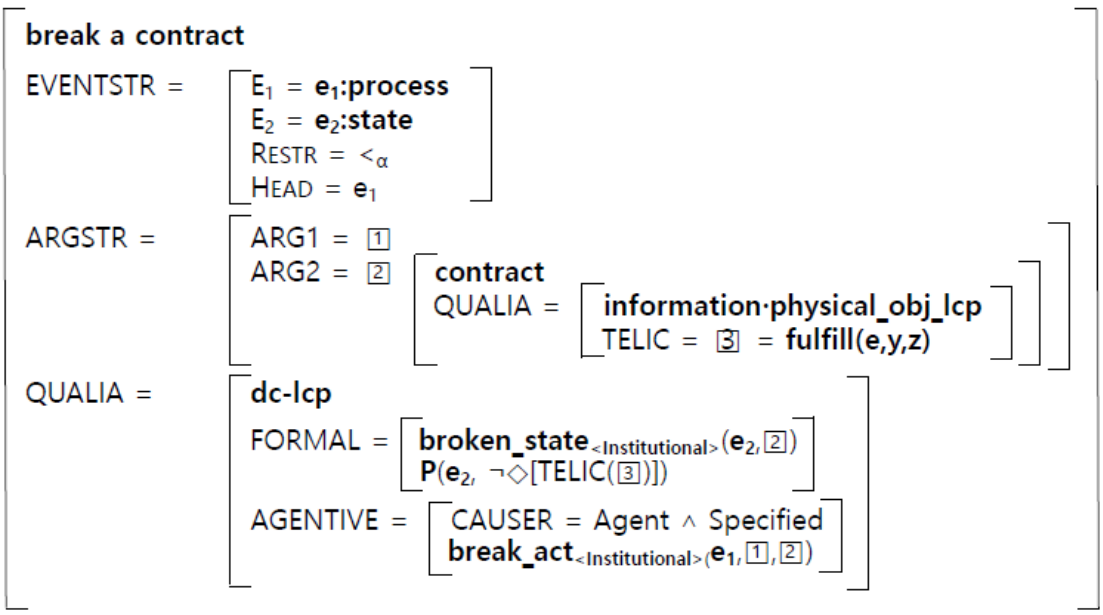
Consequently, the underspecified predicates in (25) are pragmatically enriched as follows:

- `break_act`$_{\text{<Physical>}}$ → an unobserved physical stress or force
- `broken_state`$_{\text{<Physical>}}$ → a structurally damaged state in which the toy is no longer playable

Thus, the sentence is interpreted as describing an inchoative event in which the toy's functional telos is rendered unrealizable through unknown physical rupture.

We now turn to another instantiation of *break*, this time in the causative construction *break a contract*, to illustrate how the GAPM accounts for sense layering and interpretation in context. As shown in (26), the enriched qualia structure of *break a contract* exhibits both formal and agentive roles that engage the social/institutional realization type:

(26)



**Phase 1: Co-composition and activation**

In the first phase of interpretation, co-composition occurs between the enriched qualia structures of the verb *break* and its object *contract*. As shown in (26), *break* introduces a formal qualia of `broken_state`$_{\texttt{<Institutional>}}$ that presupposes the failure or negation of the telic role of its theme argument. Meanwhile, the telic role of *contract* (as defined in (7)) is `fulfill(e, x,y,z)`, representing the successful completion of an agreement between parties. The semantic incompatibility between these two roles leads to a co-compositional clash: the verb's `broken_state` denies the realization of the noun's fulfill function.

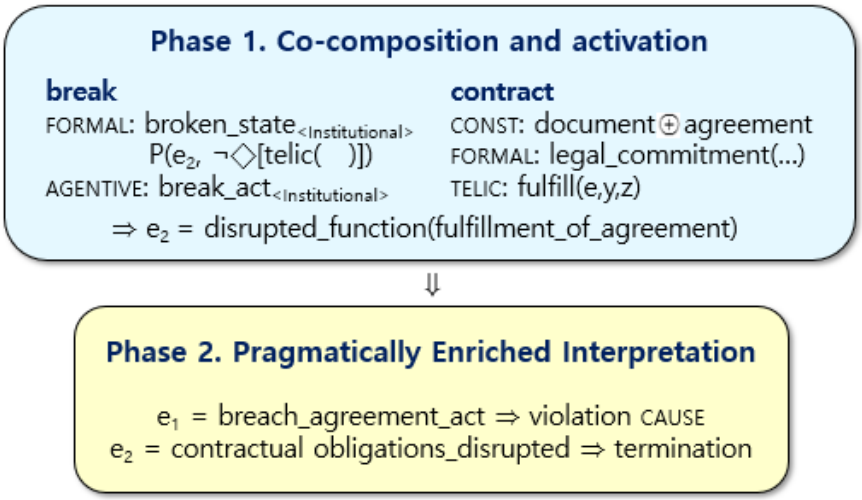
This clash activates <`Institutional`> as the realization type of both the causing event ($e_1$) and the resulting state ($e_2$). The activation is reinforced by the broader integration of *contract*'s constitutive and formal roles. The noun's constitutive structure (e.g., `document` $\oplus$ `agreement`) encodes an institutional artifact designed to bind agents within legal and social norms. Its formal role further specifies its identity as a `legal_commitment(e,x,y,z)`. These structural properties collectively establish *contract* as an institutionally grounded artifact, whose very existence presupposes the framework of fulfillment, obligation, and compliance. When this structure is inserted into the second argument position of *break*, all of these facets—including the KIND (`information:physical_obj lcp`)—contribute to the activation of institutional realization as the semantic domain in which the event is instantiated and interpreted.

Moreover, in the context of *break a contract*, the causer (typically the subject NP) is both agentive and explicitly identifiable, which corresponds to high causer salience. According to mapping principle (20), this results in a left-headed event structure in which the causing subevent ($e_1$) is foregrounded and the resulting state ($e_2$) is backgrounded. The structure thus realizes the causative alternant of *break*, with emphasis on the agent's violation act.

**Phase 2: Pragmatic enrichment and sense selection**

In Phase 2 of interpretation, the co-composed structure undergoes pragmatic enrichment informed by world knowledge and inferential reasoning, outlined in (27). The causing act ($e_1$) is interpreted as a breach of agreement, a conventionalized institutional act that presupposes the existence of an agreement and the capacity for it to be violated. This breach, in turn, causally entails the disrupted fulfillment of the contractual obligations ($e_2$), which is pragmatically extended to denote the termination or nullification of the contract. The resulting interpretation combines two linked senses—violation and termination—that are often conceptually interdependent in legal and social contexts, but grammatically distinct in the alternation patterns they license.

(27) Two-phase analysis of sense derivation in *break a contract* under GAPM




This analysis demonstrates how GAPM captures both the distributed semantic representation of *break* and its context-dependent sense selection and overlap, offering a unified account of causative alternation, sense activation, and compositional meaning construction.

We now turn to cases where semantic multiplicity arises within a single subevent. The GAPM captures this phenomenon through the co-activation of realization types, especially when different dimensions of the argument's qualia structure interact with the verb's own realization conditions.To illustrate this, consider again the example (10), repeated here as (28):

(28) A supermajority of 60 Senators can break a filibuster by invoking a cloture, the cessation on the bill, and forcing a vote.

**Phase 1: Co-composition and activation**

The expression *break a filibuster* activates a complex interpretation through the compositional interaction of the verb *break* and its object *filibuster*. As represented in the semantic representation in (12) above, the noun *filibuster* possesses a qualia structure characterized by both eventive/processual and informational properties. These include a formal role denoting resistance to progress, a telic role encoding the function of delaying or blocking a vote, and an agentive role associated with *prolonged speech*.

In this context, the qualia structure of *filibuster* enters the argument structure of *break*, and its kind, telic and formal roles are co-composed with *break*'s formal qualia—which encodes a `broken_state` realized through the negation of the object's telic function. As represented in the semantic structure of *break a filibuster* in (29), these properties of *filibuster* projected into the verb's formal role directly activate the temporal + constructive realization of *break* as an act of disrupting an ongoing obstacle/resistance.

Simultaneously, the formal and telic properties of *filibuster* (i.e., obstruction of legislative progress and delay of voting procedures) instantiate a constructive-navigational realization: that is, an act of overcoming procedural hindrance. These activate <`Constructive`> as the realization type of both the causing event ($e_1$) and the resulting state ($e_2$).

(29)

**break a filibuster**

EVENTSTR = [ $E_1$ = **$e_1$:process**; $E_2$ = **$e_2$:state**; RESTR = $<_\alpha$; HEAD = $e_1$ ]

ARGSTR = [ ARG1 = [1]; ARG2 = [2]; **filibuster** QUALIA = [ **process·information_obj_lcp**; TELIC = [3] = **delay_or_block_vote(e,y,z)** ] ]

QUALIA = [ **dc-lcp**; FORMAL = [ **broken_state**$_{\text{<Temporal + Constructive>}}$**(e**$_2$**,** [2]**)**; **P(e**$_2$**,** ¬◇[TELIC([3])]) ]; AGENTIVE= [ CAUSER = Agent ∧ Specified; **break_act**$_{\text{<Temporal + Constructive>}}$**(e**$_1$**,** [1]**,** [2]**)** ] ]

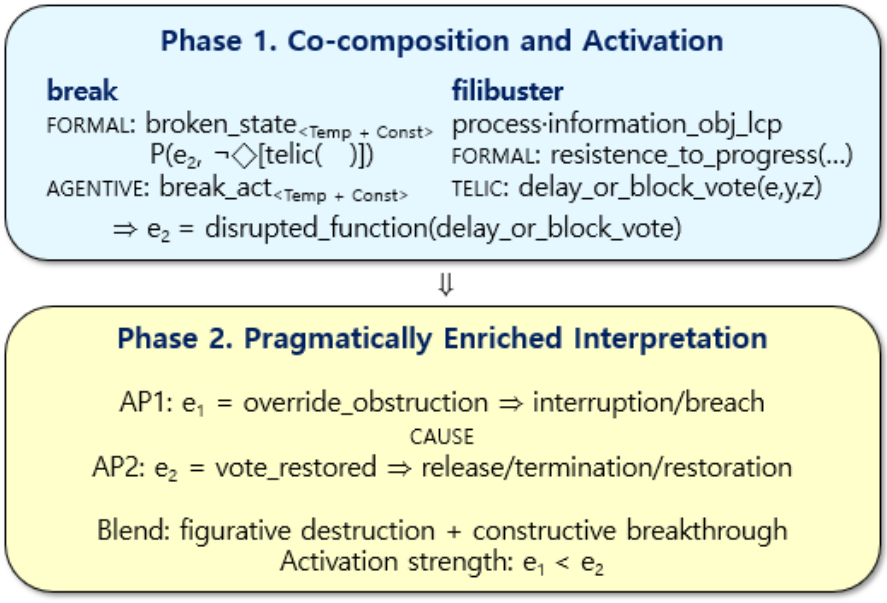
Thus, two realization packages become co-activated within the same subevent structure. Event-headedness and causative realization follow the same pattern as previously illustrated with *break a contract*.

**Phase 2: Pragmatic enrichment and sense selection**

Given this dual activation, pragmatic reasoning yields two activated packages (APs), as shown in (30):

- **AP1**: *e*₁ is interpreted as an `override_obstruction` event, leading to *interruption* or *breach* of parliamentary delay.
- **AP2**: *e*₂ is interpreted as a `vote_restored` outcome, signaling the termination of the filibuster and resumption of procedural flow.

(30) Two-phase analysis of sense derivation in *break a filibuster* under GAPM



This phase draws on both:

- The compositional constraints introduced in Phase 1 (i.e., the disrupted telic function and activated dual realizations).
- Contextual cues from (28), including expressions like *invoking a cloture* and *forcing a vote*, which strongly activate the constructive endpoint interpretation.

As shown in (30), the final interpretation in context is thus a weighted blend of the two activated packages:

- Figurative destruction of procedural obstruction ($e_1$).
- Constructive breakthrough restoring legislative function ($e_2$).

Crucially, this blend is non-symmetric: $e_2$ is pragmatically more salient than $e_1$, leading to graded activation where the constructive outcome dominates but coexists with its causal precursor.

This dual reading observed in example (11), repeated here as (31), can be explained in the same way. In this case, the coexistence of the "coercive rupture" and "systematic restoration" readings results from the co-activation of multiple realization packages within a single subevent. The relative sense dominance within these overlapping readings—here, the former (*destruction/breakthrough*) being more prominent—can be captured in terms of differences in activation strength among the simultaneously activated packages.

(31) The status quo can not continue. <We must break the gridlock> and move forward.

Finally, we revisit the challenging case in (16), repeated below as (32), to show how GAPM accounts for its interpretation through co-activation of realizations—a mechanism that extends beyond what the GL framework alone can explain.

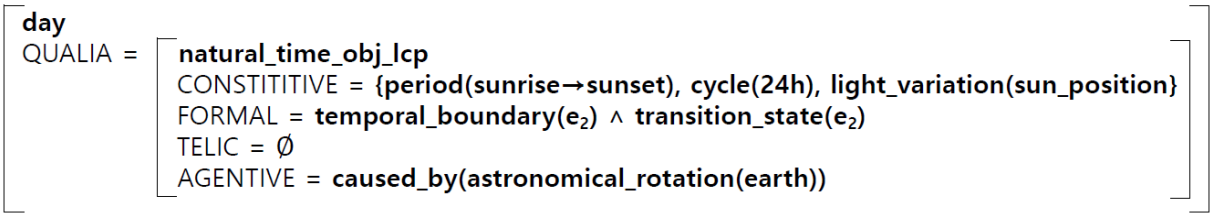
(32) The day broke over the quiet village.

This example features a non-agentive subject (*the day*) and a noncausative use of *break*, resulting in a metaphorical reading. Unlike cases involving functional artifacts (e.g., *toy*, *contract*), *day* lacks a telic role, which makes traditional co-composition inapplicable. Yet, the sentence conveys a clear, structured meaning. GAPM handles this through realization activation.

**Phase 1: Activation of realizations**

The qualia structure of *the day* is given in (33). While it lacks a telic role, it possesses rich formal and constitutive roles that encode its temporal structure and environmental correlates:

(33)

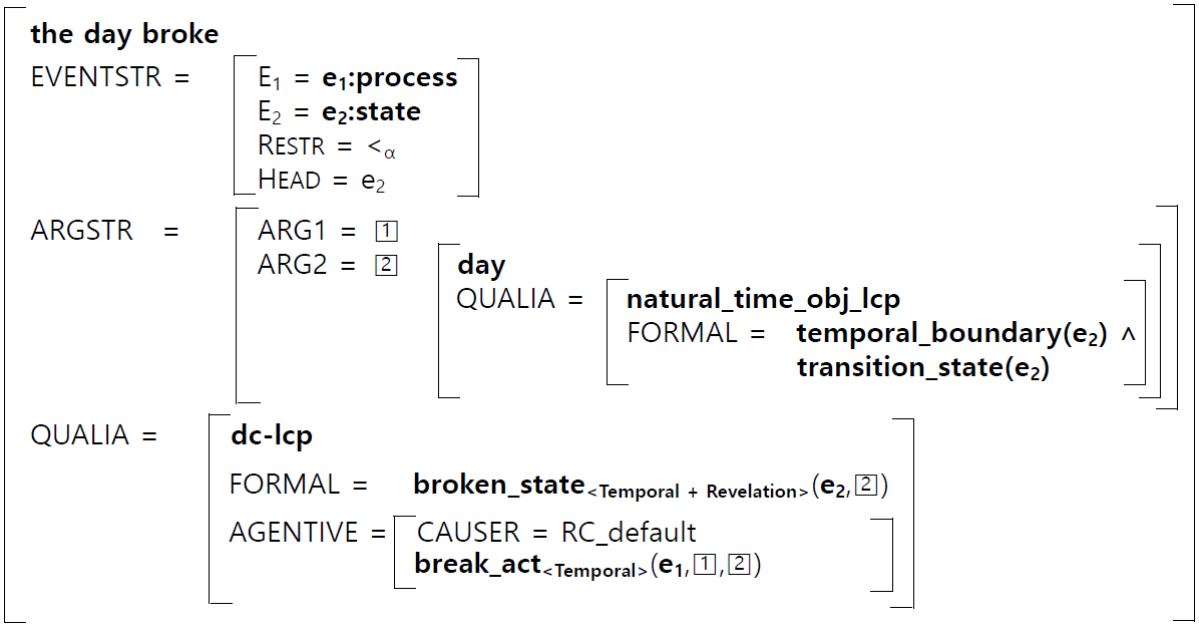



These qualia values—especially the constitutive and formal roles—support the activation of two realization types in *break*:

- Temporal realization in $e_1$ and $e_2$, highlighting the temporal shift and rupture between night and day (e.g., "day breaks" as morning begins),
  - $e_1$: `break_act`$_{\langle Temporal \rangle}$ → temporal shift from dark → light and rupture
  - $e_2$: `broken_state`$_{\langle Temporal \rangle}$ → Daylight state established
- Revelation realization in $e_2$, involving the gradual manifestation of light and visibility in the world.
  - $e_2$: `broken_state`$_{\langle Revelation \rangle}$ → Emergence of light and manifestation of visibility in the world

The semantic representation of *the day broke* is provided in (34), where the causal subevent is not a volitional act but rather a causal factor—namely, the astronomical rotation of the Earth. This type of cause corresponds to the default cause category within the recoverable causer (RC) types discussed in Chapter 2: causes that are widely known, predictable, and unremarkable in discourse unless explicitly foregrounded. That is, the process of "day breaking" is universally associated with the Earth's rotation, making the cause recoverable by default, yet discourse-irrelevant and non-salient. Given this, the resulting state of the day breaking (i.e., the emergence of light, the transition from night to morning) is foregrounded, while the causal subevent is backgrounded. This distribution of prominence in the event structure satisfies the conditions of the mapping principle (21), which states that low causer salience leads to a right-headed event structure. Accordingly, the GAPM representation results in the selection of the noncausative alternant, where the subject (*the day*) surfaces as the theme undergoing change, and the causal factor remains implicit.

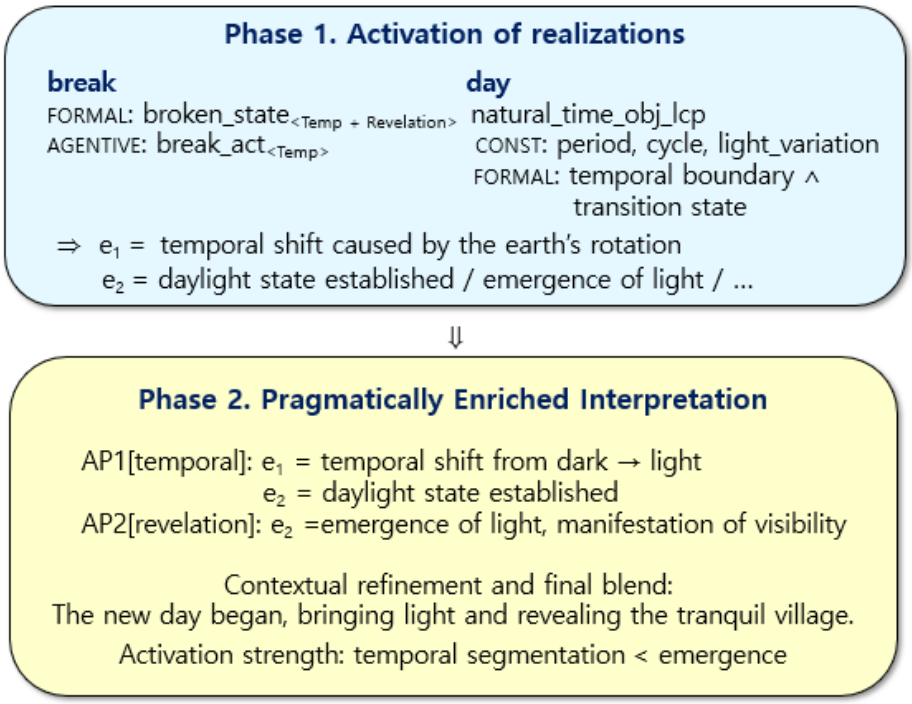
(34)




**Phase 2: Pragmatic enrichment and sense selection**

The second interpretive phase involves the refinement of meaning based on contextual information and world knowledge. In this case, the prepositional adjunct *over the quiet village* evokes a serene, early-morning atmosphere, and world knowledge about dawn supports the interpretation of *break* as the onset of morning light, as shown in (35). Between the two activated readings (temporal and revelation readings), contextual refinement favors revelation, given the emphasis on visual emergence rather than temporal segmentation. Accordingly, the final interpretation is best captured as a graded activation profile, with the revelation realization emerging as the dominant reading. Thus, GAPM effectively models the emergence of figurative meaning in *the day broke* through realization-based activation and context-sensitive sense modulation—even in contexts lacking agentivity or telic structure where the generative mechanism of co-composition does not apply.

(35) Two-phase analysis of sense derivation in *the day broke* under GAPM




To summarize, this section has demonstrated how GAPM accounts for a diverse range of contextual sense selections involving the verb *break*, spanning both causative and noncausative uses. By incorporating both realization types and causer salience into its enriched qualia structure, GAPM captures how different senses emerge through the interaction of argument qualia structures, contextual cues, and event structural mappings. Whether through co-composition (as in *break a toy* or *break a contract/filibuster*) or through realization-driven activation (as in *the day broke*), the framework successfully captures both overlap and dominance in sense derivation. These case studies collectively show that GAPM offers a cognitively and semantically motivated architecture for interpreting polysemy in context.

### 5.4. Conclusions

This chapter has proposed a new hybrid framework—the Generative Activation Package Model (GAPM)—to account for context-driven sense variation in polysemous verbs. The motivation for this model arose from both the limitations of existing lexical semantic theories and the empirical challenges observed in actual language use. Section 5.1 began by surveying major theoretical accounts of how the meanings of polysemous words—especially their underspecified or general meanings—are interpreted and represented: core meaning approaches, thin semantic theories, and rich semantic models. Section 5.2 examined how the Generative Lexicon (GL)—a leading framework among rich semantic models—accounts for sense selection, particularly in verbs like *break* that exhibit overlapping or layered readings in naturalistic usage. We have shown that while GL successfully captures many cases of polysemy via its generative mechanisms and pragmatic enrichment, it also faces limitations in capturing (i) sense dominance or salience within overlapping readings and (ii) non-agentive, naturally occurring changes.

In Section 5.3, a critical comparison of the GL and the Activation Package Model (APM) highlighted the need for a more integrative approach. While GL offers a rich, structured

account of lexical decomposition and event structure, it lacks mechanisms for dynamic sense activation. In contrast, APM models context-sensitive variation well but lacks a generative mechanism for compositional semantics and argument realization, as well as the meaning–syntax mapping principles necessary to account for alternation phenomena.

GAPM was developed to bridge these gaps by synthesizing the generative mechanisms of GL with the activation dynamics of APM. It does so by modeling two interpretive phases: (1) co-composition and activation, and (2) pragmatic enrichment and sense selection. This framework was tested in Section 5.3 through detailed analyses of the verb *break*, examining a variety of constructions that illustrate its polysemous behavior. The model accounted for causative alternation patterns (e.g., *break the toy* vs. *the toy broke*), sense chaining and overlap (*break a contract*), co-activation of multiple realizations (*break a filibuster*), and activation without co-composition (*the day broke*). In each case, GAPM successfully captured the dynamic interplay between lexical structure and context, demonstrating its ability to model both selective activation and blended interpretations across overlapping readings.

However, the proposed model is not without limitations. First, the empirical scope of GAPM in this chapter has been limited to a single verb, *break*, albeit analyzed in depth. A wider application across verb classes is needed to test the generalizability of the framework. Second, the mechanisms of pragmatic enrichment and graded sense dominance have been discussed in informal terms. For GAPM to serve as a fully formal and computationally tractable theory, these components must be translated into rigorous representational structures in future research.

Despite these limitations, the model introduced in this chapter represents a promising shift in the direction of usage-based lexical semantics. By moving beyond purely type-level representations and toward token-level meaning construction, GAPM opens a path toward modeling actually communicated meaning in context. In contrast to traditional frameworks that aim to encode static lexical entries, GAPM provides a flexible and structured way to account for how meanings are shaped by contextual pressures, argument structures, and discourse knowledge. In doing so, it not only advances the theoretical landscape of lexical semantics but also enhances our understanding of polysemy as a dynamic, context-sensitive phenomenon grounded in real-world usage.

# Chapter 6

# Conclusions and Prospects

## 6.1. Polysemy in Semantics with Contextualized Language Models

This final chapter synthesizes the findings and theoretical insights developed throughout the preceding chapters and reflects on their broader implications for the study of polysemy, lexical meaning, and semantic representation.

The purpose of this book was to explore how semantic approaches leveraging large neural language models (LLMs)—which provide semantically rich, contextualized representations—could open a path toward a context-sensitive, usage-based account of lexical meaning and offer new explanations for long-standing, unresolved challenges in explaining verbal polysemy. The analyses presented in this book demonstrate that contextualized language models move beyond traditional sense enumeration, capturing continuous and graded meaning structures that dynamically emerge in usage.

Chapter 2 examined the distributional semantics of causative alternation verbs by integrating annotated corpus variables with BERT-based analyses. The primary goal was to identify the core semantic dimensions underlying these verbs and to explain their divergent constructional preferences. The analysis showed that causer intentionality and identifiability function as robust discriminators between causative and noncausative uses—a finding supported by proportional distributions and MANOVA tests. Principal Component Analysis (PCA) revealed two major semantic dimensions: PC1, contrasting physical/material transformations with qualitative or emergent changes; and PC2, contrasting immediate with gradual changes. Regional verb distribution analyses confirmed that these dimensions consistently structure the semantic space while also predicting systematic constructional preferences: causatives cluster around externally induced, physical transformations, whereas noncausatives favor gradual, internally motivated, or emergent changes. A case study of *empty* further illustrated that BERT-derived semantic spaces capture subtle variation in verb usage across PCA regions, providing qualitative insights that complement quantitative results.

Building on this foundation, the chapter also presented an in-depth comparison of two polysemous change-of-state verbs, *break* and *freeze*, which exhibit contrasting constructional tendencies. The BERT-based analyses of sense distributions revealed systematic associations between sense types, causer properties, and constructional biases: causative-dominant senses clustered with intentional or explicitly identifiable causers, while noncausative-dominant senses aligned with nonintentional or weakly identifiable causers. Moreover, the comparison of *break* and *freeze* revealed complementary patterns: *break* is dominated by causative-oriented senses that reinforce its causative preference, whereas *freeze* is characterized by noncausative-oriented senses underpinning its bias toward the noncausative variant. Taken together, these complementary distributions illustrate a systematic divergence predicted by the model: verbs biased toward the causative variant (e.g., *break*) are dominated by senses associated with intentional and specified causers. These findings show that constructional preferences across verbs emerge from the alignment between sense-level semantics and the informativeness contributed by causer properties.

Chapter 3 shifted the focus to one of the most persistent challenges in theoretical semantics—the problem of sense boundaries. Traditional research on polysemy and lexical ambiguity has paid limited attention to this issue. Widely used sense inventories in NLP and corpus linguistics, such as WordNet**,** follow a sense-enumeration approach, treating discrete interpretations of words as distinct senses. More recent computational and psycholinguistic studies on word sense disambiguation, however, have supported a graded view of sense distinctions (e.g., Erk and McCarthy 2009; McCarthy et al. 2016; Lan et al. 2016). The latest developments, enabled by contextualized language models (Haber and Poesio 2021, 2024; Petersen and Potts 2023), provide further evidence that sense boundaries are often fuzzy rather than discrete. Yet the full range of factors that sharpen or blur the boundaries of polysemous verb senses has remained unclear.

To address this gap, Chapter 3 proposed a framework for exploring verb sense boundaries through probing and fine-tuning contextualized language models. Baseline performance was first established through layer-wise probing experiments, followed by fine-tuning for sense classification. Overall, fine-tuning produced a marked sharpening of class boundaries across the confusion matrix: while probing revealed fuzzy overlaps—especially in abstract domains—the fine-tuned model achieved clearer separations, raising overall accuracy to 87 percent. These improvements highlighted the role of supervised adaptation in reinforcing distinctions that remain opaque under probing alone.

Despite the high accuracy, systematic patterns of misclassification emerged. Error analyses identified key factors underlying sense confusability, including restricted distribution, semantic proximity, and semantic overlap. Based on the meaning-classification patterns of the fine-tuned models, the chapter derived the major distributional and semantic factors contributing to the sharpening or blurring of the boundary between the polysemous senses of *break* and *freeze*:

1. Restricted distribution — occurrence with narrowly defined theme arguments
2. Semantic proximity — shared or hierarchically close superordinate categories, often linked metaphorically
3. Semantic overlap — causal or extension-based connections among senses

Restricted distribution tended to make sense boundaries clearer, whereas semantic proximity and overlap made them more diffuse and fuzzy. The highest-performing classes in the fine-tuning experiment—decoding**,** preservation**,** disclosure, and bodily harm—shared the characteristic of restricted, well-defined theme arguments. In contrast, those senses exhibiting greater confusion were semantically closer to or overlapped with other senses.

Differences in classification performance across senses of *break* and *freeze* were systematically explained by their degree of semantic proximity and overlap. High-performing senses showed narrow distributions and clear boundaries, whereas the lowest-performing senses—such as economic freezing**,** change, and emotional/mental freezing—were deeply embedded in networks of proximity or causal/extension-based overlap. This approach integrates insights from lexical and cognitive semantics with those from distributional semantics, using contextualized models as a bridge. The findings support a view of polysemy as a continuum, ranging from near-synonymous relations (e.g., economic freezing and suspension) to homonymy-like contrasts (e.g., the emergence vs. weakening readings of *the storm broke*).

Chapter 4 extended this inquiry through three complementary experiments designed to probe the perception of polysemous sense boundaries. While Chapter 3 focused on contextualized language models—specifically BERT, RoBERTa, and ALBERT—these are Transformer-based encoder models that generate context-sensitive embeddings suitable for fine-grained analyses of distributional patterns but not for text generation. In contrast, the present chapter investigated generative AI models, represented by GPT and Gemini. These Transformer-based decoder models possess powerful generative capabilities and are accessible via APIs, making them particularly well-suited for conducting rating and judgment experiments in parallel with human participants. This contrast enabled a direct comparison between encoder-based models of contextualization and decoder-based models of generation, highlighting their respective contributions to understanding polysemous sense boundaries.

Across the three experiments—human meaning selection, sense applicability judgment, and usage similarity judgment—we systematically examined how categorical versus graded sense distinctions emerge and how they are shaped by semantic proximity and overlap.

In the meaning-selection task, human participants demonstrated that sense confusion patterns were not random but highly systematic. Senses with strong semantic connectivity—such as change, immobilization, and economic freezing—exhibited high confusion rates and clustered closely with semantically related senses in hierarchical analyses. By contrast, clearly delimited senses such as bodily harm or termination produced minimal errors. These findings confirmed that semantic overlap and proximity are key predictors of sense confusability in human judgments, paralleling the prediction patterns observed in Chapter 3.

The sense-applicability judgments further revealed that the gradedness of judgments varied systematically with the type of semantic relation between senses. Correct and dissimilar labels were evaluated categorically, whereas similar labels exhibited graded, overlapping distributions—especially when the relations involved causal or extension-based overlap or semantic subcategorization. Regression analyses confirmed that these overlapping relations significantly reduced the Correct–Similar score gap, producing gradient judgments of sense applicability. Conversely, context-dependent polysemy and contextual implicature yielded large, categorical distinctions, while metaphorical extension occupied an intermediate position. Both GPT and Gemini broadly mirrored these human patterns, indicating that generative models are sensitive to the same semantic structures that guide human sense evaluation.

The usage-similarity judgments extended the analysis to sentence-level meaning comparisons. Human and model ratings jointly showed that semantic proximity among senses produced higher similarity scores and greater overlap in distribution, particularly for causally or extensionally related senses. Literal–figurative readings of the same sense yielded intermediate similarity, aligning with subcategorical relations rather than distinct sense boundaries. Mixed-effects regression further demonstrated that similarity ratings were systematically predicted by relation type, confirming the graded continuum between categorical and overlapping senses.

Across all three experiments, a convergent pattern emerged: semantic connectivity—manifested as causal, extension-based, or subcategorical relations—consistently predicted gradient sense boundaries, whereas context-dependent or implicature-based relations sustained discrete distinctions. Human and AI judgments alike reflected this continuum, differing primarily in variability rather than in underlying structure. Together, these results provide robust empirical support for a graded, continuum-based model of polysemy, in which sense boundaries are shaped by structured patterns of semantic proximity and overlap rather than by

discrete categorical divisions. Despite architectural differences and variability in generation, GPT and Gemini aligned closely with human evaluative patterns, suggesting that large language models capture key aspects of human-like semantic organization.

The results of the experiments presented in Chapters 3 and 4 call for a fundamental reconsideration of how lexical meaning should be understood and represented. They encourage linguists to adopt a more fluid, usage-based foundation for semantic theory—one capable of accounting for the nuanced and often overlapping nature of word senses as they are realized in actual linguistic contexts. Chapter 5 explored how to integrate these empirical insights into formal theories of word meaning.

The chapter began by reviewing major theoretical approaches to the interpretation and representation of polysemous words—particularly those addressing underspecified or general meanings—including core-meaning approaches, thin semantic theories, and rich semantic models. It then examined how the Generative Lexicon (GL), a leading rich-semantic framework, accounts for sense selection, focusing on verbs like *break* that exhibit overlapping or layered readings in natural usage. The analysis showed that although GL successfully captures many cases of polysemy through its generative mechanisms and pragmatic enrichment, it struggles to represent (i) sense dominance or salience within overlapping readings and (ii) non-agentive, naturally occurring changes.

A critical comparison between the GL and the Activation Package Model (APM) highlighted the need for a more integrative framework. While GL offers a detailed account of lexical decomposition and event structure, it lacks mechanisms for dynamic sense activation. Conversely, APM models context-sensitive variation effectively but lacks a generative mechanism for compositional semantics and argument realization, as well as explicit meaning–syntax mapping principles needed to explain alternation phenomena.

To bridge these gaps, the Generative Activation Package Model (GAPM) was proposed. GAPM synthesizes the generative architecture of GL with the activation dynamics of APM, modeling meaning construction through two interpretive phases: (1) co-composition and activation, and (2) pragmatic enrichment and sense selection. The framework was tested through detailed analyses of the verb *break*, covering a range of constructions that illustrate its polysemous behavior. The model successfully accounted for causative alternation patterns (*break the toy / the toy broke*), sense chaining and overlap (*break a contract*), co-activation of multiple realizations (*break a filibuster*), and activation without co-composition (*the day broke*). In each case, GAPM captured the dynamic interplay between lexical structure and contextual cues, demonstrating its ability to model both selective activation and blended interpretations across overlapping readings.

By integrating corpus-based evidence, probing and fine-tuning experiments, and human and AI judgment experiments, this study contributes to a deeper understanding of how semantic differentiation emerges within continuous semantic spaces and how such gradience interacts with syntactic realization and contextual cues in natural language use.

### 6.2. Theoretical and Methodological Implications

The findings and the proposed Generative Activation Package Model (GAPM) carry significant implications for semantic theory and methodology. The model reconceptualizes polysemy as a context-sensitive process of activation within a structured network of semantic potentials,

shifting the focus from type-level sense enumeration to token-level meaning construction. This reorientation challenges the traditional view of lexical meaning as a discrete set of listed senses and instead posits that meanings dynamically emerge through the interplay of co-composition, activation, and pragmatic enrichment.

Theoretically, GAPM reframes polysemy as a dynamic and context-sensitive process of activation within a structured network of semantic potentials. This reconceptualization departs from traditional, type-based models that assume discrete sense inventories, advancing instead a usage-based and gradient perspective on lexical meaning. By integrating the generative architecture of the Generative Lexicon (GL) with the activation dynamics of the Activation Package Model (APM), GAPM provides a unified account of how multiple related senses are selectively realized in context.

By integrating the generative mechanisms of the GL with the activation dynamics of the APM, GAPM offers a unified account of how multiple, interrelated senses are selectively realized in context. This synthesis provides a principled explanation for how verbs like *break* can instantiate both causative and noncausative variants, or simultaneously evoke overlapping readings such as violation and termination. In doing so, GAPM bridges a key divide in theoretical semantics—between symbolic models emphasizing lexical decomposition and distributional approaches emphasizing contextual variability—demonstrating that the two perspectives are not mutually exclusive but complementary facets of meaning construction.
From a theoretical standpoint, GAPM provides new tools for addressing long-standing challenges in lexical semantics. It offers a concrete mechanism for sense individuation, capturing how distinct yet interconnected meanings compete or co-activate within the same lexical entry. It also redefines the syntax–semantics interface, showing how activation patterns interact with argument-structure realization and why specific sense clusters align with particular constructions. More broadly, it reframes contextual meaning construction as a continuous and graded process, in which lexical, compositional, and pragmatic factors jointly shape the emergent interpretation. This perspective situates polysemy within a usage-based, cognitively plausible continuum rather than a set of isolated sense categories.

Methodologically, this study underscores the potential of combining linguistic theory, experimental data, and neural modeling in a mutually reinforcing cycle. The integration of corpus analysis, probing and fine-tuning experiments, and human–AI judgment studies demonstrates that contextualized embeddings can serve as interpretive data for theoretical inquiry rather than as opaque model outputs. This approach promotes a bidirectional relationship between theory and computation: linguistic hypotheses about meaning gradience, compositionality, and contextual modulation can be operationalized within neural architectures, while model behavior can, in turn, refine and constrain theoretical assumptions. Such iterative feedback fosters a more transparent and empirically grounded understanding of how meaning is represented and constructed in both human cognition and artificial systems.

The convergence of symbolic theory and neural modeling also opens new directions for future research. Extending GAPM to multimodal contexts would allow exploration of how meaning activation operates across linguistic, visual, and perceptual inputs. Likewise, psycholinguistic and neurocognitive validation could test whether the activation dynamics predicted by GAPM correspond to human processing patterns in real-time comprehension. Cross-linguistic applications offer another promising path, examining how activation and alternation interact in languages with different morphological and constructional systems. Finally, the model's focus on graded, context-sensitive activation invites diachronic and

sociolinguistic extensions, tracing how contextual biases evolve over time or vary across speaker communities.

Taken together, the analyses presented in this book converge on a unified view of polysemy as a continuum of dynamically activated meanings grounded in usage and context. By combining corpus-based evidence, experimental validation, and neural modeling, this study advances a framework in which semantic theory and computational modeling inform each other in both explanatory power and empirical precision. The proposed GAPM represents a step toward reconciling symbolic and distributional approaches, bridging the gap between linguistic abstraction and contextual instantiation.

Ultimately, the findings invite a broader rethinking of what it means to "know" a word: not as storing a fixed set of senses, but as possessing a flexible system for activating meaning potentials in context. As large language models continue to evolve, they provide not only technical tools but also conceptual catalysts for understanding the nature of meaning itself. This book thus aims to contribute to an emerging paradigm—polysemy in semantics with contextualized language models—that unites theoretical depth, empirical rigor, and computational innovation in the ongoing exploration of lexical meaning.