Lexical Semantics with Large Language Models: A Case Study of English *break*[*]

Erika Petersen
Stanford University
epetsen@stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

Abstract

Large neural language models (LLMs) can be powerful tools for research in lexical semantics. We illustrate this potential using the English verb *break*, which has numerous senses and appears in a wide range of syntactic frames. We show that LLMs capture known sense distinctions and can be used to identify informative new sense combinations for further analysis. More generally, we argue that LLMs are aligned with lexical semantic theories in providing high-dimensional, contextually modulated representations, but LLMs' lack of discrete features and dependence on usage-based data offer a genuinely new perspective on traditional problems in lexical semantics.

1 Introduction

Pater (2019) builds a compelling case that linguistic and neural network research have great potential for common ground and common cause. His case has only grown stronger in recent years, with the arrival of large neural language models (LLMs) that provide semantically rich, contextual representations (McCann et al., 2017; Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019).

In this paper, we argue that LLMs are powerful devices for studying lexical semantics in ways that can deeply inform linguistic theory. We illustrate this with a detailed case study of the lexical semantics of the English verb *break*, building on a richly annotated dataset from Petersen (2020) and drawing on methods from prior work in this area (Tenney et al., 2019; Reif et al., 2019; Wiedemann et al., 2019; Loureiro et al., 2021). *Break* has long been central to theoretical work in lexical semantics because it has a staggering range of senses that appear to be systematically related to its argument structure. Our central empirical finding is that LLM representations capture many of these known sense distinctions and can be used to identify new sense combinations for further analysis.

We use these findings as a chance to reflect on the core theoretical commitments of lexical semantics as they pertain to LLM-based investigations. Our discussion is centered around the three tenets of lexical semantics given in Table 1: lexical representations are *high dimensional*, *contextually modulated*, and include *discrete features*.

The high dimensionality property is not phrased as a direct claim in the literature as far as we know, but it reflects the practice of linguists, who identify numerous interacting features of lexical items. Section 2 offers a summary picture for *break*. Similarly, discreteness is often assumed by linguists working in the broadly generative tradition. For our purposes, the key question is whether there are *any* features that are discrete, since LLMs do not naturally support having such features.

Contextual modulation is a direct claim. We trace the origins to Dowty (1976, 1979), who argues that aspectual analyses need to include at least the entire verb phrase (see also Kratzer 1996). Borer (2005a,b, 2013) pushes this further, arguing that open-class lexical items are "tantamount to raw material, 'stuff' which is poured into the structural mould to be assigned grammatical properties" (2005a, p. 108). On this view, lexical items are mostly unvalued discrete feature representations that are fleshed out and modulated by the environment in which they appear.

A similar view is taken by work in the Generative Lexicon of Pustejovsky (1991, 1995), which posits an extensible lexicon that is "open-ended in nature and accounts for the novel, creative, uses of words in a variety of contexts by positing procedures for generating semantic expressions for words on the basis of particular contexts" (Pustejovsky, 2006). This also aligns with Clark's (1997) rejection of the "Dogma of Sense Selection", which says "Listeners determine an enumerable set of senses for each

[*] Data and code: https://github.com/epetsen/break-llms

	Linguistics	Static vectors	LLMs
High dimensionality: Lexical semantic entries consist of many features.	Yes	Yes	Yes
Contextual modulation: A word sense will be influenced by its immediate morphosyntactic context as well as the broader context of use.	Yes	No	Yes
Discreteness: The features in lexical semantic entries are discrete and highly structured.	Yes	No	No

Table 1: Core tenets. Our focus is in particular on the relationship between 'Linguistics' and 'LLMs' in this table.

expression, and in understanding what a speaker means, they select the appropriate sense from that set." For Clark, lexical items are highly malleable and constrained only by what the discourse participants can reliably communicate with each other (see also Clark and Clark 1979). On all these views, lexical items are highly abstract computational objects that can be realized in very diverse ways.

The field of NLP has a complex relation to our tenets. Early work on symbolic grammars in NLP was clearly aligned with all the tenets. The Generative Lexicon is a prominent example and proved influential in linguistics and NLP. When distributional methods first became central to NLP, the dominant mode of lexical representation involved static vector representations. These representations align with the consensus in linguistics only regarding high dimensionality, as we discuss in Section 4.

LLMs have changed NLP's relationship to lexical semantics considerably. With LLMs, we have a strong commitment to high dimensionality and contextual modulation and a denial of discreteness (Section 5). The points of agreement present a significant opportunity for linguists and NLP researchers to collaborate, as we hope our case study shows. The points of disagreement seem also to be opportunities for people to take new perspectives. We argue in particular that the facts surrounding *break* should lead linguists to reconsider their commitment to discreteness and embrace a more fluid, usage-based foundation for semantic theory.

2 English *Break*

English *break* is one of the best studied lexical items in lexical semantics, for a few reasons. First, it is a canonical instance of a change-of-state verb that undergoes the causative alternation:

(1) The linguist broke the window

(2) The window broke.

In fact, alternating change-of-state verbs are referred to as *break*-verbs (Acedo-Matellán and Mateu, 2014; Fillmore, 1970; Levin, 2017; Majid et al., 2008). The intransitive variant of the causative alternation (2) is analyzed in terms of the *unaccusativity hypothesis* (Perlmutter, 1978; Burzio, 1986; Levin and Rappaport Hovav, 1995), which says that the subjects in these cases are underlyingly internal arguments to the verb, bearing more theme-like semantic roles, and have been promoted to subject position to fulfill a subjecthood requirement. Research on *break* has also contributed to the study of the lexical properties of unaccusative verbs (Levin and Rappaport Hovav, 1995).

Second, *break* can take on a wide array of senses. Table 2 provides a partial list; we cannot hope to be comprehensive (there may not even be a fixed stock of senses; Section 5), but our examples convey the nature of the attested variation.

Third, the sense distinctions interact with the causative alternation. Whereas senses 1–4 all alternate, senses 5–10 are all strictly transitive. The non-alternating senses of *break* have informed the debate about which variant of the causative alternation (if any) is basic and which is derived (e.g. Levin and Rappaport Hovav, 1995; Alexiadou et al., 2006; Piñón, 2001). Though the debate is still unsettled, it has evinced that participation in the causative alternation is not a property of the verb itself, but of the verb in combination with its theme argument (Petersen, 2020; Spalek, 2012), just as with telicity and other aspectual properties (Dowty, 1976, 1979; Borer, 2005b).

Prior work has sought to capture the obligatorily transitive nature of some of these senses by appeal to a thematic role requirement: *break* in combination with its internal argument determines the range of semantic roles – agent, instrument, or natural force – that the subject of a transitive

	Frame	Sense
1.	break the vase	shatter
2.	break the computer	render inoperable
3.	break the news	reveal
4.	break the silence	interrupt
5.	break the record	surpass
6.	break the code	decipher
7.	break the law	violate
8.	break the horse	tame
9.	break a $10 bill	make change
10.	break the fall	lessen
11.	the weather broke	changed
12.	the day broke	began

(a) Uses without particles/predicates.

	Frame	Sense
13.	break off the engagement	end
14.	break out	begin
15.	break out in hives	get
16.	break into the building	intrude
17.	break down the problem	analyze
18.	break down the proteins	decompose
19.	break in	enter
20.	break in	interrupt
21.	break free	escape
22.	break even	profit = loss
23.	break forth	emerge
24.	break to the right	turn

(b) Uses with particles/predicates.

Table 2: Senses for *break*. A comprehensive account of senses may not be possible (Section 5.3).

break frame may bear (Rappaport Hovav and Levin, 2012), and some frames require their subjects to be agentive (Levin and Rappaport Hovav, 1995; Piñón, 2001; Alexiadou et al., 2006; Schäfer, 2008). Since necessarily agentive subjects cannot be left unexpressed, these frames do not show intransitive variants. However, this cannot be the full story, as there are some obligatorily transitive *break* frames where the subject need not be an agent but which nonetheless do not alternate, like *the cushion broke her fall* vs. **the fall broke* (Petersen, 2020).

In addition, examples like *break the record* (sense 5) and *break the code* (sense 6) may be graded or uncertain in regard to their participation in the causative alternation. They are often assumed not to have intransitive uses (Levin and Rappaport Hovav, 1995; Piñón, 2001; Alexiadou et al., 2006; Schäfer, 2008; Rappaport Hovav and Levin, 2012), but there are attested cases like the following that suggest this is a point of variation.

(3) Almost sixty years later, Frank Rowlett, a cryptologic pioneer and head of the "Purple" team, remembered that historic day when the code broke.

(4) The Guinness World Record broke, our furniture didn't.

There are also strictly intransitive uses, as in 11–12 of Table 2a. These are analyzed in the same way as the intransitive variant of alternating frames (2), i.e., as unaccusatives. Why these *break* frames do not allow a cause subject – e.g., **the Earth's rotation broke the day* – is an open question.

As seen in Table 2b, *break* also combines with a wide range of predicates and particles to create new senses. Except for 13, 17, and 18, these uses are all intransitive, but they seem to differ from the particle-less uses in a key way: whereas intransitive particle-less *break* cases are all unaccusative, the particle cases vary in this regard. For example, *the war broke out* seems unaccusative, but *we broke into the building* has an agentive subject and so would not be analyzed as unaccusative.

A key question for lexical semantic theories is whether there is a single unifying semantic frame underlying this diverse array of senses – or, if not a single frame, then perhaps a few of them feeding into distinct sense clusters. This position is advanced, for example, by Kellerman (1978, 65), for whom "[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 position for Spanish *romper* 'break'). Another approach would be to posit a few more primitive semantic dimensions that give rise to a combinatorial space of predicted senses, which might in turn lead to predictions about argument structure realization and other structural and distributional properties.

3 Feature-based Theories

In this section, we take the somewhat unusual step of bringing together existing ideas from the linguistics literature into a feature space of the sort one is likely to encounter in NLP contexts. We do this

		Transitive	*Unaccusative*	*Agent*	*Metaphorical*	*separate*	*violate*	*end*	*appear*	*out_escape*	*out_begin*
1.	We broke the vase	1	0	1	0	1	0	0	0	0	0
2.	The vase broke	0	1	0	0	1	0	0	0	0	0
3.	We broke the law	1	0	1	1	0	1	0	0	0	0
4.	The silence broke a procedural rule	1	0	0	1	0	1	0	0	0	0
5.	We broke the silence	1	0	1	1	0	0	1	0	0	0
6.	The day broke	0	1	0	1	0	0	0	1	0	0
7.	The storm broke	0	1	0	1	0	0	0	1	0	0
8.	Sweat broke on his forehead	0	1	0	1	1	0	0	1	0	0
9.	We broke out (of jail)	0	0	1	0	0	0	0	0	1	0
10.	Fighting broke out	0	1	0	1	0	0	0	0	0	1

Table 3: Partial feature-based analysis of *break* in different syntactic contexts.

for a few reasons. First, it reveals that, though theories in linguistics and NLP often take very different forms, there is actually a lot of common ground between them: on both sides, vector representations of data can serve as a common language. Second, the feature space reveals how deeply linguistic theories are committed to our contextual modulation tenet from Table 1: to honor the insights from the literature, we have to define the feature space in terms of (at least) full sentences.

Table 3 is our (highly partial) feature-based analysis. The Transitive feature captures whether a particular *break* frame has two nominal arguments or one. Causative alternation uses can then be reconstructed by looking at shared meaning dimensions that vary in their Transitive value, as in rows 1–2. We separately define an Unaccusative feature, since the uses in Table 2b show that these can come apart. This is evident especially in rows 9–10.

The Agent feature captures whether the subject of each example is agentive or not. We mentioned in Section 2 that the obligatory transitivity of some *break* frames has been traced, unsuccessfully in our view, to the agentivity of the subject of these frames. The Agent feature in combination with the Transitive feature and the meaning dimensions reveals the incompleteness of this explanation: 'violate' examples, which are obligatorily transitive, may have subjects that are agentive (row 4) and non-agentive (row 5).

We have a column for Metaphorical, though coding this is sufficiently hard that it looks like a multidimensional category to us rather than a single feature. Due to the difficulty of classifying senses of *break* and other polysemous verbs as (non)metaphorical, previous literature that has engaged with this question (e.g. Kellerman, 1978; Piñón, 2001; McNally and Spalek, 2017, 2022) has used the heuristic of equating metaphorical senses with senses with abstract participants, like *break the silence*, and non-metaphorical ones with senses with concrete participants, such as *break the vase*. We follow this (admittedly simplifying) heuristic in our feature-based analysis. However, we agree with McNally and Spalek (2022, 6) that the "the distinction between 'literal' and 'figurative' senses can become blurred over time, and sometimes can only be diachronically reconstructed."

Following these features are a few meaning dimensions. The full class of meaning annotations we use in Section 5.3 has 72 classes, so this is just a sample. The sample was chosen to emphasize three aspects of the meanings of *break*. First, as already illustrated in Table 2, these meanings are highly diverse semantically. Second, it is difficult (and maybe even futile) to determine with confidence how many distinct (and non-overlapping) senses *break* may express. For example, does *break* express the same meaning, 'appear', in row 6 as in row 7, as we suggest in Table 3? Or should these examples be seen as expressing distinct senses? Third, we believe there are some examples where *break* simultaneously expresses more than one meaning, as shown in row 8 of Table 3, where *break* shows both an 'appear' and a 'separate' meaning.

Break with particles/predicates can sometimes express meanings that particle-less *break* cannot convey: e.g. ‘escape’ in row 9. We assume that the particles/predicates contribute an irreducible meaning and reflect this in our feature-based analysis by preceding these meaning dimensions with the corresponding particles/predicates: ‘out_escape’. The particles/predicates do not determine a unique meaning, though, as we see with the two senses of *break out*: ‘out_escape’ and ‘out_begin’.

Potentially all of the columns are actually just informal stand-ins for much more complex concepts. The labels could be natural language predicates on par with *break*, in which case the column names are really just hooks into a larger lexical web, or they could be glosses for more intricate theoretical concepts that demand further decomposition before the theory can be regarded as complete.

How many lexical items does this theory posit? The answer to this question is not clear. We could say that each attested combination of the features is a new sense, or we could select a few features and say that specific combinations of them correspond to distinct senses. Both decisions have a certain arbitrariness to them given the feature space itself, and we might infer from this that the theory does not posit distinct senses or distinct lexical items as first-class linguistic constructs. This may be a consequence of the contextual modulation tenet.

Relatedly, it is unclear to us what a *complete* analysis in these terms would look like. What would it mean to have determined all and only the correct features? Could it be that the investigation will always admit of further dimensions, or decomposition of existing dimensions?

In sum, it is easy to see how this analysis makes good on the central tenets in Table 1. The representations are high-dimensional vectors with discrete values. In addition, the representations themselves directly bring in context. The vector for *break* alone, if it exists in the theory at all, needs to be mostly unspecified values that only become values in specific syntactic or usage contexts.

4 Static Vector Modeling

The above feature-based analysis might be described as a *sparse vector representation* approach. We now contrast that with a dense vector representation approach that models individual lexical items as fixed (static) vectors. A variety of such methods have been developed. Here we look at the treatment of *break* by three prominent methods: word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Mikolov et al., 2018). A brief technical overview of these methods is included in Appendix B.1.

None of these methods use discrete features, and thus they are in conflict with our discreteness tenet from Table 1. The raw input to all of them is a matrix of co-occurrence counts, which could be viewed as a set of discrete distributional features. However, much of the power of these models derives from their ability to compress this information into a lower-dimensional space of continuous values in which the columns are unlikely to have direct interpretations as features.

The GloVe vocabulary is largely restricted to individual words from a fixed list. By contrast, the vocabulary used for our word2vec instance includes some phrase-like elements that were inferred by the authors using simple co-occurrence statistics, and our chosen fastText model includes sub-word components and so also ends up with a more expansive view of what counts as a lexical item.

All of these models have proven successful as representations of words and as components in larger systems. However, we find that these representations are disappointing for studying *break*. In Table 4, we show the top 20 nearest neighbors (according to cosine similarity) for some uses of these models. We chose what seemed to be the semantically richest instance of each model from a larger set of such results (see Appendix B.2). All of the models capture morphological variants very clearly. However, the other semantic associations generally only weakly indicate other specific senses (via associations with other words). We do see some positive benefits from the quasi-phrasal vocabulary used by word2vec and fastText, but overall these spaces look like only superficial pictures of the underlying semantic richness of *break*.

The cause for this semantic blandness likely traces to the basic design decision: every wordform has only a single representation. This means that a single vector must encode all the different senses that we see in Table 2 as well as others that we did not include there. The result is probably something like a weighted average of these senses, which seems not to be in a particularly interesting part of the embedding space.

Adherents to our central tenets (Table 1) might have predicted this negative result. While GloVe representations are high dimensional, they do not

1.	break	11.	up
2.	breaks	12.	trying
3.	breaking	13.	away
4.	end	14.	start
5.	broke	15.	get
6.	down	16.	again
7.	take	17.	’ll
8.	let	18.	back
9.	going	19.	out
10.	leave	20.	off

(a) GloVe, Common Crawl 840B tokens, 300d.

1.	breaks	11.	brief_respite
2.	breaking	12.	Nadal_netted_forehand
3.	broke	13.	loosen
4.	broken	14.	smash
5.	Break	15.	rip
6.	Breaking	16.	overhit_forehand
7.	breather	17.	miscued_forehand
8.	shatter	18.	cut
9.	crack	19.	slip
10.	breaker	20.	Breaks

(b) word2vec, GoogleNews, 300d.

1.	break	11.	breakin
2.	breaks	12.	breaked
3.	breaking	13.	broken
4.	breake	14.	legbreak
5.	re-break	15.	reak
6.	break-	16.	semi-break
7.	unbreak	17.	minibreak
8.	breakes	18.	breaker
9.	break.	19.	breaking-down
10.	broke	20.	tea-break

(c) fastTest WikiNews, with subword modeling, 300d.

Table 4: Nearest neighbors of *break* in static embedding spaces. All the methods place morphological variants of *break* next to *break* itself and seem to sporadically find different senses and near synonyms of *break*. The lists given here are, in our judgment, the best from each of the three methods. For additional lists, see Appendix B.2.

allow for contextual modulation. Each basic unit of the vocabulary is assigned exactly one representation. Contextual modulation may occur if the representations are embedded in a larger system, but it is not intrinsic to the vectors as lexical theory.

5 LLM Investigations

We come now to our primary investigative tool: LLMs. We concentrate on models that have the core structure of the Transformer (Vaswani et al., 2017) and are trained at least in part using masked language modeling, which allows for bidirectional context. In the interest of space, we will mostly presuppose familiarity with these models. However, Appendix C provides an overview of their structure to try to bridge any gaps between the linguistics and NLP literature.

In our main text, we report results for RoBERTa-large (Liu et al., 2019), which has 24 layers. Our appendices cover BERT and DeBERTa. Our RoBERTa-large results are slightly better than all of these others, but the results are generally quite comparable, suggesting that all of these models can fruitfully be used for lexical semantics.

Before turning to our experiments, let’s consider how LLMs relate to our core tenets from Table 1. First, the representations we obtain at each hidden layer are all high-dimensional and modulated by the context. Our experiments show that, for the case of *break*, this contextual modulation is rich and linguistically systematic. Thus, LLMs and traditional lexical semantic theories are aligned on these two tenets. However, the two theories part ways when it comes to the question of having discrete features. The column dimensions of LLM representations are continuous and highly abstract. Discrete linguistic features might be latently encoded in these representations, or extractable from them with some noise, but this does not detract from the fact that these representations are highly fluid and do not presuppose the existence of any particular features or dimensions. Rather, all the features are learned from data in a free-form way that is grounded entirely in distributions.

5.1 Annotated Dataset

The basis for our investigation is an annotated dataset created by Petersen (2020) and subsquentally updated by Petersen to include more examples and senses. The examples are extracted from the Corpus of Contemporary American English (CoCA; Davies 2008). We focus on a subset of 1,042 sentences that have been annotated for, among other things, the core semantic class of the reading and the construction type (‘unergative’, ‘unaccusative’, ‘causative’). Petersen assigns a single semantic class to each example. However, as mentioned in Section 3, we believe in some cases *break* can be said to simultaneously express more than one meaning. We use our experiment in Section 5.3 to identify examples with this property.

We rely primarily on the meaning class distinctions and make secondary use of the constructional annotations. Petersen’s annotation scheme uses 72 semantically rich meaning classes, which have a highly skewed distribution. The full distribution is given in Appendix E.

5.2 Probing Experiments

We want to explore the LLM representations in a fluid way that will lead us to identify new readings. Our tools for doing this are supervised probe models applied to the column of representations above the *break* token in each of our examples. We

Layer	Probe	Control	Selectivity
1	0.33	0.03	0.30
6	0.81	0.03	0.79
12	0.83	0.03	0.80
18	0.80	0.03	0.76
24	0.86	0.03	0.83

(a) Meaning-class probing results.

Layer	Probe	Control	Selectivity
1	0.50	0.33	0.17
6	0.94	0.34	0.60
12	0.96	0.33	0.63
18	0.96	0.35	0.61
24	0.97	0.32	0.65

(b) Construction-type probing results.

Table 5: RoBERTa-large probing results. We report Macro F1 and Selectivity, which is the Macro F1 score for the task minus the Macro F1 for a control task (random assignment of tokens to classes). Results for other models are similar; see Appendix D.

probe for meaning-class and construction-type (see also Papadimitriou et al. 2021). These probes serve as a quantitative evaluation of the extent to which these *break* representations encode these important properties, and they are also tools for heuristically finding new uses and readings.

For our construction-type probing work, we can use all 1,042 sentences, since there are only three classes and all have substantial representation in the data (causative: 673 examples, unaccusative: 197, unergative: 172). For the meaning-type work, there are 72 classes, many with only a few instances. Thus, we limit attention to just the classes with at least 10 examples (Appendix E).

Our probe models are L2-regularized classifiers with a cross-entropy loss. Our core metric is the macro F1 score, which assigns equal weight to each class's F1 score regardless of the class size. Following Hewitt and Liang (2019), we report *selectivity* scores, which are the probe scores minus the performance on a control task, which here is random assignment of *break* representations to meaning classes. We report selectivity scores averaged across 20 random 80%/20% train/test splits, with bootstrapped 95% confidence intervals.

The probe results show a clear pattern: the lowest layers are not very robust when it comes to this probing work, but higher layers are very robust in this sense (see also Reif et al. 2019; Ethayarajh 2019). We see similar results for other LLMs in the class we are focused on, as reported in Appendix D. For this reason, we focus on layer 24 of RoBERTa-large from now on.

5.3 Discovering New Example Types

Our primary goal is to see whether it is possible to use LLMs to gain new insights about lexical semantics. Our probing results suggest that LLM representations are systematic enough to make this plausible, but they are very high-level. We need an investigative technique that is more free-form and that can bring to our attention new kinds of theory-relevant examples.

A natural choice is visualization. We provide t-SNE visualizations (van der Maaten and Hinton, 2008) in Appendix G, and we find that they are indeed useful: where an example of meaning class a is nestled among examples of class b, the a-class example is often an interesting blend of a-class and b-class meanings, and such examples seem genuinely worthwhile to study further. However, these visualizations introduce known distortions resulting from compressing high-dimensional spaces into two dimensions (Wattenberg et al., 2016), and they can even vary in qualitatively substantive ways across models and runs.

For something more stable, we return to our probe models. The selectivity scores for both are conservative if we think of them as tools for finding new examples: the meaning-class probes achieve results above 80% macro F1, and the construction-type probes are nearly perfect in their performance (where chance is around 33%). Thus, we decided to extract and review the errors made by these models, with the expectation that many of these examples could inform lexical semantic theory itself.

Table 6 is a selection of examples that we extracted in this way for further analysis. This is a small curated set of (so-called) errors, though these examples do not look like errors to us, but rather like instances in which multiple senses and multiple construction types emerge in the same example.

Example 1 is predicted unergative, whereas the gold label is unaccusative. For us, this raises a new question: what is the role of *spontaneous* in the overall agentivity of the subject, and should this play into how we characterize the syntactic frame?

	Meaning		Construction	
Sentence	Gold	Predicted	Gold	Predicted
1. Patients will sometimes break out in a spontaneous recitation of the rosary	break_out_ start	break_out_ start	unacc.	unerg.
2. It was like you knew something, like you knew the story was getting ready to break again.	reveal	appear	unacc.	unacc.
3. People have so many problems overcoming the disputes that occur when families break up	break_up_ end_relationship	break_up_ separate_into_parts	unacc.	unacc.
4. "So why tell the whole story now? Somebody, some male, has got to be willing to break this code of silence," he says.	violate	end	unacc.	unacc.
5. Wind, naturally acidic rain, and physical processes such as freezethaw cycles also break down rock.	break_down_ separate_into_parts	break_down_ destroy	causative	causative

Table 6: A curated sample of theoretically informative examples.

Example 2 looks like a clear case where multiple senses can be activated and different utterance contexts might favor different readings. Is the breaking of a news story an agentive act of revealing information (the gold label), or can it be (or be described as) something more like a natural process of appearing (the predicted label)? Both readings seem available, and individual uses might blend them for a particular rhetorical effect.

We also find cases where multiple senses are present due to contextual entailment relations between them. In example 3, should we focus on the direct and perhaps metaphorical reading (gold) or the more literal likely consequence (predicted)? In example 4, violating the code of silence entails ending it. And example 5 reveals that the event structure of *break* examples can be very complex. When a rock is broken down by natural forces, we can think of this as a process of breaking down into smaller parts (gold meaning) with the end state being total destruction (predicted meaning). Many examples in which the breaking event leads to the fragmentation of the theme participant can show similar blends.

These are just a few examples of a much larger set of interesting cases that emerge from studying the interaction between our LLM-based probe models and our linguistic annotations. Appendix F provides a larger sample with brief annotations about potential theoretical relevance. We close this paper by reflecting on how best to incorporate these insights into the linguistic theory itself.

6 Discussion

The fact that linguistic theory and LLMs agree on the core tenets of high dimensionality and contextual modulation is a striking alignment of theoretical idea with engineering success.

The fact that LLMs do not use discrete features, but rather derive dense, real-valued representations from data seems like an opportunity for linguists to reflect on the role of discreteness. As we noted in Section 3, it seems unlikely that purely analytic work and traditional corpus work will lead to an exhaustive hand-built representation for any lexical items. With LLMs, we can mine the existing representations while considering the LLM architecture and learned parameters to be the theory.

The deep contextual modulation countenanced by linguistic theory and operationalized by LLM embeddings invites a further question: do lexical items exist outside of their tokens of use? Even for the hand-built feature representations in Table 3, the rows could in principle vary based even on usage information, which would suggest a theory that is actually more about tokens (instances of use) rather than types. Similarly, for LLMs, though they do contain type-level representations (in the form of an embedding for the vocabulary), these play a minor role, and all the representations we have considered in this paper were in terms of representations that are more like token-level representations.

Overall, then, a theory of lexical semantics that draws heavily on LLMs as investigative tools, and even as ways to state theoretical ideas, is likely to become more usage-based than traditional theories would assume. This could lead them to focus less on pure representation and more on what is actually communicated between people when they communicate. Traditional questions are likely to take on new forms in this setting, and exciting and relevant new questions – and new pieces of evidence – are likely to arise.

7 Limitations

Our general thesis is that LLMs are valuable tools both for conducting lexical semantic analyses and for providing valuable perspectives for lexical semantic theory design in general. Although we think this thesis is widely supported by prior literature, our own case study is limited to just a partial analysis of a single verb. This creates the risk that our general conclusions may be more specific to this verb, or to English, than we would like. The prior literature inherits many of our English-only biases as well (but see Papadimitriou et al. 2021).

Our main results use RoBERTa-large, and our appendices report on parallel analyses with different versions of BERT and DeBERTa. These models share core architectural features and were optimized in largely similar ways using very large – and largely uncontrolled – datasets. This means that these artifacts are certainly biased in ways that are relevant for lexical semantics. However, we are unlikely to be able to identify, isolate, and factor out these biases with the methods used in our paper. Our core methods are reasonably simple and lightweight, and we are releasing all our code. We hope that these steps allow easy reproduction of our core analyses whenever newer LLMs are released, so that we can begin to understand better how LLM biases can affect linguistic theorizing in the mode we are advocating for.

Our current approach also has an analytic limitation: though we fit probe models and use them as devices for finding potentially relevant examples, the final step in our analysis involves inspection of those examples by linguists like ourselves. This means that the final step is not as reproducible as the others, and it means that any analytic biases that the linguists involved might have are likely to make their way into the analyses. We do not see a way to avoid these analytic steps entirely, since linguistic analysis favors this kind of low-level work, but we do think that we can mitigate the concerns about analyst bias by making all our data available for others to inspect, as a way of opening up many perspectives on the data and the associated theoretical questions.

References

Víctor Acedo-Matellán and Jaume Mateu. 2014. From syntax to roots: A syntactic approach to root interpretation. In Artemis Alexiadou, Hagit Borer, and Florian Schäfer, editors, *The Syntax of Roots and the Roots of Syntax*, pages 14–32. Oxford University Press, Oxford.

Artemis Alexiadou, Elena Anagnostopoulou, and Florian Schäfer. 2006. The properties of anticausatives crosslinguistically. In Mara Frascarelli, editor, *Phases of Interpretation*, pages 187–211. Mouton de Gruyter, Berlin.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Collin F. Baker and Hiroaki Sato. 2003. The FrameNet data and software. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 161–164, Sapporo, Japan. Association for Computational Linguistics.

Hagit Borer. 2005a. *In Name Only*, volume 1 of *Structuring Sense*. Oxford University Press.

Hagit Borer. 2005b. *The Normal Course of Events*, volume 2 of *Structuring Sense*. Oxford University Press.

Hagit Borer. 2013. *Taking Form*, volume 3 of *Structuring Sense*. Oxford University Press.

Luigi Burzio. 1986. *Italian Syntax*. D. Reidel Publishing Company, Dordrecht.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55(4):767–811.

Herbert H. Clark. 1997. Dogmas of understanding. *Discourse Processes*, 23(3):567–59.

Ryan Cotterell, Adam Poliak, Benjamin Van Durme, and Jason Eisner. 2017. Explaining and generalizing skip-gram through exponential family principal component analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 175–181, Valencia, Spain. Association for Computational Linguistics.

Mark Davies. 2008. The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at `http://corpus.byu.edu/coca/`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Dowty. 1976. Montague grammar and the lexical decomposition of causative verbs. In Barbara H. Partee, editor, *Montague Grammar*, pages 201–245. Academic Press, New York.

David Dowty. 1979. *Word Meaning and Montague Grammar*. D. Reidel, Dordrecht.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Charles Fillmore. 1970. The grammar of hitting and breaking. In R.A. Jacobs and P.S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 120–133. Ginn, Waltham, MA.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Eric Kellerman. 1978. Giving learners a break: Native language intuitions as a source of predictions about transferability. *Working Papers on Bilingualism*, 15:60–92.

Angelika Kratzer. 1996. Severing the external argument from its verb. In Johan Rooryck and Laurie Zaring, editors, *Phrase Structure and the Lexicon*, pages 109–137. Kluwer, Dordrecht.

Beth Levin. 2017. The elasticity of verb meaning revisited. In *Proceedings of SALT 27*, pages 571–599.

Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax–Lexical Semantics Interface*. MIT Press, Cambridge, MA.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv:1907.11692.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Asifa Majid, James S. Boster, and Melissa Bowerman. 2008. The cross-lingusitic categorization of everyday events: A study of cutting and breaking. *Cognition*, 109:235–250.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems 30*, pages 6294–6305.

Louise McNally and Alexandra Anna Spalek. 2017. 'Figurative' uses of verbs and grammar. Unpublished manuscript.

Louise McNally and Alexandra Anna Spalek. 2022. Grammatically relevant aspects of meaning and verbal polysemy. *Linguistics*, pages 1–45.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Leon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Joe Pater. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

David M. Perlmutter. 1978. Impersonal passives and the Unaccusative Hypothesis. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 38, pages 157–189. Berkeley Linguistics Society, Linguistic Society of America.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Erika Petersen. 2020. Break + NP constraints and the causative alternation. Ms., Stanford University.

Christopher Piñón. 2001. A finer look at the causative-inchoative alternation. In *Proceedings of SALT 11*, pages 346–364.

James Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4):409–441.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.

James Pustejovsky. 2006. Introduction to Generative Lexicon. Ms., Brandeis.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Ms, OpenAI.

Malka Rappaport Hovav and Beth Levin. 2012. Lexicon uniformity and the causative alternation. In Martin Everaert, Marijana Marelj, and Tal Siloni, editors, *The Theta System: Argument Structure at the Interface*, pages 150–176. Oxford University Press, Oxford.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.

Florian Schäfer. 2008. *The Syntax of (Anti-)Causatives: External Arguments in Change-of-State Contexts*. John Benjamins, Amsterdam.

Alexandra Anna Spalek. 2012. Putting order into literal and figurative uses of verbs: *romper* as a case study. *Borealis*, 1(2):140–167.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Peter D Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning*, pages 491–502. Springer.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

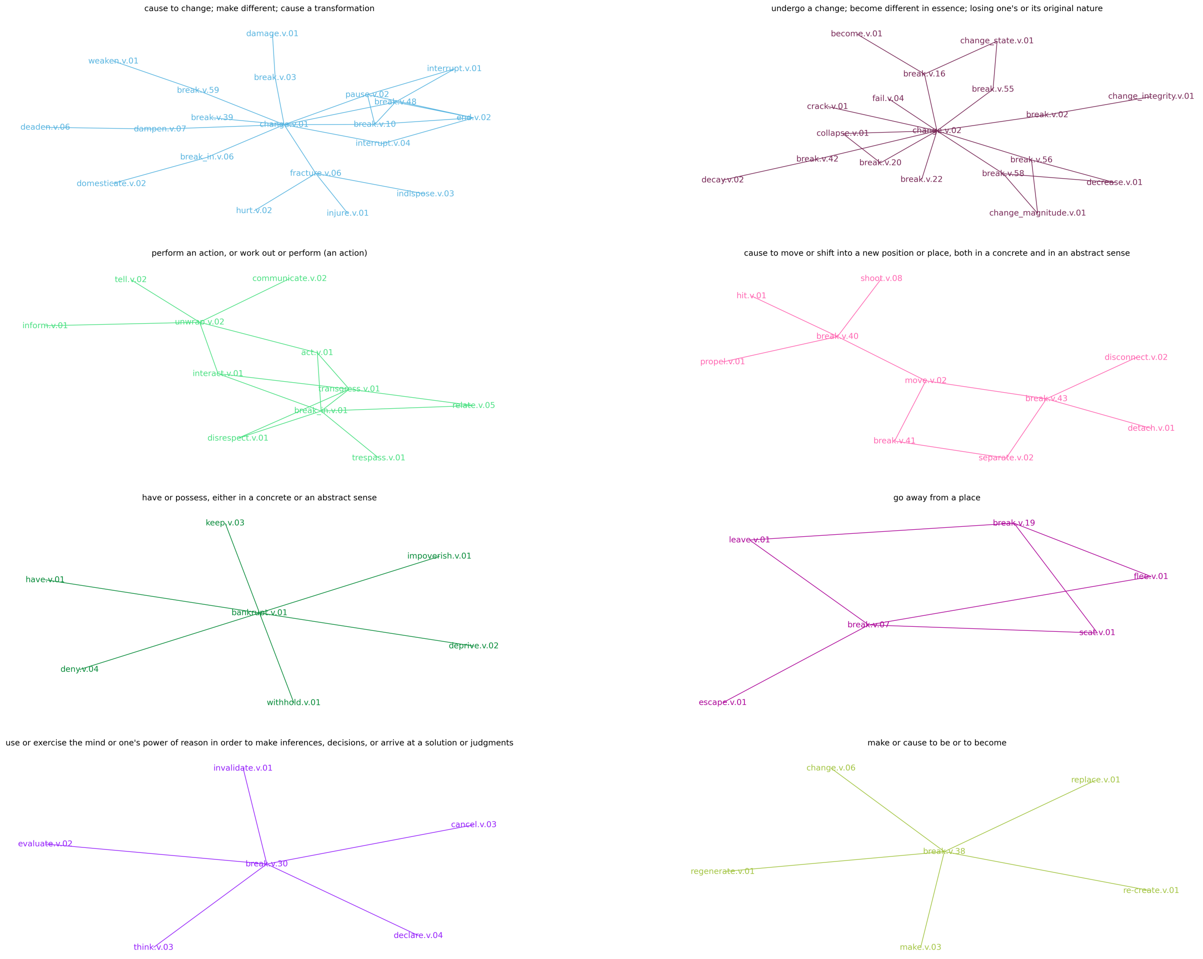
Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to use t-SNE effectively. *Distill*, 1(10):e2.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *ArXiv*, abs/1909.10430.

Supplementary Materials

A WordNet-based Features

The feature-based analyses of Section 3 are easily extended with features obtained using more approximate, data-driven techniques. To illustrate this potential, we looked to WordNet (Fellbaum, 1998), which has a very rich picture of *break*. The lemma *break* participates in 59 SynSets in WordNet. We built a graph of these SynSets based on the hypernym relation. The resulting graph has 29 connected components (29 subgraphs). Figure 1 depicts the largest connected components as subgraphs. If we label these subgraphs with their most-specific shared hypernym, we get potentially new meaning dimensions like "Cause to change; make different; cause a transformation" and "undergo a change; become different in essence; losing one's or its original nature". These are similar, but only the first conveys agency. Both seem like plausible latent semantic dimensions that we could add to Table 3, either as primitive features or as sets of more basic meanings. And of course this is only a single example of many that WordNet would support, and additional features could be extracted from FrameNet (Baker et al., 1998; Baker and Sato, 2003; Ruppenhofer et al., 2006).



cause to change; make different; cause a transformation

undergo a change; become different in essence; losing one's or its original nature

perform an action, or work out or perform (an action)

cause to move or shift into a new position or place, both in a concrete and in an abstract sense

have or possess, either in a concrete or an abstract sense

go away from a place

use or exercise the mind or one's power of reason in order to make inferences, decisions, or arrive at a solution or judgments

make or cause to be or to become

Figure 1: Largest WordNet connected components for *break* labeled with the name of their most specific shared hypernym. These hypernym labels suggest interesting abstract meaning dimensions for these senses.

B Static Vectors

B.1 Static Vector Models

The three static vector methods we consider learn representations for words based purely on co-occurrence patterns in unstructured text. The precise learning objectives are different in each case, but all are closely related to Pointwise Mutual Information (PMI; Church and Hanks 1990; Turney 2001). In PMI, we assign weights to pairs of words w_i and w_j based on whether their observed joint probability of co-occurrence is larger or smaller than what we would expect given the null hypothesis that w_i and w_j have independent distributions. All three methods learn regularized, reduced dimensional vector representations according to roughly this same goal (Levy and Goldberg, 2014; Cotterell et al., 2017).

B.2 Additional Static Vector Analyses

Table 7 extends Table 4 from the main text with additional variants of word2vec, fastText, and GloVe. The overall picture seems consistent across these variants. For the main text, we simply chose the variant of each model that looked the best to us in terms of capturing meaning dimensions of *break*.

C LLM Structure

The input to the LLMs we consider is always a sequence of tokens $[x_1, \ldots, x_n]$. Each token may correspond to a full word type or a word piece, depending on the tokenization method. For instance, whereas *the* is tokenized as a single unit, *breakage* is likely to be analyzed as two pieces, *break* and *##age*, where the *##* prefix indicates a word-internal piece. This is a detail we set aside in our analyses, since we consider only examples involving *break* and all the models we use analyze *break* as a single token.

The elements of the input sequence are looked up in a static embedding space. The result is a sequence of vectors $[\mathbf{x}_1, \ldots \mathbf{x}_n]$, where each $\mathbf{x}_j$ has dimension d. These are akin to the static representations from models like those in Section 4: there is one vector per word piece and thus no contextual modulation.

The static embeddings are additively combined with one or more separate embeddings that record aspects of each token's position in the sequence. In the simplest case, there is a single positional embedding that is used to create a sequence of vectors $[\mathbf{p}_1, \ldots, \mathbf{p}_n]$, each of dimension d, and we obtain positionally enriched representations as $H_0 = [\mathbf{x}_1 + \mathbf{p}_1, \ldots, \mathbf{x}_1 + \mathbf{p}_n]$. Thus, already at this point in the model, a single word will have different representations depending on where it appears in the input sequence.

The positionally-enriched embeddings are fed into the Transformer architecture itself. This creates numerous interactions between the representations. Each Transformer block $i > 0$ results in a sequence $H_i = [\mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,n}]$ of hidden representations, each one of dimension d. The models we consider have between 12 and 24 of these layers.

We focus on models that are trained in the manner of the BERT model. The core of that training regime is *masked language modeling*, in which elements of the input sequence are randomly masked out or replaced with randomly chosen tokens from the vocabulary, and the task of the model is to learn to assign high likelihood to the actual token, using the entire surrounding sequence. This is a very advanced form of distributional learning, but the core intuition is very similar to that of the single vector models: we are learning linguistic properties entirely from co-occurrence patterns in corpus data.

In our main text, we report results for RoBERTa (Liu et al., 2019), which is '**R**obustly **o**ptimized **BERT** **a**pproach'. We focus on the 'large' variant, which has 24 hidden layers. In Appendix D, we report parallel experiments with the case-sensitive version of the original BERT model as well as two variants of the new DeBERTa model: version 1 and version 3, which introduces some modifications to the pretraining regime. DeBERTa is potentially interesting from the perspective of lexical semantics, because it more fully separates the traditional static embeddings $[\mathbf{x}_1, \ldots \mathbf{x}_n]$ from the positional embeddings $[\mathbf{p}_1, \ldots, \mathbf{p}_n]$. This might be taken to reify word types (as separate from token occurrences) more than the other models do. Like RoBERTa, BERT and the DeBERTa variants have 'base' (12-layer) and 'large' (24-layers) instances. For our main text, we chose to focus on RoBERTa-large because it seems slightly better overall than the rest, but our findings indicate that all these models perform about the same in our evaluations, suggesting that all of them can support lexical semantic investigation.

1.	break	11.	before
2.	breaking	12.	put
3.	broke	13.	start
4.	breaks	14.	take
5.	set	15.	trying
6.	try	16.	could
7.	chance	17.	to
8.	time	18.	broken
9.	again	19.	end
10.	back	20.	finally

(a) GloVe, Wikipedia+Gigaword, 300d.

1.	break	11.	weeks
2.	time	12.	start
3.	breaks	13.	last
4.	before	14.	end
5.	then	15.	broke
6.	take	16.	again
7.	days	17.	next
8.	after	18.	maybe
9.	let	19.	leave
10.	up	20.	down

(b) GloVe, Twitter, 2B tweets, 200d.

1.	break	11.	get
2.	breaks	12.	out
3.	breaking	13.	trying
4.	broke	14.	we
5.	going	15.	broken
6.	let	16.	again
7.	away	17.	come
8.	take	18.	down
9.	up	19.	make
10.	’ll	20.	before

(c) GloVe, Common Crawl 42B tokens, 300d.

1.	break	11.	up
2.	breaks	12.	trying
3.	breaking	13.	away
4.	end	14.	start
5.	broke	15.	get
6.	down	16.	again
7.	take	17.	’ll
8.	let	18.	back
9.	going	19.	out
10.	leave	20.	off

(d) GloVe, Common Crawl 840B tokens, 300d (from Table 4).

1.	breaks	11.	brief_respite
2.	breaking	12.	Nadal_netted_forehand
3.	broke	13.	loosen
4.	broken	14.	smash
5.	Break	15.	rip
6.	Breaking	16.	overhit_forehand
7.	breather	17.	miscued_forehand
8.	shatter	18.	cut
9.	crack	19.	slip
10.	breaker	20.	Breaks

(e) word2vec, GoogleNews, 300d (from Table 4).

1.	break	11.	follow
2.	breaks	12.	smash
3.	breaking	13.	BREAK
4.	broke	14.	knock
5.	Break	15.	water-main
6.	broken	16.	miss
7.	crack	17.	tie
8.	take	18.	go
9.	shatter	19.	relax
10.	fix	20.	start

(f) fastTest WikiNews, 300d.

1.	break	11.	breakin
2.	breaks	12.	breaked
3.	breaking	13.	broken
4.	breake	14.	legbreak
5.	re-break	15.	reak
6.	break-	16.	semi-break
7.	unbreak	17.	minibreak
8.	breakes	18.	breaker
9.	break.	19.	breaking-down
10.	broke	20.	tea-break

(g) fastTest WikiNews, subword modeling, 300d (from Table 4).

1.	break	11.	break.The
2.	breaks	12.	break.I
3.	breaking	13.	break.It
4.	Break	14.	break.This
5.	broke	15.	broken
6.	break.And	16.	break.So
7.	Breaking	17.	break.In
8.	break.	18.	break-
9.	BREAK	19.	breack
10.	Breaks	20.	break.That

(h) fastTest Common Crawl 600B tokens, 300d.

1.	break	11.	take
2.	breaks	12.	broken
3.	breaking	13.	re-break
4.	Break	14.	breake
5.	broke	15.	abreak
6.	break.	16.	break.But
7.	break.And	17.	break-
8.	rebreak	18.	break.What
9.	break.So	19.	bend
10.	breack	20.	break.That

(i) fastTest, Common Crawl 600B tokens, subword modeling, 300d.

Table 7: Static embedding spaces: closest neighbors of *break*.

D Additional Probing Results

Table 8a gives meaning-class probing results for all of the models described in Appendix C, and Table 8b provides a parellel set of results for the construction-type probes. The models are very consistent with each other in terms of layer-wise trends and overall performance. Only the DeBERTa variant stands out as showing differences that may be truly substantive.

		Probe	Control	Selectivity
bert-base-cased	1	0.64	0.04	0.60
	6	0.80	0.03	0.77
	12	0.81	0.03	0.78
bert-large-cased	1	0.65	0.04	0.61
	6	0.78	0.03	0.75
	12	0.83	0.03	0.80
	18	0.83	0.03	0.81
	24	0.84	0.03	0.81
deberta-base	1	0.72	0.03	0.68
	6	0.81	0.03	0.78
	12	0.85	0.03	0.82
deberta-large	1	0.72	0.04	0.68
	6	0.84	0.03	0.81
	12	0.81	0.04	0.78
	18	0.78	0.04	0.74
	24	0.83	0.03	0.81
deberta-v3-base	1	0.70	0.04	0.65
	6	0.84	0.03	0.81
	12	0.75	0.03	0.72
deberta-v3-large	1	0.66	0.04	0.62
	6	0.84	0.04	0.80
	12	0.83	0.03	0.79
	18	0.80	0.04	0.77
	24	0.79	0.04	0.75
roberta-base	1	0.66	0.03	0.63
	6	0.81	0.04	0.78
	12	0.83	0.03	0.80
roberta-large	1	0.33	0.03	0.30
	6	0.81	0.03	0.79
	12	0.83	0.03	0.80
	18	0.80	0.03	0.76
	24	0.86	0.03	0.83

(a) Meaning class.

		Probe	Control	Selectivity
bert-base-cased	1	0.75	0.34	0.40
	6	0.93	0.34	0.60
	12	0.95	0.33	0.63
bert-large-cased	1	0.72	0.33	0.39
	6	0.91	0.34	0.57
	12	0.94	0.33	0.62
	18	0.97	0.35	0.62
	24	0.97	0.33	0.63
deberta-base	1	0.88	0.34	0.54
	6	0.96	0.34	0.62
	12	0.97	0.32	0.64
deberta-large	1	0.86	0.33	0.53
	6	0.96	0.33	0.63
	12	0.96	0.33	0.64
	18	0.95	0.34	0.61
	24	0.96	0.34	0.63
deberta-v3-base	1	0.87	0.32	0.54
	6	0.96	0.34	0.62
	12	0.94	0.32	0.61
deberta-v3-large	1	0.80	0.34	0.45
	6	0.94	0.34	0.61
	12	0.96	0.33	0.64
	18	0.97	0.32	0.65
	24	0.95	0.36	0.60
roberta-base	1	0.82	0.33	0.49
	6	0.96	0.34	0.62
	12	0.96	0.32	0.64
roberta-large	1	0.50	0.33	0.17
	6	0.94	0.34	0.60
	12	0.96	0.33	0.63
	18	0.96	0.35	0.61
	24	0.97	0.32	0.65

(b) Construction type.

Table 8: Full probing results.

E Full Meaning Class Distribution

Table 9 gives the full set of meaning classes, with their counts, from the dataset of Petersen 2020. There are 72 classes in all. Our meaning-class probing experiments use only the 27 classes with at least 10 examples. Our construction-type probing experiments use the full dataset.

Meaning class	
separate_into_parts	150
end	126
decipher	62
break_down_separate_into_parts	61
violate	59
break_up_separate_into_parts	35
surpass	34
break_down_destroy	31
break_into_intrude	28
reveal	26
appear	25
break_through_pass_through	24
render_inoperable	23
unclassified	21
break_down_render_inoperable	21
break_free_escape	19
break_down_succumb	18
cause_to_fail	17
break_up_end_relationship	17
break_up_end	16
break_out_escape	15
break_even_profit=loss	14
succumb	13
break_out_start	12
experience_sorrow	11
break_away_detach	10
break_off_end	10
break_in_enter	9
break_apart_detach	9
break_off_detach	7
break_for_pause	7
destroy	6
break_into_start	6
pioneer	6
lessen	6
break_with_end_relationship	5
break_open_open	5
break_in_interrupt	5
break_loose_detach	5
begin_construction	4
eat_with_sb	4
change	4
break_from_detach	3
cost_too_much	3
break_loose_start	3
break_through_succeed	3
break_up_destroy	3
break_down_unclassified	3
show_disagreement_with_group	3
slow_down	3
begin_to_sweat	2
break_out_unclassified	2
break_down_pause	2
break_up_unclassified	2
happen	2
break_out_separate_into_parts	2
break_out_prepare_for_consumption	2
break_in_mould_shoes	2
break_down_fail	2
dismantle_camp	2
break_loose_escape	1
break_into_unclassified	1
break_off_stop	1
go_bankrupt	1
break_away_pause	1
break_in_train	1
break_past_pass_through	1
tame	1
break_in_unclassified	1
break_with_detach	1
break_out_have_skin_eruption	1
break_beef	1

Table 9: Full meaning-class distribution.

F Examples Selected as Theoretically Relevant

Here we provide the full set of examples extracted from our dataset using the procedure described in Section 5.3 and then selected by us as interesting for lexical semantic theory. The examples in bold are those that appear in Table 6.

	Meaning		Construction		
Sentence	Gold	Predicted	Gold	Predicted	Notes
Most of this information exchange takes place through what are known as newsgroups, which essentially just break all this international online babble up into different topics and areas of interest.	break_up_ separate_ into_parts	break_ down_ separate_ into_parts	causative	causative	Both senses seem active or possible.
What happens, when groups break up that means somebody got caught stealing the money or some guy does n't like it because another guy's a bigger star- KING: Or he married someone who- Mr. GATLIN: Right@!KING.	break_up_ separate_ into_parts	break_up_ end_ relationship	unacc.	unacc.	Both senses seem active.
Wind, naturally acidic rain, and physical processes such as freezethaw cycles also break down rock.	break_ down_ separate_ into_parts	break_ down_ destroy	causative	causative	Both senses seem active.
But her husband was determined not to break up the family.	break_up_ separate_ into_parts	break_up_ end_ relationship	causative	causative	Both senses seem active.
It was like you knew something, like you knew the story was getting ready to break again.	reveal	appear	unacc.	unacc.	Both senses seem active.
"So why tell the whole story now? Somebody, some male, has got to be willing to break this code of silence," he says.	violate	end	causative	causative	Contextual entailment relation between the two labels.
Then too, stress can also work to break down the immune system, increasing the likelihood of respiratory and creating gastrointestinal and nervous disorders.	break_ down_ render_ inoperable	break_ down_ destroy	causative	causative	Contextual entailment relation between the two labels.
If you deprive yourself, you're going to break your diet and fall off it.	violate	end	causative	causative	Contextual entailment relation between the two labels.
Sen. BOB KERREY: I don't want to destroy Social Security or break a commitment.	violate	end	causative	causative	Contextual entailment relation between the two labels.
So they forwarded the pictures to Madrid, where another officer noticed some printing on a towel that helped break the case.	decipher	end	causative	causative	Contextual entailment relation between the two labels.
It's one example of how the standard model might break down.	break_ down_ render_ inoperable	break_ down_ succumb	unacc.	unacc.	Contextual entailment relation between the two labels.
Instead, crews will break down the structures over three years, releasing the water in the reservoirs at a rate that's more manageable for the animals and the people who live in the area.	break_ down_ separate_ into_parts	break_ down_ destroy	causative	causative	Contextual entailment relation between the two labels.

"The Comes would try to break the Saxon ranks with a mounted charge.	separate_into_parts	end	causative	causative	Contextual entailment relation between the two labels.
Then the troops break formation and move out to a formation and stand guard, even from above, making sure the so-called detainees are safely behind the fence.	separate_into_parts	end	causative	causative	Contextual entailment relation between the two labels.
The Soviet Union will break up into between six and twenty (or more) separate countries.	break_up_separate_into_parts	break_up_end	unacc.	unacc.	Contextual entailment relation between the two labels.
It didn't take being an ICU exec to break the code: trade secret.	decipher	violate	causative	causative	Genuine uncertainty about which sense is intended.
@(Soundbite-of-music)@!Mr-GELB: (Singing) Tell me who's going to pick up the pieces when you start to break down.	break_down_separate_into_parts	break_down_succumb	unacc.	unacc.	Gold meaning is literal; predicted meaning is metaphorical.
"People have so many problems overcoming the disputes that occur when families break up, and then to have to overcome the barriers that government puts up when they hold on to the money, literally sends children to bed hungry," says Jensen.	break_up_end_relationship	break_up_separate_into_parts	unacc.	unacc.	Gold meaning is metaphorical; predicted meaning is literal.
"I just don't want to break up such happy couples.	break_up_end_relationship	break_up_separate_into_parts	causative	causative	Gold meaning is metaphorical; predicted meaning is literal.
I had to break it up.	break_up_end	break_up_separate_into_parts	causative	causative	Gold meaning is metaphorical; predicted meaning is literal.
Will the kibbutz movement "renew its days as of old" when it has recovered from the present crisis, as did the Hutterites at several points in their history? Will it continue to exist, but in a radically revised form, like Amana and other colonies? Or will the kibbutzim simply break up, to form part of the historical heritage of the Israeli nation, and no more– like so many of the well-preserved sites that aroused such powerful feelings in Yaakov Oved? The considerations I have advanced here seem to militate against the first of these possibilities and favor one of the others– perhaps a mixture of both.	break_up_end	break_up_separate_into_parts	unacc.	unacc.	Gold meaning is metaphorical; predicted meaning is literal.
The past few days had consisted of a simple routine of drinking melted snow to stay hydrated and sleeping while waiting for the storm to break.	appear	end	unacc.	unacc.	Model prediction may be correct.
A small pair of scissors will easily break the seal, but bringing those scissors in your carry-on bag may no longer be permitted.	separate_into_parts	decipher	causative	causative	The decipher prediction seems sensible given that a seal is like a lock or (easy) code that needs to be overcome.
Patients will sometimes break out in a spontaneous recitation of the rosary	break_out_start	break_out_start	unacc.	unerg.	The modifier "spontaneous" seems to affect agentivity and perhaps also argument structure.

Millennial darlings began to break down like virus-ridden websites, from the supercharged (Qualcomm, Oracle) to the superhyped (Amazon, Yahoo!) to the just plain super (Sun, Lucent, AOL).	break_down_succumb	break_down_render_inoperable	unacc.	unacc.	There is a comparison of "millennial darlings" with "virus-ridden websites". The gold meaning may apply to "millennial darlings" and the predicted meaning to "virus-ridden websites".
I felt disappointed, but I waited, hoping the clouds would break.	separate_into_parts	appear	unacc.	unacc.	Weather events are persistently uncertain about whether they describe the start or end of something.

G Visualizations

Figure 2 uses t-SNE to visualize *break* embeddings from layer 1 of RoBERTa-large, and Figure 3 shows the embeddings from layer 24. We use color to distinguish the top 10 meaning classes (and the rest are gray). Underlined examples are unergative and boxed examples are unaccusative. The layer 24 visualization has much more structure than the layer 1 visualization. By layer 24, the model seems strikingly well-aligned with the meaning categories and construction types, as evidenced by how examples with the same color cluster together, and how the construction type annotations also cluster within those spaces. The other models we consider show effectively these same patterns.

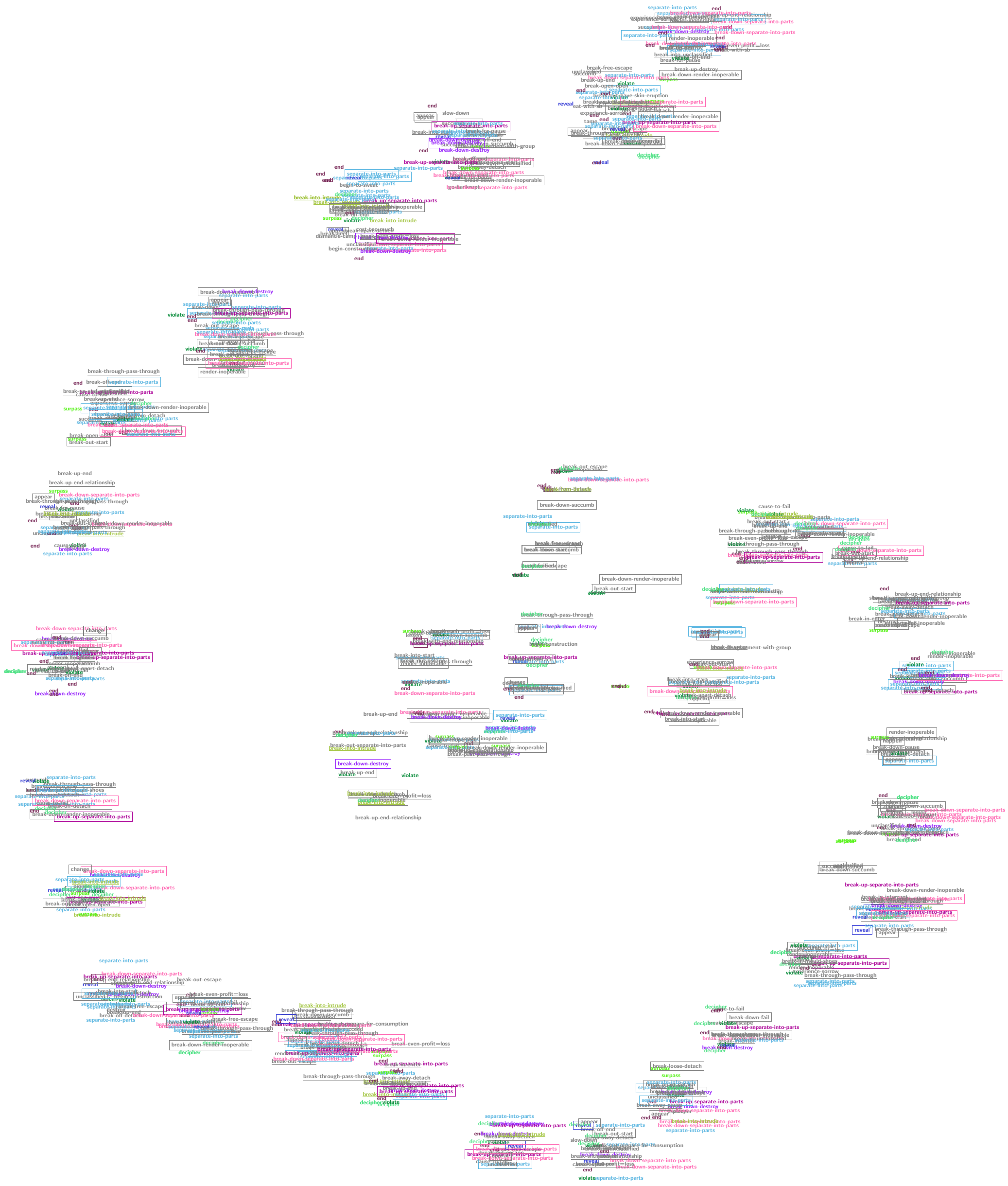


Figure 2: t-SNE of break with RoBERTa-large, layer 1

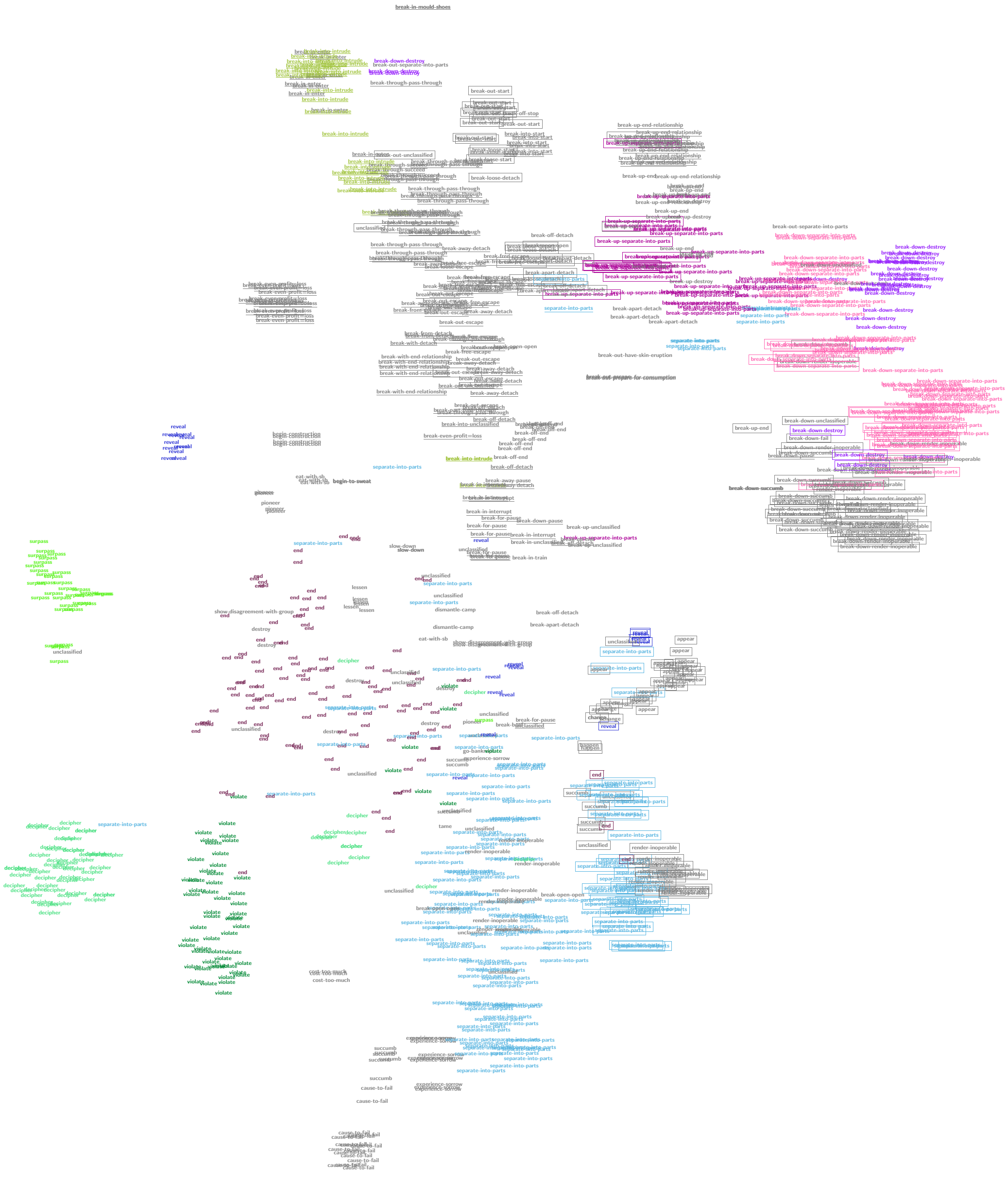


Figure 3: t-SNE of break with RoBERTa-large, layer 24