# Neural reality of argument structure constructions

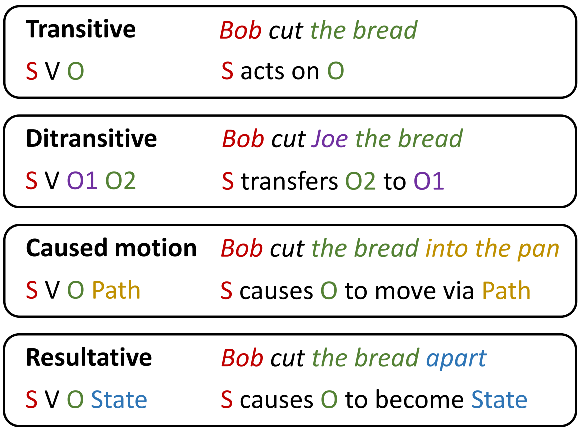
**Bai Li**[1,4], **Zining Zhu**[1,4] **Guillaume Thomas**[2], **Frank Rudzicz**[1,4,5], **Yang Xu**[1,3,4]
[1] University of Toronto, Department of Computer Science
[2] University of Toronto, Department of Linguistics
[3] University of Toronto, Cognitive Science Program
[4] Vector Institute for Artificial Intelligence [5] Unity Health Toronto
{bai, zining, frank, yangxu}@cs.toronto.edu
guillaume.thomas@utoronto.ca

## Abstract

In lexicalist linguistic theories, argument structure is assumed to be predictable from the meaning of verbs. As a result, the verb is the primary determinant of the meaning of a clause. In contrast, construction grammarians propose that argument structure is encoded in constructions (or form-meaning pairs) that are distinct from verbs. Decades of psycholinguistic research have produced substantial empirical evidence in favor of the construction view. Here we adapt several psycholinguistic studies to probe for the existence of argument structure constructions (ASCs) in Transformer-based language models (LMs). First, using a sentence sorting experiment, we find that sentences sharing the same construction are closer in embedding space than sentences sharing the same verb. Furthermore, LMs increasingly prefer grouping by construction with more input data, mirroring the behaviour of non-native language learners. Second, in a "Jabberwocky" priming-based experiment, we find that LMs associate ASCs with meaning, even in semantically nonsensical sentences. Our work offers the first evidence for ASCs in LMs and highlights the potential to devise novel probing methods grounded in psycholinguistic research.

## 1 Introduction

Pretrained Transformer-based language models (LMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) have recently achieved impressive results on many natural language tasks, spawning a new interdisciplinary field of aligning LMs with linguistic theory and probing the linguistic capabilities of LMs (Linzen and Baroni, 2021). Most probing work so far has investigated the linguistic knowledge of LMs on phenomena such as agreement, binding, licensing, and movement (Warstadt et al., 2020a; Hu et al., 2020) with a particular focus on determining whether a sentence is linguistically acceptable (Schütze, 1996). Relatively little work has attempted to determine whether the linguistic knowledge induced by LMs is more similar to a formal grammar of the sort postulated by mainstream generative linguistics (Chomsky, 1965, 1981, 1995), or to a network of form-meaning pairs as advocated by construction grammar (Goldberg, 1995, 2006).



**Transitive** *Bob cut the bread*
S V O S acts on O

**Ditransitive** *Bob cut Joe the bread*
S V O1 O2 S transfers O2 to O1

**Caused motion** *Bob cut the bread into the pan*
S V O Path S causes O to move via Path

**Resultative** *Bob cut the bread apart*
S V O State S causes O to become State

Figure 1: Four argument structure constructions (ASCs) used by Bencini and Goldberg (2000), with example sentences (top right). Constructions are mappings between form (bottom left) and meaning (bottom right).

One area where construction grammar disagrees with many generative theories of language is in the analysis of the argument structure of verbs, that is, the specification of the number of arguments that a verb takes, their semantic relation to the verb, and their syntactic form (Levin and Rappaport Hovav, 2005). Lexicalist theories were long dominant in generative grammar (Chomsky, 1981; Kaplan and Bresnan, 1982; Pollard and Sag, 1987). In lexicalist theories, argument structure is assumed to be encoded in the lexical entry of the verb: for

*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 7410 - 7423
May 22-27, 2022 ©2022 Association for Computational Linguistics

example, the verb *visit* is lexically specified as being transitive and as requiring a noun phrase object (Chomsky, 1986). In contrast, construction grammar suggests that argument structure is encoded in form-meaning pairs known as *argument structure constructions* (ASCs, Figure 1), which are distinct from verbs. The argument structure of a verb is determined by pairing it with an ASC (Goldberg, 1995). To date, a substantial body of psycholinguistic work has provided evidence for the psychological reality of ASCs in sentence sorting (Bencini and Goldberg, 2000; Gries and Wulff, 2005), priming (Ziegler et al., 2019), and novel verb experiments (Kaschak and Glenberg, 2000; Johnson and Goldberg, 2013).

Here we connect basic research in ASCs with neural probing by adapting several psycholinguistic studies to Transformer-based LMs and show evidence for the neural reality of ASCs. Our first case study is based on sentence sorting (Bencini and Goldberg, 2000); we discover that in English, German, Italian, and Spanish, LMs consider sentences that share the same construction to be more semantically similar than sentences sharing the main verb. Furthermore, this preference for constructional meaning only manifests in larger LMs (trained with more data), whereas smaller LMs rely on the main verb, an easily accessible surface feature. Human experiments with non-native speakers found a similarly increased preference for constructional meaning in more proficient speakers (Liang, 2002; Baicchi and Della Putta, 2019), suggesting commonalities in language acquisition between LMs and humans.

Our second case study is based on nonsense "Jabberwocky" sentences that nevertheless convey meaning when they are arranged in constructional templates (Johnson and Goldberg, 2013). We adapt the original priming experiment to LMs and show that RoBERTa is able to derive meaning from ASCs, even without any lexical cues. This finding offers counter-evidence to earlier claims that LMs are relatively insensitive to word order when constructing sentence meaning (Yu and Ettinger, 2020; Sinha et al., 2021). Our source code and data are available at: `https://github.com/SPOClab-ca/neural-reality-constructions`.

## 2 Psycholinguistic background

### 2.1 Construction grammar and ASCs

Construction grammar is a family of linguistic theories proposing that all linguistic knowledge consists of *constructions*: pairings between form and meaning where some aspects of form or meaning are not predictable from their parts (Fillmore et al., 1988; Kay and Fillmore, 1999; Goldberg, 1995, 2006). Common examples include idiomatic expressions such as *under the weather* (meaning  "to feel unwell"), but many linguistic patterns are constructions, including morphemes (e.g., *-ify*), words (e.g., *apple*), and abstract patterns like the ditransitive and passive. In contrast to lexicalist theories of argument structure, construction grammar rejects the dichotomy between syntax and lexicon. In contrast to transformational grammar, it rejects any distinction between surface and underlying structure.

We focus on a specific family of constructions for which there is an ample body of psycholinguistic evidence: argument structure constructions (ASCs). ASCs are constructions that specify the argument structure of a verb (Goldberg, 1995). In the lexicalist, verb-centered view, argument structure is a lexical property of the verb, and the main verb of a sentence determines the form and meaning of the sentence (Chomsky, 1981; Kaplan and Bresnan, 1982; Pollard and Sag, 1987; Levin and Rappaport Hovav, 1995). For example, *sneeze* is intransitive (allowing no direct object) and *hit* is transitive (requiring one direct object). However, lexicalist theories encounter difficulties with sentences like *"he sneezed the napkin off the table"* since intransitive verbs are not permitted to have object arguments.

Rather than assuming multiple implausible senses for the verb *"sneeze"* with different argument structures, Goldberg (1995) proposed that ASCs operate on an arbitrary verb, altering its argument structure while at the same time modifying its meaning. For example, the *caused-motion* ASC adds a direct object and a path argument to the verb *sneeze*, with the semantics of causing the object to move along the path. Other ASCs include the transitive, ditransitive, and resultative (Figure 1), which specify the argument structure of a verb and interact with its meaning in different ways.

### 2.2 Psycholinguistic evidence for ASCs

**Sentence sorting.** Several psycholinguistic studies have found evidence for argument structure con-

| | **Transitive** | **Ditransitive** | **Caused-motion** | **Resultative** |
|---|---|---|---|---|
| **Throw** | Anita threw the hammer. | Chris threw Linda the pencil. | Pat threw the keys onto the roof. | Lyn threw the box apart. |
| **Get** | Michelle got the book. | Beth got Liz an invitation. | Laura got the ball into the net. | Dana got the mattress inflated. |
| **Slice** | Barbara sliced the bread. | Jennifer sliced Terry an apple. | Meg sliced the ham onto the plate. | Nancy sliced the tire open. |
| **Take** | Audrey took the watch. | Paula took Sue a message. | Kim took the rose into the house. | Rachel took the wall down. |

Table 1: Stimuli from Bencini and Goldberg (2000), consisting of a 4x4 design, with 4 different verbs and 4 different argument structure constructions.

structions using experimental methods. Among these, Bencini and Goldberg (2000) used a sentence sorting task to determine whether the verb or construction in a sentence was the main determinant of sentence meaning. 17 participants were given 16 index cards with sentences containing 4 verbs (*throw, get, slice*, and *take*) and 4 constructions (*transitive, ditransitive, caused-motion*, and *resultative*) and were instructed to sort them into 4 piles by overall sentence meaning (Table 1). The experimenters measured the deviation to a purely verb-based or construction-based sort, and found that on average, the piles were closer to a construction sort.

**Non-native sentence sorting.** The same set of experimental stimuli was used with L2 (non-native) English speakers. Gries and Wulff (2005) ran the experiment with 22 German native speakers, who preferred the construction-based sort over the verb-based sort, showing that constructional knowledge is not limited to native speakers. Liang (2002) ran the experiment on Chinese native speakers of 3 different English levels (46 beginner, 31 intermediate, and 33 advanced), and found that beginners preferred a verb-based sort, while advanced learners produced construction-based sorts similar to native speakers (Figure 2). Likewise, Baicchi and Della Putta (2019) found the same result in Italian native speakers with B1 and B2 English proficiency levels. Overall, these studies show evidence for ASCs in the mental representations of native and L2 English speakers alike, and furthermore, preference for constructional over verb sorting increases with increasing English proficiency.

**Multilingual sentence sorting.** Similar sentence sorting experiments have been conducted in other languages, with varying results. Kirsch (2019) ran a sentence sorting experiment in German with 40 participants and found that they mainly sorted by verb but rarely by construction. Baicchi and Della Putta (2019) ran an experiment with non-native learners of Italian (15 participants of B1 level and 10 participants of B2 level): both groups preferred the constructional sort, and similar to Liang (2002), the B2 learners sorted more by construction than the B1 learners. Vázquez (2004) ran an experiment in Spanish with 16 participants, and found approximately equal proportions of constructions and verb sort. In Italian and Spanish, some different constructions were substituted as not all of the English constructions had an equivalent in these languages; see the appendix for the complete set of stimuli in each language.

**Priming.** Another line of psycholinguistic evidence comes from priming studies. Priming refers to the condition where exposure to a (prior) stimulus influences the response to a later stimulus (Pickering and Ferreira, 2008). Bock and Loebell (1990) found that participants were more likely to produce sentences of a given syntactic structure when primed with a sentence of the same structure; Ziegler et al. (2019) argued that Bock and Loebell (1990) did not adequately control for lexical overlap, and instead, they showed that the construction must be shared for the priming effect to occur, not just shared abstract syntax.

**Novel verbs.** Even with unfamiliar words, there is evidence that constructions are associated with meaning. Kaschak and Glenberg (2000) constructed sentences with novel denominal verbs and found that participants were more likely to interpret a transfer event when the denominal verb was used in a ditransitive sentence (*Tom crutched Lyn an apple*) than a transitive one (*Tom crutched an apple*).

Johnson and Goldberg (2013) used a “Jabberwocky” priming task to show that abstract constructional templates are associated with meaning. Participants were primed with a nonsense sentence of a given construction (e.g., *He daxed her the norp* for the ditransitive construction), followed by a lexical decision task of quickly deciding if a string of characters was a real English word or a non-word. The word in the decision task was semantically congruent with the construction (*gave*) or incongruent (*made*); furthermore, they experimented with target words that were high-frequency (*gave*), low-frequency (*handed*), or semantically related but not associated with the construction (*transferred*). They found priming effects (faster lexical decision times) in all three conditions, with the strongest effect for the high-frequency condition, followed by the low-frequency and the semantically nonassociate conditions.

We adapt several of these psycholinguistic studies to LMs: the sentence sorting experiments in Case study 1, and the Jabberwocky priming experiment in Case study 2. We choose these studies because their designs allow for thousands of stimuli sentences to be generated automatically using templates, avoiding issues caused by small sample sizes from manually constructed sentences.

# 3 Related work in NLP

## 3.1 Linguistic probing of LMs

Many studies have probed for various aspects of syntax in LSTMs and Transformer-based LMs. Linzen et al. (2016) tested LSTMs on their ability to capture subject-verb agreement, using templates to generate test data. This idea was extended by BLiMP (Warstadt et al., 2020a), a suite encompassing 67 linguistic phenomena, including filler-gap effects, NPI licensing, and ellipsis; Hu et al. (2020) released a similar test suite. Template generation is a convenient method to construct stimuli exhibiting specific linguistic properties, but alternative approaches include CoLA (Warstadt et al., 2019), which compiled an acceptability benchmark of sentences drawn from linguistic publications, and Gulordava et al. (2018), who perturbed natural sentences to study LMs' knowledge of agreement on nonsense sentences. We refer to Linzen and Baroni (2021) for a comprehensive review of the linguistic probing literature.

So far, relatively few papers approached LM probing from a construction grammar perspective. Madabushi et al. (2020) probed for BERT's knowledge of constructions via a sentence pair classification task of predicting whether two sentences share the same construction. Their probe was based on data from Dunn (2017), who used an unsupervised algorithm to extract plausible constructions from corpora based on association strength. However, the linguistic validity of these automatically induced constructions is uncertain, and there is currently no human-labelled wide-coverage construction grammar dataset in any language suitable for probing. Other computational work focused on a few specific constructions, such as identifying caused-motion constructions in corpora (Hwang and Palmer, 2015) and annotating constructions related to causal language (Dunietz et al., 2015). Lebani and Lenci (2016) is the most similar to our work: they probed distributional vector space models for ASCs based on the Jabberwocky priming experiment by Johnson and Goldberg (2013).

## 3.2 Psycholinguistic treatment of LMs

Some recent probing studies adapted methods and data from psycholinguistic research, treating LMs as psycholinguistic participants. Using a cloze completion task, Ettinger (2020) found that BERT was less sensitive than humans at commonsense inferences and detecting role reversals, and fails completely at understanding negation. Michaelov and Bergen (2020) compared LM surprisals with the N400 (a measure of human language processing difficulty) across a wide range of conditions; Li et al. (2021) used psycholinguistic stimuli and found that LMs exhibit different layerwise surprisal patterns for morphosyntactic, semantic, and commonsense anomalies. Wilcox et al. (2021) compared LM and human sensitivities to syntactic violations using a maze task to collect human reaction times. Prasad et al. (2019); Misra et al. (2020) investigated whether LMs are sensitive to priming effects like humans. The advantage of psycholinguistic data is that they are carefully constructed by expert linguists to test theories of language processing in humans; however, their small sample size makes it challenging to make statistically meaningful conclusions when the (oft-sparse) experimental stimuli are used to probe a language model.

# 4 Case study 1: Sentence sorting

This section describes our adaptation of the sentence sorting experiments to Transformer LMs.

## 4.1 Methodology

**Models.** To simulate varying non-native English proficiency levels, we use MiniBERTa models (Warstadt et al., 2020b), trained with 1M, 10M,

[1] Bencini and Goldberg (2000) ran the sentence sorting experiment twice, so we take the average of the two runs.

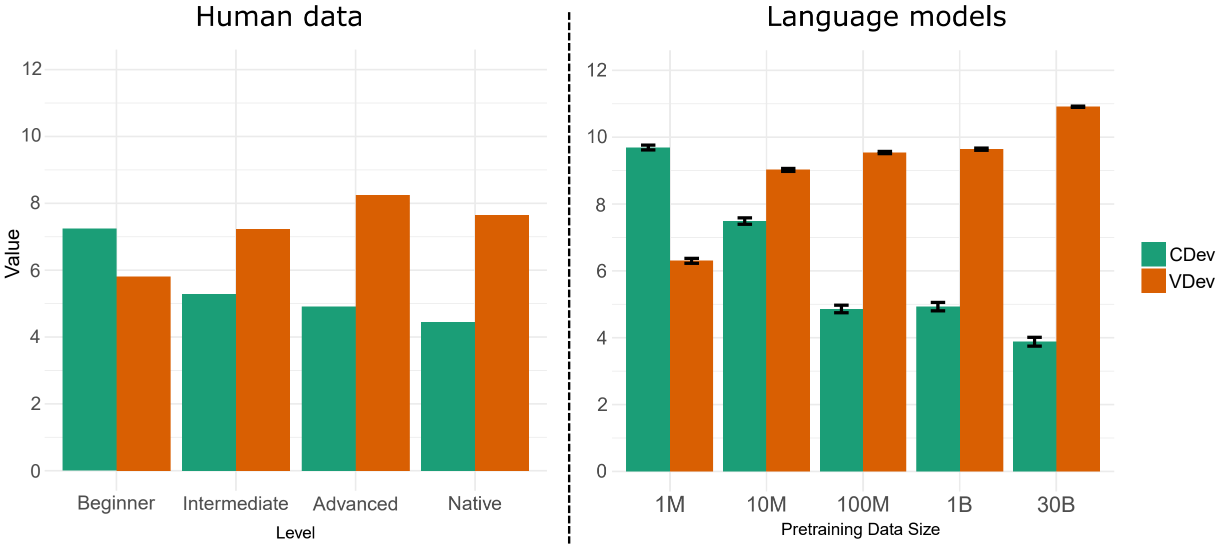


Figure 2: Sentence sorting results for humans and LMs, measured by deviation from pure construction and verb sort (CDev and VDev). Non-native human results are from Liang (2002); native human results from Bencini and Goldberg (2000).[1]LM results are obtained using MiniBERTas (Warstadt et al., 2020b) and RoBERTa (Liu et al., 2019b) on templated stimuli. The MiniBERTa models use between 1M to 1B tokens for pretraining, while RoBERTa uses 30B tokens. Error bars indicate 95% confidence intervals.

100M, and 1B tokens. We also use the base RoBERTa model (Liu et al., 2019b), trained with 30B tokens. In other languages, there are no available pretrained checkpoints with varying amounts of pretraining data, so we use the mBERT model (Devlin et al., 2019) and a monolingual Transformer LM in each language.[2] We obtain sentence embeddings for our models by taking the average of their contextual token embeddings at the second-to-last layer (i.e., layer 11 for base RoBERTa). We use the second-to-last because the last layer is more specialized for the LM pretraining objective and less suitable for sentence embeddings (Liu et al., 2019a).

**Template generation.** We use templates to generate stimuli similar to the 4x4 design in the Bencini and Goldberg (2000) experiment. To ensure an adequate sample size, we run multiple empirical trials. In each trial, we sample 4 random distinct verbs from a pool of 10 verbs that are compatible with all 4 constructions (*cut, hit, get, kick, pull, punch, push, slice, tear, throw*). We then randomly fill in the slots for proper names, objects, and complements for each sentence according to its verb, such that the sentence is semantically coherent, and there is no lexical overlap among the sentences of any construction. Table 3 in the appendix shows a set of template-generated sentences. In English, we generate 1000 sets of stimuli using this procedure; for other languages, we use the original stimuli from their respective publications.

**Evaluation.** Similar to the human experiments, we group the sentence embeddings into 4 clusters (not necessarily of the same size) using agglomerative clustering by Euclidean distance (Pedregosa et al., 2011). We then compute the deviation to a pure construction and pure verb sort using the Hungarian algorithm for optimal bipartite matching. This measures the minimal number of cluster assignment changes necessary to reach a pure construction or verb sort, ranging from 0 to 12. Thus, lower construction deviation indicates that constructional information is more salient in the LM's embeddings.

### 4.2 Results and interpretation

Figure 2 shows the LM sentence sorting results for English. All differences are statistically significant ($p < .001$). The smallest 1M MiniBERTa model is the only LM to prefer verb over construction sorting, and as the amount of pretraining data grows, the LMs increasingly prefer sorting by construction instead of by verb. This closely mirrors the trend observed in the human experiments.

The results for multilingual sorting are shown in Figure 3. Both mBERT and the monolingual LMs consistently prefer constructional sorting over verb

[2] We use monolingual German and Italian models from `https://github.com/dbmdz/berts`, and the monolingual Spanish model from Cañete et al. (2020).

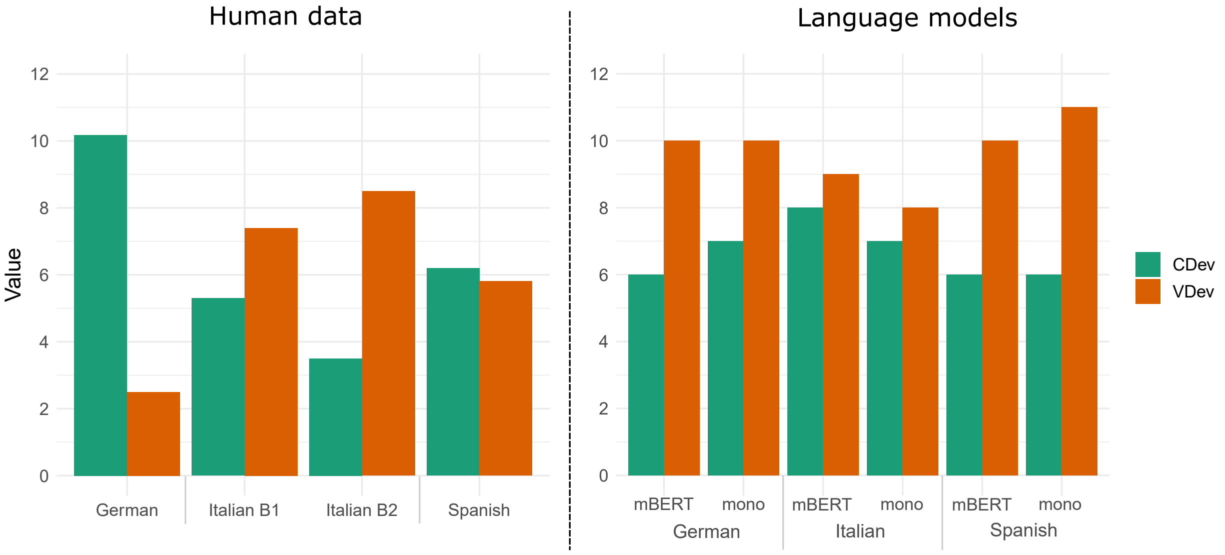


Figure 3: Multilingual sentence sorting results for German (Kirsch, 2019), Italian (Baicchi and Della Putta, 2019), and Spanish (Vázquez, 2004). LM results are obtained using the same stimuli; we use both mBERT and a monolingual LM for each language.

sorting in all three languages, whereas the results from the human experiments are less consistent.

Our results show that RoBERTa can generalize meaning from abstract constructions without lexical overlap. Only larger LMs and English speakers of more advanced proficiency are able to make this generalization, while smaller LMs and less proficient speakers derive meaning more from surface features like lexical content. This finding agrees with Warstadt et al. (2020b), who found that larger LMs have an inductive bias towards linguistic generalizations, while smaller LMs have an inductive bias towards surface generalizations; this may explain the success of large LMs on downstream tasks. A small quantity of data (10M tokens) is sufficient for LMs to prefer the constructional sort, indicating that ASCs are relatively easy to learn: roughly on par with other types of linguistic knowledge, and requiring less data than commonsense knowledge (Zhang et al., 2021; Liu et al., 2021).

We note some limitations in these results, and reasons to avoid drawing unreasonably strong conclusions from them. Human sentence sorting experiments can be influenced by minor differences in the experimental setup: Bencini and Goldberg (2000) obtained significantly different results in two runs that only differed on the precise wording of instructions. In the German experiment (Kirsch, 2019), the author hypothesized that the participants were influenced by a different experiment that they had completed before the sentence sorting one. Given this experimental variation, we cannot attribute differences across languages to differences in their linguistic typology. Although LMs do not suffer from the same experimental variation, we cannot conclude statistical significance from the multilingual experiments, where only one set of stimuli is available in each language.

## 5 Case study 2: Jabberwocky constructions

We next adapt the "Jabberwocky" priming experiment from Johnson and Goldberg (2013) to LMs, and make several changes to the original setup to better assess the capabilities of LMs. Priming is a standard experimental paradigm in psycholinguistic research, but it is not directly applicable to LMs: existing methods simulate priming either by applying additional fine-tuning (Prasad et al., 2019), or by concatenating sentences that typically do not co-occur in natural text (Misra et al., 2020). Therefore, we instead propose a method to probe LMs for the same linguistic information using only distance measurements on their contextual embeddings.

### 5.1 Methodology

**Template generation.** We generate sentences for the four constructions randomly using the templates in Table 2. Instead of filling nonce words like *norp* into the templates as in the original study, we take an approach similar to Gulordava et al. (2018) and generate 5000 sentences for each construction

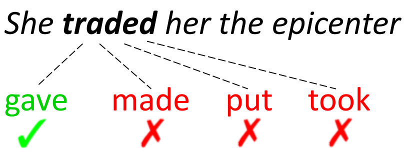


Figure 4: In our adapted Jabberwocky experiment, we measure the Euclidean distance from the Jabberwocky verb (*traded*) to the 4 prototype verbs, of which 1 is congruent (✓) with the construction of the sentence, and 3 are incongruent (✗).

| Construction | Template / Examples |
|---|---|
| Ditransitive | S/he V-ed him/her the N.<br>*She traded her the epicenter.*<br>*He flew her the donut.* |
| Resultative | S/he V-ed it Adj.<br>*He cut it seasonal.*<br>*She surged it civil.* |
| Caused-motion | S/he V-ed it on the N.<br>*He registered it on the diamond.*<br>*She awarded it on the corn.* |
| Removal | S/he V-ed it from him/her.<br>*He declined it from her.*<br>*She drove it from him.* |

Table 2: Templates and example sentences for the Jabberwocky construction experiments. The templates are identical to the ones used in Johnson and Goldberg (2013), except that we use random real words instead of nonce words.

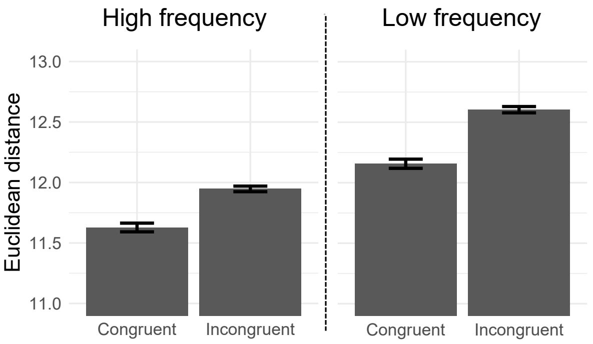


Figure 5: Euclidean distance between Jabberwocky and prototype verbs for congruent and incongruent conditions. Error bars indicate 95% confidence intervals.

by randomly filling real words of the appropriate part-of-speech into construction templates (Table 2). This gives nonsense sentences like *"She traded her the epicenter"*; we refer to these random words as *Jabberwocky words*. By using real words, we avoid any potential instability from feeding tokens into the model that it has never seen during pretraining. We obtain a set of singular nouns, past tense verbs, and adjectives from the Penn Treebank (Marcus et al., 1993), excluding words with fewer than 10 occurrences.

**Verb embeddings.** Our probing strategy is based on the assumption that the contextual embedding for a verb captures its meaning in context. Therefore, if LMs associate ASCs with meaning, we should expect the contextual embedding for the Jabberwocky verb to contain the meaning of the construction. Specifically, we measure the Euclidean distance to a *prototype* verb for each construction (Figure 4). These are verbs that Johnson and Goldberg (2013) selected whose meaning closely resembles the construction's meaning: *gave*, *made*, *put*, and *took* for the ditransitive, resultative, caused-motion, and removal constructions, respectively.[3] We also run the same setup using lower frequency prototype verbs from the same study: *handed*, *turned*, *placed*, and *removed*.[4] As a control, we measure the Euclidean distance to the prototype verbs of the other three unrelated constructions.

The prototype verb embeddings are generated by taking the average across their contextual embeddings across a 4M-word subset of the British National Corpus (BNC; Leech (1992)). We use the second-to-last layer of RoBERTa-base, and in cases where a verb is split into multiple subwords, we take the embedding of the first subword token as the verb embedding.

### 5.2 Results and interpretation

We find that the Euclidean distance between the prototype and Jabberwocky verb embeddings is significantly lower ($p < .001$) when the verb is congruent with the construction than when they are incongruent, and this is observed for both high and low-frequency prototype verbs (Figure 5). Examining the individual constructions and verbs (Figure 6), we note that in the high-frequency scenario, the lowest distance prototype verb is always the congruent one, for all four constructions. In the low-frequency scenario, the result is less consis-

[3] The reader may notice that the four constructions here are slightly different from Bencini and Goldberg (2000): the transitive construction is replaced with the removal construction in Johnson and Goldberg (2013).

[4] Johnson and Goldberg (2013) also included a third experimental condition using four verbs that are semantically related but not associated with the construction, but one of the verbs is very low-frequency (*ousted*), so we exclude this condition in our experiment.

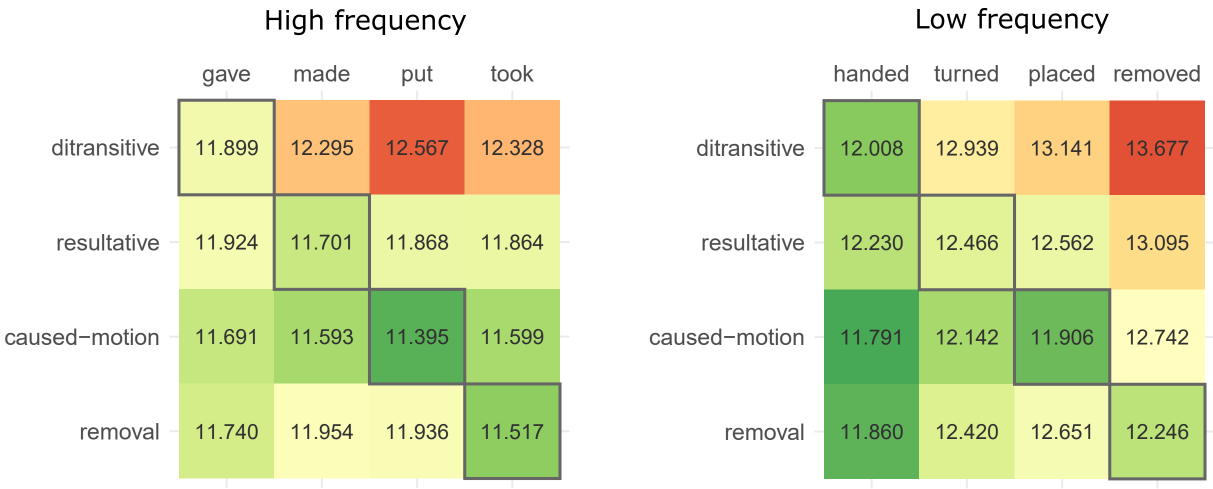


Figure 6: Mean Euclidean distance between Jabberwocky and prototype verbs in each verb-construction pair. Diagonal entries (gray border) are the congruent conditions; off-diagonal entries are incongruent.

tent: the congruent verb is not always the lowest distance one, although it is always still at most the second-lowest distance out of the four.

The main result holds for both high and low-frequency scenarios, but the correct prototype verb is associated more consistently in the high-frequency case. This agrees with Wei et al. (2021), who found that LMs have greater difficulty learning the linguistic properties of less frequent words. We also note that the Euclidean distances are higher overall in the low-frequency scenario, which is consistent with previous work that found lower frequency words to occupy a peripheral region of the embedding space (Li et al., 2021).

### 5.3 Potential confounds

In any experiment, one must be careful to ensure that the observed patterns are due to the phenomenon under investigation rather than confounding factors. We discuss potential confounds arising from lexical overlap, anisotropy of contextual embeddings, and neighboring words.

**Lexical overlap**. The randomized experiment design ensures that the Jabberwocky words cannot be lexically biased towards any construction, since each verb is equally likely to occur in every construction. Technically, the lexical content in the four constructions are not identical: i.e., words such as *"from"* (occurring only in the removal construction) or *"on"* (in the caused-motion construction) may provide hints to the sentence meaning. However, the ditransitive and resultative constructions do not contain any such informative words, yet RoBERTa still associates the correct prototype verb for these constructions, so we consider it unlikely to be relying solely on lexical overlap. There is substantial evidence that RoBERTa is able to associate abstract constructional templates with their meaning without lexical cues. This result is perhaps surprising, given that previous work found that LMs are relatively insensitive to word order in compositional phrases (Yu and Ettinger, 2020) and downstream inference tasks (Sinha et al., 2021; Pham et al., 2021), where their performance can be largely attributed to lexical overlap.

**Anisotropy**. Recent probing work have found that contextual embeddings suffer from anisotropy, where embeddings lie in a narrow cone and have much higher cosine similarity than expected if they were directionally uniform (Ethayarajh, 2019). Furthermore, a small number of dimensions dominate geometric measures such as Euclidean and cosine distance, resulting in a degradation of representation quality (Kovaleva et al., 2021; Timkey and van Schijndel, 2021). Since our experiments rely heavily on Euclidean distance, anisotropy is a significant concern. Following Timkey and van Schijndel (2021), we perform standardization by subtracting the mean vector and dividing each dimension by its standard deviation, where the mean and standard deviation for each dimension is computed from a sample of the BNC. We observe little difference after standardization: in both the high and low frequency scenarios, the Euclidean distances are lower for the congruent than the incongruent conditions, by a similar margin compared to the original experiment without standardization. We also run standardization on the first case study, and find that the results remain essentially unchanged: smaller LMs still prefer verb sorting while larger LMs prefer construction sorting. Thus, neither of our experiments appear to be affected by anisotropy.

**Neighboring words.** A final confounding factor is our assumption that RoBERTa's contextual embeddings represent word meaning, when in reality, they contain a mixture of syntactic and semantic information. Contextual embeddings are known to contain syntax trees (Hewitt and Manning, 2019) and linguistic information about neighboring words in a sentence (Klafka and Ettinger, 2020); although previous work did not consider ASCs, it is plausible that our verb embeddings leak information about the sentence's construction in a similar manner. If this were the case, the prototype verb embedding for *gave* would contain not only the semantics of transfer that we intended, but also information about its usual syntactic form[5] of *"S gave NP1 NP2"*, and both would be captured by our Euclidean distance measurement. Controlling for this syntactic confound is difficult – one could alternatively probe for transfer semantics without syntactic confounds using a natural language inference setup (e.g., whether the sentence entails the statement *"NP1 received NP2"*), but we leave further exploration of this idea to future work.

## 6 Conclusion

We find evidence for argument structure constructions in Transformer language models from two separate angles: sentence sorting and Jabberwocky construction experiments. Our work extends the existing body of literature on LM probing by taking a constructionist instead of generative approach to linguistic probing. Our sentence sorting experiments identified a striking resemblance between humans' and LMs' internal language representations as LMs are exposed to increasing quantities of data, despite the differences between neural language models and the human brain. Our two studies suggest that LMs are able to derive meaning from abstract constructional templates with minimal lexical overlap. Both sets of experiments were inspired by psycholinguistic studies, which we adapted to fit the capabilities of LMs – this illustrates the potential for future work on grounding LM probing methodologies in psycholinguistic research.

[5] Bresnan and Nikitina (2003) estimated that 87% of usages of the word *"give"* occur in the ditransitive construction.

## Acknowledgements

YX is funded through an NSERC Discovery Grant, a SSHRC Insight Grant, and an Ontario ERA award.

## References

Annalisa Baicchi and Paolo Della Putta. 2019. Constructions at work in foreign language learners' mind: A comparison between two sentence-sorting experiments with English and Italian learners. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, 17(1):219–242.

Giulia ML Bencini and Adele E Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, 43(4):640–651.

Kathryn Bock and Helga Loebell. 1990. Framing sentences. *Cognition*, 35(1):1–39.

Joan Bresnan and Tatiana Nikitina. 2003. The gradience of the dative alternation. *Unpublished manuscript, Stanford University*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris Publications, Dordrecht.

Noam Chomsky. 1986. *Knowledge of Language*. Praeger, New York.

Noam Chomsky. 1995. *The Minimalist Program*. The MIT Press, Cambridge, Massachusetts.

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.

Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196.

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and Cognition*, 9(2):254–292.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, 64(3):501–538.

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

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford.

Stefan Th Gries and Stefanie Wulff. 2005. Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, 3(1):182–200.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. Association for Computational Linguistics.

Jena D Hwang and Martha Palmer. 2015. Identification of caused motion construction. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 51–60.

Matt A Johnson and Adele E Goldberg. 2013. Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Language and Cognitive Processes*, 28(10):1439–1452.

R. M. Kaplan and Joan Bresnan. 1982. Lexical functional grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–282. MIT Press, Cambridge, MA.

Michael P Kaschak and Arthur M Glenberg. 2000. Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of memory and language*, 43(3):508–529.

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

Simon Kirsch. 2019. The psychological reality of argument structure constructions: A visual world eye tracking study. *Unpublished MSc thesis, University of Freiburg*.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt Transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405.

Gianluca E Lebani and Alessandro Lenci. 2016. "beware the Jabberwock, dear reader!" Testing the distributional reality of construction semantics. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)*, pages 8–18.

Geoffrey Neil Leech. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research*, 28:1–13.

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

Beth Levin and Malka Rappaport Hovav. 2005. *Argument Realization*. Cambridge University Press, Cambridge, UK.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? Layerwise detection of linguistic anomalies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4215–4228, Online. Association for Computational Linguistics.

Junying Liang. 2002. Sentence comprehension by Chinese learners of English: Verb-centered or construction-based? *Unpublished MA thesis, Guangdong University of Foreign Studies*.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does RoBERTa know and when? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Findings*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094.

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

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank.

James Michaelov and Benjamin Bergen. 2020. How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 652–663.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4625–4635.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine Learning research*, 12:2825–2830.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Martin J Pickering and Victor S Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin*, 134(3):427.

Carl Pollard and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics*, volume 13. CSLI, Stanford.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546.

Montserrat Martínez Vázquez. 2004. Learning argument structure generalizations in a foreign language. *Vigo International Journal of Applied Linguistics*, (1):151–165.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020a. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952, Online. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Jayden Ziegler, Giulia Bencini, Adele Goldberg, and Jesse Snedeker. 2019. How abstract is syntax? Evidence from structural priming. *Cognition*, 193:104045.

## A Visualization of sentence sorting

We use principal components analysis (PCA) to visualize the sentence sorting experiment for the MiniBERTa models (trained with 1M and 100M tokens) and RoBERTa-base (trained with 30B tokens). In RoBERTa, there is strong evidence of clustering based on constructions; the effect is unclear in the 100M model and nonexistent in the 1M model (Figure 7). This visually confirms our quantitative evaluation based on the construction and verb deviation metrics (Figure 2).

## B Additional experimental stimuli

Table 3 shows an example set of template-generated stimuli for sentence sorting: we generate 1000 similar sets of 16 sentences to increase the sample size. We also present the sentence sorting stimuli for German (Table 4), Italian (Table 5), and Spanish (Table 6). German uses the same four constructions as English. Italian does not have the ditransitive construction but instead uses the prepositional dative construction to express transfer semantics. Spanish has no equivalents for the caused-motion and resultative constructions, so the authors in that experiment instead used the unplanned reflexive (expressing accidental or unplanned events), and the middle construction (expressing states pertaining to the subject).

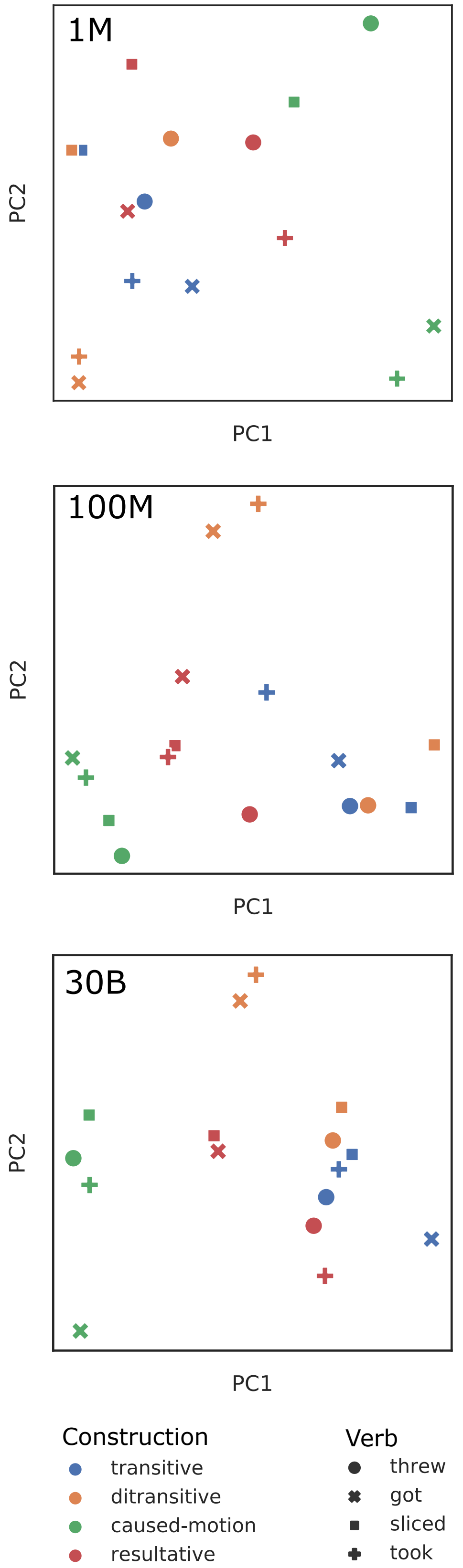


Figure 7: PCA plots of Bencini and Goldberg (2000) sentence sorting using the 1M and 100M MiniBERTa models and RoBERTa-base (30B). Figure best viewed in color.

| | Transitive | Ditransitive | Caused-motion | Resultative |
|---|---|---|---|---|
| **Slice** | Harry sliced the bread. | Henry sliced Eric the box. | Sam sliced the ball onto the bed. | John sliced the book apart. |
| **Kick** | Thomas kicked the box. | Mike kicked Frank the ball. | Michael kicked the wall into the house. | James kicked the door open. |
| **Cut** | George cut the ball. | Adam cut Paul the tree. | Bill cut the box into the water. | Bob cut the bread apart. |
| **Get** | Tom got the book. | Andrew got Steve the door. | Jack got the fridge onto the elevator. | David got the ball stuck. |

Table 3: Example of our 4x4 sentence sorting stimuli, similar to those by Bencini and Goldberg (2000) in Table 1, but generated automatically using templates.

| | Transitive | Ditransitive | Caused-motion | Resultative |
|---|---|---|---|---|
| **Werfen** | Anita warf den Hammer. | Berta warf Linda den Bleistift. | Erika warf den Schlüsselbund auf das Dach. | Laura warf die Kisten auseinander. |
| **Bringen** | Michelle brachte das Buch. | Simone brachte Lydia eine Einladung. | Emma brachte den Ball ins Netz. | Leonie brachte die Stühle zusammen. |
| **Schneiden** | Karolin schnitt das Brot. | Luisa schnitt Paula einen Apfel. | Jennifer schnitt die Wurst auf den Teller. | Doris schnitt den Reifen auf. |
| **Nehmen** | Maria nahm die Uhr. | Sophia nahm Jasmin das Geld. | Helena nahm die Rosen in das Haus. | Theresa nahm das Plakat herunter. |

Table 4: German sentence sorting stimuli, obtained from Kirsch (2019).

| | Transitive | Prepositional Dative | Caused-motion | Resultative |
|---|---|---|---|---|
| **Dare** | Lauda dà un esame. | Carlo dà una mela a Maria. | Luca dà una spinta a Franco. | Paolo dà una verniciata di verde alla porta. |
| **Fare** | Mario fa una torta. | Luigi fa un piacere a Giovanna. | Fabio fa entrare la macchina in garage. | Stefano fa bruciare il sugo. |
| **Mettere** | Annalisa mette la giacca. | Riccardo mette il cappello al bambino. | Silvia mette la penna nel cassetto. | Filippo mette la casa in ordine. |
| **Portare** | Linda porta lo zaino. | Laura porta la pizza a Francesco. | Michele porta il libro in biblioteca. | Irene porta l'esercizio a termine. |

Table 5: Italian sentence sorting stimuli, obtained from Baicchi and Della Putta (2019).

| | Transitive | Ditransitive | Unplanned Reflexive | Middle |
|---|---|---|---|---|
| **Romper** | Carlos rompió el cristal. | Alfonso le rompió las gafas a Pepe. | A Juan se le rompieron los pantalones. | La porcelana se rompe con facilidad. |
| **Doblar** | Felipe dobló el periódico. | Pablo le dobló el brazo a Lucas. | A Pedro se le dobló el tobillo. | El aluminio se dobla bien. |
| **Acabar** | Leonardo acabó su tesis. | Tomás le acabó la pasta de dientes a Santi. | A Luis se le acabaron los cigarrillos. | Las carreras de 10 km se acaban sin problemas. |
| **Cortar** | Isidro cortó el pan. | Jorge le cortó el paso a Yago. | A Ignacio se le cortó la conexión. | Esta tela se corta muy bien. |

Table 6: Spanish sentence sorting stimuli, obtained from Vázquez (2004).