Construction Identification and Disambiguation Using BERT: A Case Study of NPN

Wesley Scivetti **Nathan Schneider**
Georgetown University
{wss37, nathan.schneider}@georgetown.edu

Abstract

Construction Grammar hypothesizes that knowledge of a language consists chiefly of knowledge of form–meaning pairs ("constructions") that include vocabulary, general grammar rules, and even idiosyncratic patterns. Recent work has shown that transformer language models represent at least some constructional patterns, including ones where the construction is rare overall. In this work, we probe BERT's representation of the form and meaning of a minor construction of English, the NPN (noun–preposition–noun) construction—exhibited in such expressions as *face to face* and *day to day*—which is known to be polysemous. We construct a benchmark dataset of semantically annotated corpus instances (including distractors that superficially resemble the construction). With this dataset, we train and evaluate probing classifiers. They achieve decent discrimination of the construction from distractors, as well as sense disambiguation among true instances of the construction, revealing that BERT embeddings carry indications of the construction's semantics. Moreover, artificially permuting the word order of true construction instances causes them to be rejected, indicating sensitivity to matters of form. We conclude that BERT does latently encode at least some knowledge of the NPN construction going beyond a surface syntactic pattern and lexical cues.[1]

1 Introduction

The "black box" nature of Language Models (LMs) like has spawned a great deal of research investigating the extent to which these LMs are able to represent and understand a variety of linguistic phenomena (Linzen and Baroni, 2021; Rogers et al., 2021; Chang and Bergen, 2024). There has been substantial work focusing on many aspects of linguistic knowledge, including hierarchical structure (Clark et al., 2019; Hewitt and Manning, 2019; Jawahar et al., 2019), lexical semantics (Chang and Chen, 2019; Vulić et al., 2020), negation (Ettinger, 2020), agreement phenomena (Linzen et al., 2016; Weissweiler et al., 2023), and filler-gap dependencies (Wilcox et al., 2018, 2024). Broadly, these results show that even relatively modest sized LSTMs and transformer models are able to demonstrate nontrivial (though far from perfect) linguistic knowledge. However, there is some indication that these models are sometimes reliant on more surface level heuristics, and fail in situations which are straightforward to humans (McCoy et al., 2019; Ettinger, 2020). More generally, language models have been generally shown to struggle in out-of-domain situations (McCoy et al., 2024) and have some difficulty applying linguistic paradigms to nonce words (Weissweiler et al., 2023) and rare syntactic constructions (Scivetti et al., 2025).

Thus, there is need to evaluate language models on a range of linguistic tasks which go beyond the more studied "core" linguistic phenomena. Such work serves to provide a more complete picture of how language models succeed and fail across the broad spectrum of phenomena in language. Indeed, beyond the more mainstream notions of linguistic structure and information, there is also work on investigating LM knowledge of more idiosyncratic *constructions*, as defined by Construction Grammar. Construction Grammar is broadly a family of linguistic theories which consider all parts of language to be made up of constructions, which are pairings of linguistic *forms* with *meaning* or function (Goldberg 1995; Croft 2001, *inter alia*). It remains unclear the extent to which LMs may implicitly view constructions as distinct units. Because of their emphasis on pairing form with meaning, CxG theories provide possibilities for testing language model capabilities at the interface of form and meaning for different aspects of language, in contrast to past work which has focused on either syntax (e.g. Hewitt and Manning 2019) or seman-

[1] Code and Data: https://github.com/WesScivetti/NPN_probing

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 365–377
July 31–August 1, 2025, 2025© 2025 Association for Computational Linguistics

tics (e.g. Vulić et al. 2020) in isolation. A substantial and growing amount of research has recently focused on the intersection of LM knowledge and Construction Grammar (Tayyar Madabushi et al., 2020; Tseng et al., 2022; Pannitto and Herbelot, 2023; Veenboer and Bloem, 2023, *inter alia*), with a particular focus on argument structure constructions (Li et al., 2022), the English Comparative Correlative (Weissweiler et al., 2022), and the English AANN construction (Chronis et al., 2023; Mahowald, 2023). While these studies have provided valuable insight into LM processing of constructions with varying levels of schematicity, there remain many constructions which have not been addressed at all in previous work. Furthermore, while Zhou et al. (2024) do test model understanding of constructions which are similar in form, no past work has focused on individual constructions as polysemous units. We argue this is a gap in past work, as constructions, like words, can have related but distinct meanings that must be properly disambiguated in context in order for correct interpretation. We address this gap by providing experiments which pair formal sensitivity with semantic disambiguation in a controlled manner for a single construction.

This work is the first to study whether language models capture the NPN construction (Jackendoff, 2008), an infrequent yet productive pattern exhibited in expressions like *face to face* and *day to day*. Even for the subset where two instances of the same noun are linked by the preposition *to*, the pattern is polysemous, and sequences matching this pattern on the surface are not always instances of the construction (§2). Guided by CxG theory, we separate our inquiry in terms of the construction's *form* and *meaning* in context. To investigate language modeling of NPN, we:

- Construct and annotate a novel dataset of natural NPN examples from COCA (§3).
- Probe BERT's ability to distinguish true constructional instances from related constructions and artificial orders (§4 and §5).
- Introduce the task of construction sense disambiguation and perform experiments using our dataset (§6).

To summarize our findings, we show that probes using BERT embeddings are able to both identify correct instances of NPN and disambiguate the construction within context at respectable accuracy. Overall, these findings indicate that BERT latently encodes relevant information to the NPN construction, leading to strong sensitivity to both the construction's form and its meaning.

2 The NPN Construction

The NPN construction (Jackendoff, 2008) follows the general pattern of Noun + Preposition + Noun. Below are 2 examples of the NPN construction. These examples, along with all others, are taken from the Corpus of Contemporary American English (COCA, Davies 2010).

(1) There is a rebellious quality to your **day to day** responses which have not gone unnoticed.

(2) I need you to get this **word for word**.

Given the general rules of English, the NPN construction has several unique properties, which we argue separate it from more "core" linguistic phenomena. Firstly, the nouns almost always lack determiners, which is unusual for count nouns like "day". Secondly, the construction can occur in a variety of syntactic positions, including as an adverbial modifier (as in (2)) and as a prenominal modifier (as in (1)). Finally, it conveys a meaning which is not entirely predictable from its components, and varies considerably depending on the preposition. Common meanings of the NPN construction are the SUCCESSION meaning (shown in (1)) and the MATCHING/COMPARISON meaning (shown in (2)). See Jackendoff (2008) for an overview of the NPN construction and the common meanings associated with various prepositional lemmas.

While it is conceptually and intuitively appealing to think of NPN as a single construction, some work has argued in favor of viewing NPN as a group of related constructions, which are linked within the mind but not necessarily dominated by a single overarching abstract NPN construction (Sommerer and Baumann, 2021). Due to the wide variety of meanings and distributions of the different NPN constructions, we choose to limit our focus to a single subtype of NPNs, which all share the lemma "to" as their preposition, which we refer to as the N*to*N construction. There is still considerable semantic variation even within the N*to*N construction, with 2 broad meanings that we highlight: SUCCESSION (shown in (3)) and JUXTAPOSITION (shown in (4)).

(3) I was living **moment to moment**.

(4) You can preserve core warmth by huddling with a buddy, **chest to chest**.

While there are additional meanings of NPN that do not occur with "to" as the preposition, it is one of the only prepositions that is ambiguous in the NPN construction. By not considering examples of NPN with other prepositions, we remove the prepositional lemma as a potential shallow cue that models could learn to predict the construction's semantics. While there are arguably examples of NPNs where the two nouns are not identical, we limit our analysis to cases where the two nouns in the construction match exactly. This allows us to easily gather examples of the construction from corpus data.

3 Dataset

3.1 Corpus Gathering and Cleaning

In this work, we endeavor to use natural corpus data to the extent that it was possible. First, we use a simple pattern matching query to extract instances of the sequence Noun + "to" + Noun from COCA. We extract the examples from the corpus in a fixed window of +/- 50 tokens from the construction, and then used Stanza (Qi et al., 2020) to segment the results into sentences and extract the sentences which contained N*to*Ns. We automatically exclude sentences which contained "from" preceding the construction, because *from* N *to* N does not have exactly the same distribution as the more general N*to*N (Jackendoff, 2008), and is sometimes studied as a separate (but closely related) construction (Zwarts, 2013).

After extracting all sentences which contained a possible instance of N*to*N, we then manually clean the data, removing sentences that were either too short (<5 tokens) or contained too many typos. We annotate all instances of the construction for their semantic subtype, and double annotate roughly 25% of the dataset, achieving an agreement of 84% and a Cohen's kappa value of .754 between the two annotators, indicating strong agreement.[2] The final dataset has 6599 instances of N*to*N, of which 1885 were double annotated.

3.2 Near Minimal Pairs

In addition to true instances of the N*to*N construction, we also find grammatical corpus instances of Noun + "to" + Noun patterns, which are not instances of the construction. These patterns often occur when a verb licenses a direct object and a "to" prepositional phrase, and the direct object and the object of the preposition happen to have the same lemma. Three examples are shown below in (5), (6), and (7).

(5) Then there's the problem of sticking plastic to plastic.

(6) In Rome largesse was doled out by individuals to individuals.

(7) I don't have time to time travel ...

We do not consider such cases to be examples of the N*to*N construction because the surface pattern of Noun + Preposition + Noun clearly arises from a different syntactic context (e.g. a verb licensing a direct object and a PP modifier). Furthermore, the meanings of these examples do not evoke the unique semantics that accompany the N*to*N construction. While these cases are not instances of the N*to*N construction, they do provide a set of negative examples which we can use to probe the model's ability to recognize true N*to*N constructions. Throughout this paper, we refer to this set of examples as instances of the N*to*N *distractors*, since we test of if the model is "distracted" by the shallow similarity of the examples to the NPN construction. We refer to true examples of N*to*N as instances of the N*to*N *construction*. Since these N*to*N examples exhibit the same surface form as the N*to*N *construction*, we consider them to be near minimal pairs, following Weissweiler et al. (2022) who extract near minimal pairs from corpus data based on part-of-speech patterns. While these sentences inevitably contain more lexical biases than a true minimal pair dataset, they are completely natural, and provide a good comparison point for a construction where creating true minimal pairs is otherwise difficult (because there is no obvious minimal change that can be made to result in a grammatical sentence that is not an example of the construction, similar to the struggles of Weissweiler et al. (2022) regarding the Comparative Correlative construction). In total, we collect 456 total instances of N*to*N *distractors* from COCA.

3.3 Train/Test Split

The resulting dataset contains many instances of very common N*to*N constructions, such as "day to day". We control for the effect of these frequent

[2]Disagreements between the two annotators were resolved through discussion and a gold label was chosen jointly.

	SUCCESSION	JUXTAPOSITION	Distractors
train	289	287	287
test	731	678	72

Table 1: Number of noun–*to*–noun sequences: two meanings of the NPN Construction, as well as *distractors*. Train sets are balanced to be equal between the categories. The remaining examples are left for testing.

lemmas in two ways. Firstly, we artificially shrink the dataset by randomly sampling 20 sentences for each noun lemma which occurs more than 20 times, and discard the remaining sentences for the purposes of model training and testing. This is to make sure that no overly common lemmas have an overstated impact on the probing classifier performance.

Secondly, we generate random train/test splits based on lemma of the noun in the N*to*N, meaning that there are no lemmas that are seen in both the training set and the testing set. In other words, if an example with "day to day" is seen during training, a sentence with "day to day" will never be seen during testing (but a sentence with "week to week" might be). Each sentence in the dataset has one target instance of the N*to*N construction.

In Table 1, we report the final dataset sizes, split by semantic subtype for the construction examples. N*to*N *constructions* are much more frequent than the N*to*N *distractor* patterns which serve as their near minimal pairs. We choose to balance the sizes of the two types of examples during training. We take 80 percent of the N*to*N *distractor* patterns for training and withhold twenty percent. We take a similar number of N*to*N *constructions* for training and then test on the remainder, ensuring training sets are balanced between *constructions* and *distractors*.

4 Experiment 1: Constructions vs. Distractors

4.1 Methodology

We probe the ability for BERT to distinguish natural instances of the N*to*N *construction* from natural examples of the N*to*N *distractor* pattern. To address the issue of lexical overlap, we control for the lexical cue of the nouns in N*to*N by making sure there is no overlap of nouns in the training and testing data splits, as described in §3.3. However, it is still entirely possible that the classifier learns to utilize lexical similarity of the nouns in the construction, or even other words beyond the

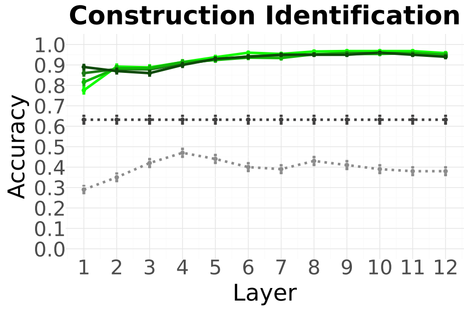

Figure 1: Accuracy of N*to*N *construction* across layers of BERT-base, averaged across 5 random seeds. Maximal accuracy in the mid to late layers. Reducing the number of training examples does not drastically harm performance. The light grey line represents control probe (Hewitt and Liang, 2019) accuracy, which hovers around chance. The dark grey line represents accuracy of the lexical semantic GloVe baseline. Darker lines indicate larger amounts of training examples, with possible values of 10, 25, 100, and 287. Reducing the amount of training examples for the probes does not lead to drastically changed performance. Error Bars indicate 95% confidence intervals over the mean accuracies across the 5 runs.

construction. We address this by providing two baseline systems which give perspective on performance based on lexical cues: a *control classifier* (Hewitt and Liang, 2019) and a non-contextual baseline based on GloVe embeddings (Pennington et al., 2014).

Control classifiers involve training new classifiers based on data where the labels are randomized and correspond deterministically to word type, ideally leading to chance performance. Following Hewitt and Liang (2019), who deterministically assign each word a POS tag for their probing experiments, we assign a random positive or negative label deterministically based on the first noun word type in the construction. The performance of these control classifiers should be near chance, in the absence of any spurious correlations which allow the classifier to solve the task given arbitrary labels.

We provide an additional, non-contextual baseline by training a linear classifier on GloVe embeddings for the nouns in the construction as input. It is well known that the NPN construction is biased towards certain lexical types of nouns, such as temporal phrases and body parts (Jackendoff, 2008). Thus, we expect that a classifier trained on the static embedding of the noun alone will achieve nontrivial performance. We argue that if a BERT-based classifier substantially outperforms this baseline, the difference in performance is an indication of nontrivial contextual understanding of the construction as a whole, beyond the lexical semantics of the present nouns.

Following previous probing work which tracks performance layer by layer Liu et al. (2019); Weissweiler et al. (2022), we train a separate probe based on embeddings from each layer of BERT and track performance across layers. We use the BERT-base-cased model, available through the Huggingface transformers library (Wolf et al., 2020), and choose logistic regression as our linear classification architecture.[3] For all experiments and data settings, we run probes with 5 random seeds and report the average results.

4.2 Results

For the probing classifier results, we graph accuracy on the N*to*N *construction* in Figure 1. As we can see, the classifier is relatively strong at distinguishing the N*to*N *construction* from *distractors* even in the early layers, with an accuracy over .90 by layer 5 with full training examples. Additionally, the classifiers are robust to sharp reductions in the number of training examples (shown in lighter shades of green in Figure 1), showing strong performance even with as few as 10 per-class training examples, echoing similar findings for other constructions (Tayyar Madabushi et al., 2020). The control classifier achieves roughly chance performance, meaning that our trained probes have high *selectivity* (Hewitt and Liang, 2019). The lexical semantic baseline using GloVe achieves performance well above chance (≈68%), though its performance lags far behind the BERT-based probes, regardless of how many training example those BERT-based probes receive. This shows that overall, the probing classifier seems to be picking up on some sort of information in BERT which can reliably distinguish the N*to*N *construction* from its near minimal pair N*to*N *distractor* counterparts, beyond what is possible through lexical semantic clues alone. However, the *distractor* examples generally have syntactic structure which is divergent from the *construction* examples. To provide another comparison point, we now test if the existing probes can distinguish true instances of the N*to*N construction from examples with artificially altered word orders.

[3] We take the embedding of "to" as the input into the classifier, as some past work has considered it the "head" of the overall construction (Jackendoff, 2008).

5 Experiment 2: Perturbing Word Order

As we have seen in §4.2, a BERT-based probe can generally distinguish the N*to*N *distractor* patterns from the N*to*N *construction*. However, we wish to further test how robust the model is at distinguishing the construction from related patterns. While we have compared to naturally occuring near minimal pairs, we now test the classifier on a set of examples with artificially perturbed word order. If the classifier is robust at recognizing the N*to*N *construction*, it should be able to correctly distinguish *construction* instances from artificial sentences with altered non-NPN word orders. To illustrate this point, consider the following two sentences:

(8) I need you to get this **word for word**.

(9) I need you to get this **for word word**.

Example (8) is a copy of (2) and is a true NPN construction. On the other hand, (9) is not an instance of the construction (because it does not follow the NPN word order), and is a generally ungrammatical sentence. We hypothesize that if the probe trained in §4 is not robust to the actual word order pattern of N*to*N, it will be unable to distinguish sentences like (8) from those like (9). If indeed the lexical cues are influencing classifier performance independent of word order, we expect that the classifier will predominantly classify examples like (9) as positive instances of the N*to*N *construction*.

To test this hypothesis, we manipulate the test set of the probe by creating 4 perturbed orderings of each test example sentence: ***PNN***, ***PN***, ***NNP***, ***NP***. A true N*to*N example is shown in (10) the corresponding 4 different perturbed orderings are shown below in (11), (12), (13), and (14).

(10) Go **room to room** removing anything you don't need and selling it. (Original N*to*N)

(11) Go **to room room** removing anything you don't need and selling it. (PNN Perturbed Order)

(12) Go **to room** removing anything you don't need and selling it. (PN Perturbed Order)

(13) Go **room to** removing anything you don't need and selling it. (NP Perturbed Order)

(14) Go **room room to** removing anything you don't need and selling it. (NNP Perturbed Order)

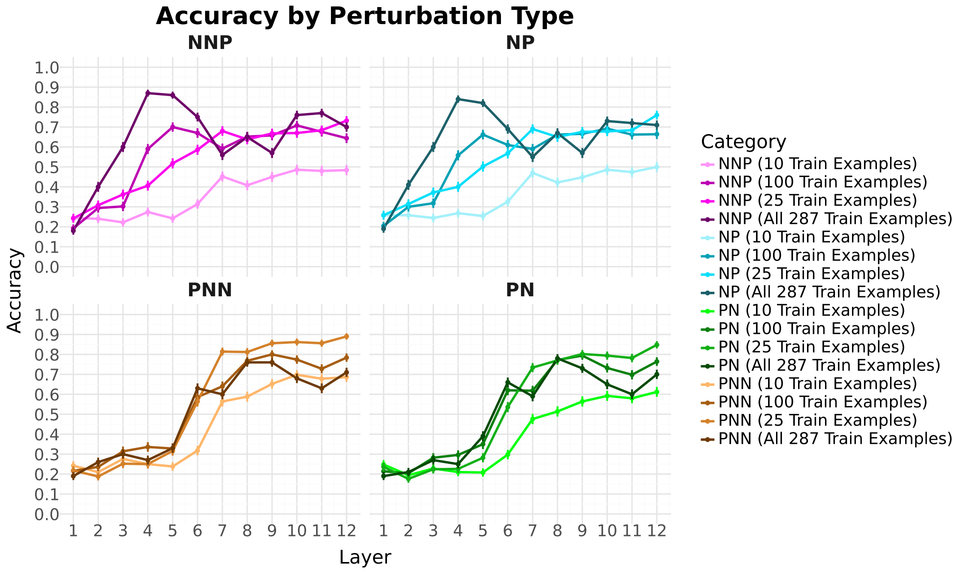

Figure 2: Accuracy of perturbed orderings of original N*to*N *constructions*. Since the perturbed word orders are not true instances of the construction, the true class is negative for all instances. High accuracy indicates that probes are rejecting the validity of the artificial orderings. Lighter colors represent fewer training examples for the probings. Error bars indicate 95% confidence intervals over the average of 5 random seeds.

Crucially, we do not retrain the linear probe on this perturbed data. This means that during training, the classifier only saw instances with the correct N + *to* + N ordering, either positive instances of the N*to*N *construction* (like in (1) and (2)), or near minimal pairs of the N*to*N *distractor* patterns (like in (5), (6), and (7)). Thus, this experiment tests the robustness of the original probing classifier when it is confronted with out of domain word orders that contain the same lexical cues as positive instances of the construction.

5.1 Results

Figure 2 shows the probe's performance on the perturbed test sets for the N*to*N *construction*. We see that in the very early layers (1–3), the probe often predicts the N*to*N *construction* despite the word order shifts, leading to relatively low accuracy. This possibly means that the classifier is biased by the lexical cues in the sentence early on. Interestingly, performance on *PN* and *PN* perturbations is substantially worse than performance on *NP* and *NNP* in the early layers. Accuracy on all perturbations trends upwards in the later layers, with reduction in training examples leading to drops in performance especially for *NP/NNP*.

5.2 Analysis

Overall, we find that classifier probes are able to distinguish instances of the N*to*N *construction* from both near minimal pairs (N*to*N *distractor* patterns) and artificial examples (perturbed word orderings). This finding aligns with the strong performance on form-based recognition that has been observed in previous work on other constructions (Li et al., 2022; Weissweiler et al., 2022; Mahowald, 2023). The peak in performance in the late-middle layers is consistent with much previous work on linguistic probing, which show that the middle and late-middle layers perform best for a variety of linguistic tasks (Goldberg, 2019; Hewitt and Manning, 2019; Lin et al., 2019; Liu et al., 2019).

The differences in the performance between the *NP/NNP* and the *PN/PNN* perturbed orderings is an unexpected finding. According to Rogers et al. (2021), the earlier layers of BERT encode "word order", while the middle layers are where syntactic capabilities emerge. Based on this logic, it is unsurprising that the classifier's ability to distinguish *PN/PNN* emerges in the middle and later layers. Why might the *NP/NNP* instances be distinguished so much quicker? Our intuition is that in general, preposition tokens probably attend more to their immediately following word than their immediately preceding word. This is because prepositions are often immediately followed by objects, while their syntactic governor may or may not be directly adjacent to them. Perhaps in the early layers of the model (before hierarchy is as explicitly represented) prepositions attend to their following token more quickly because this is a surface word order pattern that feeds quite well into syntax.

One alternative explanation is that *PN/PNN* may produce generally more grammatical sounding sentences than *NP/NNP*. For instance, (12) sounds much closer to a real sentence than (14). It could be that the classifier probe takes into account the ungrammaticality of *NP/NNP*, even though it was not explicitly trained to do this, since the classifier probe is only trained on grammatical sentences. How exactly the ungrammaticality is represented in these embedding representations is unknown, but provides one possible explanation for the differential performance of the perturbed word ordering patterns.

Having established that performance on identifying the N*to*N construction is strong, we now turn to the task of disambiguating the meaning of the construction within context.

6 Experiment 3: Semantic Disambiguation

6.1 N*to*N Subtypes

We have established that classifier performance is strong at identifying instances of the N*to*N construction relative to similar patterns. However, the construction itself is ambiguous, and can have different meanings in context. The two primary meanings are SUCCESSION and JUXTAPOSITION, which are shown in (3) and (4) respectively.

The two types co-occur with different nouns at different frequencies. The SUCCESSION subtype most often occurs with spatiotemporal nouns (e.g. *day to day* or *coast to coast*). On the other hand, the JUXTAPOSITION subtype most often occurs with body parts or humans (e.g. *face to face* or *friend to friend*). However, the noun meaning is not determinative, and within context some noun lemmas occur with the less common meaning. Furthermore, both constructions occur with rare noun lemmas for which it is not clear what type would be more common.

6.2 Methodology

In this section, we train a classifier to distinguish semantic subtypes of N*to*N. We focus on the two main subtypes that are well attested in the data: SUCCESSION and JUXTAPOSITION. We also include examples of the N*to*N *distractor* patterns which are not examples of the *construction*. Thus, the probe is faced with a 3-class classification problem: it must distinguish between the SUCCESSION subtype, the JUXTAPOSITION subtype, and non-examples of the construction (*distractors*). Following Hewitt and Liang (2019), we train *control classifiers* with a random label assigned to each lemma. If the probes are properly selective, the control classifiers should have accuracies of around 33 percent.

6.3 Results

Figure 3 shows the precision and recall scores of the semantic probing experiments. Across all semantic types, performance is generally high for the classifiers trained on the full split of data, with recall on all 3 classes near 80%, and strong performance even in the early layers. This is in contrast to some other semantic tasks, for which probes only reach their peaks in the mid to late layers of BERT.

Across all layers, both SUCCESSION and JUXTAPOSITION perform worse with only 10 training examples, but performance stabilizes after only 25 examples for the probe. The relatively low recall for JUXTAPOSITION and SUCCESSION when the classifiers are only trained with 10 examples indicates that the probe has not fully learned to correctly distinguish the two main semantic subtypes. It is somewhat striking that there is not a larger difference between SUCCESSION and JUXTAPOSITION in performance, given that SUCCESSION accounts for roughly 68% of all instances of the construction in our dataset. While probes are trained with balanced training sets, the relative frequency of these semantic subtypes within our dataset (and by extension COCA) is a strong indication that SUCCESSION is the more frequent meaning. Nevertheless, performance is roughly comparable between the two semantic subtypes. In all cases, the *distractor* class is overpredicted, leading to a relatively low precision compared to the subtypes of the construction. As expected, the control classifiers achieve roughly chance performance across layers, indicating that our probes have high selectivity. The GloVe-based baseline achieves an average recall of around .54 across the subtypes, but has widely variable performance depending on the semantic subtype. In general, the GloVe based classifier is much more likely to underpredict SUCCESSION, leading to very high precision and very low recall for this class.[4]

[4] We report GloVe and control results using the full training set. Performance of the GloVe baselines degrades with fewer examples, while the control classifiers remain near chance.

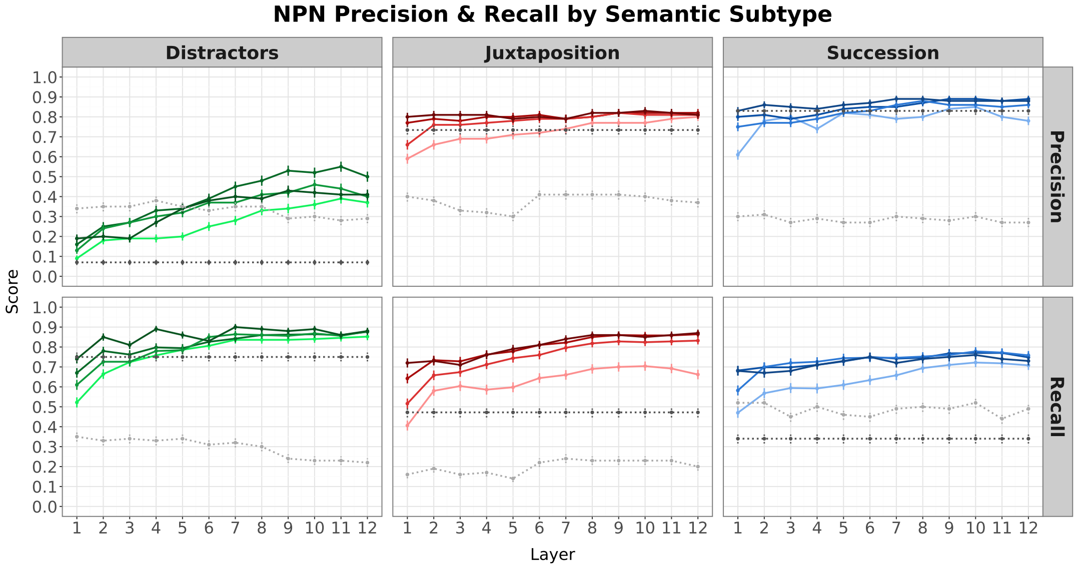

Figure 3: Precision and Recall of different semantic subtypes of NPN in 3-way classification. Lighter colors indicate fewer training examples, with possible values of 10, 25, 100, and 287 training examples per class. Classifiers trained with at least 25 per-class training examples begin to show strong performance across classes. JUXTAPOSITION takes substantially more training examples for classifiers to learn compared with SUCCESSION. Each line represents the average of 5 random seeds. Dotted lines represent baselines: GloVe (black) and control (gray). Error Bars indicate 95% confidence intervals over the average of the random seeds.

7 Related Work

There has been substantial research on investigating the linguistic information that is encoded by BERT. Much of this work has focused on syntactic structure (Hewitt and Manning, 2019; Jawahar et al., 2019; Liu et al., 2019; Hu et al., 2020), agreement phenomena (Lin et al., 2019) and semantics (Vulić et al., 2020; Chang and Chen, 2019; Ettinger, 2020), with the BLiMP (Warstadt et al., 2020) and SyntaxGym (Gauthier et al., 2020) providing key evaluation datasets. Belinkov (2022) and Elazar et al. (2021) provide critiques of the probing classifier methodology for its indirectness and susceptibility to spurious correlations. Various improvements on the methodology have been suggested, with a general focus on providing more controlled probing environments (Pimentel et al., 2020; Kim et al., 2022) and causal claims through counterfactuals (Ravfogel et al., 2021; Elazar et al., 2021). Of particular relevance to this work is Hewitt and Liang (2019), who propose the control classifier methodology as one methodology for controlling for spurious correlations in classifier performance. We believe our use of control classifiers and non-contextual baselines provide proper context for our probing results.

Earlier computational linguistic work on English trained classifiers for such grammatico-semantic phenomena as identifying argument structure constructions (Hwang and Palmer, 2015) and disambiguating functions of tense and definiteness (Reichart and Rappoport, 2010; Bhatia et al., 2014), as well as generally to disambiguate the senses of prepositions (Litkowski and Hargraves, 2007; Schneider et al., 2018). Tayyar Madabushi et al. (2020) were the first to investigate BERT's performance on learning constructions, finding that BERT is able to identify a large set of hundreds of automatically identified constructions. Regarding well-established argument structure constructions, Li et al. (2022) find that RoBERTa implicitly contains abstract knowledge of the constructions beyond specific lexical cues. Weissweiler et al. (2022) find that BERT-scale models are able to correctly distinguish the COMPARATIVE-CORRELATIVE construction from similar looking patterns, but find that the models fail on reasoning tests related to the construction's semantics. Mahowald (2023) finds that the larger GPT-3 model can provide acceptability judgments for the Article+Adjective+Numeral+Noun (AANN) construction which generally align with human judgements, and find that the model is sensitive to constraints on the slots in the construction. Chronis et al. (2023) test BERT's knowledge of the same AANN construction by projecting tokens in the construction into an interpretable embedding space, finding that features aligning with measure-words are evoked by tokens in the construction. Beyond BERT-scale models, Zhou et al. (2024), Bonial and Tayyar Madabushi (2024) and Scivetti et al. (2025) all test LLM knowledge of constructions in more complex scenarios, finding that their performance generally lags behind humans regarding construction understanding, though there is variation depending on the construction. Zhou et al. (2024) test a range of LLMs on understanding the CAUSAL-EXCESS constructions in comparison to constructions with highly similar forms, showing that the model is often misled by form-based cues. Their experiments most closely mirror our inquiries into construction sense disambiguation, though they disambiguate between similar but distinct constructions while we focus on a single polysemous construction. While Zhou et al. (2024) find that LLMs largely are unsucessful at meaning-based disambiguation, and Weissweiler et al. (2022) also find negative results regarding the semantics of the COMPARATIVE-CORRELATIVE, our relatively positive results on construction disambiguation in this present work demonstrate that for N*to*N, models may possess more robust models of constructional semantics than would be previously expected.

While NPN has not been the major focus of past analysis, Weissweiler et al. (2024) do consider it as one of the constructions which they include in their UCxn dataset, which is compiled by automatically using Universal Dependencies (de Marneffe et al., 2021) graphs to find indications of constructions across 10 languages. We do not use this dataset due to its limited size (it contains under 50 total examples of the NPN construction in English).

8 Conclusion

In this work, we constructed a novel dataset of N*to*N construction by extracting all instances of the construction which we found in COCA. Using our dataset, we have probed BERT's knowledge of the N*to*N construction by training a linear probe to distinguish instances of the construction from near minimal pairs from corpus data. We show that a linear probe is largely able to distinguish true instances construction from naturally occurring *distractor* patterns, as well as from artificially perturbed versions of the construction, though the probe is more robust to recognizing the effect of some word order changes than others. Furthermore, we show that a BERT-based classifier can disambiguate the sense of the N*to*N construction in context, beyond the lexical semantic cues that are present. For both form- and meaning-based experiments, we show that the classifier results are robust even in the face of dramatic reductions in the number of training examples. This indicates that constructional knowledge is likely latently encoded within BERT and not due to spurious correlations learned by the classifiers. Overall, these results contribute to the growing body of evidence that LMs have some ability to acquire grammatical properties of rare and idiosyncratic constructions.

9 Limitations

This work is limited in several ways. Due to natural relative frequencies of various constructions, the dataset used for N*to*N is unbalanced between the N*to*N *construction* and *pattern*. This means that the training set for the classifier was quite small, because we ensured that training was balanced between the different classes. While the probing classifiers do achieve high accuracy, it is unclear how much accuracy is being capped by the limited data available. However, this fact, alongside our experiments with reduced training set sizes, indicate that the probes can learn with relatively little training signal.

This is experiment is also limited in only considering N*to*N, as opposed to the broader NPN construction. This is an intentional choice, as "to" has the most semantic subtypes of NPN associated with it. Future work is needed to see if the results here are robust to the inclusion of additional NPN examples with other lemmas into the dataset. We also only consider the English NPN construction, though the construction has been observed in a range of languages, including Dutch, English, French, German, Norwegian, Japanese, Mandarin, Polish, and Spanish (Weissweiler et al., 2024). We also limit our experiments to cases where the nouns match. This choice greatly simplifies our process of detecting true constructions as well as distractors, but also excludes some interesting examples of the construction, as pointed out by Jackendoff (2008).

Finally, this work utilizes the probing classifier methodology, which has been criticized for providing indirect/correlational evidence of linguistic information in LM representations (Belinkov, 2022). Future work is needed to broaden the analysis to include causal probing methodologies (e.g. AlterRep, Ravfogel et al. 2021; MaPP, Karidi et al. 2021; Reconstruction Probing, Kim et al. 2022).

Acknowledgments

We thank CoNLL reviewers and NERT lab members for their thoughtful comments. This research was supported in part by NSF award IIS-2144881.

References

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Archna Bhatia, Chu-Cheng Lin, Nathan Schneider, Yulia Tsvetkov, Fatima Talib Al-Raisi, Laleh Roostapour, Jordan Bender, Abhimanu Kumar, Lori Levin, Mandy Simons, and Chris Dyer. 2014. Automatic classification of communicative functions of definiteness. In *Proc. of COLING*, pages 1059–1070, Dublin, Ireland.

Claire Bonial and Harish Tayyar Madabushi. 2024. A Construction Grammar Corpus of Varying Schematicity: A Dataset for the Evaluation of Abstractions in Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 243–255, Torino, Italia. ELRA and ICCL.

Ting-Yun Chang and Yun-Nung Chen. 2019. What Does This Word Mean? Explaining Contextualized Embeddings with Natural Language Definition. 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)*, page 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language Model Behavior: A Comprehensive Survey. *Computational Linguistics*, 50(1):293–350.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A Method for Studying Semantic Construal in Grammatical Constructions with Interpretable Contextual Embedding Spaces. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 242–261, Toronto, Canada. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 276–286, Florence, Italy. Association for Computational Linguistics.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

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.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, page 70–76, Online. Association for Computational Linguistics.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. (arXiv:1901.05287). ArXiv:1901.05287 [cs].

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)*, page 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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)*, page 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, page 1725–1744, Online. Association for Computational Linguistics.

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

Ray Jackendoff. 2008. "construction after Construction" and Its Theoretical Challenges. *Language*, 84(1):8–28.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 3651–3657, Florence, Italy. Association for Computational Linguistics.

Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, and Vivek Srikumar. 2021. Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 10300–10313, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Najoung Kim, Jatin Khilnani, Alex Warstadt, and Abed Qaddoumi. 2022. Reconstruction Probing. (arXiv:2212.10792). ArXiv:2212.10792 [cs].

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 241–253, Florence, Italy. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7(Volume 7, 2021):195–212. Publisher: Annual Reviews.

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.

Ken Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proc. of SemEval*, pages 24–29, Prague, Czech Republic.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. 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)*, page 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Kyle Mahowald. 2023. A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121. Publisher: Proceedings of the National Academy of Sciences.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 3428–3448, Florence, Italy. Association for Computational Linguistics.

Ludovica Pannitto and Aurélie Herbelot. 2023. CALaMo: a Constructionist Assessment of Language Models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, page 21–30, Washington, D.C. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4609–4622, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, page 101–108, Online. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, page 194–209, Online. Association for Computational Linguistics.

Roi Reichart and Ari Rappoport. 2010. Tense sense disambiguation: a new syntactic polysemy task. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 325–334, Cambridge, MA.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL*, pages 185–196, Melbourne, Australia.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025. Assessing Language Comprehension in Large Language Models Using Construction Grammar. (arXiv:2501.04661). ArXiv:2501.04661 [cs].

Lotte Sommerer and Andreas Baumann. 2021. Of absent mothers, strong sisters and peculiar daughters: The constructional network of English NPN constructions. *Cognitive Linguistics*, 32(1):97–131.

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*, page 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A Construction and Context-aware Language Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, page 6361–6369, Marseille, France. European Language Resources Association.

Tim Veenboer and Jelke Bloem. 2023. Using Collostructional Analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 12937–12951, Toronto, Canada. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 7222–7240, Online. Association for Computational Linguistics.

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

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. (arXiv:1910.03771). ArXiv:1910.03771 [cs].

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 3804–3811, Torino, Italia. ELRA and ICCL.

Joost Zwarts. 2013. From N to N: The anatomy of a construction. *Linguistics and Philosophy*, 36(1):65–90.