Using Collostructional Analysis to evaluate BERT's representation of linguistic constructions

Tim Veenboer
University of Amsterdam
tim.veenboer@student.uva.nl

Jelke Bloem
University of Amsterdam
j.bloem@uva.nl

Abstract

Collostructional analysis is a technique devised to find correlations between particular words and linguistic constructions in order to analyse meaning associations of these constructions. Contrasting collostructional analysis results with output from BERT might provide insights into the way BERT represents the meaning of linguistic constructions. This study tests to what extent English BERT's meaning representations correspond to known constructions from the linguistics literature by means of two tasks that we propose. Firstly, by predicting the words that can be used in open slots of constructions, the meaning associations of more lexicalized constructions can be observed. Secondly, by finding similar sequences using BERT's output embeddings and manually reviewing the resulting sentences, we can observe whether instances of less lexicalized constructions are clustered together in semantic space. These two methods show that BERT represents constructional meaning to a certain extent, but does not separate instances of a construction from a near-synonymous construction that has a different form.

1 Introduction

Neural language models have proven to be immensely successful in processing and generating language – especially transformer-based models such as BERT have displayed phenomenal results (Devlin et al., 2018). However, there is a glaring lack of transparency in the linguistic representations that these models rely on. While previously, language would be parsed according to specific formalisms grounded in linguistic theory, recent approaches rely on models inferring structures in an unsupervised way. This impedes our understanding of what structures a model should learn to accurately carry out natural language processing tasks, and dissuades theoretical linguists from researching them (Baroni, 2021). Recently there has been an increased interest in opening the black box of these language models (Rogers et al., 2020) showing that they capture various aspects of syntax, semantics and morphology, but these efforts are not always grounded in specific theories.

One line of theoretical research that lends itself to these efforts is Construction Grammar (CxG). Grounded in language acquisition research, Construction Grammar claims that human language is structured into learned pairs of forms and meaning, called constructions, ranging from specific words to abstract grammatical patterns. As long as a linguistic pattern carries semantic meaning restricted exclusively to that pattern, it's considered a construction. These constructions are acquired by human learners based on distributional data (Boyd and Goldberg, 2011). This theory has proven itself to be compatible with data-driven, quantitative and distributional approaches to language (Levshina and Heylen, 2014; van Trijp, 2017) as we can view each construction as having distributional meaning (Rambelli et al., 2019).

It is not obvious that transformer models would learn constructions at all. Their tokenizers already have certain built-in assumptions about the nature of linguistic units that may differ from those of Construction Grammar. Furthermore, BERT's training objectives are aimed at learning lexical-semantic information for these tokens, rather than for (partially abstracted) combinations of tokens or categories of tokens. Lastly, the BERT training process of course differs in significant ways from human language learning (Warstadt et al., 2023). Nevertheless, BERT's prediction accuracy indicates that it represents plenty of contextual information about the ways in which a token can be used. This raises the question, do BERT's contextual meaning representations correspond to CxG constructions?

A key idea within Construction Grammar is that specific words or phrases are more closely attracted to certain constructions than others, based on mean-

Findings of the Association for Computational Linguistics: ACL 2023, pages 12937–12951

ing associations. The method of collostructional analysis (Gries and Stefanowitsch, 2003) is used to quantitatively study such associations. Such associations might be learned by BERT models as well, leading to clusters that we might call constructions. An evaluation of BERT output contrasted with a reproduced collostructional analysis results could reveal evidence of whether BERT represents meaning beyond tokens in this way. This in turn will lead to more linguistically plausible intrinsic embedding evaluation and increased model transparency.

We propose two tasks that are aimed at detecting meaning associations of specific constructions in an English-language BERT model. Firstly, by predicting the words within the open slots of constructions, we aim to evaluate to what extent BERT represents the meaning of constructions. Such associations between words and constructions are indicative of a construction's meaning according to the distributional hypothesis of meaning and studying them "provides an objective approach to identifying the meaning of a grammatical construction" (Gries and Stefanowitsch, 2003, p. 211). Secondly, by finding similar sentences using BERT's output embeddings, we evaluate whether those output embeddings contain the information necessary to distinguish constructions by both form and meaning. To ground our approach in the linguistic literature, we evaluate our results by comparing them to our reproduction of the results of Gries and Stefanowitsch (2003), whose method involves selecting exemplars based on part-of-speech annotation and manual filtering and can thus serve as a gold standard.

2 Construction Grammar

Construction Grammar (CxG), stemming from cognitive linguistics, deems constructions to be the fundamental components of written and spoken human language, and claims that the knowledge of an individual regarding language is solely defined by a complex set of constructions, labeled the *constructicon* (Goldberg, 2003). These constructions are form-meaning pairs, specifically, $\boldsymbol{C}$ is a construction *iff*$_{def}$ $\boldsymbol{C}$ is a form–meaning pair <F_i, S_i> such that some aspect of F_i or some aspect of S_i is not strictly predictable from $\boldsymbol{C}$'s component parts or from other previously established constructions. (Goldberg, 1995, p. 4)

So, rather than separating syntax and semantics like many other linguistic theories, CxG inextricably links them. A construction is only a construction when it is impossible to deduce the form or semantic meaning entirely from the other elements of the construction. Constructions range from very abstract to very specific – the past tense construction, for example is a highly abstract construction into which any verb can enter, while words are fully lexical, specific constructions.

We discuss two constructions in more detail, and they are the constructions we will test BERT for. We chose to analyse these two English-language constructions because they were among the constructions on which the method of collostructional analysis was first demonstrated (Gries and Stefanowitsch, 2003), allowing us to compare our results to established results from the literature. Furthermore, they have been thoroughly studied from various theoretical perspectives. The two greatly differ in their level of abstractness, providing different methodological challenges.

2.1 X waiting-to-happen construction

First, we test the X waiting-to-happen construction, e.g. *This is a disaster waiting to happen*, a rather lexicalized and idiomatic construction where only the word *disaster* (the open slot X) can be substituted. It is classified as a *lexically open idiom*, since there is an alternating variable which is not predetermined (Fillmore et al., 1988). It has a specific constructional meaning, as not every noun can easily fill the variable slot X:

(1) It was **an accident** waiting to happen.

(2) ? It was **a door** waiting to happen.

Gries and Stefanowitsch (2003, p. 220) analysed this construction using collostructional analysis and argue that it is used to refer to something that "will almost certainly occur and that this is already obvious at the present point in time (often used with a negative connotation)".

2.2 Ditransitive construction

Second, we test the ditransitive construction, e.g. *I give him the ball*, a more abstract construction where any of 4 elements could be substituted. This construction denotes a transfer of a direct object (O_d) between two entities, often the subject (S) and an indirect object (O_i). The transfer is indicated by a verb (V). Thus, the ditransitive construction consists of four open slots which correspond to the four aforementioned constituents [S, V, O_d, O_i].

We can get a sense of the meaning of this construction by using nonce words, as explained by Hilpert (2014, p. 29). If English speakers see *Henry flinked Eve the wug*, they might still get the impression that a transfer takes place, even though they do not know the meaning of two of the words. In Gries and Stefanowitsch's (2003) analysis of this construction, they show that it is also used with many verbs that extend the literal transfer meaning, such as *teach* or *say*.

This construction is also interesting because it can have multiple forms: the double object construction (DOC) and the prepositional dative construction (PDC). Together, these two constructions engage in the dative alternation, where in many contexts, one can be substituted for the other:

(3) Henry flinked Eve the wug. (DOC)

(4) Henry flinked the wug to Eve. (PDC)

Whether there is a semantic difference between the two has long been debated (Krifka, 2004), although there are some senses expressed by only one of the two constructions:

(5) I brought a glass of water to the table.

(6) ?? I brought the table a glass of water.

In this example from Partee (2015, p. 60), the double object construction is infelicitous while the prepositional dative can be used, indicating some meaning difference. Nevertheless, both constructions appear to have greatly overlapping meanings. This means that by testing for one of the variants of the ditransitive construction, we can tease apart whether BERT's contextual embedding space separates constructions predominantly by meaning or also by form. In the former case, testing for the double object construction would yield both constructions, and in the latter case, it would yield mainly DOCs.

3 Related work

Prior research seems to support the idea that BERT learns some underlying structure beyond lexical semantics (Lin et al., 2019). Such a structure might be Construction Grammar. However, the combination of Construction Grammar and large language models has not received much attention. Pannitto and Herbelot (2022) present an overview of the use of neural networks to test usage-based theories of language acquisition, which sometimes use construction grammar representations. Fonteyn et al. (2020) apply BERT embeddings to study constructional change in the BE-about futurate construction.

A few works directly address Construction Grammar in BERT. Madabushi et al. (2020) investigated whether BERT is able to tell if two sentences contain the same construction. They took over 22000 different automatically identified constructions and let BERT classify whether examples of them contain the same construction, obtaining 94% accuracy. However, due to automatic construction identification, their dataset may have contained many patterns that would not be considered constructions in CxG, and no comparison to results of linguistic analysis is carried out.

Tseng et al. (2022) state that a Chinese BERT-based model represents the difference between lexical elements and open slots of constructions by showing lower prediction probabilities for open slots of constructions. They tuned a MLM for the task of predicting the content of open slots to better learn the probability distributions of words in open slots (comparable to collexeme strength in CxG).

Two recent studies do focus on closely analysing specific construction types in BERT. Weissweiler et al. (2022) syntactically probe several BERT-based models for the English comparative correlative construction in a CxG framework using minimal pair exemplars. They also perform an inferencing task and find that the models are not able to make inferences based on the meaning of the construction in a zero-shot setup. However, this may be due to the models' inadequate performance on inferencing tasks and logic in general rather than an inadequate intrinsic semantic representation.

Li et al. (2022) study argument structure constructions in transformer LMs, including the ditransitive. In a sentence sorting task, they find that sentence embeddings of these constructions are clustered by their construction (ditransitive, resultative, caused-motion or removal) rather than by their verb in embedding spaces of several BERT-based models. In a random word experiment, they fill the construction's slot with random real words and test whether the resulting context embeddings are close to averaged context embeddings of a verb prototypical of that construction. In this work we take the opposite approach, following Gries and Stefanowitsch's (2003) linguistic analysis method, focusing on what meanings appear in the slots to reveal the meaning of a construction.

These studies conclude that constructions are

or are not represented in BERT models to varying degrees. We contribute novel evidence to this debate by 1) direct comparison to previous corpus-based linguistic analyses by Gries and Stefanowitsch (2003) of two specific constructions, 2) taking the content of open slots as indicative of the presence of constructional meaning in the embedding space using two different tasks and 3) investigating near-synonymous constructions to decide whether, if constructions are represented, they are clustered just by meaning or also by form.

4 Method

To examine a BERT model for the presence of constructional information and to perform collostructional analysis on these constructions we first need to obtain a set of representative exemplars of both constructions. To have a valid comparison with Gries and Stefanowitsch's (2003), we use the British National Corpus (BNC, BNC Consortium, 2007) which is the corpus they derived their results from.

4.1 Data

There are no large corpora annotated with a CxG formalism that can be queried syntactically. Even relying on part-of-speech tags is nontrivial, as they do not map directly onto constructions. The examples below are both instances of the double object construction, but in the BNC's C5-tagset[1] most words in these two sentences have different tags:

(7) John gave Mary the balls
NP0 VVD NP0 AT0 NN2

(8) He gives me a ball
PNP VVZ PNP AT0 NN1

We chose the approach of matching sets of POS-tags to the BNC data. We tried the Knuth-Morris-Pratt (KMP) algorithm (Knuth et al., 1977), the Aho-Corasick algorithm (Aho and Corasick, 1975) and template matching as used in computer vision (Brunelli, 2009) in the OpenCV implementation (Bradski, 2000). The latter was most efficient. After a preprocessing step turning POS tags into numbers, it takes about a second per POS pattern to search the BNC on a standard consumer laptop. We performed such searches for both target constructions and manually filtered the results. This yielded 1147 instances of the ditransitive construction and 35 instances of the X waiting-to-happen construction (the latter listed in Appendix B).

[1] http://ucrel.lancs.ac.uk/claws5tags.html

4.2 Collostructional Analysis

In collostructional analysis (Gries and Stefanowitsch, 2003), associations between a construction and the words that occur in its open slot(s) are computed, using corpus data and a statistical association metric. These associated words are called collexemes and their association value is called collexeme strength. It is an extension of collocational analysis to the notion of constructions.

The procedure is as follows. First, one particular construction that has one or more open slots to be filled by lexical items is chosen to be analyzed. Next, all the lexemes occurring in the slot are extracted from a text corpus, preferably a syntactically annotated one (Gries and Stefanowitsch, 2003, p. 214-215). Manual checks and filtering should be performed to get gold standard-quality data.

Over this corpus data, the strength of association between the lexemes and the construction is then calculated. Stefanowitsch & Gries chose the Fisher's Exact Test (FET) association measure since it is able to handle low-frequency data and it does not make any distributional assumptions. The input to FET are single and joint frequencies of the construction and the given lexeme, i.e. the frequencies of the lexeme in the construction, the lexeme in other constructions, the construction with other lexemes and finally all other constructions with all other lexemes. The output is a p-value according to which the collostructions can be ranked: the smaller the p-value, the more strongly associated the construction and the collexeme are (Gries and Stefanowitsch, 2003, p. 218-219).[2] Finally, by way of linguistic analysis, the first ten to thirty ranks of the collexemes are examined, and Stefanowitsch & Gries classify them according to their semantic and sometimes also syntactical properties.

The work was later extended to a family of methods, including distinctive collexeme analysis (Gries and Stefanowitsch, 2004), which provides a measure of the preference of a collexeme for one of two different constructions. In this method, collexeme lists of two constructions are compared directly. Distinctive collexeme analysis is typically used for identifying meaning differences between grammatical constructions that express similar meanings, such as the dative alternation. Another variation

[2] Subsequent studies have used measures of effect size such as the odds ratio for comparison and ranking, since comparing p-values directly is a controversial practice.

is covarying collexeme analysis, in which associations between words in multiple different open slots in the same construction are computed.

Linguists use these methods to discover relationships between lexemes and constructions from distributional information, in order to describe the meaning associations and thereby meaning of a construction. We use it as a supervised method of obtaining constructional meaning associations, as only the manually corrected "gold standard" instances of the construction enter into the analysis. Our collostructional analyses are meant to be exact reproductions of Gries and Stefanowitsch's (2003) results, so this is why we do not try to innovate despite the limitations of the method and the corpus that was used. We use Fisher's exact test (FET) as our association measure following Gries and Stefanowitsch (2003), though many alternatives are possible (Wiechmann, 2008). We use Flach's (2017) R implementation to conduct the analyses.

4.3 Task 1: Masked Language Modeling

BERT encodes associations between words and its contexts, so we might expect that collexeme strength is also reflected in BERT models. The difference is that in collostructional analysis, exemplars of the construction are selected for analysis, while BERT has no access to annotated data. To represent constructions it would have to detect these patterns in an unsupervised way. For constructions with a single open slot that are otherwise highly lexicalized, the most obvious approach is to obtain collexeme strength values from BERT through a masked language modeling (MLM) task.

For the X waiting-to-happen construction, we replace the X with a single mask token in the exemplar sentences obtained from the BNC. This is then used as input for the MLM task, for which we use the pretrained BERT-base-cased model, after tokenization by BERT's WordPiece tokenizer. The outcome of the MLM will always be a probability distribution over the vocabulary of the model (28996 types), so we create an average probability distribution over the 35 sentences. The choice for a single mask was made for comparability to the single open slot in Gries and Stefanowitsch's (2003) analysis.

The averaging of all probability distributions over the 35 mask predictions is done to subdue the influence of other words in the natural corpus sentences on the construction. This also brings us to the limitations of the MLM task for analysing construction meaning. Firstly, only items that exist in the model vocabulary as a single token can be predicted, which excludes rare words. Secondly, many constructions have more than one open slot, such as the double object construction which has four. When examining a multiple-slot construction with a single mask, the lexical content of the other open slots would heavily impact the desired masked slot. We can see this if we try to mask only the verb slot of the DOC:

(9) Mary [MASK] John the ball.

(10) Mary [MASK] John the story.

Here, the verb prediction would be influenced mainly by the direct object and the selection restrictions it implies, rather than by the DOC as a whole – different things happen to stories than to balls. In collostructional analysis, this issue is addressed through covarying collexeme analysis (Stefanowitsch and Gries, 2005), but there is no equivalent established task for contextual word embeddings that we are aware of. A model would not have enough context to predict a DOC like *I gave you the keys yesterday* from a multiple masked language modeling prompt such as "[MASK] [MASK] [MASK] [MASK] *yesterday*". Issues also arise when it is possible for a single open slot to contain multiple tokens as tokenized by BERT. We therefore note that the established MLM approach only works for highly lexicalized constructions and propose a novel task which makes use of sentence transformers to create a semantic space of sentence vectors where we can calculate the average vector of our selected constructions.

4.4 Task 2: Sentence transformers

SBERT (Reimers and Gurevych, 2019) is a Siamese BERT model optimized for creating sentence context vectors quickly. It uses mean pooling over BERT output embeddings of each word in a sentence to assemble these vectors, also called sentence embeddings. Instead of directly comparing one context vector with another vector using the Siamese networks, we store the resulting embedding of a sentence in a FAISS-index (Johnson et al., 2017) to create a semantic space of sentence context vectors. This allows us to quickly search for similarity between these same vectors. This approach was inspired by Hoover et al. (2020), who created a sentence similarity search for large numbers of context vectors.

We first convert all 6026276 sentences from the British National Corpus into a 768-dimensional context vector and store it in a FAISS-index. The computations were performed with a NVidia 2080 Ti GPU, yielding 18.5 GB of vectors. We then convert a collection of exemplars of the double object construction into context vectors and average these same vectors. The remaining vector is consequently an average of the sentences, which itself are already averages of the words contained in those sentences. The input sentences are minimal examples of the DOC such as *You showed me some cards* but with varying part-of-speech sequences in terms of the C5 tagset, and they are listed in Appendix A. We use these simple sentences rather than the 1147 exemplars of the DOC from the BNC to minimize the effect of lexical elements outside of the construction, although it may have the consequence that exemplars of the DOC inside complex and long sentences will have lower similarity to the average vector. The average vector of these sentences is then our query for the DOC and is presented to the FAISS index for a similarity search with cosine distance as the distance metric. If this search yields corpus exemplars of the DOC as nearest neighbours out of the roughly six million context vectors from the BNC, this would show that exemplars of the DOC cluster together in SBERT, and that constructional information is represented in BERT output embeddings of words.

4.5 Evaluation method

Both collostructional analysis and the MLM task yield ranked lists of results (collexemes) which should be compared. These lists are likely to be non-conjoint; they will not contain the exact same words because BERT was not (only) trained on the BNC. To accurately measure the similarity of both rankings with the top-k outputs from both collostructional analysis and BERT, we apply Ranked Biased Overlap (RBO, Webber et al., 2010, p. 21). RBO is based on the concept of Average Overlap, which uses set intersection. The idea is to consistently intersect the two lists, now represented as sets, with an extra element added each time, up to the length of the shortest list. These steps are called the depth d of the intersection. For each intersection the overlap is computed, and over that, the average overlap is computed. This metric of list similarity also has a weighting factor p, where a lower value of p means more emphasis is placed on the similarity among elements at the top of the list. We use this to check to what extent larger coherent clusters of similar exemplars appear as predictions.

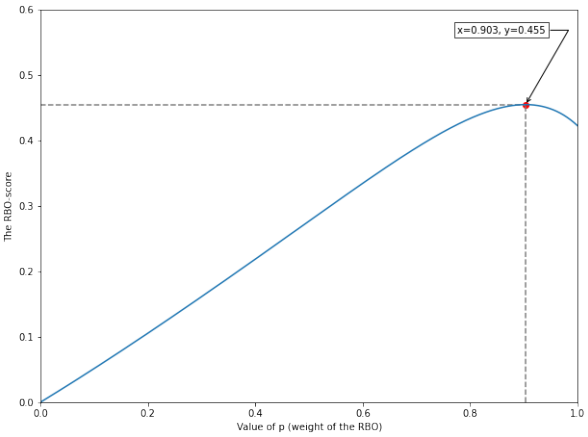

Figure 1: The RBO list comparison for Task 1 for different values of p.

The evaluation of task 2 on the double object construction is done manually. The collostructional analysis for the verb slot of this construction will produce a list of verbs; the SBERT query will output a list of similar sentences. We can only manually verify whether these similar sentences contain instances of the target construction.

5 Results

5.1 X waiting-to-happen construction

The collostructional analysis for the X waiting-to-happen construction is shown in Table 1a. This is a replication of Gries and Stefanowitsch's (2003) analysis of the same construction.[3] The small sample size (35) is apparent. BERT's top 11 predictions for the masked open slot in the construction are shown in table 1b, with their probabilities and normalized probability over this list. We compare the lists by computing their RBO with different values of the parameter p, shown in Figure 1.

The optimal value for the parameter p is 0.903, producing a score of 0.455 out of 1. This signifies that the RBO attributes higher similarity to the lists when the weight is more evenly distributed over the elements in a list. The rising slope towards this high point appears because four out of the first six elements of table 1b overlap, then it descends because the tails of the lists differ greatly. The RBO score of 0.455 indicates moderate similarity.

[3]Gries & Stefanowitsch show the regular p-value of the Fisher's Exact Test, not the negative natural log.

Collexeme	Freq.	Collex. Strength
accident	14	77.54
disaster	12	75.68
welkom	1	10.02
earthquake	1	6.01
invasion	1	4.95
recovery	1	4.33
revolution	1	4.09
crisis	1	3.81
dream	1	3.71
sex	1	3.56
event	1	2.67

(a) Simple collexeme analysis for the X waiting-to-happen construction. Collexeme strength is shown as the negative natural log of the Fisher's Exact Test.

Prediction	Prob.	Prob. Top-K
event	0.097763	0.2703
disaster	0.064560	0.1611
accident	0.059664	0.1394
explosion	0.049361	0.1390
invasion	0.016694	0.0486
earthquake	0.016525	0.0478
action	0.016206	0.0422
emergency	0.014662	0.0417
attack	0.013799	0.0403
miracle	0.013404	0.0371
adventure	0.011491	0.0324

(b) The top eleven words from the average distribution of BERT's predictions over a series of X waiting to happen construction sentences from the BNC.

Table 1: Results for the X waiting-to-happen construction

We see that both *disaster* and *accident*, the top two collexemes of Gries and Stefanowitsch (2003), constitute about thirty percent of the probability mass of the list. It is interesting that *event* is BERT's top prediction, while being ranked low by collostructional analysis. This may reflect the fact that high-frequency words are weighted differently by the two methods, and the small sample size. Overall, the items unique to the BERT list do reflect the semantics assigned to the X waiting-to-happen construction by Gries and Stefanowitsch (2003): the words describe events that are imminent, obvious and often but not always negative, and all of them are felicitous in the context of the construction.

5.2 Double object construction

Table 2a displays the top 20 strongest collexemes from our collostructional analysis, therefore accounting for 960 out of 1147 found constructions. These results differ from Gries and Stefanowitsch's (2003), as they used the comparable but smaller ICE-GB corpus, which is syntactically annotated. The prototypical verb of transfer *give* is clearly the most strongly associated collexeme. Table 2b shows the top 20 sentences discovered to be similar to an average of sentence context vectors with simple double object constructions by BERT. The emphasized words are the verbs that signify a transfer or an extension of the transfer sense as described by Gries and Stefanowitsch (2003, Table 9).

There are 14 sentences (70%) that contain the DOC. Listing by descending frequency, the verbs that appear as main verbs within the sentences are *send* (4), *give* (3), *buy* (3), *sell* (2), *say*, *tell*, *throw*, *show*, *thrust*, *bring*, and *pay* (1). Fourteen out of the twenty sentences have verbs that also appear in the collostructional analysis, with *sell*, *throw*, *thrust* and *pay* not occurring even though they are in a DOC. Out of the seven verbs used in the input sentences seen in Appendix A, four return in the sentences of table 2b (*give*, *bring*, *buy*, *show*). These four verbs are present in a total of nine sentences, meaning that eleven of the twenty other sentences contain novel verbs (verbs not included within the input). The DOC is found eight times in the eleven sentences containing novel verbs.

Of the six non-DOC sentences, three contain the prepositional dative construction, which is the DOC's near-synonymous counterpart in the dative alternation. Sentences 6 and 12 contain reported speech constructions, which have a few lexical semantic connections to ditransitives. Firstly, the verbs *say* and *tell* express extensions of the transfer sense (extension F, Communication as transfer, of Gries and Stefanowitsch 2003, Table 9). Secondly, the verbs inside the reported speech (*take* and *send*) are semantically similar to the verbs of transfer, but do not signify transfer. Sentence 18 is superficially similar to a prepositional dative construction, but with a prepositional phrase instead of a prepositional object, and the verb is a verb of transfer. All sentences are of similar length and have pronouns in the subject slot. This is probably because 7 of the 9 prompt sentences followed this pattern, dominating the average vector.

The fact that eleven of the output verbs were not present in the prompts, and the fact that seven of

Collexeme	Freq.	Collex. Strength
give	403	inf
tell	149	636.76
hand	90	578.16
show	108	449.51
offer	36	136.55
call	35	111.22
send	21	72.769
pass	19	67.992
teach	11	41.254
leave	16	36.962
bring	12	29.115
cost	8	28.109
save	7	22.687
lend	5	21.916
read	8	21.645
buy	8	20.677
reach	7	18.295
envy	3	17.762
deny	5	17.144
ask	9	17.123

(a) Simple collexeme analysis performed for the double object construction. Strength is measured with the negative natural log of Fisher's Exact Test.

	Sentence	Distance
1	I **gave** him the camera. ✓	4.19e-06
2	He **sold** me the caravan. ✓	4.3e-06
3	I **sent** them the recycling. ✓	4.34e-06
4	He has **bought** me a drink. ✓	4.37e-06
5	He **sent** me his work. ✓	4.44e-06
6	I **said** take the bus. ✗	4.47e-06
7	I **threw** him my matches. ✓	4.48e-06
8	She **showed** him the music. ✓	4.58e-06
9	He **bought** me a brandy. ✓	4.6e-06
10	She **gave** it to her Samantha. ✗**PD**	4.68e-06
11	He **sold** me a car. ✓	4.69e-06
12	**Tell** the landlord I **sent** you. ✗	4.69e-06
13	He **sent** you a letter. ✓	4.7e-06
14	He **thrust** the money at her. ✗**PD**	4.72e-06
15	I **sent** him the originals. ✓	4.73e-06
16	You **bring** me a file. ✓	4.78e-06
17	I **buy** them for him. ✗**PD**	4.81e-06
18	I **paid** those to see you. ✗	4.83e-06
19	She **gives** him a filthy look. ✓	4.9e-06
20	He **gives** me the creeps. ✓	4.96e-06

(b) The most similar sentences in the BNC in comparison to the sum of output embeddings of known double object constructions, found in *Appendix B*. The marks indicate ditransitivity.

Table 2: Results for the double object construction.

them are used in the DOC in the output, shows us that the results of this average embedding query draw from a generalized representation that goes beyond remembering words. The verb *send* is especially interesting because it is the most frequently found verb in the results of the similarity search, although it did not occur in the prompt sentences. It is a high ranking verb in the collostructional analysis and is apparently closely associated with our average double object construction context vector by SBERT.

6 Discussion

In construction grammar, constructions are defined as pairs of form and meaning. To claim that BERT represents a construction, we need more than to find its constructional meaning in the model, as Li et al. (2022) have done. It should be distinguishing form as well as meaning in its representations.

As for meaning, our masked language modeling task showed that BERT accurately represents the meaning of a highly lexicalized construction, matching a previous analysis from the linguistics literature. All eleven resulting words are nouns, and are felicitous in the X waiting-to-happen construction. *Accident* and *disaster* are attributed high probability, and Gries and Stefanowitsch's (2003) collostructional analysis clearly shows that these are the strongest ranking collexemes of the construction. This is in line with Madabushi et al.'s (2020) finding that BERT is better at identifying semantically specific constructions. However, this method cannot be used to study more abstract constructions or constructions where similar meanings may be expressed in different forms.

Our sentence embedding querying task showed that sentence context vectors derived from BERT output embeddings yield both the target construction and its near-synonymous variant, providing an approximation of at least the meaning of the construction. This somewhat goes against Madabushi et al.'s (2020) conclusion that constructional information is not explicitly available in the output layers and should be brought out by tuning on a construction identification task. Based on verb semantics, seventeen of the twenty nearest neighbour sentences to our prompt match the basic sense of the ditransitive or its extensions, with the remaining three only matching in terms of the lexical semantics of the verb(s). This similarly shows that BERT

clusters together sequences that signify a transfer in semantic space, though not perfectly.

This leaves us with the question of whether constructional information in the embedding space is also separated by form. It is interesting that the near-synonymous, but distinct in form, prepositional dative construction was found only thrice using our double object construction prompt, compared to 14 DOCs. However, this appears to be an input frequency effect: Sánchez (2018) found that in the British National Corpus, of the 17081 instances of the dative alternation they identified through automated querying, 13921 (81.5%) are DOCs as opposed to PDs, and in our top 20 result the proportion is 82%. This provides evidence that both of the near-synonymous ditransitive constructions participating in the dative alternation are clustered together in SBERT embedding space. This means we found no evidence that BERT clusters by constructional form when meaning is similar.

The fact that the similarity search provided multiple ditransitive sentences with verbs not present in the input indicates that BERT generalizes over instances of a construction. However, based on our evidence, we cannot claim that BERT represents CxG constructions, defined as pairs of form and meaning, because the model does not fully distinguish constructions with similar meaning but different form.

There are some caveats with this method. There is no evident explanation of why exactly some sentences are found to be similar. It is impossible to deduce from BERT's output embeddings why phrases such as *I said take the bus* and *He sent me his work* are considered to be similar to each other. The two sentences clearly contain different constructions, and both semantically and syntactically they do not appear to be that similar besides both containing a verb of transfer. This black box problem extends to the waiting-to-happen construction. The predictions for the open slot are consistently nouns that fit the input exemplars of the construction used in the experiment but why BERT predicts these exact nouns is unclear, as links to features of the training data are not preserved by transformer models. We also cannot fully exclude the influence of the sentence contexts that surround a construction, even when averaging the output distribution for multiple input sequences.

7 Conclusion

Overall, we have shown that the two methods we propose can retrieve substantial amounts of interpretable constructional information from BERT. Unlike previous work, taking the content of open slots as indicative of constructional meaning enabled us to make comparisons between the position in semantic space of embeddings potentially representing the construction, and supervised data from the linguistics literature. It appears that BERT represents more lexicalized constructions better than more abstract constructions, and that BERT output embeddings do contain constructional information. However, to claim that BERT contains representations of CxG constructions it would be necessary to find evidence for separation in contextual embedding space of near-synonymous constructions. This was not tested in previous work and we did not find it in this study.

In future work, it would be interesting to perform a contrastive analysis of two alternating constructions in SBERT embedding space, while controlling for context, to further investigate the question of form. It would also be useful to compare methods for obtaining sentence context vectors for a construction that might target the construction more accurately as evaluated against traditional collostructional analyses. Some further analyses might be insightful, such as viewing all instances of a construction as a cluster in embedding space and analysing the cluster coherence or the properties of its outliers. We could consider averaging vectors of larger or smaller parts of sentences containing the construction, or with more variation in terms of sentence lengths, lexical diversity or other factors that might affect the use of the target construction. It would also be interesting to scale up the analysis to larger corpora, although the results would be less comparable to those from the linguistic literature. It would also be interesting to see where BERT positions ungrammatical exemplars, which of course did not occur in previous corpus studies.

The methods we propose might be of interest to linguists, for example as a less transparent but more data-driven way of obtaining construction exemplars or studying collexemes, requiring less annotation effort. Lastly, we hope to have shown that Construction Grammar can serve as a framework for studying meaning representations beyond the token in large language models, even for those not interested in Construction Grammar specifically.

Limitations

Although the work is aimed at better understanding BERT's internal representations, there is no transparent way to know on the basis of what features of the training data some particular sentences are found to be similar. For task 2, the representations may have been affected in unexpected ways by the process of creating averaged sentence embeddings. There is no way to fully exclude the effect of lexical context and thus get a representation of the meaning of a construction without noise in unsupervised transformer models, which may affect the extent to which we can accurately probe for a construction. In task 1, the set of potential words that could be predicted is limited by the BERT tokenizer's vocabulary.

Some limitations are caused by our choice to compare to Stefanowitsch & Gries's results. We used the relatively small corpus that they used and we have demonstrated the methods for a limited set of two English constructions. We also limited the analysis of the results to the top 20 most strongly associated collexemes, as they did. Using a larger corpus would probably yield more than 35 instances of the X-waiting-to-happen construction that S&G found and that our reproduction yielded. We also did not experiment with ungrammatical or perturbated input as such results cannot be compared to the original corpus study, which only uses natural language data. The scope of our study was also limited by to the construction-specific data collection, preprocessing and manual annotation required. For modern web-scale corpora, task 2 would require significant GPU resources.

As the way in which BERT is trained clearly differs in many ways from how humans acquire language, also according to the Construction Grammar framework, this BERT-based work does not warrant any claims about how human language works besides extremely broad ones and findings are limited to conclusions about transformer-based language models.

References

Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic Systems and the Representation of Linguistic Knowledge*, Collective Volume:1–18.

BNC Consortium. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Jeremy K Boyd and Adele E Goldberg. 2011. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, pages 55–83.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Roberto Brunelli. 2009. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Susanne Flach. 2017. *collostructions: An R Implementation for the Family of Collostructional Methods*. R package version 0.1.0.

Lauren Fonteyn, F Karsdorp, B McGillivray, A Nerghens, and M Wevers. 2020. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. *Computational Humanities Research CEUR-WS*, pages 257–268.

Adele Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. Cognitive theory of language and culture. The University of Chicago Press, Chicago, Ill., [etc.

Adele Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Stefan Th. Gries and Anatol Stefanowitsch. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, 9(1):97–129.

Martin Hilpert. 2014. *Construction Grammar and its Application to English*. Edinburgh textbooks on the English language. Advanced. Edinburgh University Press, Edinburgh.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A visual analysis tool to explore learned representations in Transformer models. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics: System Demonstrations, pages 187–196, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734.

Donald E Knuth, James H Morris, Jr, and Vaughan R Pratt. 1977. Fast pattern matching in strings. *SIAM journal on computing*, 6(2):323–350.

Manfred Krifka. 2004. Semantic and pragmatic conditions for the dative alternation. *Korean Journal of English Language and Linguistics*, 4(1):1–31.

Natalia Levshina and Kris Heylen. 2014. A radically data-driven construction grammar: Experiments with dutch causative constructions. *Extending the scope of Construction Grammar*, 54:17.

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)*, pages 7410–7423.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Harish Madabushi, Tayyar, 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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ludovica Pannitto and Aurelie Herbelot. 2022. Can recurrent neural networks validate usage-based theories of grammar acquisition? *Frontiers in Psychology*, 13.

Barbara H Partee. 2015. *Subject and Object in Modern English*. Routledge Library Editions: The English Language. Taylor and Francis.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, Chu-Ren Huang, and Alessandro Lenci. 2019. Distributional semantics meets construction grammar. towards a unified usage-based model of grammar and meaning. In *First International Workshop on Designing Meaning Representations (DMR 2019)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Juan Lorente Sánchez. 2018. 'give it him and then i'll give you money for it.' The dative alternation in Contemporary British English. *Research in Corpus Linguistics*, pages 15–28.

Anatol Stefanowitsch and Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1):1–43.

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*, pages 6361–6369.

Remi van Trijp. 2017. A computational construction grammar for english. In *2017 AAAI Spring Symposium Series*.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers–the BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20:1–20:38.

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. *arXiv preprint arXiv:2210.13181*.

Daniel Wiechmann. 2008. On the computation of construction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Lingustic Theory*, 4:253–290.

A Ditransitive construction input sentences for BERT

(1) I give him the ball.
PNP VVB PNP AT0 NN1

(2) He gives me the remote.
PNP VVZ PNP AT0 NN1

(3) He brought me the coat.
PNP VVD PNP AT0 NN1

(4) Mary gave John a pencil.
NP0 VVD NP0 AT0 NN1

(5) She bought him a pair of socks.
PNP VVD PNP AT0 NN0 PRF NN2

(6) Sam told Harry a lie.
NP0 VVD NP0 AT0 NN1

(7) You showed me some cards.
PNP VVD PNP DT0 NN2

(8) I transferred Carrie the money.
PNP VVD NP0 AT0 NN1

(9) She left me her number.
PNP VVD PNP DPS NN1

B X waiting-to-happen input sentences for BERT

(1) So the East German events of this week were an [MASK] waiting to happen

(2) ‘ Just Cause ’ was a carefully planned [MASK] just waiting to happen , poised at the starting gate for the kind of justification that the macho thug in Panama City was bound to provide sooner or later

(3) Roger Bootle , chief economist at stockbrokers Greenwell Montagu , said yesterday : ‘ I think the [MASK] has been waiting to happen for the last couple of months

(4) Part two develops this theme , identifying ‘ [MASK] waiting to happen ’ associated with liquified natural gas , oil and gas , power stations and grids , and nuclear power

(5) Any one of these may be a [MASK] waiting to happen

(6) ‘ We must stop this motoring madness ’ RESIDENTS CALL FOR ACTION SPEEDING motorists are putting lives in danger at Holybourne , and worried residents are certain that ‘ an [MASK] is waiting to happen ’

(7) Unless , of course , it was an [MASK] waiting to happen

(8) Bands like that are [MASK] waiting to happen in a world where 99 per cent of groups are casualties of their own blatant ambition

(9) Learn baby-swop lesson WHILE my heart goes out to the parents in the baby-swop drama , I have to agree with the midwife interviewed on TV who said that it was ‘ a [MASK] waiting to happen ’

(10) Only once before has this riveting axis started a game and the first-half goal rush was an [MASK] waiting to happen

(11) The arguments that a new industrial [MASK] is waiting to happen in space are , for now , unconvincing

(12) We have been warning ever since the company was formed of the [MASK] at the heart of the company waiting to happen : now IBM ’s signalling of the death of the mainframe coincides with the German economy heading into the same kind of structural — rather than cyclical — recession that is busy laying waste to IBM itself

(13) Cartoon accident In your December issue (page 1330) , you had a cartoon about an [MASK] waiting to happen

(14) ‘ Why ? ’ ‘ Because Stud ’s like an [MASK] waiting to happen , that ’s why

(15) Every sixth or seventh day or so , in the morning , as we prepare to sack out , and go through the stunned routines of miring , of mussing (we derange each eyebrow with a fingerstroke against the grain) , Tod and I can feel the [MASK] just waiting to happen , gathering its energies from somewhere on the other side

(16) The Sony voice-activated machine was used to record the conversation with DEA attaché Micheal Hurley seven months before Lockerbie in which Coleman warned him of the ‘ [MASK] waiting to happen ’

(17) ‘ This is a [MASK] waiting to happen , ’ he added , in a prophecy that would come back to haunt him

(18) — ‘ Well — for a business [MASK] waiting to happen , you seem to have come off remarkably unscathed

(19) A [MASK] was waiting to happen

(20) For them , last Saturday was an [MASK] that had been waiting to happen

(21) First , there is the utter incompetence of the Government ’s management of the economy ; secondly , there are its housing policies , described in The Independent as ’ a [MASK] waiting to happen ’

(22) as if [MASK] ’s just over the horizon , waiting to happen to me , as weird and wonderful as all the things that happened last autumn

(23) All went so well after that that there had just had to be one monumental [MASK] waiting to happen , Leith later realised

(24) An [MASK] waiting to happen

(25) People living near the site say it was an [MASK] waiting to happen

(26) Male speaker The state of the building means it was an [MASK] waiting to happen

(27) You are absolutely right to condemn their actions which are little more than [MASK] waiting to happen

(28) It was an [MASK] waiting to happen

(29) It was an [MASK] waiting to happen

(30) The explicit confirmation that the Commons really does not matter was the real constitutional [MASK] waiting to happen , vindication to all those Euro-sceptics who argue that Maastricht rides roughshod over parliamentary sovereignty

(31) This latest [MASK] is a graphic illustration of the disaster that 's waiting to happen out there

(32) Mr Stewart said that there was an [MASK] waiting to happen and he feared lives would be lost

(33) Councillors will tell Lord Donaldson that the grounding of the ship on Stroma nine days ago is a graphic example of a [MASK] waiting to happen

(34) But then , he 's not the only drinker with that problem ... Be safe not sorry : Page 13 Jagger urged to rebuild marriage ROLLING Stone Keith Richard today said Mick Jagger and Jerry Hall should get back together and that Bill Wyman 's marriage to Mandy Smith was a [MASK] waiting to happen

(35) The government 's fear is that there may be many more [MASK] waiting to happen , and if racial conflict does spread in South Africa , it could seriously unsettle a delicate process of change which is underway

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6 (Discussion) and final Limitations statement

☒ A2. Did you discuss any potential risks of your work?
We do not foresee any noteworthy risks associated with looking for particular types of linguistic structures in BERT beyond what is known about probing LLMs already

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 4 and 5

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All artifacts we used are widely used and their licenses are known.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We used academic tools for research purposes, not too noteworthy.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We used a widely used corpus made use of public data that is also old.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 2 and 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.