Form and Meaning in Contextual Embeddings: Analyzing the Constructional Representation of Complementizer Omission in Dutch and Multilingual Encoder Language Models

Marije Kouyzer
University of Amsterdam

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

Abstract

We investigate the presence of Construction Grammar-like representations in encoder language models beyond English by examining how a Dutch allostruction is encoded in embedding space. We compare embeddings of sentences that differ in form but not in meaning (having or lacking a complementizer) with embeddings of sentences that differ in form and meaning (having or lacking a negation marker). We use both monolingual and multilingual models, expecting multilingual models to distinguish this allostruction less well than monolingual models, as this complementizer optionality does not exist in English. We find that monolingual models do distinguish the two allostructions in embedding space, contrasting with previous work on the English dative alternation. Multilingual models show mixed results. This shows that evidence of constructional representations can also be found in language models beyond English, even for a language that is not the dominant one of a multilingual model.

1 Introduction

With the rising popularity of the use of Large Language Models (LLMs) in many different contexts, it is increasingly important to have a better understanding of how they work. Recently, the field has turned to linguistic experimental and theoretical work to find potential explanations for structures that might underpin the linguistic capabilities of large language models. Psycholinguistic studies of human behaviour provide benchmarks and norms against which LLM behaviour can be compared (Conde et al., 2025), and linguistic theory can provide accounts of what sorts of features or structures might be encoded in the models (Baroni, 2022).

A recent body of work specifically focuses on construction grammar theory as a potential explanatory framework (Bonial and Madabushi, 2024), as this line of theoretical inquiry is particularly compatible with how modern large language models work. Goldberg (2024) discusses these parallels from a theoretical linguisitic perspective. The theory of Construction Grammar (CxG) defines linguistic constructions as learned pairs of forms and meaning (Goldberg and Suttle, 2010). These can vary from individual words to those of complex syntactical patterns.

Meaning representations in large language models rely (at least in part) on the distributional assumption (Boleda, 2025) that a word's meaning can be derived from its distribution in natural language, and is encoded in continuous rather than categorial representations. This aligns well with the usage-based learning fundamentals of construction grammar, where constructions are acquired by human learners based on distributional data (Boyd and Goldberg, 2011).

Based on experiments with recent (mainly encoder) language models, there is an ongoing debate regarding whether and to what extent large language models encode CxG constructions. Madabushi et al. (2020) showed that the BERT encoder model can detect whether sentences contain the same construction, and Li et al. (2022) found that sentence embeddings of argument structure construction are clustered by their construction rather than by their verb in BERT-based models. However, there are also limits on the extent to which constructional information is encoded. Veenboer and Bloem (2023) note that BERT models do not distinguish constructions that have the same meaning but different form (allostructions) in embedding space, and Bonial and Madabushi (2024) find that higher-level, more schematic constructions with fewer fixed lexical elements are very difficult to distinguish for GPT models.

Most of this work is on English constructions, while most large language models are dominated by English-language training data. In multilingual models, constructional information might be less effectively encoded for languages that are less

prevalent in the training data. It is also not clear whether these findings apply when an English-dominated language model is fine-tuned for another language, as often happens with modern decoder language models.

We explore these issues by assessing Dutch monolingual and multilingual encoder models for their representations of a Dutch allostruction, with two different forms and very little meaning difference, contrasting it to a condition that is similar in form but has a large meaning difference. We take naturally occurring instances of these three different constructions as clusters in sentence embedding space and use internal cluster analysis methods to analyze the encodings of sentences containing these linguistic constructions.

2 Background

We focus on complementizer optionality as instantiated by Dutch sentences with a verbal 'te'-infinitival complement clause with an optional complementizer 'om' (IJbema, 2002). In a corpus study by Bouma (2017), this construction has been well-documented as a case of optionality that encodes little to no meaning distinction, and its occurrence is largely accounted for by the specific governing verb that is used. In construction grammar, this type of alternating constructions with little meaning distinction have been termed allostructions by Cappelle (2006).

The following Dutch example sentences contain this infinitival complement clause where the complementizer 'om' can be included, but is also grammatical and has the same meaning without this complementizer. The English translation of these sentences is included below each sentence.

1. *Zij beloven (om) te helpen.*
   They promise to help.

2. *Ik ben vergeten (om) eieren te kopen.*
   I forget to buy eggs.

3. *Jij besluit (om) naar Zuid Amerika te reizen.*
   You decide to travel to South America.

These patterns contrast with German and English, where no equivalent complementizer is used in infinitival contexts. English does have complementizer optionality with 'that' being optional in certain context, while the German and Dutch equivalents of this are not optional.

2.1 Linguistic Constructions and Large Language Models

Construction Grammar (Goldberg and Suttle, 2010), often abbreviated to CxG, is a theory within the field of linguistics that focuses on the existence of linguistic constructions as the foundation of which language is built. The constructions are any pairings of form and function, which includes individual words and patterns. The patterns can be largely predefined with one or two open slots, or consist of many open slots. According to CxG theory, the entire grammar of a language consists of combinations of different constructions.

Various studies have aimed to establish whether language models encode such constructions, either through post-hoc explainability (e.g. testing whether a model can differentiate constructions, which would probably require knowledge of constructions) or intrinsic explainability (e.g. testing whether constructions theorized by linguists can be detected in the internal representations of a model). This area of research was surveyed by Bonial and Madabushi (2024), who found that the presence of constructional knowledge in LLMs may be limited to the more substantive constructions lower in the hierarchy of the 'constructicon' (constructional lexicon), with more lexically fixed elements.

2.1.1 Post-hoc explainability

In an initial post-hoc explainability experiment, Madabushi et al. (2020) use sentences from featured articles from Wikipedia and automatically label them with constructions within each sentence from a list of 22 thousand constructions. It is not entirely clear if all of the 22 thousand constructions used for this study would be considered constructions by linguists. This study shows that BERT models have access to constructional information and that BERT is capable of distinguishing between different linguistic constructions.

Weissweiler et al. (2022) used CxG to investigate syntactic and semantic encoding in three BERT models. The construction in question is the comparative correlative in English (e.g. *the more, the merrier*). Using minimal pairs, where both sentences look similar but one is an instance of the construction and one is not, they tested if the models were capable of distinguishing the constructions syntactically. They found that while the models were able to distinguish the construction syntactically, they did not show a semantic understanding of the construction.

Subsequently, Zhou et al. (2024) observe that GPT-4 and Llama 2 fail to distinguish between a group of constructions with three classes of adjectives that cannot be distinguished by surface features, in a natural language inference task. Scivetti et al. (2025b) also document limited generalization ability of GPT-4o and Llama 3 models for English phrasal constructions.

In another post-hoc experiment, Weissweiler et al. (2025) investigate the understanding of various decoder language models of the caused-motion construction ("She sneezed the foam off her cappuccino") in an experiment where non-prototypical verbs are substituted by a prototypical motion verb. They find that all models struggle with understanding the characteristic motion component of this construction, indicating that decoder models have limited generalization capabilities regarding this type of less frequent CxG constructions that humans are able to robustly generalize with.

Scivetti and Schneider (2025) study the English version of the noun-preposition-noun construction ('day to day') in BERT, by training a probe to distinguish natural language instances of this construction from similar-looking distractors. They show that BERT is also able to disambiguate the sense of the construction in context, indicating representation of constructional meaning for this construction. This construction also exists in Dutch.

2.1.2 Intrinsic explainability

In a more intrinsic explainability experiment, Li et al. (2022) investigate linguistic constructions in several BERT models. They use a sentence sorting task and compare embeddings of sentences with the same constructions with embeddings of sentences that contain the same verb. They find that the sentences with the same constructions are closer together in the embedding space than the sentences with the same verbs.

Veenboer and Bloem (2023) evaluate to what extent instances of the dative alternation, an alternation of two constructions with very similar meaning, are mixed or distinct in embedding space. They use sentence transformers to embed instances and also obtain an average vector for each variant, then examine the nearest neighbours of the average vector. They find that the model does not distinguish between constructions that have the same meaning but different forms.

Ramezani et al. (2025) investigate the processing of argument structure constructions in BERT, where they quantify the degree of clustering of these constructions using Generalized Discrimination Values. With a controlled set of sentences, they use CLS tokens as construction representations to investigate which layers of the model best cluster instances of the construction, finding good results with early layers but not the first layer.

Rozner et al. (2025) criticize previous work for studying models that have been trained on far more data than what human children use to acquire language. They study constructional learning with developmentally plausible amounts of data with masked language models from the 2024 BabyLM challenge, and find that various types of CxG constructions can be learned also under these data constraints, based on model output distributions.

Relatedly, Scivetti et al. (2025a) use an experiment with the English LET-ALONE construction to show that BabyLM models are able to generalize the form of this construction (showing sensitivity to formal constraints), but do not make correct inferences about its meaning, revealing a similar asymmetry as was observed by Veenboer and Bloem (2023) and Weissweiler et al. (2022).

Dunn and Eida (2025) highlight another side of LLM's potential construction learning capabilities – that of fake constructions. Two experiments, metalinguistic probing of GPT-4 and behavioural probing of contextual embeddings following the method of Li et al. (2022), show that models also hallucinate constructions that do not exist, over-learning patterns that do not exist for human speakers.

2.1.3 Beyond English

CxG-LLM studies have strongly focused on English, which is also well-represented in linguistic studies of construction grammar. We are only aware of three studies that relate to constructional knowledge of LLMs beyond English. First, Huang et al. (2025) perform an evaluation using minimal pairs of Mandarin verb-resultative Complement Constructions, introducing the ZhVrcMP benchmark of minimal pairs. Second, Bunzeck et al. (2025) experiment with German BabyLMs to investigate whether the constructional profile of child-directed speech is beneficial to acquisition, by manipulating the relative frequencies of constructions in training data. Third, Li et al. (2026) investigate the representation of information structure constructions through an experiment with minimal pairs of Mandarin constructions of this type.

However, these studies do not address questions

related to model multilinguality.

2.2 Sentence embeddings

Different aspects of the linguistic capabilities of encoder language models such as BERT (Devlin et al., 2019) and EuroBERT (Boizard et al., 2025) have been and continue to be studied in order to gain insight on what sorts of structures can be learned from unsupervised learning of textual data (Rogers et al., 2020). The use of embeddings for meaning representation aligns well with the construction grammar idea that constructions have their own semantics. Ideas from sentence embeddings are therefore used in the CxG-LLM literature to represent constructions.

There are several ways to get sentence embeddings from word embeddings. Firstly, the CLS-token is used within BERT models explicitly to represent the entire sentence. Secondly, mean pooling can be used to create sentence embeddings from word embeddings. Huang et al. (2021) observe that this better represents sentence semantics than the CLS token. Third, max pooling can be used in a similar way. Sentence-BERT (Reimers and Gurevych, 2019) is a model fine-tuned on BERT models. It applies the strategies mentioned above for creating sentence embeddings, while also being tuned on the objective of distinguishing sentence triplets to make the vectors more distinctive.

In this study, we use mean pooling to create sentence embeddings, because it tends to perform better than the CLS-token on textual similarity benchmarks (Reimers and Gurevych, 2019). We decided not to use Sentence-BERT because a pretrained model was not available for all models used in this study, but more importantly because it is specifically tuned to give semantically relevant outputs, while we are interested in the model's actual representations without tuning for specifically that.

2.3 Internal Cluster Analysis

A few previous studies viewed constructions as labeled groupings of sentence embeddings (representing instances of a construction) in embedding space (Veenboer and Bloem, 2023; Ramezani et al., 2025) We take this idea a step further by applying general cluster quality metrics to gain insight into how instances of a construction are distributed in embedding space. Specifically, we take labeled instances of constructions as predefined clusters and use these metrics to find out how coherent and well separated these clusters are. A similar idea was applied by Zhou and Bloem (2021), but they used domain-specific semantic categories as cluster labels for word embeddings rather than constructions as cluster labels for sentence embeddings.

Two of these internal cluster analysis methods are the Silhouette Score (Rousseeuw, 1987) and the Davies-Bouldin Index (Davies and Bouldin, 1979). Both of these methods have been used extensively as they are model-agnostic and usable in unsupervised settings. Usually they are used to measure the effectiveness of clustering methods, as opposed to the pre-existing clusters we use here. Due to this setting, we cannot use cluster metrics that are tied to particular clustering algorithms.

The Silhouette Score uses inter- and intra-cluster distances for each pair of members to calculate how well separated the clusters are. For each member of the clusters a Silhouette Coefficient is calculated and the average of these is the Silhouette Score. The Davies-Bouldin Index is a relative cluster analysis method that can be used to compare cluster coherence. The Davies-Bouldin Index calculates the maximum ratio between the intra-cluster distances and the inter-cluster distances.

We use both of these metrics, because the Silhouette Score gives us an objective idea of how well separated the clusters are, while the Davies-Bouldin Index can be used to compare the cluster coherence of one group with another.

Ramezani et al. (2025) also use cluster analysis to investigate how constructions are clustered in embedding space, but they use the Generalized Discrimination Value, another metric that compares within-cluster distances to between-cluster distances. We did not include this metric as it has a less intuitive interpretation and the two aforementioned metrics are more commonly used.

3 Methodology

3.1 Data

The data used for this study has been extracted from the SoNaR portion (Dutch Language Institute, 2015) of the Lassy Groot corpus (Dutch Language Institute, 2023). This is a 500 million token treebank of Dutch sentences from the Netherlands and Flanders with automatically generated syntactic annotations.

3.1.1 Data Extraction

Sentences containing three different linguistic structures, the verbal 'te'-infinitival complement

clause both with and without the complementizer 'om' as well as sentences containing the word 'niet' followed by the verbal 'te'-infinitival complement clause without the word 'om', are extracted from the corpus using the python library lxml and XPath queries. The full queries to identify these sentences can be found in appendix A. To create these queries, GrETEL (Odijk et al., 2017), a search engine for syntactically annotated corpora, was used.

From the corpus, 803,119 sentences were found containing this construction with 'om', 722,619 without 'om' (the form contrast), and 14,588 with 'niet' and without 'om' (the meaning contrast).

3.1.2 Data Preprocessing

Hits without 'om' are not necessarily instances of the target construction. To ensure that we only capture cases that can engage in this alternation, a list of verbs was created that optionally take the word 'om' as a complementizer, following the methodology used by Bouma (2017).

A few sentences of each of the verbs for both categories were manually checked to see if they indeed contained an optional 'om' complementizer. 33 verbs were excluded because not all their senses can engage in the alternation, and 91 verbs were excluded because we found non-alternating instances, such as idioms where 'om' cannot be inserted. This left us with a list of 117 verbs, which includes verbs such as 'proberen' ('to try'), 'toelaten' ('to allow'), 'dienen' ('to serve'), 'vragen' ('to ask'), 'oproepen' ('to summon'), 'weigeren' ('to refuse') and 'besluiten' ('to decide'). The full list can be found in Appendix B. The sentences in all three categories were then filtered to only include sentences with verbs from this final list.

This means that all selected sentences will contain a clause of one of the following forms:

- a verb from the list + the word *om* + the word *te* + an infinitive verb
- a verb from the list + the word *te* + an infinitive verb
- a verb from the list + the word *niet* + the word *te* + an infinitive verb

We filtered out 5,668 sentences that had both the construction with and without 'om', and we filtered out 3,015 sentences from the category without 'om' that also had the 'niet' negation in it and were matches for both these construction types. 55 sentences in both the 'om' and the 'niet' group were filtered out. We also filtered out sentences shorter than the first quartile minus 1.5 times the interquartile range and longer than the third quartile plus 1.5 times the interquartile range (49 words). Finally, all three construction groups were downsampled to 2,645 sentences each (the size of the smallest 'niet' category), for a total of 7,935 sentences.

3.1.3 Synthetic Data

We created a synthetic dataset by removing the word 'om' from the sentences in the 'om' condition, creating an equivalent to the condition without 'om'. This enables a more direct comparison of embeddings of the two constructions, but with sentences that are not naturally occurring.

3.2 Encoder models

We used two Dutch and two multilingual encoder models to be able to assess whether form and meaning differences are clustered differently in these types of models (Table 1).

BERTje (de Vries et al., 2019) is a monolingual Dutch BERT model that has been shown to perform well on intrinsic semantic (Brans and Bloem, 2026) and extrinsic (de Vries et al., 2023) evaluations. It uses the same architecture and size as the original BERT. RobBERT-2023 (Delobelle and Remy, 2023), another monolingual Dutch model, is a new version of the RobBERT model (Delobelle et al., 2020) based on RoBERTa (Liu et al., 2019). We use the large version, which has 355 million parameters and uses a Tik-to-Tok tokenizer (Remy et al., 2023).

As for the multilingual models, mBERT (Devlin et al., 2019) also uses the same architecture as the original BERT and was trained on a dataset consisting of the Wikipedia texts of the 104 biggest languages. EuroBERT (Boizard et al., 2025) is a more recent model trained on 15 different languages, including Dutch. The EuroBERT model is trained on a multilingual dataset with 5 trillion words. However, it was observed to not have great embedding representations for Dutch without further tuning (Vlantis and Bloem, 2025).

For all models, sentences in our dataset are turned into sentence embeddings through mean pooling,

3.3 Cluster Analysis

To be able to compare the influence of meaning and form of the linguistic constructions, the data

Model	Tokenizer	Parameters	Data	Language	Based on	Dimensions
RobBERT-2023	Tik-to-Tok	355M	3.3B	Dutch	RoBERTa	1024
BERTje	WordPiece	110M	2.4B	Dutch	BERT	768
mBERT	WordPiece	179M		>100 languages	BERT	768
EuroBERT	LlaMa 3	210M	5T	15 languages	BERT	768

Table 1: Overview of encoder models.

is divided into two comparisons of two constructions that each only differ by one word. The first comparison is the groups with and without 'om', corresponding to the 'om'-optionality. These constructions have the same meaning, and this comparison is used to measure how well the clusters separate based on form.

The second comparison is the group with the negation marker 'niet' and the group without 'om'. Since the addition of the negation 'niet' changes the meaning of the sentences about as much as is possible with a single short word, this comparison is used to measure how well the clusters separate based on meaning as well as form.

We use cluster quality metrics to assess to what extent both sides of each comparison are mixed in embedding space. The Silhouette Score calculates the average Silhouette Coefficient (Rousseeuw, 1987) for all sentences:

$$s = \frac{b - a}{max(a, b)}$$

where a is the mean distance between a sentence embeddings and the other embeddings within the same group and b is the mean distance between a sentence embedding and the embeddings in the other group. Scores closer to 1 indicate better cluster separation. A score near 0 means that the clusters are overlapping, while negative values point to sentences that might be closer to sentences of the other cluster.

The Davies-Bouldin Index (Davies and Bouldin, 1979) measures cluster coherence, and a lower value indicates that the clusters are better separated. It divides the sum of the average distance within each cluster by the distance between the centroids of the clusters:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{d_{ij}}$$

where k is de number of clusters, i and j are different clusters, s_i and s_j are the average distance between each point of that cluster and its centroid and d_{ij} is the distance between the centroids of clusters i and j.

	Form	Meaning	Difference
RobBERT-2023	0.016	0.032	-0.016
BERTje	0.014	0.031	-0.017
EuroBERT	0.012	0.026	-0.014
mBERT	0.012	0.028	-0.017

Table 2: Silhouette Score results for the two comparisons: difference in form (with/without 'om') and difference in meaning (with/without 'niet'). Values closer to 1 mean the cluster coherence is better.

	Form	Meaning	Difference
RobBERT-2023	7.876	5.468	2.408
BERTje	8.427	5.518	2.909
EuroBERT	8.856	6.068	2.789
mBERT	9.074	5.796	3.278

Table 3: Davies-Bouldin Index results for the two comparisons: difference in form (with/without 'om') and difference in meaning (with/without 'niet'). A lower value indicates better cluster coherence.

Following Veenboer and Bloem (2023), we also create prototypes for each construction by averaging their embeddings and examine whether the ten nearest neighbours (by Euclidian distance) of the prototypes are instances of the target construction or of the other construction in the comparison. With a random baseline of 50%, higher percentages of the target construction indicate better separation in embedding space.

4 Results

4.1 Form and Meaning

We start with the comparison of the constructions that differ only in form, and those that differ in form and meaning. The results can be found in Table 2 for the Silhouette Scores and Table 3 for the Davies-Bouldin Index. As expected with related constructions in high-dimensional embedding space, the Silhouette Scores close to 0 indicate a lot of cluster overlap. Nevertheless, the observed

	Top 10	Top 100
RobBERT-2023	90%	72%
BERTje	80%	66%
EuroBERT	90%	67%
mBERT	30%	61%

Table 4: The percentage of 'om te' construction instances in the top 10 and top 100 sentences closest to the prototype of the 'om te' construction in the form comparison.

	Silhouette	D-B Index
RobBERT-2023	0.0032	16.630
BERTje	0.0028	17.574
EuroBERT	0.0030	17.567
mBERT	0.0017	21.884

Table 5: Silhouette scores and Davies-Bouldin Indices for the synthetic data containing the sentences with 'om te' and the same sentences with 'om' removed.

differences are consistent across models, with the constructions that differ in form and meaning being better separated than those that differ only in form. The DBI, a relative metric, shows that the comparison of constructions that differ in form and meaning have better cluster coherence (lower scores). These results are what we would expect of good semantic representations, and show that our metrics are sensitive to meaning differences between constructions.

We also split the results by sentence length, with short sentences being defined as containing less than 20 words and long sentences containing 20 or more words, and find that the constructions in our dataset are better separated for the long sentences than for the short sentences, according to both metrics (Appendix C).

4.2 Model comparison

To compare the different BERT models, we look at Table 2 for the Silhouette Scores and Table 3 Davies-Bouldin Index. Table 2 shows that for the monolingual models (RobBERT-2023 and BERTje), the clusters seem to be slightly better separated than for the multilingual models (EuroBERT and mBERT). The DBI (Table 3) gives similar results, with the monolingual models outperforming the multilingual models. This difference is bigger in the form comparison. The model that best separates the constructions in both groups is RobBERT-2023, followed by BERTje. In the form comparison mBERT performs the worst, while in the meaning comparison EuroBERT performs the worst.

4.3 Prototype nearest neighbours

Next, we focus on the more difficult task of separating the two allostructions. Table 4 shows the types of nearest neighbours of the 'om te' constructional prototype in the form comparison. The results are shown for 10 and 100 nearest neighbours. Higher percentages show that more of the nearest neighbours are the target construction, with a baseline of 50%.

From the top ten closest sentences to the average for both the RobBERT-2023 model and the EuroBERT model, nine sentences are the target construction, while for the BERTje model eight sentences are. This suggests that in these three models, the two allostructions are quite separated in sentence embedding space. The mBERT model does less well at this separation. Only three of the sentences closest to the average are the target, less than random chance. For the top 100 nearest neighbours, we still see notable separation (72%-61%), though it is far from complete.

4.4 Synthetic Data

Our comparison between the two forms using natural language data isn't controlled for factors such as sentence length, word frequency or other contextual properties that could potentially affect the results. The separation metrics could be artefacts of other properties of the test sentences besides the use of 'om'. To have a more controlled comparison, we repeat the comparison with synthetic data where 'om' was removed from the 'om te' construction sentences.

This should lead to lower separation scores as the comparison will consist of pairs of very similar sentences where each sentence has a corresponding synthetic sentence that is close to it in embedding space. Indeed, Table 5 shows that Silhouette Scores are clearly lower than in the original comparison (Table 2), and the DBI scores are also higher, indicating worse separation. Some separation remains, but finer-grained methods that target particular tokens of the construction rather than entire sentence embeddings are likely needed to assess whether allostructions are distinct in embedding space.

5 Discussion

In general the Silhouette Scores showed that none of the clusters are well separated. This makes sense,

because the sentences are natural language sentences about a wide variety of topics and do not have that much in common semantically. With controlled sentences that only contain the target construction and nothing else, we might expect somewhat clearer separation. So we focus mainly on the relative cluster coherence compared to the other conditions.

As expected, we find that the sentences that differ in meaning as well as form have better cluster coherence than the sentences that differ only in form. An unexpected result is that, generally, the longer sentences have a better cluster coherence than the shorter sentences. One possible explanation for this is that there were slightly more long sentences than short sentences included in the study. Another explanation could be that it is easier for BERT models to distinguish between linguistic constructions in longer sentences than it is in shorter sentences.

The monolingual models perform better than the multilingual models. This difference is bigger for the condition that differs only in form than in the condition that differs in meaning as well as form, where the Davies-Bouldin Indices were close together. This indicates that it is easier for monolingual BERT models to pick up on differences in linguistic constructions purely on form than it is for multilingual models. The clearest comparison is between monolingual BERTje and multilingual mBERT, which have the same dimensionality. BERTje shows better separation between the two. Overall, the RobBERT-2023 model performed best, though this is also the model with a higher dimensionality than the others.

Vlantis and Bloem (2025) also found that monolingual models performed better than multilingual models on a semantic similarity task. In their study EuroBERT tended to perform worse than both the monolingual models, BERTje and RobBERT, and than the mBERT model, while here we see that EuroBERT performs comparable or slightly better than mBERT. The results we find here are also similar to the Dutch Model Benchmark (de Vries et al., 2023) which compares the results of different language models on Dutch Natural Language Processing tasks, such as part-of-speech tagging and sentiment analysis, using BERTje as the baseline. The Dutch Model Benchmark also shows better results for RobBERT-2023 than BERTje and worse results for mBERT than BERTje, as we found here as well. EuroBERT is not included in this benchmark.

Similar results were found when dividing the sentences based on their length, but there was an effect of sentence length for the BERTje model specifically.

The results from nearest neighbours to construction prototypes show that the models have separated these allostructions in embedding space to some extent. This contrasts with the findings of Veenboer and Bloem (2023) regarding the English dative alternation. Furthermore, the RobBERT-2023, BERTje and EuroBERT models are more capable of separating the allostructions than the mBERT model. We do see that the differences between the models are a lot smaller when considering the 100 closest sentences instead of the ten closest sentences.

It is notable that EuroBERT, a model where Dutch data is a minority of the training data, patterns similarly here to two monolingual Dutch models. An equivalent to the 'om'-optionality does not exist in English, and these models have seen more English training data than Dutch training data, yet EuroBERT still separates these Dutch allostructions. This suggests that we might be able to find other constructional representations in multilingual models where the target language is not the dominant one. This provides further evidence for a central role of constructional information in language model representations, aligning with the BabyLM model findings.

Overall, the monolingual RobBERT-2023 model performed best in separating the clusters based on linguistic constructions.

6 Conclusion

With this study, we show that BERT-based encoder language models better separate constructions that differ in form and meaning than allostructions, in construction comparisons where just one word differs. This confirms that the models are sensitive to localized meaning differences also when using sentence embeddings. We furthermore see that these models distinguish allostructions in their embedding spaces to some extent, though more so in natural occurrences of the allostructions than in synthetic data. Allostructions are constructions with the same meaning (though not necessarily the same pragmatic and functional properties).

We also find that monolingual BERT models differentiate between these linguistic constructions

more easily than their multilingual counterparts. RobBERT-2023 performed best overall, though this could also be due to model size. BERTje vs mBERT is the cleanest comparison between a monolingual and a multilingual model and BERTje showed better cluster separation between the two.

Future work could include extending this intrinsic line of work to decoder models and correlating the results to behavioural generalization experiments with the constructions.

Lastly, further extensions of this line of work beyond English would be very welcome. While our work suggests that constructional representations can also exist in a non-dominant language of a language model, even when the construction does not exist in the dominant language, it remains to be seen whether this is also the case for lesser-resourced languages or for constructions of higher or lower schematicity than ours. Other studies suggest that constructions of higher schematicity are more difficult for language models to learn (Bonial and Madabushi, 2024), and this difficulty should increase for constructions in under-resourced languages. Studies with constructions in more languages that do not exist in English are needed to validate this.

7 Limitations

For the meaning dimension, we only used construction pairs that are minimally different (allostructions) or maximally different (a negation). Some options in between could yield more detailed insight. However, it is difficult to control for degrees of meaning difference with natural language data, at least when keeping similarity in form.

The dataset contained automatically annotated sentences. Any form of automatic annotation is going to contain more mistakes than manual annotation, which could effect the results. Mistakes can especially occur for naturally occurring sentences, as found in the used dataset, because they can contain a lot of spelling and grammatical mistakes. An advantage of using automatically annotated data is that there is a lot more of it available (Bloem et al., 2014).

We rely mainly on naturally occurring sentences in this study. An advantage is that this aligns with the real-world distribution of the optionality as observed in the corpus, but it makes direct comparison more difficult. Our experiment with synthetic data showed that this choice affects the magnitude of the result.

Our study is limited to encoder models. While most intrinsic CxG explainability studies use encoder models, it is also possible to extract embedding representations from decoder models and examine their separation in embedding space. As discussed previously, several prompting studies have shown that decoder models struggle with generalizing constructions, perhaps pointing to limitations in constructional representation. However, we are not aware of any large decoder language models primarily trained on Dutch data without being adapted from English, although recently a dataset was released for this purpose in the context of the GPT-NL project (Oort et al., 2026). For now, we are not able to investigate a monolingual condition with decoder models and leave this for future work.

The RobBERT-2023 model that was used constructs word embeddings with 1024 dimensions, while the other three models construct word embeddings with 768 dimensions. It might have been fairer to compare them directly if they all had the same dimensionality. While we chose models of somewhat similar parameter sizes, this factor was also not fully controlled, with RobBERT-2023 being the largest model. The equivalently sized multilingual counterpart would be the XLM-RoBERTa-Large model for a more direct comparison, though it uses a different tokenizer. BERTje and mBERT are a very comparable pair of a monolingual and multilingual model in our study in terms of dimensionality, but it may still be the case that multilingual models inherently show less separation due to having to separate a large variety of languages in an embedding space of the same dimensionality, so it is difficult to have a fully controlled comparison between multilingual and monolingual models.

References

Marco Baroni. 2022. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language*, pages 1–16. CRC Press.

Jelke Bloem, Arjen Versloot, and Fred Weerman. 2014. Applying automatically parsed corpora to the study of language variation. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1974–1984.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte Miguel Alves, Andre Martins, Ayoub Hammal, Caio Corro, Celine Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan,

Gabriel Hautreux, João Alves, Kevin El Haddad, Manuel Faysse, Maxime Peyrard, Nuno M Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. EuroBERT: Scaling multilingual encoders for European languages. In *Second Conference on Language Modeling*.

Gemma Boleda. 2025. LLMs as a synthesis between symbolic and distributed approaches to language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9365–9379.

Claire Bonial and Harish Tayyar Madabushi. 2024. Constructing understanding: On the constructional information encoded in large language models. *Language Resources and Evaluation*, pages 1–40.

Gosse Bouma. 2017. Om-omission. *Mining for Parsing Failures*, pages 65–74.

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.

Lizzy Brans and Jelke Bloem. 2026. Multi-SimLex for Dutch: Benchmarking embedding- and prompt-based model performance on semantic similarity. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4846–4860, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? In *The SIGNLL Conference on Computational Natural Language Learning*, pages 169–186.

Bert Cappelle. 2006. Particle placement and the case for 'allostructions'. *Constructions*.

Javier Conde, Miguel González Saiz, María Grandury, Pedro Reviriego, Gonzalo Martínez, and Marc Brysbaert. 2025. Psycholinguistic word features: A new approach for the evaluation of LLMs alignment with humans. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 8–17, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2023. DUMB: A benchmark for smart evaluation of Dutch models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7221–7241, Singapore. Association for Computational Linguistics.

Pieter Delobelle and François Remy. 2023. RobBERT-2023: Keeping Dutch language models up-to-date at a lower cost thanks to model conversion.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Jonathan Dunn and Mai Mohamed Eida. 2025. LLMs learn constructions that humans do not know. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 13–23.

Dutch Language Institute. 2015. SoNaR-corpus (version 1.2.1) [data set]. Available at the Dutch Language Institute.

Dutch Language Institute. 2023. Lassy Groot-corpus (version 7.0) [data set]. Available at the Dutch Language Institute.

Adele Goldberg and Laura Suttle. 2010. Construction Grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4):468–477.

Adele E. Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyao Huang, Yue Pan, Stefan Hartmann, and Yang Yanning. 2025. Assessing minimal pairs of Chinese verb-resultative complement constructions: Insights from language models. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 144–150.

Aniek IJbema. 2002. *Grammaticalization and infinitival complements in Dutch*. Netherlands Graduate School of Linguistics.

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.

Shihui Li, Xiaojuan Tan, and Jelke Bloem. 2026. Examining large language models' form-meaning mappings of information structure constructions in Mandarin Chinese. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pages 613–625, San Diego, California, USA. Association for Computational Linguistics.

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

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jan Odijk, Martijn van der Klis, and Sheean Spoel. 2017. Extensions to the GrETEL treebank query application. In *Proceedings of the 16th international workshop on treebanks and linguistic theories*, pages 46–55.

Jesse J. Van Oort, Frank Brinkkemper, Erik de Graaf, Bram Vanroy, and Saskia Lensink. 2026. GPT-NL public corpus: A permissively licensed, Dutch-first dataset for LLM pre-training. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6893–6903, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Pegah Ramezani, Achim Schilling, and Patrick Krauss. 2025. Analysis of argument structure constructions in the large language model BERT. *Frontiers in Artificial Intelligence*, 8:1477246.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

François Remy, Pieter Delobelle, Bettina Berendt, Kris Demuynck, and Thomas Demeester. 2023. Tik-to-Tok: Translating language models one token at a time: An embedding initialization strategy for efficient language adaptation. *Preprint*, arXiv:2310.03477.

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.

Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025. Constructions are revealed in word distributions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2138, Suzhou, China. Association for Computational Linguistics.

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025a. Unpacking *let alone*: Human-scale models generalize to a rare construction in form but not meaning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27503–27514, Suzhou, China. Association for Computational Linguistics.

Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using BERT: A case study of NPN. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376, Vienna, Austria. Association for Computational Linguistics.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025b. Assessing language comprehension in large language models using construction grammar. *arXiv preprint arXiv:2501.04661*.

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*, pages 12937–12951.

Daniel Vlantis and Jelke Bloem. 2025. Intrinsic evaluation of mono- and multilingual Dutch language models. *Computational Linguistics in the Netherlands Journal*, 14:525–553.

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*, pages 10859–10882.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2025. Hybrid human-LLM corpus construction and LLM evaluation for the caused-motion construction. *Northern European Journal of Language Technology*, 11(1):27–57.

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)*, pages 3804–3811.

Wei Zhou and Jelke Bloem. 2021. Comparing contextual and static word embeddings with small data. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 253–259.

A XPath Queries

The following XPath queries are used to extract the sentences from the corpus.

1. For the sentences containing the verbal 'te'-infinitival complement clause with the word 'om':
   '//node[@cat="oti"]'
2. For the sentences containing the verbal 'te'-infinitival complement clause without the word 'om':
   '//node[@cat="ti" and parent::node[@cat="ssub" or @cat="smain" or @cat="sv1"]]'
3. For the sentences containing the word 'niet' followed by the verbal 'te'-infinitival complement clause without the word 'om':
   '//node[node[@word="niet"]
   and node[@cat="ti" and parent::node[@cat="ssub" or @cat="smain" or @cat="sv1"]]]'

B Verbs that take optional 'om'

1. aanbevelen
2. aanbieden
3. aandringen
4. aandurven
5. aankondigen
6. aankunnen
7. aanmanen
8. aanmoedigen
9. aanraden
10. aansporen
11. aanzetten
12. aarzelen
13. adviseren
14. afraden
15. ambiëren
16. beijveren
17. bekendmaken
18. belemmeren
19. beletten
20. beloven
21. beogen
22. bepleiten
23. beslissen
24. besluiten
25. bevelen
26. bezielen
27. bezweren
28. bidden
29. dienen
30. dwingen
31. eisen
32. engageren
33. gebieden
34. gelieven
35. haasten
36. helpen
37. hopen
38. inslagen
39. inspannen
40. inspireren
41. instrueren
42. interesseren
43. kiezen
44. kosten
45. leren
46. lonen
47. lukken
48. machtigen
49. manen

50. meehelpen
51. meevallen
52. motiveren
53. nalaten
54. noodzaken
55. nopen
56. opdragen
57. opkomen
58. opleggen
59. opmaken
60. opperen
61. oproepen
62. overeenkomen
63. overhalen
64. overtuigen
65. overwegen
66. permitteren
67. plachten
68. pleiten
69. prefereren
70. presteren
71. prikkelen
72. proberen
73. resten
74. riskeren
75. schamen
76. schromen
77. schuwen
78. smeken
79. sommeren
80. spijten
81. stimuleren
82. streven
83. toelaten
84. toestaan
85. toezeggen
86. uitdagen
87. uitkijken
88. uitkomen
89. uitnodigen
90. uitsluiten
91. verbieden
92. verdienen
93. verdommen
94. vergeten
95. verheugen
96. verhinderen
97. verkiezen
98. verlangen
99. verleiden
100. vermijden
101. veronderstellen
102. verplichten
103. vertikken
104. verwachten
105. verzoeken
106. verzuimen
107. volhouden
108. volstaan
109. voornemen
110. voorstellen
111. vragen
112. vrezen
113. vrijstaan

114. wagen

115. weerhouden

116. weigeren

117. wensen

C Cluster Metric Results Sentence Length

This appendix contains cluster metric results split for shorter and longer sentences.

	Form short	Meaning short	Difference	Form long	Meaning long	Difference
RobBERT-2023	0.010	0.029	-0.020	0.018	0.047	-0.028
BERTje	0.011	0.031	-0.020	0.014	0.041	-0.026
EuroBERT	0.009	0.022	-0.013	0.015	0.044	-0.029
mBERT	0.008	0.025	-0.017	0.014	0.043	-0.029

Table 6: The Silhouette Score results separated by sentence length (cutoff = 20 words), showing difference in form (with/without 'om') and difference in meaning (with/without 'niet'). Values closer to 1 mean the cluster coherence is better.

	Form short	Meaning short	Difference	Form long	Meaning long	Difference
RobBERT-2023	8.010	5.914	2.095	7.285	4.355	2.930
BERTje	8.076	5.651	2.425	8.199	4.711	3.487
EuroBERT	9.172	6.797	2.375	7.541	4.444	3.097
mBERT	9.872	6.263	3.609	7.781	4.603	3.178

Table 7: The Davies-Bouldin Index results separated by sentence length (cutoff = 20 words), showing difference in form (with/without 'om') and difference in meaning (with/without 'niet'). A lower value indicates better cluster coherence.