Form and Meaning in BERT Embeddings

Analyzing the verbal 'te'-infinitival complement clause with optional complementizer 'om' in Dutch and multilingual BERT models

submitted in partial fulfillment for the degree of master of science

Marije Kouyzer
15858103

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 27.06.2025

	UvA Supervisor
Title, Name	Jelke Bloem
Affiliation	UvA Supervisor
Email	j.bloem@uva.nl

ABSTRACT

The encoding of different linguistic constructions by Dutch BERT models (RobBERT-2023 and BERTje) and multilingual BERT models (mBERT and EuroBERT) are examined. We compare embeddings of sentences with the verbal 'te'-infinitival complement clause with and without the optional complementizer 'om' and the verbal 'te'-infinitival complement clause without 'om' preceded by the word 'niet', which means 'not'. We divide this into two groups, one group in which the constructions differ only in form and one group in which the constructions differ in meaning as well as form. Cluster coherence is measured by the Silhouette Score and the Davies-Boudin Index, and the sentences closest to the average embedding are examined. A lot of overlap was found between the different clusters, suggesting it is hard for these models to differentiate between these constructions. However, better results were found in the group that differed in meaning as well as form than in the group that differed only in form. Better cluster coherence was also found for longer sentences than shorter sentences. The monolingual models performed better than the multilingual models. RobBERT-2023 performed best, suggesting a potential advantage for models based on the RoBERTa architecture when it comes to differentiating between linguistic constructions.

KEYWORDS

Large Language Models, BERT, explainability, Sentence Embeddings, Cluster Analysis, Construction Grammar, 'te'-Infinitival Complement Clauses, BERTje, RobBERT, EuroBERT, mBERT, Silhouette Score, Davies-Bouldin Index, sentence length

GITHUB REPOSITORY

https://github.com/Marije-Kouyzer/Thesis-Data-Science

1 INTRODUCTION

With the rising popularity of the use of Large Language Models (LLMs) in many different contexts, it is more important than ever that we have a better understanding of how they work. These LLMs are generally 'black box' models. This means that researchers, while they understand the inputs and outputs of these models, ultimately do not know how they work internally. They do not know why the models make certain decisions and not others. The growing field that is attempting to learn more about how these models work is called Explainable Artificial Intelligence (XAI). This paper specifically focuses on post-modeling explainability [20], which is the approach where we try to learn more about how existing black box models work internally, of Dutch and multilingual BERT models.

One of the aspects in explainability of LLMs we are interested in is how well linguistic concepts found in human language are encoded in LLMs, to see how similar it is to the way humans use language. In this paper we will research how well linguistic constructions that differ in meaning and form and linguistic constructions that differ only in form, but have the same meaning, are encoded by Dutch and multilingual BERT models.

The theory of Construction Grammar defines linguistic constructions as learned pairs of forms and meaning [10]. These can vary from the meaning of individual words to those of complex syntactical patterns. In this paper we focus on Dutch sentences with a verbal 'te'-infinitival complement clause with an optional complementizer 'om' [3] [13]. 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.

To contrast the use of these constructions that only differ in form, another construction will be included as well. This construction includes the word 'niet' (meaning 'not') before the verbal 'te'-infinitival complement clause without the complementizer 'om'. Like the constructions described above, this only adds one word to the sentence, but in this case the meaning of the sentence changes as well, the complement clause preceded by the word 'niet' changes to the opposite meaning. By including this construction we can compare the encoding of form of linguistic constructions to the encoding of meaning.

In this paper, we will take these three different constructions as pre-existing clusters and use internal cluster analysis methods to analyze the encodings of sentences containing these linguistic constructions in two of the most used Dutch BERT models BERTje [5] and RobBERT [8]. These are both Dutch versions of the well-known pre-trained Bidirectional Encoder Representations from Transformers (BERT) model [9]. We also analyze the encoding of the sentences containing these linguistic constructions by the multilingual models EuroBERT [2] and mBERT [9].

This leads us to the following research questions:

- How does the encoding of linguistic constructions in Dutch and multilingual BERT models compare when they only differ in form and when they differ in form and meaning in terms of cluster coherence?
- How well do sentence embeddings from the Dutch Large Language Models BERTje and RobBERT and the multilingual models EuroBERT and mBERT encode the verbal te-infinitival complement clause with and without the complementizer 'om'?
- What is the effect of sentence length on the encoding of these linguistic constructions by BERT models?
- How do the RobBERT, BERTje, EuroBERT and mBERT models differ from each other in encoding these different linguistic constructions?

2 RELATED WORK

The following sections will describe previous research that has been done about linguistic constructions in Large Language Model outputs, word and sentence embeddings, and high-dimensional cluster analysis methods.

2.1 Linguistic Constructions and Large Language Models

Construction Grammar [10], 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 the CxG theory, the entire grammar of a language consists of combinations of different constructions.

The encoding of linguistic constructions by Large Language Models has not been widely studied yet, especially for Dutch. Previous work in analyzing linguistic constructions in the output of Large Language Models includes the work by Veenboer and Bloem [26]. They used collostructional analysis. This technique uses correlations between linguistic constructions and individual words in these constructions, called collexemes, to find how strongly they are semantically associated with each other, measuring the collexeme strength. This research looked at the X waiting-to-happen construction (e.g. *an accident waiting to happen*) and the ditransitive construction (e.g. *I gave her the envelope*) in English. The X-waiting to happen construction has one open slot for collexemes, while the ditranstive construction has four. All the different words that these slots are filled with are extracted from a corpus, these are then used to calculate the collexeme strengths between the words and the constructions. This study examined if these different collexeme strengths were found in the output of BERT models, using Masked Language Modeling and Sentence Transformers. The ranked lists of collexemes, based on their respective collexeme strength, were compared using Ranked Biased Overlap (RBO). RBO uses the intersections between two lists. The overlap at each intersection is calculated, and the average overlap of these is calculated. Using this method they were able to see if the MLM task and the collostructional analysis led to similar results. This study showed that it is possible to find information on specific linguistic constructions in English in the output from BERT. This makes it likely that it is possible in other languages, such as Dutch, as well.

Weissweiler et al. [28] used Construction Grammar 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. To test the semantic knowledge of the models about these constructions, an example sentence with the comparative correlative was given to the model, followed by a sentence with a masked word to see if the model predicted the correct word for the mask. They found that while the models were able to distinguish the construction syntactically, they did not show an semantic understanding of the construction.

Li et al. [17] investigate linguistic constructions in several BERT models. They use a sentence sorting task and compare embeddings of sentences with the same constructions with embeddigns 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 same verbs.

Madabushi et al. [25] uses sentences from featured articles from Wikipedia and automatically labels 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 in the theory of Construction Grammar. This study shows that BERT models have access to constructional information and that BERT is capable of distinguishing between different linguistic constructions.

2.2 Word and Sentence embeddings

2.2.1 BERT Models and Contextualized Word Embeddings. BERT [9] and models based on it, such as RoBERTa [18], BERTje [5], mBERT [9], EuroBERT [2] and RobBERT [8] output contextualized word embeddings. Contextualized word embeddings are representations of words based on words that are found in the context of this word. Each word that is input into a BERT-based model gets a word embedding in the form of a vector as output.

These word embeddings can be used for many downstream Natural Language Processing tasks, such as text classification, sentiment analysis, named entity recognition, and many more. They can also be used to get sentence embeddings.

2.2.2 Sentence Embeddings. There are several ways to get sentence embeddings from word embeddings. Sentence embeddings combine the contextualized word embeddings gotten from the language model for each word in a sentence to represent the entire sentence. Even though word embeddings are contextualized, they only represent the individual word, whereas a sentence embedding is a representation of all the words in the sentence together. Three of the methods to create sentence embeddings will be discussed here: the CLS-token, mean pooling which is also known as average pooling and using a pre-trained model such as Sentence-BERT.

The CLS-token is used within BERT models explicitly to represent the entire sentence. The token is added at the beginning of each sentence in the input for BERT models. CLS stands for classification. This token is used in the training of BERT models for sentence classification and next sentence prediction.

Mean pooling, or average pooling, takes the average of all the word embeddings to create a sentence embedding. This gives each word the same influence over the resulting sentence embedding. Huang et al. [11] find that the method of mean pooling the word embeddings to create sentence embeddings performs better than using the CLS-token when it comes to sentence semantics.

Sentence-BERT [22] is a model fine-tuned on BERT models. Instead of outputting a vector for each individual word in the input it outputs one vector for the entire input sentence. It does this by first combining the word embeddings, and is trained to give semantically relevant output, meaning that more semantically similar sentences will be closer to each other in the vector space. To combine the word embeddings, Sentence-BERT combines three strategies.

The first strategy is to use the output of the aforementioned CLS-token. The output belonging to this token is representative of the entire sentence. The second strategy is to use mean pooling, as described before. The final strategy Sentence-BERT uses is to take the maximum of each output vector, also known as max pooling. The resulting sentence embeddings can be used similarly to word embeddings, but represent the entire sentence instead of individual words. Sentence-BERT was originally trained on the BERT and RoBERTa models, but now Sentence-BERT models pretrained on many different BERT models are available.

For this study mean pooling is chosen to create sentence embeddings from the word embeddings, because it tends to perform better than the CLS-token. It was decided not to use Sentence-BERT because a pretrained model was not available for all the BERT models used in this study, but more importantly because it is specifically trained to give semantically relevant outputs and the goal of this study is to find how and if the linguistic constructions are encoded directly into the word embeddings output by the models. By using mean pooling all word embeddings are represented equally in the sentence embeddings.

2.3 Internal Cluster Analysis

This research takes existing groups as predefined clusters and then applies internal cluster analysis methods to find out how coherent and well separated these clusters are.

Two of these internal cluster analysis methods are the Silhouette Score [24] and the Davies-Bouldin Index [4]. Both of these methods have been used extensively. Usually they are used to measure the effectiveness of clustering methods, as opposed to the pre-existing clusters we use here.

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. Layton et al. [16] uses the Silhouette Coefficient to evaluate the effect of the choice of authorship distance method on their clustering algorithm to label documents by authorship. Lovmar et al. [19] uses the Silhouette Score to evaluate how well their single nucleotide polymorphism genotyping clustering method works.

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. Idrus et al. [12] use the average Davies-Bouldin Index for different measure types to find the optimal number of clusters using clustering methods. Ashari et al. [1] also used the Davies-Bouldin Index to find the optimal number of clusters in their study on the most popular films using data from IMDb.

We use both of these internal cluster analysis methods, 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.

3 METHODOLOGY

In this section first the used dataset as well as the data extraction and preprocessing process will be described. The BERT models used to get the word embeddings and the process of getting the sentence embeddings is described next. Finally, the different cluster analysis methods are explained.

3.1 Data

The data used for this research has been extracted from the Lassy Groot corpus [15]. This is a corpus of Dutch sentences with automatically generated syntactic annotations. It is comprised of about 700 million words. The annotations of the sentences contain the syntactical relations between words in each sentence and the part-of-speech (PoS) tags of individual words as well as other linguistic features of the words such as the lemma.

Because of computational limitations, only part of the Lassy Groot corpus is included. This was comprised of the folders WR-P-E-A up to and including WR-P-E-L, and the folders WR-P-P-B up to and including WR-P-P-G, which contain texts from online discussions and magazines, newsletters, press releases, subtitles, teletext pages, websites including Wikipedia, blogs, tweets, books, brochures, manuals, legal texts and newspapers. All of these originated from the SoNaR corpus [14]. The SoNaR includes Dutch sentences from a diverse range of texts from both Dutch and Flemish authors.

Similarly for computational limitations, files that exceeded 1 MB were not included in the data extraction process.

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 [21], a search engine for syntactically annotated corpora, is used.

From the corpus 803,119 sentences were found that contained the verbal 'te'-infinitival complement clause with the complementizer 'om'. 722,619 sentences were found containing the verbal 'te'-infinitival complement clause without the complementizer 'om'. And 14,588 sentences were found that contained the word 'niet' in combination with the verbal 'te'-infinitival complement clause without the complementizer 'om'.

3.1.2 Data Preprocessing. To ensure that the sentences with and without 'om' are actually sentences that can optionally take the complementizer 'om', a list of verbs was created that optionally take the word 'om' as a complementizer. From the sentences containing the verbal 'te'-infinitival complement clause with and without 'om', the verbs found right before this clause were counted. Verbs that were used in at least ten sentences in both categories were put in a list, similar to how it was done by Bouma [3]. A few sentences of each of the verbs for both categories were manually checked to see if they indeed contained an optional 'om' complementizer. Some sentences contained verbs that only take an optional 'om' in specific constructions. For example, 'hebben' ('to have') takes an optional 'om' in the construction 'het plan hebben om' ('to have the plan to'), but not in most other constructions. 33 verbs were dismissed from the list for this reason. This included verbs such as 'hebben',

'zijn' ('to be'), 'krijgen' ('to get'), 'maken' ('to make') and 'staan' ('to stand'). 91 other verbs, such as 'lijken' ('to seem'), 'beginnen' ('to begin'), 'komen' ('to come') and 'blijken' ('to turn out') were dismissed as well. They were found in both kinds of constructions, but it was often not possible to add or remove the complementizer 'om' from these sentences, and thus a lot of these sentences did not actually contain the optional 'om' construction. 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. The category containing the word 'niet' and the verbal 'te'-infinitival complement clause, was also filtered by this list of verbs to make the comparisons more directly comparable.

This means that all the 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

After filtering the sentences to only contain verbs on the list, the remaining sentences in each group were checked for overlap, to see if the same sentences might be part of the multiple groups. 5,668 sentences were found to be part of both the verbal 'te'-infinitival complementizer clause without 'om' group and the verbal 'te'-infinitival complementizer clause with 'om' group. These sentences contained both constructions in different parts of the sentences and were removed from both groups.

3,015 sentences in the group containing the word 'niet' in combination with the verbal 'te'-infinitival complement clause without the word 'om' were also found in the group containing verbal 'te'-infinitival complement clause without the word 'om'. These were removed from the latter group, because they did contain the word 'niet', and thus were more appropriately placed in the former group.

55 sentences were found in both the group containing sentences with the verbal 'te'-infinitival complement clause with the complementizer 'om' and the group containing the word 'niet' in combination with the 'te'-infinitival complement clause without the complementizer 'om'. These sentences also contained both kinds of constructions and were removed from both groups.

Also, any sentences that were found multiple times within a group were removed.

The length of each sentence, defined as the number of words in the sentence, was calculated. The interquartile range was used to detect and remove outliers. Any sentences shorter than the first quartile minus 1.5 times the interquartile range and any sentences longer than the third quartile plus 1.5 times the interquartile range were considered outliers and removed from the dataset. In practice, this meant that only sentences containing 49 words or more were dismissed from the dataset.

The median of the sentence lengths was a length of 20 words. This was used to divide the sentences in two groups: short sentences, defined as sentences containing less than 20 words, and long sentences, defined as sentences containing equal to or more than 20 words.

	short sentences	long sentence	total
'om' group	960	1,685	2,645
'te' group	1,287	1,358	2,645
'niet' group	1,413	1,232	2,645
total	3,660	4,275	

Table 1: Number of sentences in each group. The groups containing the different constructions are the same size, while there are slightly more long sentences (with a length equal to or of more than 20 words) than short sentences (with a length of less than 20 words).

Finally, the three groups containing sentences with the different constructions are made the same size. Since the smallest group (that consists of sentences with the word 'niet' in combination with the verbal complement clause without the word 'om') contains 2,645 sentences, the other two groups are reduced to 2,645 sentences each as well. This is done to better be able to compare the cluster coherence between the groups.

3.1.3 Data Description. In total, 7,935 sentence were included in the results. As described in the previous section, each of the three construction groups contains 2,645 sentences. The distribution of sentences over the groups based on sentence length can be found in Table 1.

These sentences are tokenized with the tokenizers of the respective BERT models. The BERT models will then be used to obtain word embeddings, which will then be transformed into sentence embeddings. This will be described in the following sections.

3.2 Dutch and Multilingual BERT models.

After these sentences have been extracted from the corpus, the following BERT-based Large Language Models, RobBERT-2023, BERTje, EuroBERT and mBERT, were used to get word embeddings for each sentence. An overview of the models is found in table 2. After obtaining the word embeddings, they are transformed into sentence embeddings.

3.2.1 RobBERT-2023. RobBERT-2023 [7] is a new version of the RobBERT model [8]. This is a Dutch BERT model, but instead of being directly based on BERT, it is based on the RoBERTa model [18]. RoBERta is an optimized version of BERT, trained with more data and a slightly different process than the original BERT [9]. RobBERT-2023 uses the same way of training as RoBERTa. It has 355 million parameters and uses a Tik-to-Tok tokenizer [23]. This tokenizer uses tokens from a language with more resources and maps these to semantically similar tokens in the language with less resources, in this case Dutch. The model was trained on a Dutch corpus of 6.6 billion words, only half of which was used for training this large variant of the model we use here [7]. The word embeddings produced by this model have 1024 dimensions.

The specific version that is used is called *robbert-2023-dutch-large* on Hugging Face.

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

Table 2: An overview of the differences in the architecture of the used RobBERT-2023, BERTje, EuroBERT and mBERT models. Inlcuded are the tokenizer that is used with the model, the number of parameters within the model, how much training data the model was trained on, if the model is monolingual or multilingual, what architecture the model is based on and how many dimensions the word embeddings resulting from the model have.

3.2.2 BERTje. BERTje [5] is also a Dutch BERT model. It uses the same architecture as the original BERT. The original BERT model has 110 million parameters [9], as does this model. BERTje is trained on a diverse dataset of Dutch texts, which includes in total 2.4 billion words. The dataset contained, among others, news sources, Wikipedia pages and books. The BERTje model is used with a WordPiece tokenizer. The word embeddings from this model have 768 dimensions.

The specific version that is used is called *bert-base-dutch-cased* on Hugging Face.

3.2.3 EuroBERT. EuroBERT [2] is a newly developed multilingual BERT model. It can be used for 15 different languages, which include both European languages as well as languages that are widely spoken globally. One of the included languages is Dutch. The EuroBERT model is trained on a multilingual dataset with 5 trillion words, covering texts in these 15 languages. The model is used with a LLaMa 3 tokenizer. The word embeddings from this model have 768 dimensions.

The specific version that is used is called *EuroBERT-210m* on Hugging Face. This version has 210 million parameters.

3.2.4 mBERT. Multilingual BERT [9], or mBERT, is a multilingual BERT model. It uses the same architecture as the original BERT model and is available for over 100 languages. mBERT has 179 million parameters and is used with a WordPiece tokenizer, like the original BERT model. Multilingual BERT is trained on a dataset consisting of the Wikipedia texts of the 104 biggest languages. The word embeddings from this model have 768 dimensions.

The specific version that is used is called *bert-base-multilingual-cased* on Hugging Face.

3.2.5 Sentence Embeddings. The initial output from all the these BERT models are contextualized word embeddings. To be able to compare the different sentences with each other, we need to transform these into sentence embeddings. There are different ways to do this. For example, all the word embedding vectors from a sentence could be summed to get the sentence embedding. Other ways are max pooling, for which the the maximum of is taken for each vector position, min pooling, for which the minimum is taken, and mean pooling, which calculates the average over the word embeddings in the sentence. BERT also provides a [CLS] token at the start of each sentence to represent the entire sentence. This can also be used as a sentence embedding. Another option is the Sentence-BERT model [22]. The Sentence-BERT model is created specifically to fine-tune existing BERT models to output sentence embeddings instead of word embeddings. There are many pretrained versions of Sentence-BERT available, trained on different BERT models. Unfortunately, there is no pretrained sentence-BERT model available for some of the models used here.

Because the goal is to evaluate the output of the BERT models as directly as possible, a simpler method is preferred. For this reason, mean pooling is chosen as the method to get sentence embeddings. Mean pooling tends to perform better than the CLS-token for semantic tasks and gives each word embedding equal influence over the final sentence embedding. Using mean pooling gives a more direct insight into how linguistic constructions are encoded by the different BERT models compared to using a model such as Sentence-BERT which has been pre-trained explicitly to represent sentences semantically.

So mean pooling was used to get sentence embeddings for each sentence from each of the different BERT models. The resulting sentence embeddings have the same dimensions as the word embeddings from each BERT model. So the sentence embeddings from the RobBERT-2023 model are 1024-dimensional, while the sentence embeddings from the BERTje, EuroBERT and mBERT models are 768-dimensional.

3.3 Cluster Analysis

The high-dimensional sentence embedding vectors are assigned to their respective linguistic construction cluster: the verbal 'te'-infinitival complement clause with complementizer 'om', the verbal 'te'-infinitival complement clause without complementizer 'om', and the word 'niet' followed by the verbal 'te'-infinitival complement clause without 'om'.

To be able to compare the influence of meaning and form of the linguistic constructions, the data is divided into two groups of two constructions that each only differ one word. The first group consists of the verbal 'te'-infinitival complement clause with the complementizer 'om' and the verbal 'te'-infinitival complement clause without the complementizer 'om'. These constructions have the same meaning and their only difference is the inclusion of the word 'om'. Because the difference between these two groups is only in form and not in meaning, this group is used to measure how well the clusters separate based on form.

The second group consists of the verbal 'te'-infinitival complement clause without complementizer 'om', like the first group does, and this construction preceded by the word 'niet'. The verbal 'te'-infinitival complement clause with the complementizer 'om' is not included in this group. The difference in form between the constructions in this group is only the negation word 'niet'. Since the

addition of the word 'niet' changes the meaning of the sentences, this group will be used to measure how well the clusters separate based on meaning as well as form.

These different groups are used to compare the cluster coherence of these pre-existing clusters on the basis of form and meaning. The aforementioned short and long sentences are used to investigate the effect of sentence length on the encoding of linguistic structures. The different BERT models are included to examine how the encoding of these constructions differs per model, as well as differences between monolingual and multilingual models, and differences between RoBERTa-based models and BERT-based models.

The cluster coherence of these groups are evaluated by the cluster evaluation methods, the Silhouette Coefficient and the Davies-Bouldin Index. The following sections will describe these cluster analysis methods.

3.3.1 Silhouette Score. The second evaluation method is the Silhouette Score. The Silhouette Score calculates the average Silhouette Coefficient [24] for all sentences.

The Silhouette Coefficient is calculated with the following formula:

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

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.

The Silhouette Coefficient is calculated per sentence. For each sentence embedding vector, the Silhouette Coefficient is calculated, comparing the distance between this vector and the members within its cluster and the distance between this vector and members of other clusters. The value of the Silhouette Score can vary between -1 and 1. If the score is closer to 1, the clusters are better separated. A score near 0 means that the clusters are overlapping, while negative values can point to sentences that might be closer to sentences of the other cluster than the sentences within its cluster. The silhouette score function from the scikit-learn library is used.

3.3.2 Davies-Bouldin Index. The next evaluation method is the Davies-Bouldin Index [4]. Like the Silhouette Coefficient, the Davies-Bouldin Index measures how good cluster coherence is. To calculate this, it divides the sum of the average distance within each cluster by the distance between the centroids of the clusters. A lower value for the Davies-Bouldin Index indicates that the clusters are better separated.

The following formula is used to calculate the Davies-Bouldin Index:

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

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. The davies bouldin score function from the scikit-learn library is used.

	'om te' and 'te'	'te' and 'niet te'	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 3: Silhouette Score Results. For the Silhouette Score, values closer to 1 mean the cluster coherence is better. The values are all close to 0, which means that there is a lot of overlap between the clusters and the clusters are not well separated. The group that differs in meaning as well as form ('te' and 'niet te') is slightly better separated than the group that only differs in form ('om te' and 'te'). The difference column shows the difference between the group that differs only in form and the group that differs in meaning as well as form for each model.

3.3.3 Sentences closest to the average. For this method, for each of the four BERT models the average of all the sentence embeddings of the sentences containing the verbal 'te'-infinitival complement clause with the complementizer 'om' is calculated. Then the Euclidean distance between this average and each of the sentence embeddings containing the verbal 'te'-infinitival complement clause both with and without the complementizer 'om' are calculated. The ten sentences that are closest to this average embedding are defined as the sentences with the smallest Euclidean distance between them and the average. Of these sentences the percentage of which are sentences that come from the group containing the verbal 'te'-infinitival complement clause with the complementizer 'om' is calculated. The rest of the sentences are from the group containing the verbal 'te'-infinitival complement clause without the complementizer 'om'. A higher percentage of sentences from the group with the complementizer 'om' closest to the average of the sentences containing the verbal 'te'-infinitival complement clause with the word 'om' suggests that the groups are better separated then a lower percentage. For comparison, this is calculated with the ten closest sentences, as described, and also with the 100 closest sentences. This way we can see if the trends for the first few sentences that are closest to the average hold up when you look at more of the closest sentences.

4 RESULTS

4.1 Form and Meaning

The results from comparing the two groups, that differ only in form and differ in form and meaning, respectively, can be found in table 3 for the Silhouette Scores and table 4 Davies-Bouldin Index. The Silhouette Scores are all close to 0, meaning that there is a lot of overlap between the clusters and none of the clusters are well separated. You can see that the group of sentences that differs in meaning as well as form, which is the group containing sentences with the verbal 'te'-infinitival complement clauses without the complementizer 'om', preceded by 'niet' and not preceded by 'niet', has slightly higher silhouette scores than the group that differs only in form, which is the group with the verbal 'te'-infinitival complement clauses with and without the complementizer 'om'.

	'om te' and 'te'	'te' and 'niet te'	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 4: Davies-Bouldin Index Results. A lower value of the Davies-Bouldin Index indicates better cluster coherence. For all the models the group that differs in meaning as well as form ('te' and 'niet te') has a better cluster coherence than the group that only differs in form ('om te' and 'te'). The difference column shows the difference between the group that differs only in form and the group that differs in meaning as well as form for each model.

This indicates that the clusters in the group that differs in meaning as well as form are slightly better separated than the group that differs only in form. The Davies-Bouldin Index shows that the group that differs in meaning as well as form has lower scores than the group that differs only in form. A lower Davies-Bouldin Index indicates better cluster coherence. So the clusters of the group that differs in meaning as well as form have better cluster coherence than the clusters of the group that differs only in form.

4.2 Effect of Sentence Length

The results from comparing the different groups for short and long sentences can be found in table 5 for the Silhouette Scores and table 6 for the Davies-Bouldin Index. As was the case with all the sentences together, the Silhouette Scores are again all close to zero, indicating a lot of overlap between the clusters for both group. This means the clusters are not well separated. We do see that the Silhouette Scores for the long sentences are slightly higher than they are for the short sentences, though they are still close to zero. This indicates that for the long sentences, the clusters are slightly better separated than they are for the short sentences. For both short and long sentences, we can see that the clusters in the group that differs in meaning as well as form, which contains sentences with the verbal 'te'-infinitival complement clauses without the complementizer 'om', with and without the preceding word 'niet', is slightly better separated than the group that differs only in form, containing the verbal 'te'-infinitival complement clauses with and without the complementizer 'om'.

The Davies-Bouldin Index shows that the groups with long sentences have better cluster coherence than the groups with short sentences. It also shows consistently that the group that differs in meaning as well as form has better cluster coherence than the group that differs only in form.

4.3 Different BERT models

To compare the different BERT models, we look at table 3 for the Silhouette Scores and table 4 Davies-Bouldin Index.

The Silhouette Scores show, by all being close to zero, that none of the models can separate the clusters very well. 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 Davies-Bouldin Index gives similar results, with the monolingual models performing better than the multilingual models. This difference is bigger in the group that differs in only in form than in the group that differs in form as well as meaning. In the group that differs in meaning as well as form, the Davies-Bouldin Indices are all close together. The best performing model in both groups is RobBERT-2023, followed by BERTje. In the group that differs only in form mBERT performs the worst, while in the group that differs in meaning as well as form, EuroBERT performs the worst.

We look at table 5 and table 6 to see the effects of sentence length on the different models. When the groups are divided based on sentence length (short sentences and long sentences), the Silhouette Scores have similar results, with the monolingual generally performing better than the multilingual results. Interestingly, for the long sentences, in the group that differs in meaning as well as form, BERTje actually performs slightly worse than both the multilingual models, while for the short sentences it performs best of all the models.

The same can be seen in the Davies-Bouldin Indices in the groups divided on sentence length. While the monolingual models still generally perform better than the multilingual models, BERTje again performs worse than both multilingual models on the long sentences. The Davies-Bouldin Index also shows that BERTje performs better than RobBERT-2023 on the short sentences for the group that differs in meaning as well as form, while RobBERT-2023 performs slightly better for the group that differs only in form.

4.4 Sentences closest to the average

The results for the sentences closest to the average of the sentence embeddings of the sentences containing the verbal 'te'-infinitival complement clause with the complementizer 'om' can be found in table 7. The results are shown both for the ten sentences closest to this average and the 100 closest sentences. The percentages in this table represent the amount of sentences of these ten or 100 that come from the group of sentences containing the verbal 'te'-infinitival complement clause with 'om', while the rest of the sentences come from the group containing the verbal 'te'-infinitival complement clause without 'om'. A higher percentage suggests better separation between these two groups of sentence embeddings. From the top ten closest sentences to the average for both the RobBERT-2023 model and the EuroBERT model, nine sentences were from the group containing the complementizer 'om', while for the BERTje model eight sentences were from this group. This suggests that these three models are quite capable of separating the sentence embeddings from the sentences containing these two constructions. The mBERT model does less well at this separation. Only three of the sentences closest to the average were sentences that contained the verbal 'te'-infinitival complement clause with 'om'. This is less than the five sentences you would expect from random chance alone. The results of the different models are closer together when we look at the top 100 closest sentences, with RobBERT-2023 performing best at 72%, or 72 of the 100 closest sentences, and mBERT performing worst at 61%, or 61 of the 100 closest sentences.

	'om te' and 'te' short	'te' and 'niet te' short	difference	'om te' and 'te' long	'te' and 'niet te' 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 5: The Silhouette Score results with only short and long sentences. For the Silhouette Score, values closer to 1 mean the cluster coherence is better. The values are all close to 0, which means that there is a lot of overlap between the clusters and the clusters are not well separated. The long sentences have slightly higher values than their short counterparts. This shows that the long sentences are better separated than the short sentences. For both the short and the long sentences, the sentences that differ in meaning as well as form ('te' and 'niet te') are slightly better separated than the sentences that only differ in form ('om te' and 'te'). The difference columns show the differences between the group that differs only in form and the group that differs in meaning as well as form for both the short and the long sentences.

	'om te' and 'te' short	'te' and 'niet te' short	difference	'om te' and 'te' long	'te' and 'niet te' 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 6: The Davies-Bouldin Index results with only short and long sentences. A lower value of the Davies-Bouldin Index indicates better cluster coherence. For both the short and the long sentences, the group that differs in meaning as well as form ('te' and 'niet te') shows better cluster coherence than the group that differs only in form ('om te' and 'te'). The long sentences show better cluster coherence than the short sentences. The difference columns show the differences between the group that differs only in form and the group that differs in meaning as well as form for both the short and the long sentences.

Here all the models perform better than 50% but the separation is definitely not flawless.

5 DISCUSSION

In general the Silhouette Scores showed that none of the clusters are well separated. This makes sense, because the sentences are about a wide variety of topics and do not have that much in common semantically. 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 is 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 group that differs only in form than in the group that differs in meaning as well as form, where the Davies-Bouldin Indices were pretty 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. Overall, the RobBERT-2023 model performed best. Vlantis and Bloem [27] 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 [6] 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 surprisingly, BERTje actually performed worse than the multilingual models for the long sentences, while performing the best overall for the short sentences. This shows an effect of sentence length for the BERTje model specifically.

The results from calculating the ten sentences closest to the average of the embeddings of the sentences containing the verbal 'te'-infinitival complement clause with the complementizer 'om' show that the RobBERT-2023, BERTje and EuroBERT models are more capable of separating sentences with the verbal 'te'-infinitival complement clause with the complementizer 'om' from those without the complementizer 'om' than the mBERT model. The worse performance of mBERT is in line with the results of the Dutch Model Benchmark [6]. 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. Here, we see that all the models can somewhat separate the two construction, but not perfectly.

Overall, the RobBERT-2023 model performed best in separating the clusters based on linguistic constructions. This is the only included model based on the RoBERTa architecture. So it could be that

	Percentage 'om te' sentences in top 10	Percentage 'om te' sentences in top 100
RobBERT-2023	90%	72%
BERTje	80%	66%
EuroBERT	90%	67%
mBERT	30%	61%

Table 7: The top 10 and top 100 sentences closest to the average sentence embedding of the sentences containing the verbal 'te'-infinitival complement clause with the complementizer 'om' from all the sentences with both the aforementioned construction and the sentences containing the verbal 'te'-infinitival complement clause without the complementizer 'om'. The percentages show how many of these sentences where from the group containing the verbal 'te'-infinitival complement clause with 'om'. These results show that RobBERT-2023, BERTje and EuroBERT are all quite capable of separating the two constructions, though not perfectly. The results also show that the mBERT model is less capable of separating the two constructions.

the RoBERTa architecture is better at distinguishing linguistic constructions. However, RobBERT-2023 also used word embeddings, and consequently sentence embeddings as well, with more dimensions than the other three models, which could also be part of the reason why it performed better. Other possibilities are that this is because RobBERT-2023 includes the most parameters of all four models, or because of the use of the Tik-to-Tok tokenizer.

5.1 Limitations

There are several limitations of the way the research project was designed. The choice of using the verbal 'te'-infinitival complement clause with optional complementizer 'om' as the linguistic construction may not have been the best choice because they are so close together in form, differing only by one word. Other linguistic constructions that have the same meaning but differ more in form, may lead to better or clearer results. An example of such a construction could be the passive and the active voice. Similarly, using a construction with a negation in it was potentially not the best choice, as BERT models are known to struggle with negation sometimes. On the other hand, choosing constructions that differ only one word and include negations is interestingly precisely because this is relatively hard for language models to deal with and this way we can see how these constructions are affecting the encoding of the sentences by the language models.

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.

We use naturally occurring sentences in this study. An advantage is that this is exactly how these sentences occur naturally in the real world. Another option is to artificially add or remove the word 'om' from existing sentences. This could have lead to potentially more data and more direct comparisons between the sentences, but the sentences would not have been naturally occurring.

The manual process of checking which verbs do and don't include an optional complementizer 'om' is subjective. A more objective way of conducting this process might have been better. The choice to not include verbs that are part of more complicated constructions (e.g. 'van plan zijn (om) te') means we might be missing out on the effects of these constructions in the results.

Using mean pooling could lead to a potentially big effect of the length of sentences. But since the long sentences are separated better than the short sentences this does not seem to be the case. It also could have been interesting to see the effects of the different sentence embedding methods by using multiple and comparing them. The use of sentence embeddings made the analysis of the BERT output less direct. BERT models output word embeddings, not sentence embeddings, so it would have been a more direct analysis if we had found a way to directly analyze the word embeddings.

Two of the groups of sentences with linguistic constructions were cut off to be the same size as the smallest group. While having all three groups be the same size is good, the cutting off of the groups could effect the results. Also, the sentence length groups differed slightly in size. This could also affect the results of the internal cluster analysis methods.

The sentences were grouped by length to be able to see the influence of sentence length on the results. Only two sentence length groups were included here. It could also be interesting to see the effects of sentence length better by dividing the sentences in more than two groups. Furthermore, the very long sentences were filtered out. It might also be interesting to see the effect of these very long sentences by including them.

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 dimensions.

6 CONCLUSION

This study shows that sentences containing linguistic constructions that differ in meaning as well as form are separated better by BERT models than sentences containing linguistic constructions that differ only in form. We also find that the sentences containing linguistic constructions are separated more easily in longer sentences compared to shorter sentences and that monolingual BERT models differentiate between these linguistic constructions more easily than their multilingual counterparts. RobBERT-2023 performed best overall, suggesting a potential advantage for language models following the RoBERTa architecture when it comes to separating based on linguistic constructions. A surprising result is that the Dutch model BERTje specifically performs worse than even the multilingual models when presented with long sentences while it performs best overall for the short sentences, showing an effect of sentence length for this model specifically. Calculating

the closest sentence embeddings to the average of the sentences containing a verbal 'te'-infinitival complement clause with the complementizer 'om' showed that, while not perfect, the RobBERT-2023, BERTje and EuroBERT models are better at separating this construction from the verbal 'te'-infinitival complement clause without the complementizer 'om' than the mBERT model.

6.1 Future Work

Future work could include using different BERT models, both monolingual and multilingual. It could also be interesting to investigate the encoding of linguistic constructions in decoder models, as opposed to encoder models used in this study. Generative Artificial Intelligence models like ChatGPT are increasing in popularity every day and it would be good to learn more about how they encode linguistic constructions.

Different ways of constructing sentence embeddings could be interesting in the context of research similar to this one as well. Mean pooling is used as the sentence embedding method here, but other methods such as max pooling, using the [CLS] token and pretrained models such as Sentence-BERT, are also good methods that could be considered and investigated. It could even be potentially interesting to use multiple sentence embedding methods within one study to directly compare the effects of using the different methods.

The effects of sentence length can be more fully investigated by including more fine-grained groups of different sentence lengths. For example, by dividing the sentences into five groups or more.

To more directly compare the two constructions, the verbal 'te'-infinitival complement clause with and without the complementizer 'om', it is also possible to create synthetic data. This can be done by taking existing sentences that include the verbal 'te'-infinitival complement clause with the complementizer 'om' and remove the word 'om' from the sentence to create a sentence with the verbal 'te'-infinitival complement clause without the complementizer 'om'. This creates pairs of sentences that can be compared one on one directly.

In addition to the verbal 'te'-infinitival complement clause with optional complementizer 'om', which differs only one word in form and nothing in meaning, it is also interesting to look at different linguistic constructions that differ more in form. Potential future work could use cluster analysis methods to study, for example, similar constructions in the passive and active voice.

Other linguistic aspects of the constructions within sentences can also be considered, such as if the construction is found within the main clause or a subordinate clause of the sentence, to see the effects of these aspects on the encoding of the constructions by the models.

Of course, there are also many other languages in the world with their own exciting linguistic constructions and it would also be interesting to study the encoding of these constructions in those languages' respective BERT models.

REFERENCES

[1] Ilham Firman Ashari, Romantika Banjarnahor, Dede Rodhatul Farida, Sicilia Putri Aisyah, Anastasia Puteri Dewi, Nuril Humaya, et al. 2022. Application of data mining with the K-means clustering method and Davies Bouldin index for grouping IMDB movies. *Journal of Applied Informatics and Computing* 6, 1 (2022), 07–15.

[2] Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M Alves, André FT Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, et al. 2025. EuroBERT: Scaling Multilingual Encoders for European Languages. *CoRR* (2025).

[3] Gosse Bouma. 2013. Om-omission in Dutch verbal complements. *Manuscript in preparation* (2013).

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

[5] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *ArXiv* (2019), 1912–09582.

[6] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7221–7241. https://aclanthology.org/2023.emnlp-main.447

[7] P Delobelle and F Remy. 2023. RobBERT-2023: Keeping Dutch Language Models Up-To-Date at a Lower Cost Thanks to Model Conversion. https://clin33.uantwerpen.be/abstract/robbert-2023-keeping-dutch-language-models-up-to-date-at-a-lower-cost-thanks-to-model-conversion/

[8] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2021. RobBERT: a Dutch RoBERTa-based Language Model. In *BNAIC/BENELEARN 2021, Date: 2021/11/09-2021/11/12, Location: Esch-sur-Alzette, Luxemburg*. 1–14.

[9] 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)*. 4171–4186.

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

[11] 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. *arXiv e-prints* (2021), arXiv–2104.

[12] Ali Idrus. 2022. Distance analysis measuring for clustering using k-means and davies bouldin index algorithm. *TEM Journal* 11, 4 (2022), 1871–1876.

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

[14] Dutch Language Institute. 2015. SoNaR-corpus (Version 1.2.1) [Data set]. http://hdl.handle.net/10032/tm-a2-h5 Available at the Dutch Language Institute.

[15] Dutch Language Institute. 2023. Lassy Groot-corpus (Version 7.0) [Data set]. https://hdl.handle.net/10032/tm-a2-w8 Available at the Dutch Language Institute.

[16] Robert Layton, Paul Watters, and Richard Dazeley. 2013. Evaluating authorship distance methods using the positive Silhouette coefficient. *Natural Language Engineering* 19, 4 (2013), 517–535.

[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)*. 7410–7423.

[18] 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 e-prints* (2019), arXiv–1907.

[19] Lovisa Lovmar, Annika Ahlford, Mats Jonsson, and Ann-Christine Syvänen. 2005. Silhouette scores for assessment of SNP genotype clusters. *BMC genomics* 6 (2005), 1–6.

[20] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review* (2022), 1–66.

[21] 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*. 46–55.

[22] 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)*. 3982–3992.

[23] 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. arXiv:2310.03477 [cs.CL] https://arxiv.org/abs/2310.03477

[24] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[25] Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. *arXiv e-prints* (2020), arXiv–2011.

[26] 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*. 12937–12951.

[27] Daniel Vlantis and Jelke Bloem. in press. Intrinsic evaluation of Mono- and Multilingual Dutch Language Models. *Computational Linguistics in the Netherlands Journal* 14 (in press).

[28] 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*. 10859–10882.

Appendix 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"]]]'

Appendix 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