

# Explaining Contextualization in Language Models using Visual Analytics

Rita Sevastjanova\* and Aikaterini-Lida Kalouli† and Christin Beck† and  
Hanna Schäfer\* and Mennatallah El-Assady\*

University of Konstanz

firstname.lastname@uni-konstanz.de

## Abstract

Despite the success of contextualized language models on various NLP tasks, it is still unclear what these models really learn. In this paper, we contribute to the current efforts of explaining such models by exploring the continuum between function and content words with respect to contextualization in BERT, based on linguistically-informed insights. In particular, we utilize scoring and visual analytics techniques: we use an existing similarity-based score to measure contextualization and integrate it into a novel visual analytics technique, presenting the model’s layers simultaneously and highlighting intra-layer properties and inter-layer differences. We show that contextualization is neither driven by polysemy nor by pure context variation. We also provide insights on why BERT fails to model words in the middle of the functionality continuum.

## 1 Introduction

The rise of contextualized language models (LM), i.e., contextualized word and sentence representations, such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019), has brought many well-known NLP tasks to a tremendous breakthrough. Contextualized embeddings have replaced earlier static embeddings (Mikolov et al., 2013; Pennington et al., 2014; Conneau et al., 2017), creating new standards for the state-of-the-art. LMs have learned highly transferable and task-agnostic properties of language (e.g., Belinkov, 2018; Conneau et al., 2018; Peters et al., 2018), even to a degree of imitating the classical NLP pipeline (Tenney et al., 2019a). Despite these research efforts, it remains yet unclear as to what extent LMs like BERT capture complex linguistic phenomena and whether different linguistic properties are learned

across the different layers of the model’s architecture: the existing evidence is conflicting and in some cases even contradictory (Rogers et al., 2020). One recent line of work (Ethayarajh, 2019) explores the actual contextualization captured in these models, i.e., the degree to which a word is modeled as context-specific. This sheds light on the context-specificity of individual words and the degree of contextualization of different word groups.

This paper contributes to this line of work by examining the degree of contextualization of function vs. content words. We treat functionality as a continuum, comparing and contrasting BERT’s (Devlin et al., 2019) modeling of categories of words within this continuum with the expected modeling according to the theoretical linguistic literature. It has been repeatedly shown that LMs fail to generalize and capture the compositionality of language because they struggle with words of high functionality, e.g., quantifiers, prepositions, modals, conjunctions (Dasgupta et al., 2018; Naik et al., 2018; McCoy et al., 2019, to name only a few). Thus, our linguistically-informed analysis sheds light on the peculiarities of these phenomena and contributes to our better understanding of BERT.

This paper utilizes the *self-similarity* contextualization score of Ethayarajh (2019) for better comparability. The exploration of the scores and phenomena is enabled by *LMExplorer*, a visual analytics (VA) technique for the layer-wise explanation of contextualized word embeddings. *LMExplorer* contributes a new perspective on the learned patterns of the model, and shows clusters and score developments in the model’s layers simultaneously.

Overall, the contribution of this paper is two-fold: (1) we generate insights as to how BERT captures function vs. content words (Sections 4 and 5), and (2) present a novel visual analytics technique that facilitates such insights by explaining LMs through contextualization scoring (Section 3).

\* Contribution to the visualization part.

† Equal contribution to the computational linguistics part.

## 2 Interpretability of Language Models

Research on the interpretability of LMs has been pursued in two main directions, mainly focusing on BERT. For one, probing tasks are used to investigate the linguistic properties learned by the LM by training a linear model on the basis of the corresponding contextualized embeddings for the prediction of specific linguistic properties. For another, the interpretability of LMs has been explored via adversarial datasets to assess the performance of an LM with respect to challenging linguistic phenomena. To further explore the interpretability of LMs, we see work coming from the field of VA as promising. VA techniques have been used extensively for exploring and interpreting different deep learning models (Hohman et al., 2019), incl. LMs.

**Probing** – Probing experiments have shown that BERT’s transformer architecture encodes semantic information such as word senses and semantic roles (Reif et al., 2019; Tenney et al., 2019b; Ettinger, 2020; Zhao et al., 2020), syntactic information in the form of constituents and hierarchical structure (Goldberg, 2019; Hewitt and Manning, 2019; Warstadt and Bowman, 2020; Chi et al., 2020), morphosyntactic and morphological features (Edmiston, 2020; Tenney et al., 2019b), and discourse-related information necessary for tasks such as coreference resolution (Tenney et al., 2019b). Moreover, the traditional NLP pipeline sequence of POS tagging, syntactic parsing, named entity recognition, semantic role labeling and coreference resolution can be mapped onto BERT’s transformer layers from lower to higher (Tenney et al., 2019a). Accordingly, several probing studies have shown that BERT captures a hierarchy of linguistic information (e.g., Jawahar et al., 2019; Lin et al., 2019; Edmiston, 2020): surface features are represented best in the lower layers, while syntactic features are captured best in the middle layers. The middle to higher layers represent morphological features best, and semantic information is captured best in the higher layers.

**Adversarial Testing** – Adversarial testing has shown that LMs struggle in making generalizations on basic lexical relations (Glockner et al., 2018), identifying ungrammaticality (Marvin and Linzen, 2018), efficiently capturing challenging linguistic phenomena, such as negation (Dasgupta et al., 2018; Richardson et al., 2020), modals, quantifiers and monotonicity (Richardson et al., 2020), passives (Zhu et al., 2018), conditionals (Richardson

et al., 2020), conjunctions (McCoy et al., 2019), implicatives and factives (McCoy et al., 2019), and modeling human reasoning patterns, such as numerical or common-sense reasoning (Naik et al., 2018). Overall, the evidence from adversarial testing contradicts the results of the probing studies: if the LM indeed is able to acquire ‘deep’ linguistic knowledge (e.g., about syntactic hierarchies), it should be able to deal with the phenomena present in the adversarial test sets.

**Contextualization** – Despite the conflicting evidence about the linguistic capacities of LMs like BERT, it is widely acknowledged that the word embeddings generated by such models are contextualized, i.e., there is no finite number of word sense representations and a word has different vector representations across different contexts. Particularly, by assessing a word’s contextualization on the basis of *self-similarity* scores, Ethayarajh (2019) shows that the embeddings become more contextualized, i.e., more context-specific, in the upper layers of BERT. Moreover, it has been shown that contextualized embeddings generally cluster with one another with respect to word senses (Reif et al., 2019; Wiedemann et al., 2019).

**Visual LM Explanations** – Approaches for visual LM explanations can be grouped into two main categories. One strand of research focuses on transformer-based LMs and explains *how* they learn through visualizing attentions (e.g., NLIZE (Liu et al., 2018), Seq2Seq-Vis (Strobel et al., 2018), BertViz (Vig, 2019), exBERT (Hoover et al., 2020), SANVis (Park et al., 2019), and Attention Flows (DeRose et al., 2021)). Another strand of research explains *what* the model learns by visualizing word embeddings. Although most existing work on embedding explanation is based on probing tasks, visualization of embedding characteristics has emerged as an active research topic. The first tools were related to the exploration of static embeddings, e.g., by Liu et al. (2017), who visualize word2vec and Glove embeddings, focusing on analogy exploration. Heimerl and Gleicher (2018) explain the same models and present visualizations that support analysis of multiple tasks, among others, the analysis of local word neighborhoods. Also, Boggust et al. (2019) explain static embeddings of word2vec, Glove, and fastText. Their explanations focus on local neighborhoods visualized using small multiples by applying a dimensionality reduction. Berger (2020) has recently presented a

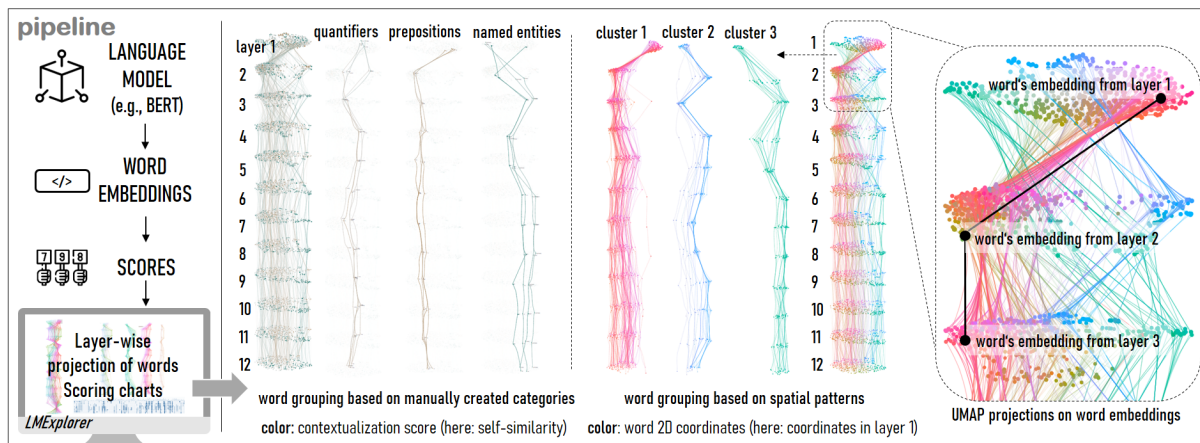


Figure 1: The main visualization of our technique uses *layer-wise interlinked-projections* that show embeddings from the layers of an LM in a 2D space; here, BERT’s 12 layers. The words in each layer are depicted as points and connected to their corresponding position throughout layers. By selectively mapping the colors of the links to computed scores or identified clusters, this visualization provides a global overview of the analyzed corpus.

visual approach for exploring correlations between embedding clusters in BERT for a single model’s layer at a time. The novelty of our approach is the explanation of contextualized word embeddings through contextualization scores that are visualized for all of the model’s layers simultaneously.

### 3 LMExplorer: Visual Analytics Technique

To support the analysis of word contextualization within the functionality continuum, we have developed a VA technique called *LMExplorer*. This technique discloses layer-wise spatial and score-based patterns in the learned embedding representations. Using interlinked embedding projections, we show the spatial relations of the high-dimensional embedding space. To provide further insights into the word contextualization, the technique utilizes scoring functions (i.e., word *self-similarity*) as a contextualization explanation. The scores are used to explore and navigate the embedding space, which is facilitated by supporting views and interactions. The technique is integrated into the *lingvis.io* framework (El-Assady et al., 2019a).

**Task Analysis** – The technique is designed to support model analysts in gaining insights into the word contextualization. The proposed design is informed by a set of tasks that were obtained through investigating the analysts in their typical analysis workflow. These are: (T1) Analyze spatial structure of the embedding space; (T2) Gain a global overview of the corpus; (T3) Conduct interactive pattern analysis; (T4) Create user-defined word groupings for detailed inspection; and (T5) Conduct a focused analysis of contextualization.

#### 3.1 Layer-wise Interlinked Projections

The main visual components of our technique are *layer-wise interlinked projections* (Figure 1) – a novel visualization displaying layers of the LM simultaneously for effective *spatial* pattern analysis.

**Motivation** – The design of this visualization was informed by T1 and T2, i.e., corpus level exploration of embedding spatial patterns in different layers of the LM. Projection-based visualizations are the most common methods to visualize word embeddings (e.g., Smilkov et al., 2016; Liu et al., 2017; van Aken et al., 2019; Aken et al., 2020) and although some approaches have enabled the exploration of embeddings in different layers (e.g., Smilkov et al., 2016; van Aken et al., 2019; Aken et al., 2020), they typically visualize only one layer of the LM at a time. However, changes in embedding positions and their neighborhoods across layers can be an indicator of the model capturing new context information. To support such analyses, our technique displays the embeddings for all layers of the LM simultaneously and visually highlights changes in their neighborhoods.

**Design Rationale** – To implement the exploration of such spatial patterns, we use a dimensionality reduction technique on the computed embedding vectors from each layer of the LM. In particular, we reduce the 768-dimensional embedding vectors to two dimensions, used as *x* and *y* coordinates to visualize words in one layer. Using this technique, words with similar embeddings are represented by similar coordinates in the 2D space. In total, 12 projections are created, each representing one layer of the BERT-base model. The projections



are ordered vertically underneath each other, starting from layer one at the very top and ending with the last layer at the bottom. The words in the projection are visualized as shapes. By default, they are displayed as circles and colored according to the word’s position in the 2D space, cf., [El-Assady et al. \(2019b\)](#). After displaying the projections, we add connecting lines between layers to support the analysis of word position changes in the visualized space. To reduce the number of crossing edges, we additionally apply an edge-bundling technique that combines neighboring edges in a more coherent representation. An example of the visualization is shown in [Figure 1](#). In our approach, both contextualized word embeddings and aggregated word embeddings (i.e., average or median embedding of all contexts of a word) can be visualized.

The words in each projection (i.e., layer) are represented by different embedding vectors. Hence, although we visualize the same words, the consecutive projections differ and may even get rotated or flipped due to artefacts that are common for most of the dimensionality reduction techniques (e.g., *UMAP* ([McInnes et al., 2018](#)), *t-SNE* ([Van der Maaten and Hinton, 2008](#))). Even if words maintained their neighborhoods, the rotation of the projections would prevent the users from easily comprehending on embedding positional changes. Thus, to prevent such artifacts, we apply an extension of *UMAP* called *AlignedUMAP*. It reduces the rotation artifacts by using the already projected data as an *anchoring*. Hence, we project the embeddings from layer 2 by specifying relations to the projection of embeddings from layer 1, and iterate this alignment process up to the last layer.

This spatialization concept enables an effective layer comparison as well as the detection of word groups with similar spatial patterns (T1, T2). The interlinked projections benefit the analysis of word functionality across layers, especially in the exploratory phase of the analysis. The user can brush neighboring words in the projection to gain an overview of word groups that are relevant to observe in detail. To support hypothesis generation and testing, we provide multiple interaction techniques that help explore the analyzed corpus. When hovering over a word in the projection, the word and its path through the different layers gets highlighted (T3) and its contexts are displayed for close-reading. To ease the analysis of words with common spatial patterns, the user can brush a group of

neighboring words in the projection and drag them aside. This reduces the displayed information and supports a more detailed pattern analysis (T4).

### 3.2 Explaining Contextualization

We employ common approaches in explaining contextualization and compute multiple word-level contextualization scores. These are integrated into the *interlinked-projection view* as an overlay (T5).

**Scoring Functions** – To explain the contextualization of a word’s representation, [Ethayarajh \(2019\)](#) introduces three metrics: *self-similarity*, *maximum explainable variance*, and *intra-sentence similarity*. In this paper, we focus on the word *self-similarity*, which [Ethayarajh](#) describes as “the average cosine similarity of a word with itself across all the contexts in which it appears, where representations of the word are drawn from the same layer of a given model.” Although the analysis in this paper is solely based on the *self-similarity* score, the technique can be effortlessly extended to further explanation scores. For instance, we have explored the word’s contextualization also by defining a *baseline* embedding and obtaining its similarity to the contextualized one. It is possible to create multiple baselines by either reducing the context size (e.g., extracting embedding from a word without a surrounding context) or selecting a specific layer of the LM for reference. [Ethayarajh \(2019\)](#) describes the 0<sup>th</sup> layer as an appropriate baseline. However, for specific hypothesis testing, one could even select one of the upper layers as a reference layer.

**Score Overlay** – The scores are mapped to the words in the *interlinked-projection view* to provide further insights into the embedding contextualization. In particular, we use three visual design elements: (a) color, (b) shape, and (c) size. First, we use a diverging color scale that maps the scores from **brown (min value)** to **green (max value)** colors. Second, we highlight words having extreme values (i.e., one standard deviation above the min value and below the max value of the score’s distribution *in the particular layer*) by displaying them as rectangles instead of the default circles. Third, we map the score’s range *across all layers* of the model to the shape’s size, supporting layer comparison (shown in [Figure 4](#)).

### 3.3 Supporting Visualizations & Interactions

To support the exploration of words with common characteristics (e.g., spatial patterns), we provide supporting visualizations and interactions.

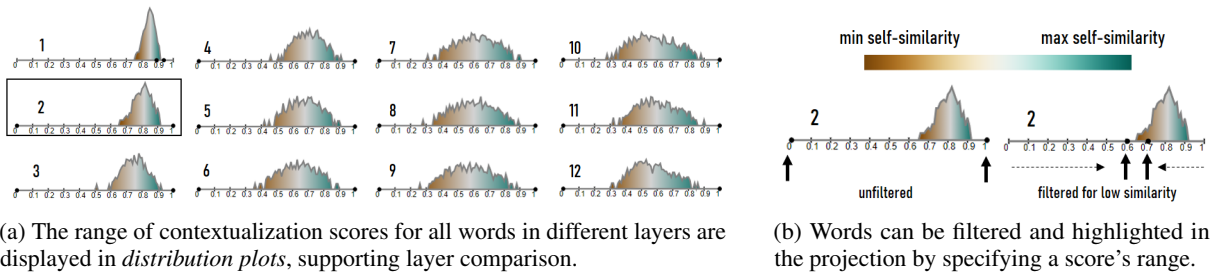


Figure 2: The *distribution plots* show that the average *self-similarity* of words decreases and, hence, word contextualization increases with increasing layers of BERT, which replicates the findings by [Ethayarajh \(2019\)](#).

The *distribution plots* provide an overview of the embedding contextualization scores (i.e., *self-similarity*) and are placed next to the corresponding layer projection. They enable the analysis of score changes through the model's layers. As shown in [Figure 2a](#), the *self-similarity* score decreases in upper layers, and the standard deviation increases accordingly. The *distribution plots* can be further used for filtering words by specifying a range in the contextualization score (shown in [Figure 2b](#)). Words that fit within the range are highlighted in the *interlinked-projection view*.

For tailored score-pattern analysis, we display the score changes in an additional, more compact *matrix plot* visualization (shown in [Figure 3](#)). The columns of the matrix represent words in the corpus, and rows show the layer-wise contextualization scores. The user can define a *query* by selecting a word in the *matrix plot* and the words with similar patterns (i.e., the *response* of the query) are highlighted in the *interlinked-projection view*. To obtain similar patterns, we first represent each word by a vector of 12 score values corresponding to each layer for BERT-base. We then compute the cosine similarity on these vectors to retrieve words with similar score patterns.

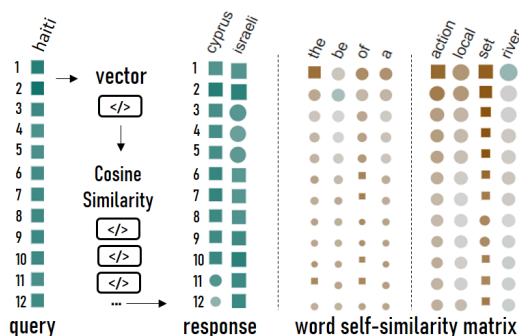


Figure 3: The *matrix plot* gives an overview of the *self-similarity* score changing over layers. By clicking on a column, the matrix is queried for similar score-patterns.

## 4 Exploring Contextualization in BERT

While [Ethayarajh \(2019\)](#) initially found that the increase in contextualization across the different BERT layers (i.e., the decreasing *self-similarity*) seems to be driven by polysemy, ‘stopwords’ such as *and*, *of*, *the* and *to* seem to contradict this conclusion. Stopwords, which in essence are function words, also become increasingly contextualized in the upper layers. Thus, contextualization seems not to be entirely driven by polysemy, but rather the variety of contexts a word appears in ([Ethayarajh, 2019](#)). However, function words are not a homogeneous class, and some function words indeed have semantic content in addition to having a grammatical function. Thus, we decided to investigate function and content words in more detail, using the *LMExplorer* to explore contextualization in BERT with respect to the functionality continuum.

### 4.1 Functional and Content Words

In theoretical linguistics, there is a traditional distinction between function and content words. Several criteria have been proposed to distinguish between the two groups, e.g., semantic content, membership openness, flexibility of syntactic attachment, separability from complements ([Corver and van Riemsdijk, 2001](#)). While content words comprise a specific semantic content and contribute to the principal meaning of a sentence, function words are rather ‘non-conceptual’ and mainly fulfill some grammatical function (e.g., expressing modality or definiteness), gluing content words together. Furthermore, content words are open-class because new members can freely be added. In contrast, function words are closed-class, i.e., they are members of a fixed set. Additionally, content words are flexible with respect to the syntactic phrase they attach to, e.g., the verb *think* can be complemented by an NP or a clause, while function words typically only combine with a specific syntactic phrase, e.g., a determiner with an NP. Also,





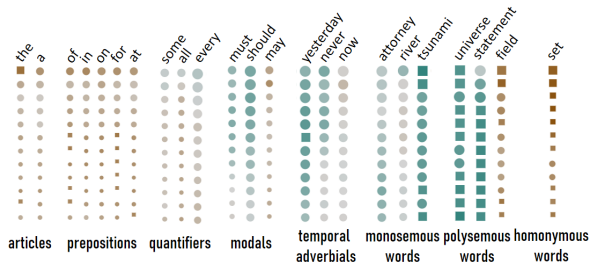


Figure 5: Layer-wise *self-similarity* scores for word samples/groups across the functionality continuum.

*home*, and homonymous words, e.g., *set*, occupy the space in the upper layers (e.g., layer 10, see Figure 4), and can also be found across the preceding layers. Prepositions (e.g., *of*, *for*) occur in the min range from the middle layers onwards. Moreover, the determiner *the* occurs in the min range at layer 11 and generally shows a low *self-similarity* (see Figure 3). In the mid range, we find temporal adverbials, e.g., *today* and *now*, modal verbs (*must*, *should*) as well as polysemous and monosemous words; see Figure 4. To shed light on these contextualization patterns, we explore the functionality continuum in more detail by looking at different groups of words across the layers.

**Word-based Selection** – We discern the following groups of words for our further explorations: 1) articles, 2) prepositions, 3) quantifiers, 4) modal verbs, 5) temporal adverbials, 6) monosemous words, 7) polysemous words and 8) homonymous words. Each group demonstrates a different pattern of *self-similarity* across layers, as shown in Figure 5. First, we observe that, before (almost) ending up in the min range, the determiners *the* and *a* start off in the mid range of the distribution with a decreasing *self-similarity* across the layers. Prepositions such as *of*, *in*, *on*, *for*, *at* are found in the mid-min area until layer 6 but from then on, they are grouped under min. Quantifiers like *some*, *all*, *every* remain in the mid range across all layers. Modal verbs such as *must*, *should*, *may* follow an inconsistent pattern: while *must* and *should* start off in the upper ends of the mid area (max-mid) and end up in the mid range from layer 9 on, *may* is at first in the min area and after layer 5 in the mid range. Temporal adverbials such as *yesterday*, *never*, *now* are also inconsistent. Some of them (e.g., *yesterday*) belong to the max group in the lower layers, but slowly move towards the mid area as the layers increase – without ever entering the exact mid area. Others (e.g., *now*, *never*) are constantly within the mid range, starting at the

higher end of mid and moving towards the middle. Monosemous words like *attorney*, *river*, *tsunami* are mostly found in the max range, with a decreasing tendency across layers, but remain in the upper ends of the max area. Polysemous words whose senses are very closely related, e.g., *universe*, *statement*, are also mostly found in the max area, while highly polysemous words whose senses are loosely related, e.g., *field*, are located in the min area in the lower layers and although their *self-similarity* increases, they remain in the min-mid area across layers. Finally, homonymous words, e.g., *set*, are in the min area across layers. These observations lead to new insights into how BERT captures contextualization, see Section 5.

## 5 Insights: The Functionality Continuum

During our exploration, we came across patterns that fit to the theory of the functionality continuum and others that were contrary to our expectations. Above all, we observed that contextualization is neither triggered merely by polysemy nor by variation in context. To explain the observed patterns, a) we positioned the defined categories within the functionality continuum<sup>2</sup> based on the inherent linguistic properties of the words and on insights from lexical semantics, and b) we identified three criteria as potential triggers of contextualization, as shown in Table 1. The first criterion refers to the sense variation (*Sense Var.*), i.e., whether a word has multiple senses (high variation), or only one or multiple but very closely related senses (low variation). The second criterion captures syntactic context variation (*SynCtx. Var.*), i.e., whether a word needs to be part of a specific syntactic structure (low) or is flexible in terms of attachment and can be found in different kinds of syntactic structures (high). Another potential trigger we identified is that of variation of semantic context (*SemCtx. Var.*). This captures whether the contexts in which a word can occur are semantically similar (low) or different (high) to one another. Based on these triggers and previous findings on contextualization by Ethayarajh (2019), we derive the *expected contextualization* (*Exp. Contextual.*) of each of the predefined categories. We can then compare this to BERT’s actual behavior (*BERT*) and shed light on BERT’s abilities to capture the functionality continuum. Note that here the expected contextualization coincides with the *SemCtx.Var.* for the categories investi-

<sup>2</sup>See also *semantic proximity continuum* by Blank (1997).


Functionality Continuum	Sense Var.	SynCtx. Var.	SemCtx. Var.	Exp. Contextual.	BERT
	homonymous	high	high	high	high ✓
	polysemous	low/high	high	low/high	low/high ✓
	monosemous	low	high	low	low ✓
	temp. adverbials	low	low	high	low ✗
	modals	high	low	high	low ✗
	quantifiers	high	low	high	low ✗
	prepositions	high	low	high	high ✓
	articles	none	low	high	high ✓

Table 1: Expected contextualization (Exp. Contextual.) and contextualization in BERT (BERT) on the basis of sense variation (Sense Var.), syntactic context variation (SynCtx. Var.) and semantic context variation (SemCtx. Var.), ordered based on the functionality continuum, from content (blue, top) to function words (yellow, bottom).

gated, but might deviate for others. Additionally, differences between the expected contextualization and the SemCtx.Var. might currently be absorbed by our binary encoding (low/high). We envision a more fine-grained Exp. Contextual. measure, accounting in detail for the relative positioning of words in the middle of the continuum.

**Homonymy** – Homonymous words, being on the ‘more content-like’ end of the continuum, have a high sense variation due to their multiple (unrelated) senses, a high syntactic variation (flexible attachment as content words) and a high semantic context variation as, due to their multiple senses, they can occur in semantically very different contexts. This means that we expect a high contextualization, i.e., the embeddings of homonymous words are highly context-specific. This is indeed confirmed with our findings since these words generally occur in the min area.

**Polysemy** – Polysemous words, mostly with ‘content-like’ properties, exhibit a low/high sense variation, depending on whether they are highly polysemous, i.e., have loosely related senses, or not, i.e., have semantically related senses. As it is typical of content words, polysemous words show high syntactic variation. Concerning their semantic context variation, they are again in a ‘grey’ area depending on the degree of polysemy: highly polysemous words mostly appear in semantically different contexts, while plain polysemy is mostly found in semantically similar contexts since the senses are closely related. With this, the expected contextualization is respective to the degree of the polysemy. Indeed, BERT meets these expectations: highly polysemous words like *field*, *home* are in the min area across layers (high contextualization), while plain polysemous words are rather found in the max area (low contextualization).

**Monosemy** – Monosemous words also seem to be correctly captured by BERT. Such words have low sense variation, high syntactic variation (as

content words) and low semantic context variation (due to their low sense variation). According to this, they are also expected to have low contextualization. We find this low contextualization in BERT as well, where monosemous words have max *self-similarity* across layers.

**Temporal Adverbials** – At the middle of the functionality continuum, temporal adverbials have a low sense variation, e.g., *yesterday* has only one meaning,<sup>3</sup> as well as low syntactic variation. On the other hand, their semantic context variation is high because they can occur in semantically very different contexts. Thus, the expected contextualization is high, i.e., their embeddings should be context-specific to match the semantically different contexts they can appear in. BERT fails to learn this: temporal adverbials are either found within the mid area across all layers or end up in this range in the upper layers, contrary to the expected min.

**Modals & Quantifiers** – BERT also struggles in capturing the functionality continuum with modals and quantifiers. These are comparable to words with high ‘sense’ variation: modals can not only have a deontic or an epistemic flavor, but also express variation through their variable quantificational force; similarly, quantifiers exhibit variation via their variable scope interpretation (wide or narrow). Both modals and quantifiers have low syntactic variation; they can only attach with specific syntactic phrases. The contexts they appear in can be semantically very different and thus they have a high semantic context variation. Based on this half-functional-half-content nature, modals and quantifiers are expected to have high contextualization, i.e., have context-specific embeddings based on the modal flavor they express, the quantificational force they capture, the scope resolution, etc. However, we can see that BERT fails to meet this expectation.

<sup>3</sup>It should be noted that such adverbials have one meaning, even if their extension is always a different one due to different reference points.



Modals and quantifiers mostly occur in the mid range – instead of the expected min.

**Prepositions** – At the functional end of the continuum, we find prepositions and articles. Prepositions are comparable to words with a high ‘sense’ variation, capturing the fact that the same preposition can, for example, be locative or temporal, depending on the context. Prepositions have low syntactic variation, as most functional words. Still, their semantic context variation matches their multiple ‘senses.’ Therefore, we expect the preposition embeddings to be highly context-specific: this is indeed the case in BERT, where prepositions are mostly found in the min area.

**Articles** – Last, we investigate articles and particularly the determiners *the* and *a*. We take them to have no sense,<sup>4</sup> low syntactic variation and high semantic context variation – the contexts they appear in do not have any semantic similarity in most cases. Thus, we expect them to demonstrate high contextualization with highly context-specific embeddings. BERT is able to model this through low *self-similarity*, which is more prominent for *the* than for *a*, nonetheless consistent for both.

**Discussion** – Summing up, we see that BERT struggles to efficiently capture the functionality continuum. While BERT manages to model the ends of the continuum, i.e., the mostly content and mostly functional words, it fails to create expressive embeddings for categories with content as well as functional properties. This finding is in line with previous literature that has shown that current LMs cannot efficiently capture hard linguistic phenomena (e.g., Dasgupta et al. (2018); McCoy et al. (2019); Richardson et al. (2020)), with modals, quantifiers and temporal reasoning belonging to these phenomena. Our work suggests that the BERT embeddings are not specific enough to capture the inherent functionality of certain word types, i.e., BERT does not learn the relevant generalizations. Additionally, we show that contextualization is neither entirely driven by polysemy nor context variation. Rather, contextualization can be explained via the harmonical combination of functionality, sense variation, syntactic variation and semantic context variation: BERT can efficiently model polysemy, homonymy and mononymy, i.e., it can efficiently capture words that appear in semantic contexts of high variation and low variation and

<sup>4</sup>We treat determiners as definiteness markers, rather than as quantifiers or discourse markers, to be in-line with their treatment in popular NLP tasks such as NLI.

independently of their polysemy. What it cannot model are words that have a semi-functional/semi-content nature (modals, quantifiers, temporal adverbials), see Table 1. Concerning modals and quantifiers, BERT cannot learn the inherent functionality from the context alone and thus treats the words as simple monosemous words. Concerning temporal adverbials, BERT cannot deal with the combination of low sense variation and high semantic context variation – a rather unusual combination – and is unable to conclude a single word meaning. Although prepositions have the same triggers as modals and quantifiers, BERT follows our expectations with respect to contextualization. This could be due to their higher syntactic flexibility or their close semantic relatedness with their content complements, but this needs to be explored as part of future work. Overall, BERT seems to follow findings of psycholinguistics and language acquisition: children learn content words easier and earlier than function words (Bates et al., 1994; Caselli et al., 1995). Drawing from language acquisition research, we see an opportunity for explainable methods to inspect BERT’s inner-workings and improve its linguistic understanding, raising LMs from their infantile state to a more linguistically-mature one.

## 6 Conclusion and Future Work

This paper presented new insights on the contextualization of the functionality continuum, showing that BERT fails to capture the nature of semi-functional-semi-content words. These insights were generated through a novel visual analytics technique for contextualized word embedding exploration and analysis. For a deeper understanding of the weaknesses of BERT, our technique can be extended with scores that model common linguistic properties of words and their nearest neighbors, e.g., WordNet semantic similarity or POS similarity scores. Hence, they could serve as means of explanation and bring added value to the eXplainable Artificial Intelligence (XAI) research field. More information about the project can be found under: <https://embeddings-explained.lingvis.io>.

## Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding within project BU 1806/10-2 “Questions Visualized” of the FOR2111 and project D02 “Evaluation Metrics for Visual Analytics in Linguistics” (Project ID: 251654672 – TRR 161).

## Broader Impact Statement

In the following, we describe the two main points with respect to the broader impact statement.

### Impact

With regard to the broader impact of our work, we are going beyond just measuring scores by revealing and explaining the inner-workings of language models. We put the measured scores in context through visual analytics, in combination with probing and adversarial testing methods, for the exploration, explanation, and analysis. With our work, we aim to open new perspectives on measuring and obtaining the model performance, which go beyond typically used performance metrics.

### Reproducibility

With regard to reproducibility concerns, we would like to note that the contextualization scores calculated in this paper rely on the word frequencies and, thus, may differ depending on the analyzed corpus. Future work should investigate the exact effect of word frequency and account for its impact.

## References

- Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. [How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations](#). In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, page 1823–1832, New York, NY, USA. Association for Computing Machinery.
- Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2020. [VisBERT: Hidden-State Visualizations for Transformers](#). In *Companion Proceedings of the Web Conference 2020, WWW '20*, page 207–211, New York, NY, USA. Association for Computing Machinery.
- Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop*, pages 1–9, Venice, Italy.
- Elizabeth Bates, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J. Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. [Developmental and stylistic variation in the composition of early vocabulary](#). *Journal of Child Language*, 21(1):85–123.
- Yonatan Belinkov. 2018. [On internal language representations in deep learning: An analysis of machine translation and speech recognition](#). Ph.D. thesis, Massachusetts Institute of Technology.
- M. Berger. 2020. [Visually Analyzing Contextualized Embeddings](#). In *2020 IEEE Visualization Conference (VIS)*, pages 276–280, Los Alamitos, CA, USA. IEEE Computer Society.
- Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.
- Angie Boggust, Brandon Carter, and Arvind Satyanarayan. 2019. [Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples](#). *arXiv e-prints*, page arXiv:1912.04853.
- Maria Cristina Caselli, Elizabeth Bates, Paola Casadio, Judi Fenson, Larry Fenson, Lisa Sanderl, and Judy Weir. 1995. [A cross-linguistic study of early lexical development](#). *Cognitive Development*, 10(2):159 – 199.
- Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. [Finding Universal Grammatical Relations in Multilingual BERT](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. [Supervised Learning of Universal Sentence Representations from Natural Language Inference Data](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single  \$\&\!#\ast\$  vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Norbert Corver and Henk van Riemsdijk. 2001. *Semilexical Categories: The Function of Content Words and the Content of Function Words*. De Gruyter Mouton, Berlin, New York.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. [The PASCAL recognising textual entailment challenge](#). In *Proceedings of the Machine Learning Challenges Workshop*, pages 177–190, Southampton, UK. Springer.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. [Evaluating Compositionality in Sentence Embeddings](#). *CoRR*, abs/1802.04302.

- Joseph F DeRose, Jiayao Wang, and M. Berger. 2021. Attention Flows: Analyzing and Comparing Attention Mechanisms in Language Models. *IEEE Transactions on Visualization and Computer Graphics*, 27:1160–1170.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Daniel Edmiston. 2020. A Systematic Analysis of Morphological Content in BERT Models for Multiple Languages. *arXiv preprint arXiv:2004.03032*.
- Mennatallah El-Assady, Wolfgang Jentner, Fabian Sperrle, Rita Sevastjanova, Annette Hautli, Miriam Butt, and Daniel Keim. 2019a. lingvis.io – A Linguistic Visual Analytics Framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18.
- Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. 2019b. Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. *IEEE transactions on visualization and computer graphics*, 26(1):1001–1011.
- Joseph E. Emonds. 1985. *A Unified Theory of Syntactic Categories*. De Gruyter Mouton, Berlin, Boston.
- Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic. Association for Computational Linguistics.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655. Association for Computational Linguistics.
- Yoav Goldberg. 2019. Assessing BERT’s Syntactic Abilities. *arXiv preprint arXiv:1901.05287*.
- Florian Heimerl and Michael Gleicher. 2018. Interactive Analysis of Word Vector Embeddings. In *Computer Graphics Forum*, volume 37, pages 253–265. Wiley Online Library.
- John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. 2019. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693.
- Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, System Demonstrations*. Association for Computational Linguistics.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT’s Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.
- Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2017. Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):553–562.
- Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. 2018. NLIZE: A Perturbation-Driven Visual Interrogation Tool for Analyzing and Interpreting Natural Language Inference Models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):651–660.



- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11):2579–2605.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted Syntactic Evaluation of Language Models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. [UMAP: Uniform Manifold Approximation and Projection](#). *The Journal of Open Source Software*, 3(29):861.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. [Stress Test Evaluation for Natural Language Inference](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. 2019. [SANVis: Visual Analytics for Understanding Self-Attention Networks](#). In *2019 IEEE Visualization Conference (VIS)*, pages 146–150. IEEE.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global Vectors for Word Representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep Contextualized Word Representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. [Visualizing and Measuring the Geometry of BERT](#). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8594–8603. Curran Associates, Inc.
- Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. [Probing Natural Language Inference Models through Semantic Fragments](#). In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 8713–8721. AAAI Press.
- 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.
- John R Ross. 1972. The category squish: Endstation hauptwort. In *Chicago Linguistic Society*, volume 8, pages 316–328.
- Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. 2016. [Embedding Projector: Interactive Visualization and Interpretation of Embeddings](#). *arXiv e-prints*, page arXiv:1611.05469.
- Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. [Seq2seq-vis: A visual debugging tool for sequence-to-sequence models](#). *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. [BERT Rediscovered the Classical NLP Pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. [What do you learn from context? Probing for sentence structure in contextualized word representations](#). In *International Conference on Learning Representations*.
- Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. [Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.
- Jason Utt and Sebastian Padó. 2011. [Ontology-based Distinction between Polysemy and Homonymy](#). In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.
- Jesse Vig. 2019. [A Multiscale Visualization of Attention in the Transformer Model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

- Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society*, Online.
- Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of KONVENS 2019*, Erlangen, Germany.
- Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. [Quantifying the Contextualization of Word Representations with Semantic Class Probing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.
- Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. [Exploring Semantic Properties of Sentence Embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia. Association for Computational Linguistics.