UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

A Psycholinguistic Analysis of BERT's Representations of Compounds

Buijtelaar, L.; Pezzelle, S.

DOI
10.48550/arXiv.2302.07232
10.18653/v1/2023.eacl-main.163

Publication date
2023

Document Version
Final published version

Published in
The 17th Conference of the European Chapter of the Association for Computational Linguistics

License
CC BY

Link to publication

Citation for published version (APA):
Buijtelaar, L., & Pezzelle, S. (2023). A Psycholinguistic Analysis of BERT's Representations of Compounds. In A. Vlachos, & I. Augenstein (Eds.), *The 17th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2023 : proceedings of the conference : May 2-6, 2023* (pp. 2230–2241). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2302.07232, https://doi.org/10.18653/v1/2023.eacl-main.163

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

A Psycholinguistic Analysis of BERT's Representations of Compounds

Lars Buijtelaar
University of Amsterdam
lars.buijtelaar@student.uva.nl

Sandro Pezzelle
ILLC, University of Amsterdam
s.pezzelle@uva.nl

Abstract

This work studies the semantic representations learned by BERT for compounds, that is, expressions such as *sunlight* or *bodyguard*. We build on recent studies that explore semantic information in Transformers at the word level and test whether BERT aligns with human semantic intuitions when dealing with expressions (e.g., *sunlight*) whose overall meaning depends—to a various extent—on the semantics of the constituent words (*sun*, *light*). We leverage a dataset that includes human judgments on two psycholinguistic measures of compound semantic analysis: lexeme meaning dominance (LMD; quantifying the weight of each constituent toward the compound meaning) and semantic transparency (ST; evaluating the extent to which the compound meaning is recoverable from the constituents' semantics). We show that BERT-based measures moderately align with human intuitions, especially when using contextualized representations, and that LMD is overall more predictable than ST. Contrary to the results reported for 'standard' words, higher, more contextualized layers are the best at representing compound meaning. These findings shed new light on the abilities of BERT in dealing with fine-grained semantic phenomena. Moreover, they can provide insights into how speakers represent compounds.

1 Introduction

Compounds such as *sunlight* or *bodyguard* are an interesting benchmark to probe the semantic representations learned by any NLP models. On the one hand, compounds that are part of a language lexicon (i.e., lexicalized compounds; Gagné and Spalding, 2006) have their own (sets of) established meaning(s). As such, they are lexical items just like any other word. On the other hand, the semantic status of compounds is special since their meaning is the result of the combination of the meaning of two words (hence, the constituents). According to psycholinguistic evidence, this semantic relation does not disappear with lexicalization. Indeed, speakers actively combine constituent meanings when processing both novel and lexicalized compounds (Gagné and Spalding, 2009; Ji et al., 2011; Marelli and Luzzatti, 2012; Marelli et al., 2014).

In this work, we argue that NLP systems capable of faithfully representing word meanings should account for these aspects. For example, to acknowledge that the meaning of *handgun* relies more on the semantics of *gun* than of *hand* (indeed, a *handgun* is a type of *gun*). Or, that the meaning of *sunlight* is more directly recoverable from the semantics of its constituents (it is more transparent) than is the meaning of *muskrat* (which is very opaque).

Transformer-based encoders such as BERT (Devlin et al., 2019) are shown to produce word representations that align well with human semantic intuitions, particularly at their lower layers (Bommasani et al., 2020; Vulić et al., 2020). This suggests that these models are effective in encoding the meaning of a word, without any additional task-specific fine-tuning. However, these conclusions are based on evaluations that explore semantic relations between words, such as pairwise similarity patterns—not between words and their parts.

In parallel, BERT's contextualized embeddings have been leveraged for tasks that involve lexical composition. For example, to learn compound representations that are effective in predicting the *literality* of a compound or its semantic interpretation (Shwartz and Dagan, 2019). In this case, mixed results were reported. While BERT-based models are effective to judge, e.g., that *market* (but not *flea*) has a literal meaning in *flea market*, they are far behind humans in predicting, e.g., that *body part* stands for *part that makes up a body*. Crucially, these results were obtained by training a binary classifier on the top of BERT's embeddings. Since the encoder parameters were updated during learning, no conclusions can be drawn on the ef-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2230–2241
May 2-6, 2023 ©2023 Association for Computational Linguistics

fectiveness of BERT's embeddings in dealing with these and similar fine-grained semantic aspects.

In this work, we leverage a dataset that includes human judgments on two psycholinguistic measures of compound semantic analysis: lexeme meaning dominance (LMD) and semantic transparency (ST). The former quantifies the semantic weight of each constituent toward the compound meaning. For example, that *gun* has more semantic weight in *handgun* than *hand* does. The second evaluates the extent to which the compound meaning is recoverable from the semantics of the constituents. For example, that *handgun* is very transparent, while *muskrat* is much less so.

We test whether, and to what extent, the measures of LMD and ST that we obtain from BERT's representations of compounds and compound constituents align with human judgments. We carry out comprehensive experiments on model versions, contexts, pooling methods, layers.[1] We show that:

- BERT is moderately aligned with human intuitions on both measures, which confirms the effectiveness of the model in accounting for the fine-grained semantic aspects captured by LMD and ST. At the same time, LMD is substantially more predictable than ST;
- only representations extracted from words in a context (in a sentence), but not without a context (in isolation), are aligned with human intuitions, which reflects BERT's struggle to handle out-of-context words. Moreover, the highest correlations are achieved in higher, deeply contextualized layers. This could be due to the nature of the semantic evaluation subtending LMD and ST, which likely requires relying on a specific semantic interpretation of the compound rather than on abstract lexico-semantic information;
- both $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ outperform the GloVe (Pennington et al., 2014) baseline, with $\text{BERT}_{\text{large}}$ achieving the best results overall. This confirms the effectiveness of BERT models to represent word-level semantics, in line with previous work (Bommasani et al., 2020; Vulić et al., 2020);
- BERT accounts for the left and right constituents equally when representing the semantics of a compound. Moreover, these representations appear to encode the complex semantic and syntactic relation tying the constituents.

2 Related Work

2.1 Compound Semantics in Psycholinguistics

Compounds are one of the favorite subjects of psycholinguistic research. One of the reasons is that they are extremely productive: a new combination of two (or more) words can be generated at any time and get lexicalized through language use (Gagné and Spalding, 2006). Indeed, compounds have been considered to serve as a "backdoor into the lexicon" (Downing, 1977). While understanding novel compounds clearly involves accessing both the meaning of the constituents and the semantic relation tying them together, recent psycholinguistic evidence has shown that an active combination of the meaning of the constituent words is routinely in place also for lexicalized compounds (Gagné and Spalding, 2009; Ji et al., 2011; Marelli and Luzzatti, 2012; Marelli et al., 2014). Indeed, most psycholinguistic research in this field focuses on the constituents and their relation with compounds. For example, to study and quantify the role of frequency, semantic transparency, or headedness (Gagné and Spalding, 2009; Marelli et al., 2009; Marelli and Luzzatti, 2012; Juhasz et al., 2015).[2]

Recently, a few studies leveraging methods from NLP have been carried out to either reproduce or quantify some of these aspects. By typically using *static* embeddings (Mikolov et al., 2013; Pennington et al., 2014) and compositional models of distributional semantics (Mitchell and Lapata, 2010; Guevara, 2010), these approaches have proven successful in building compound representations that approximate, e.g., the different semantic and syntactic role of a compound's modifier and head, semantic transparency, plausibility of a novel combination or syntax-based categorizations (Günther and Marelli, 2016; Marelli et al., 2017; Günther and Marelli, 2019; Pezzelle and Marelli, 2020).

Though powerful, these methods have one crucial limitation, namely, they require training a set of parameters via supervised learning to obtain representations for compounds that are novel or simply not present in the corpus. Transformer-based encoders such as BERT (Devlin et al., 2019) lift this limitation. Without any additional training or

[1] Data and code are made available at https://github.com/lars927/compounds-analysis-bert

[2] Headedness refers to the property of having a head—in English, typically the right constituent. The left is the modifier.

fine-tuning, in fact, they can represent any novel or unseen word—provided that it can be divided into known subwords. Since compounds involve a meaningful combination of two constituents, they represent an interesting benchmark to test the representations by these models.

2.2 Word Representation in Transformers

A recent line of work started to investigate the type of semantic information encoded in the embeddings by pre-trained Transformer encoders. While this was a classical benchmark to evaluate static, word-level embeddings learned by previous generation models, e.g., word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), the problem appears less trivial for current state-of-the-art NLP models (Westera and Boleda, 2019; Mickus et al., 2020; Lenci et al., 2022). Indeed, the embeddings by Transformer-based encoders are *contextualized*, i.e., affected by both the surrounding context and the position within a sentence. Moreover, they often represent subwords rather than whole words.

By means of a simple method to pool the various contextualized embeddings learned for a word into a single, static embedding, Bommasani et al. (2020) showed that these representations align with human judgments of semantic similarity better than how previous-generation ones do. In particular, lower layers perform the best, which reveals that these layers encode abstract, lexico-semantic information. Similar findings were reported by Vulić et al. (2020), who extended the investigation to other five languages than English. Taken together, these results are complementary to the findings that representations in higher layers tend to become more context-specific (Ethayarajh, 2019) and to better encode word senses (Reif et al., 2019).

Recent work (Shwartz and Dagan, 2019) leveraged BERT embeddings to obtain representations for compounds by means of (trained) lexical composition models—similarly to how it was done for compositional distributional semantic models. However, to the best of our knowledge, no work to date has explored how Transformer-based encoders represent compounds. The most relevant study in this direction is the one by Pinter et al. (2020), which focused on BERT's representations for blends (i.e., words such as *shoptics*, resulting from the merging of *shop* and *optics*) and included a comparison with novel compounds. They reported an overall high similarity between the compound and the constituents, slightly increasing through the layers.

By focusing on lexicalized compounds from a psycholinguistic angle, we are the first to study how BERT represents these complex expressions.

compound	LMD [0,10]	ST [1,7]
handgun	8.13 →	6.29 ↑
bodyguard	7.27 →	5.64 ↑
policeman	3.07 ←	6.13 ↑
wartime	3.47 ←	6.31 ↑
muskrat	7.53 →	2.80 ↓
primrose	7.93 →	2.00 ↓
milestone	3.36 ←	2.21 ↓
cheapskate	2.00 ←	2.00 ↓

Table 1: A few examples from the dataset with either high ↑ or low ↓ ST and either low ← or high → LMD. E.g., the meaning of *handgun* is deemed highly transparent and based more on the right than the left constituent.

3 Data

We use a psycholinguistic dataset of human judgments on compound LMD and ST (Juhasz et al., 2015). The dataset includes 629 lexicalized English compounds annotated by 189 participants for various variables.[3] LMD is a score that captures which of the two constituents of a compound is semantically dominant for the compound meaning. It ranges in [0,10], where 0 means totally dependent on the left constituent and 10 means totally dependent on the right constituent. In Table 1, we report a few examples from the dataset. As can be seen, compounds such as *handgun* or *muskrat* have a high LMD, i.e., the right constituent is semantically dominant. In contrast, compounds such as *policeman* or *milestone* have a low LMD, i.e., the left constituent is semantically dominant.

ST is defined as a score that quantifies the degree to which the meaning of a compound can be inferred or recovered from the meaning of the constituents: the higher the ST, the more transparent the compound. The compounds *handgun* and *wartime* in Table 1, for example, are fully transparent: both the constituents contribute to their meaning. In contrast, compounds such as *primrose* or *cheapskate* are fully opaque: neither of the two constituents contributes to its meaning.

Since only the compounds, but not the constituents, are provided in the dataset, we manually

[3] Such as LMD, ST, age of acquisition, and imageability.

annotate each compound (e.g., *handgun*) with its left (*hand*) and right (*gun*) constituents, so to obtain a dataset of ⟨compound, left, right⟩ triplets. While doing so, we decided to discard the pseudo-compound *mushroom*. We were left with 628 triplets, that we use in our experiments.

4 Method

We test whether, and to what extent, BERT's representations of compounds and compound constituents approximate human judgments on LMD and ST. To do so, we obtain word-level representations using two versions of BERT. We experiment with representations obtained by feeding the word either in isolation or in the context of a sentence. Moreover, building on previous work, we explore various pooling methods over BERT outputs.

4.1 Models

BERT (Devlin et al., 2019) is a Transformer-based model pre-trained on a large number of English texts. It is pre-trained using two learning objectives, i.e., Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is about predicting some words that have been masked in the input. NSP is about predicting whether two concatenated sentences follow each other (or not).

We experiment with two versions of BERT, i.e., $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$. The former has 12 encoder layers stacked on top of each other, 12 attention heads, and 110M parameters. At each layer, it learns 768-d embeddings. The latter has 24 layers, 16 attention heads, and 340M parameters. It learns 1024-d embeddings. For both models, we use HuggingFace (Wolf et al., 2020) implementations.[4]

4.2 Word-Level Representations

For each triplet in the dataset, we employ BERT models to obtain representations for the compound, the left constituent, and the right constituent. We henceforth use the general term *word* to refer to any of the items in a triplet. We obtain representations for words in two conditions, *no-context* (`NC`) and *in-context* (`C`), that we describe below.

No-Context (`NC`) In this condition, we obtain a single, static representation for a word (e.g., *snowboard*) by feeding it into the model in isolation, i.e., without any surrounding context. When fed with a single word, BERT outputs embeddings for the tokens which make it up (that result from the tokenization process), as well as for the special tokens [CLS] and [SEP] at the beginning and end of the sequence, respectively. Following previous work (Vulić et al., 2020), we explore 3 methods for obtaining a word representation. These methods differ with respect to what embeddings are taken into account when building such representation:

- `nospec` This method ignores the special tokens [CLS] and [SEP]. A word representation is built by averaging the embeddings of the tokens that make up the word (*snow*, *##board*);
- `withcls` This method builds a word representation by averaging the embedding for the special token [CLS] with the embeddings of the tokens making up the word (*snow*, *##board*);
- `all` This method builds a word representation by averaging all the embeddings that are output by BERT for the sequence, i.e., [CLS], [SEP], and the tokens making up the word.

In-Context (`C`) In this condition, we follow the method by Bommasani et al. (2020) to obtain a single, static representation of a word from the N contextualized embeddings produced by BERT for that word in context. First, we average the representations of the tokens that make up a given word—as in the `NC_nospec` setting. Second, we consider all the contextualized representations for a given word and aggregate them to obtain a single representation that is not dependent on a specific context. We do this by averaging the N contextual representations of a word $w_1, \ldots, w_N$:

$$w = mean(w_1, \ldots, w_N) \quad (1)$$

To obtain contextualized vectors, we sample sentences containing items from our 628 triplets from a cleaned English Wikipedia corpus.[5] For each word, we sample all the sentences in the corpus that contain it, up to a maximum of 100 unique instances per word. The average number of instances per word in our sample is 89.3 (min 1, max 100).

We henceforth simply refer to this setting as `C`.

Experimental details Within each setting, we therefore obtain a single 768-d ($\text{BERT}_{\text{base}}$) or 1024-d ($\text{BERT}_{\text{large}}$) embedding for each compound and

[4] https://huggingface.co/bert-base-uncased
https://huggingface.co/bert-large-uncased

[5] https://www.lateral.io/resources-blog/the-unknown-perils-of-mining-wikipedia

model	setting	metric (best layer)	
		MAE ↓	Spearman ρ ↑
GloVe	–	**0.945**	0.541
$\text{BERT}_{\text{base}}$	NC_nospec	1.095 (11)	0.375 (11)
	NC_all	1.072 (11)	0.384 (11)
	NC_withcls	1.071 (11)	0.385 (11)
	C	0.991 (11)	*0.563 (10)*
$\text{BERT}_{\text{large}}$	NC_nospec	1.130 (21)	0.247 (21)
	NC_all	1.105 (21)	0.247 (21)
	NC_withcls	1.107 (22)	0.244 (21)
	C	*0.966 (21)*	**0.586 (21)**

Table 2: LMD. Results in **bold** and *italic* are the best and second-best in the column, respectively. Results are from a model's best-performing layer (in parentheses).

constituent in our dataset. This operation is performed for each layer of each model—i.e., 12 layers in $\text{BERT}_{\text{base}}$ (1-12) and 24 layers in $\text{BERT}_{\text{large}}$ (1-24). All representations are obtained by running a pre-trained BERT in inference mode, i.e., without fine-tuning or updating the model's weights.

Baseline As a baseline, we employ static embeddings by GloVe (Pennington et al., 2014). We use the 300-d embeddings from the version of the model trained with 6B tokens.[6] Since four compounds[7] were not found in GloVe's vocabulary, for this baseline we obtain results for 624 triplets.

4.3 Predicting Psycholinguistic Measures

LMD is a scalar in [0,10] that quantifies the relative semantic role of each constituent toward the meaning of the whole compound: The higher the value, the more the compound's semantics depends on the right constituent. Using BERT's representations for a ⟨compound, left, right⟩ triplet, we therefore operationalize LMD as follows:

$$LMD(c) = 5(R - L) + 5 \quad (2)$$

where c is the compound, L is the cosine similarity in [0,1] between the left constituent and the compound, $cos(left, compound)$, and R is the cosine similarity between the right constituent and the compound, $cos(right, compound)$. The scaling and addition operations make the values range in [0,10]. If $L = 0$ and $R = 1$, then $LMD(c) = 10$. *Vice versa*, if $L = 1$ and $R = 0$, $LMD(c) = 0$.

ST is a scalar in [1,7] that quantifies the degree to which the meaning of a compound can be inferred from the meaning of the constituents. The higher the value, the more the compound semantics can be inferred from the two constituents' meanings. Using BERT's representations for ⟨compound, left, right⟩, we operationalize ST as follows:

$$ST(c) = \frac{6(L + R)}{2} + 1 \quad (3)$$

where c, L, and R are defined as above. The scaling and addition operations make the values range in [1,7]. If $L = 1$ and $R = 1$, then $ST(c) = 7$. *Vice versa*, if $L = 0$ and $R = 0$, $ST(c) = 1$.

4.4 Evaluation

We evaluate the effectiveness of each model in approximating human LMD and ST by means of two metrics: mean absolute distance (MAE) and Spearman correlation (ρ) between the predicted and human values. For MAE, the lower the distance, the better. For ρ, the higher the correlation, the better.

In the next section, we report results by all models in all settings in approximating LMD and ST.

5 Results

5.1 Lexeme Meaning Dominance

In Table 2, we report the results by (the best layer of) each model on LMD in the various settings. Several key observations can be made. First, both BERT models achieve moderate positive correlation[8] (close to 0.6) with human judgments, with $\text{BERT}_{\text{large}}$ outperforming $\text{BERT}_{\text{base}}$ by some margin. On the one hand, this indicates that BERT's representations do a fairly good job in accounting for the relative semantic weight of each constituent in a (lexicalized) compound. On the other hand, it suggests that more data and parameters play a role in approaching human intuitions.

Second, BERT models outperform GloVe in terms of correlation. This indicates that BERT's embeddings not only encode sensible semantic information (in line with previous findings; see Bommasani et al., 2020; Vulić et al., 2020) but also align with human semantic intuitions to a greater extent than the previous-generation GloVe model. However, it is worth noting that BERT models outperform GloVe only in C, but not in NC. This clearly shows that BERT's embeddings have an advantage over GloVe's ones only when leveraging information in the surrounding context and reveals

[6] https://nlp.stanford.edu/data/glove.6B.zip

[7] Namely, *livelong*, *sunlamp*, *dunghill*, and *handclasp*.

[8] As per standard interpretation (Prion and Haerling, 2014), we consider ρ correlations from ±0.41 to ±0.60 as moderate.

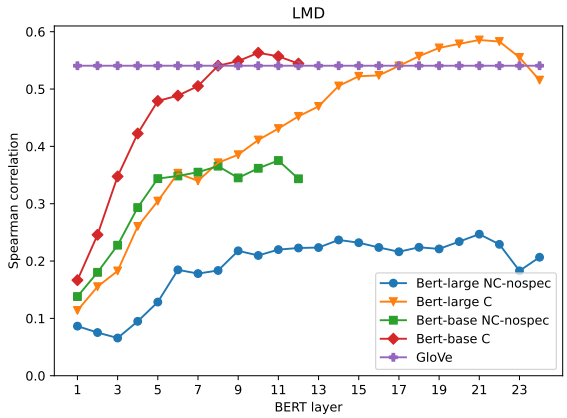

Figure 1: LMD. ρ against model layers. For out-of-context BERT models, we only report the best-performing nospec setting. Best viewed in color.

that BERT struggles to represent out-of-context words, likely due to its architecture and training regime. In A.1, we report that the lack of *any* surrounding context in NC is indeed detrimental to model representations, though sensible contextual information in C is needed to properly approximate LMD. Moreover, GloVe achieves the lowest MAE, which shows that these embeddings are effective in approximating the raw LMD values—though contextualized C BERT representations are better at capturing the overall pattern of similarity.[9]

Third, the pattern of correlation values over the model's layers highlights the importance of contextualization for compound representation. As can be seen in Figure 1, best-performing C BERT models show an almost constant increasing trend, with the highest correlation being achieved in high layers—layer 9 and 20 in $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$, respectively. This is an opposite pattern compared to what was observed in previous work, where lower layers were found to encode most lexical semantic information (Bommasani et al., 2020; Vulić et al., 2020). These patterns do not necessarily contradict each other. Indeed, we argue that judging the semantic relationship between a compound (e.g., *handgun*) and its constituent words (*hand*, *gun*) might involve relying on a specific interpretation of the compound rather than on abstract lexico-semantic information. Since previous work showed that word senses are better encoded in deeper layers (Reif et al., 2019), this could explain why these layers are also good at capturing LMD. A similar—though flatter—trend is observed for NC models.

[9]Recall that MAE measures the distance between the target and predicted values, while Spearman ρ quantifies strength and direction of association between the two ranked variables.

model	setting	metric (best layer)	
		MAE ↓	Spearman ρ ↑
GloVe	–	2.657	0.304
$\text{BERT}_{\text{base}}$	NC_nospec	0.953 (6)	0.316 (5)
	NC_all	1.129 (10)	0.234 (1)
	NC_withcls	0.989 (1)	0.275 (3)
	C	*0.899 (9)*	*0.415 (9)*
$\text{BERT}_{\text{large}}$	NC_nospec	0.989 (9)	0.195 (6)
	NC_all	1.118 (24)	0.113 (1)
	NC_withcls	1.024 (6)	0.139 (1)
	C	**0.876 (19)**	**0.476 (20)**

Table 3: ST. Results in **bold** and *italic* are the best and second-best in the column, respectively. Results are from a model's best-performing layer (in parentheses).

5.2 Semantic Transparency

In Table 3, we report the results by (the best layer of) each model on ST in the various settings. Several key observations can be made. First, both BERT models achieve moderate positive correlation with human judgments, with $\text{BERT}_{\text{large}}$ outperforming $\text{BERT}_{\text{base}}$. This indicates that BERT's representations are moderately effective in predicting the extent to which a compound meaning is recoverable from the meaning of its constituents and that more data and parameters help—indeed, the gap between the two BERT models is higher here than in LMD (0.06 vs 0.02). At the same time, the highest correlation achieved by $\text{BERT}_{\text{large}}$ (0.476) on ST is substantially lower than on LMD (0.586), which indicates that ST is more challenging to approximate compared to LMD.

Second, BERT models outperform GloVe on

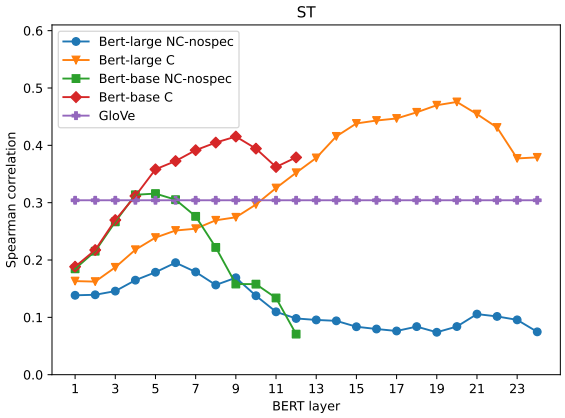

Figure 2: ST. ρ against model layers. For out-of-context BERT models, we only report the best-performing nospec setting. Best viewed in color.

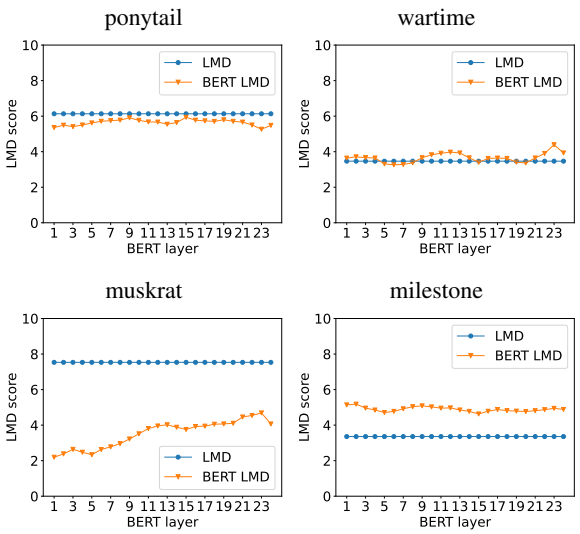

Figure 3: Examples where C $BERT_{large}$ is good (top) and bad (bottom) in approximating human LMD. From top left, clockwise: *ponytail*, *wartime*, *milestone*, *muskrat*.

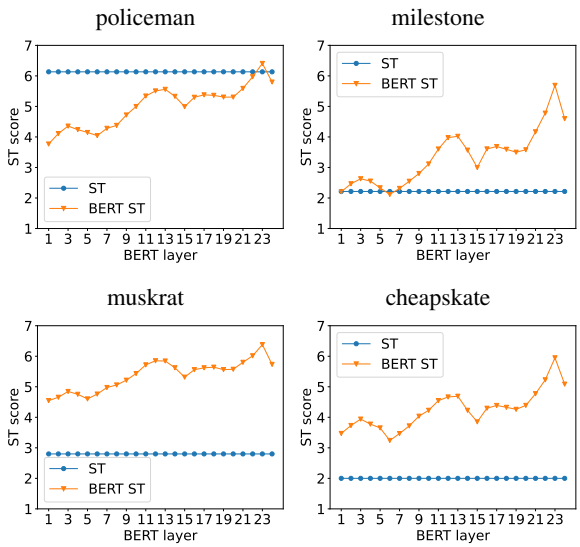

Figure 4: Examples where C $BERT_{large}$ is good (top) and bad (bottom) on ST. From top left, clockwise: *policeman*, *milestone*, *cheapskate*, *muskrat*.

both metrics. Though correlations are generally lower than in LMD, the gap between BERT models and GloVe is much more pronounced here. That is, BERT's contextualized embeddings have an even clearer advantage over previous-generation ones in modeling ST compared to LMD.

Third, as can be seen in Figure 2, the overall best results are achieved by C embeddings in high layers—layer 8 and 19 for $BERT_{base}$ and $BERT_{large}$, respectively—which replicates the findings for LMD. This confirms the role of context and contextualization for obtaining better representations of compounds.[10] Interestingly, in NC settings, BERT models show a different pattern compared to LMD, with correlation reaching a 'peak' within the first layers and then constantly decreasing. This suggests that decontextualized lexico-semantic information encoded in lower layers accounts for ST to some extent, on par with or even outperforming GloVe (this is the case for $BERT_{base}$ NC_nospec).

5.3 Examples

In Figure 3 we report some examples where the best-performing C $BERT_{large}$ is good (top) and bad (bottom) in approximating LMD. As can be seen, the distance between the predicted and human LMD is extremely low for *ponytail* and *wartime*, namely, a high-LMD (6.13) and a low-LMD (3.47) compound, respectively. In contrast, the distance is very large for high-LMD (7.53) *muskrat* and low-LMD (3.36) *milestone*. While the predicted LMD values are fairly stable over the layers for *ponytail*, *wartime*, and *milestone*, for *muskrat* the higher layers are better to approximate the real LMD by assigning an increasingly higher semantic weight to the *rat* lexeme. This could be due to the representation of *muskrat* becoming more 'aware'—through contextualization—of the semantic traits related to the animal domain, apparently less present in earlier layers. Also, it is interesting to note that for the compound *milestone*, which is 'exocentric' (i.e., the head is neither *mile* nor *stone*), BERT keeps predicting a conservative LMD value, which similarly weights the two constituents, over the layers. That is, contextualization does not make *mile* or *stone* become dominant in the compound (while, interestingly, human speakers consider *mile* as slightly dominant over *stone*).

In Figure 4, we report some good (top) and bad (bottom) cases for the same model in approximating ST. As can be seen, the distance between the predicted and human ST is very low for high-ST (6.31) *policeman* and low-ST (2.21) *milestone*. In contrast, the distance is large for low-ST (2.80) *muskrat* and low-ST (2.00) *cheapskate*. Moreover, it can be noted that higher layers are better than lower layers in approximating ST for *policeman*, which is highly transparent, while they are worse for *milestone*, *cheapskate*, and *muskrat*, which are very opaque. This goes hand in hand with a seemingly general trend observed in these examples:

[10] See the analysis in A.1 for further evidence of the role of sensible semantic context in approximating ST.

the more contextualization, the higher the ST. This could reflect a generalized increase of cosine similarity values through BERT's layers (as reported by Ethayarajh, 2019), which would lead in turn to a higher ST by virtue of how ST was operationalized; see Eq. 3. However, our results (see Figure 2) show that this increase does not correspond to a higher correlation with human judgments: indeed, correlation steadily decreases in the last 4 layers of $\text{BERT}_{\text{large}}$, where cosine similarities are highest.

Finally, while BERT struggles on both LMD and ST for *muskrat*, for *milestone* it struggles on LMD but does a good job on ST. This confirms that LMD and ST capture different semantic aspects: a good performance on one measure does not necessarily guarantee a good performance on the other.[11]

5.4 Which Factors Drive the Prediction?

To more formally investigate which factors contribute to higher predicted values of LMD and ST by the best performing model `C` $\text{BERT}_{\text{large}}$, we run two linear regression models in `R`—one for LMD, one for ST—using, for each compound, the predicted LMD/ST value by the best layers (21/20, respectively) as the dependent variable, and the following independent variables: (1) the number of tokens into which the compound was split by the tokenizer, e.g., 2 for *snowboard* (*snow*, *##board*); (2) the frequency of the compound in our dataset, i.e., the number of instances on the top of which the average representation was computed; (3) the compound concreteness; (4) the modifier (left constituent) concreteness; (5) the head (right constituent) concreteness. Concreteness values are extracted from Brysbaert et al. (2014).[12]

For LMD, both the concreteness of the head and the modifier—but not other variables—have a statistically significant role, though in the opposite direction: the higher the former, the higher the LMD (i.e., more weight to the head); the higher the latter, the lower the LMD (i.e., more weight to the modifier). This makes intuitive sense and shows that BERT assigns more 'weight' to concrete constituents. For ST, three variables have a statistically significant role in predicting higher values, and all in the same direction: the higher the number of tokens, the compound concreteness, and the modifier concreteness, the higher the ST. As for concreteness, this generally shows that BERT assigns higher similarities to concrete words. As for the effect of the number of tokens, this is an interesting finding, which reveals that BERT considers as more transparent those compounds than can be routinely broken into parts. The full tables reporting all the effects and corresponding coefficients and p-values can be found in A.2.

6 Analysis

6.1 LMD: Reversed Compounds

From the results reported in section 5.1, it appears that BERT is capable of obtaining sensible representations of compounds that encode the relation between the constituents and their respective semantic 'weight'. However, it might still be that the reported moderate correlations result from the model assigning a default high/low similarity to the constituents while being no or little aware of the semantic and syntactic (i.e., the modifier/head) relation which ties them. If that is the case, the model would consider, e.g., the contribution of *war* in *wartime* and *timewar* to be identical—though the meaning of the *reversed* compound would be intuitively very different. As such, we might expect a similar/same LMD value assigned by the model to these two compounds, and therefore a similar correlation with human judgments. Otherwise, if BERT represents a compound by genuinely accounting for the relationship between its constituents, the predicted LMD for the reversed compound (*timewar*) is likely to be different. As such, we might expect a much lower correlation with human intuitions.

In this analysis, we test this issue by re-running the LMD experiment on the *reversed* version of the compounds in our dataset, i.e., *wartime* > *timewar*, *bodyguard* > *guardbody*, etc. LMD is computed exactly as above, except that we replace the representation of the compound with that of its reversed version. Since (most of) these reversed compounds are unlikely to occur in standard *corpora* of texts, we experiment with `NC` representations. Moreover, we experiment with $\text{BERT}_{\text{base}}$ since it outperformed $\text{BERT}_{\text{large}}$ on this setting. As can be seen in Figure 5, the correlation for reversed compounds is much lower than the original one (0.11 vs 0.39). This is a good sign, which confirms that BERT does not rely on shortcuts when representing compound meanings and the relative weight of each constituent. Instead, it appears that the model can account for the semantic and syntactic relation

[11] Computing the correlation between LMD and ST human values in the dataset (-0.013) confirms this intuition.

[12] 23 compounds out of 628 were not present in the concreteness database and therefore excluded from the analysis.

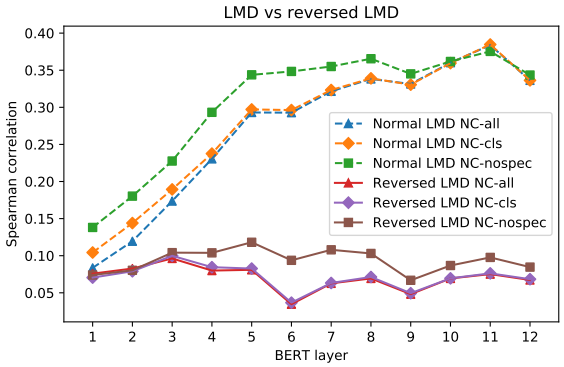

Figure 5: ρ for LMD vs LMD *reversed* values by out-of-context $\text{BERT}_{\text{base}}$ across layers. Best viewed in color.

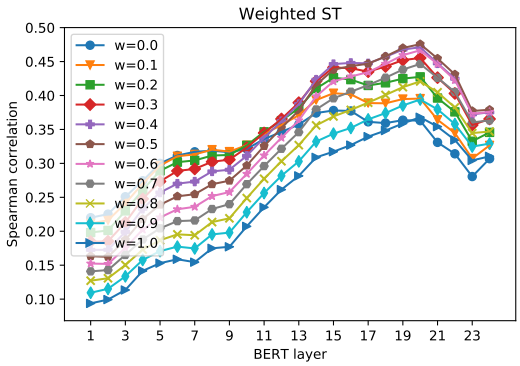

Figure 6: ρ for *weighted* ST by in-context $\text{BERT}_{\text{large}}$ across layers. Each weight stands for the weight assigned to the left constituent. Best viewed in color.

between the constituents, even when information from the surrounding context is not available.

6.2 ST: Weighted Constituents

From the results in section 5.2, it appears that BERT can approximate, to some extent, the degree to which the meaning of a compound is recoverable from the semantics of the constituents. When operationalizing ST we assumed that the two constituents are equally responsible for the overall ST—i.e., we computed the unweighted average between the two pairwise similarities. This way, a high (low) similarity between the compound and one of its constituents would not determine a high (low) ST on its own. This operationalization is in line with psycholinguistic literature, according to which we have a fully transparent compound when both lexemes contribute to its meaning and a fully opaque compound when neither of the two contribute (see, e.g., Libben, 1998). However, it is an open question whether BERT's compound representations do encode both constituents equally, or whether they disproportionately encode one constituent over the other. If the latter is the case, weighing one more than the other when computing ST would possibly result in a higher correlation with human judgments. If both constituents are equally represented in the compound embedding, instead, the unweighted version—our main experiment—would lead to the highest correlation.

In this analysis, we test this issue by computing a *weighted* version of ST where the left and right constituents are assigned different weights. For example, we assign weights 0.0 and 1.0 to the left and right constituent, respectively (in this case, only the right one, the compound's *head*, but not the left one, the *modifier*, will be responsible for ST); 0.1 and 0.9; 0.2 and 0.8; and so on. We experiment with all combinations of weights (including 0.5 and 0.5), which sums up to 11 versions of *weighted* ST.

Figure 6 reports the results of this analysis for the best-performing C $\text{BERT}_{\text{large}}$ model—note that the weight refers to the weight assigned to the left constituent. As can be seen, the highest correlation is achieved by the unweighted ST (weight 0.5) at layer 20. Since these are the results reported in Table 3, this finding confirms that both constituents are equally accounted for in the BERT's compound representation. Interestingly, weights that are close to 0.5 perform reasonably well (0.4 and 0.6 rank second and third, respectively), while correlation decreases the more we move away from this value.

Overall, the results of these two analyses confirm that BERT accounts for the left and right constituents equally when representing the semantics of a compound. Moreover, these representations seem to encode the complex semantic and syntactic relationship tying the compound constituents.

7 Conclusion

In this paper, we study how BERT represents the meaning of lexicalized compounds. We take a psycholinguistic angle and show that the model does a reasonably good job of making semantic judgments in line with those of human speakers. Since higher, more contextualized layers are shown to correlate best with human intuitions, we propose that speakers may access a specific, context-dependent representation when making these judgments. In future work, recent approaches to multimodal word representation in Transformers (Pezzelle et al., 2021) could be leveraged to test the role of visual grounding in compound semantics (Günther et al., 2020).

Limitations

Impact of the corpus To build contextualized representations for compounds, we sample sentences from a corpus of texts. A limitation of our work lies in the use of encyclopedic data only, which limits the number and variety of contextualized meanings a compound can have. Further attention should be paid to this aspect.

Operationalization of the measures While defining LMD and ST based on the cosine similarity between a compound and its constituent makes intuitive sense, it may not be the only (nor the best) way to operationalize the two measures. Further exploration on how to formally define them based on the model embeddings should be carried out.

Ethics Statement

Broader impact We do not see any serious ethical problem connected to this research. At the same time, we are aware of the risks associated with the development and use of large NLP models that we use in this research. Such risks include the environmental impact of the computational resources required for training and the encoding and possible amplification of biases present in the massive amounts of un-curated data the models learn from.

Acknowledgements

We would like to thank Marco Marelli for the idea of testing different weights when computing ST. We are grateful to the anonymous EACL reviewers and meta-reviewer for the insightful feedback.

References

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

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.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

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.

Christina L. Gagné and Thomas L. Spalding. 2006. Conceptual combination: Implications for the mental lexicon. In *The representation and processing of compound words*, chapter 7, pages 145–168. Oxford University Press Oxford, New York.

Christina L. Gagné and Thomas L. Spalding. 2009. Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, 60(1):20–35.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Fritz Günther and Marco Marelli. 2016. Understanding karma police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PloS one*, 11(10):e0163200.

Fritz Günther and Marco Marelli. 2019. Enter sandman: Compound processing and semantic transparency in a compositional perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(10):1872.

Fritz Günther, Marco Alessandro Petilli, and Marco Marelli. 2020. Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language*, 112:104104.

Hongbo Ji, Christina L. Gagné, and Thomas L. Spalding. 2011. Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language*, 65(4):406–430.

Barbara J Juhasz, Yun-Hsuan Lai, and Michelle L Woodcock. 2015. A database of 629 english compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior research methods*, 47(4):1004–1019.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, pages 1–45.

Gary Libben. 1998. Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and language*, 61(1):30–44.

Marco Marelli, Davide Crepaldi, and Claudio Luzzatti. 2009. Head position and the mental representation of nominal compounds: A constituent priming study in italian. *The Mental Lexicon*, 4(3):430–454.

Marco Marelli, Georgiana Dinu, Roberto Zamparelli, and Marco Baroni. 2014. Picking buttercups and eating butter cups: Spelling alternations, semantic relatedness, and their consequences for compound processing. *Applied Psycholinguistics*, pages 1–19.

Marco Marelli, Christina L Gagné, and Thomas L Spalding. 2017. Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224.

Marco Marelli and Claudio Luzzatti. 2012. Frequency effects in the processing of italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of memory and language*, 66(4):644–664.

Timothee Mickus, Mathieu Constant, Denis Paperno, and Kees van Deemter. 2020. What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. *Proceedings of the Society for Computation in Linguistics*, 3.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

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.

Sandro Pezzelle and Marco Marelli. 2020. Do semantic features capture a syntactic classification of compounds? insights from compositional distributional semantics. *The role of constituents in multiword expressions*, page 33.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. Will it unblend? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.

Susan Prion and Katie Anne Haerling. 2014. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10(10):535–536.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Matthijs Westera and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 120–133.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

A Appendix

A.1 Templated Linguistic Contexts

In this analysis, we investigate the extent to which the disadvantage of NC compared to C in approximating LMD and ST is due to the lack of *any* linguistic context surrounding the compound and the constituent words.[13] Since BERT has probably seen very few examples of words out of context during training, it could be that the model is poor at handling words in isolation, which would have an impact on the resulting representations—and LMD/ST values. To test this issue, we consider the best-performing $\text{BERT}_{\text{large}}$ and use it to obtain a single, contextualized representation for words (either compounds or constituents) by embedding them in the following templated sentence: *This is a*

[13]We thank an anonymous reviewer for suggesting this analysis.

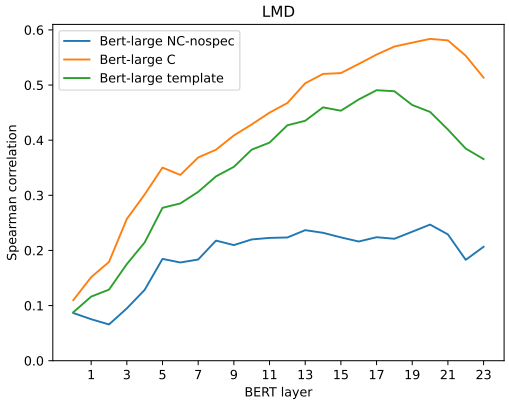

Figure 7: ρ for LMD by BERT$_{\text{large}}$ NC (blue), C (orange), and templated (green). Best viewed in color.

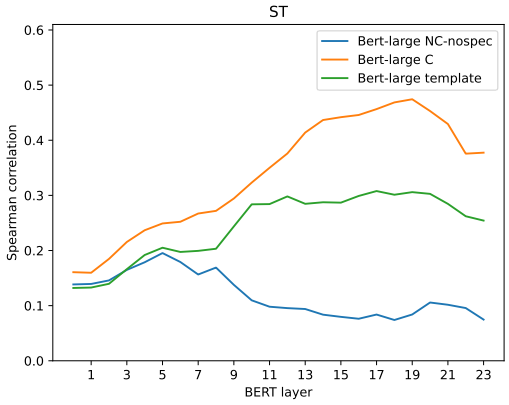

Figure 8: ρ for ST by BERT$_{\text{large}}$ NC (blue), C (orange), and templated (green). Best viewed in color.

<word>. This setting bears similarities with both NC and C. On the one hand, we compute a single representation for each word, similarly to NC. On the other hand, the representation of each word is contextualized (i.e., embedded in a linguistic context), though the surrounding context does not contain any meaningful semantic information. As such, we expect these representations to be better than NC by virtue of their higher similarity with standard training samples, but worse than C since they lack any sensible semantic information coming from the context surrounding the word at inference time.

The results follow the expected pattern. As can be seen in Figure 7 and 8, the correlation values obtained in the templated setting (best ρ for LMD: 0.491; best ρ for ST: 0.308) lie somehow in between C and NC. While this shows that embedding words in a sentence leads to a representational advantage over the out-of-context presentation (they clearly outperform NC), these representations are still far behind C and either on par with (ST) or neatly below (LMD) the GloVe baseline. This confirms that leveraging the meaningful linguistic context where compounds and constituents occur is crucial for obtaining sensible word representations that encode information on LMD and ST in line with human intuitions.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.03711 -0.37943  0.04126  0.38335  2.24977

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.4857018  0.1759300  25.497  < 2e-16 ***
Number.of.tokens       -0.0384983  0.0375281  -1.026  0.30538
Freq.Compound           0.0010187  0.0008464   1.204  0.22925
Concreteness.Compound   0.0405699  0.0334697   1.212  0.22594
Concreteness.Head       0.0952494  0.0333439   2.857  0.00443 **
Concreteness.Modifier -0.0802175  0.0334971  -2.395  0.01694 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.623 on 599 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.03101,   Adjusted R-squared:  0.02292
F-statistic: 3.834 on 5 and 599 DF,  p-value: 0.001991
```

Figure 9: LMD. Linear regression model predicting LMD values by the best-performing C BERT$_{\text{large}}$ layer.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.60126 -0.32514 -0.01245  0.31058  1.66230

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.4107078  0.1404585  24.283  < 2e-16 ***
Number.of.tokens        0.1544043  0.0299616   5.153 3.48e-07 ***
Freq.Compound          -0.0007155  0.0006758  -1.059   0.2902
Concreteness.Compound   0.1743992  0.0267214   6.527 1.43e-10 ***
Concreteness.Head       0.0406109  0.0266210   1.526   0.1277
Concreteness.Modifier   0.0611921  0.0267433   2.288   0.0225 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4974 on 599 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.2344,    Adjusted R-squared:  0.228
F-statistic: 36.68 on 5 and 599 DF,  p-value: < 2.2e-16
```

Figure 10: ST. Linear regression model predicting ST values by the best-performing C BERT$_{\text{large}}$ layer.

A.2 Linear Regression Model

Figure 9 and 10 report all the effects and corresponding coefficients and p-values of the linear regression models described in Section 5.4 for LMD and ST, respectively.