
Proceedings Paper:
Garcia, M., Vieira, T.K., Scarton, C. et al. (2021) Probing for idiomaticity in vector space models. In: Merlo, P., Tiedemann, J. and Tsarfaty, R., (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), 19-23 Apr 2021, Virtual conference. Association for Computational Linguistics (ACL), pp. 3551-3564.

Probing for idiomaticity in vector space models

Marcos Garcia
CiTIUS Research Centre
Universidade de Santiago de Compostela
Galiza, Spain

Tiago Kramer Vieira
Federal University
of Rio Grande do Sul, Brazil

Carolina Scarton
University of Sheffield, UK

Marco Idiart
Federal University
of Rio Grande do Sul, Brazil

Aline Villavicencio
University of Sheffield, UK
Federal University
of Rio Grande do Sul, Brazil

marcos.garcia.gonzalez@udc.gal, tiagokv@hotmail.com,
c.scarton@sheffield.ac.uk, marco.idiart@gmail.com,
a.villavicencio@sheffield.ac.uk

Abstract

Contextualised word representation models have been successfully used for capturing different word usages, and they may be an attractive alternative for representing idiomaticity in language. In this paper, we propose probing measures to assess if some of the expected linguistic properties of noun compounds, especially those related to idiomatic meanings, and their dependence on context and sensitivity to lexical choice, are readily available in some standard and widely used representations. For that, we constructed the Noun Compound Senses Dataset, which contains noun compounds and their paraphrases, in context neutral and context informative naturalistic sentences, in two languages: English and Portuguese. Results obtained using four types of probing measures with models like ELMo, BERT and some of its variants, indicate that idiomaticity is not yet accurately represented by contextualised models.

1 Introduction

Contextualised word representation models, like BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), seem to represent words more accurately than static word embeddings like GloVe (Pennington et al., 2014), as they can encode different usages of a word. In fact, representations of a word in several contexts can be grouped in different clusters, which seem to be related to the various senses of the word (Schuster et al., 2019), and they can be used to match polysemous words in context to specific sense definitions (Chang and Chen, 2019). However, multiword expressions (MWEs) fall into a continuum of idiomaticity[1] (Sag et al., 2002; Fazly et al., 2009; King and Cook, 2017) and their meanings may not be directly related to the meanings of their individual words (e.g., *graduate student* vs. *eager beaver* as a hardworking person). Therefore, one question is whether and to what extent idiomaticity in MWEs is accurately incorporated by word representation models.

In this paper, we propose a set of probing measures to examine how accurately idiomaticity in MWEs, particularly in noun compounds (NCs), is captured in vector space models, focusing on some widely used representations. Inspired by the semantic priming paradigm (Neely et al., 1989), we have designed four probing tasks to analyse how these models deal with some of the properties of NCs, including non-compositionality (*big fish* as an important person), non-substitutability (*panda car* vs. *bear automobile*), or ambiguity (*bad apple* as either a rotten fruit or a troublemaker), as well as the influence of context in their representation. To do so, we have created the new Noun Compound Senses (NCS) dataset, containing a total of 9,220 sentences in English and Portuguese. This dataset includes sentence variants with (i) synonyms of the original NCs; (ii) artificial NCs built with synonyms of each component; or (iii) either the head or the modifier of the NC. Moreover, it is composed of naturalistic and controlled sense-neutral sentences, to minimise the possible effect of context words.

We compare five models (one static, GloVe, and four contextualised, ELMo and three BERT-based models) in English and Portuguese. The probing measures suggest that the standard and widely adopted composition operations display a limited ability to capture NC idiomaticity.

Our main contributions are: (i) the design of novel probes to assess the representation of id-

[1] We understand idiomaticity as *semantic opacity* and its continuum as different *degrees of opacity* (Cruse, 1986).

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3551–3564
April 19 - 23, 2021.

iomaticity in vector models, (ii) a new dataset of NCs in two languages, and (iii) their application in a systematic evaluation of vector space models examining their ability to display behaviors linked to idiomaticity.

The remainder of this paper is organized as follows: First, Section 2 presents related work. Then, we describe the data and present the probing measures in Section 3. In Section 4, we discuss the results of our experiments. Finally, the conclusions of our study are drawn in Section 5.

2 Related Work

Priming paradigms have been traditionally used in psycholinguistics to examine how humans process language. For compounds, some findings suggest that idiomatic expressions are processed more slowly than semantically transparent ones, as processing the former may involve a conflict between the non-compositional and the compositional meanings (Gagné and Spalding, 2009; Ji et al., 2011). However, studies using event-related potential (ERP) data showed that idiomatic expressions, especially those with a salient meaning (Giora, 1999), have processing advantages (Laurent et al., 2006; Rommers et al., 2013). In NLP, probing tasks have been useful in revealing to what extent contextualised models are capable of learning different linguistic properties (Conneau et al., 2018). They allow for more controlled settings, removing obvious biases and potentially confounding factors from evaluations, and allowing both the use of artificially constructed but controlled sentences and naturally occurring sentences (Linzen et al., 2016; Gulordava et al., 2018). In priming tasks, related stimuli are easier to process than unrelated ones. One assumption is that, for models, related stimuli would achieve greater similarity than unrelated stimuli. These tasks have been used, for instance, to evaluate how neural language models represent syntax (van Schijndel and Linzen, 2018; Prasad et al., 2019), and the preferences that they may display, such as the use of mainly lexical information in a lexical substitution task even if contextual information is available (Aina et al., 2019).

Concerning pre-trained neural language models, which produce contextualised word representations, analyses about their abilities have shown, for instance, that they can encode syntactic information (Liu et al., 2019) including long-distance subject–verb agreement (Goldberg, 2019). Regarding semantic knowledge, the results of various experiments suggest that BERT can somewhat represent semantic roles (Ettinger, 2020). However, its improvements appear mainly in core roles that may be predicted from syntactic representations (Tenney et al., 2019). Moreover, from the representations generated by BERT, ELMo and Flair (Akbik et al., 2018) for word sense disambiguation, only the clusters of BERT vectors seem to be related to word senses (Wiedemann et al., 2019), although in cross-lingual alignment of ELMo embeddings, clusters of polysemous words related to different senses have also been observed (Schuster et al., 2019).

The use of contextualised models for representing MWEs has been reported with mixed results. Shwartz and Dagan (2019) evaluated different classifiers initialised with contextualised and non-contextualised embeddings in five tasks related to lexical composition (including the literality of NCs) and found that contextualised models, especially BERT, obtained better performance across all tasks. However, for capturing idiomaticity in MWEs, static models like *word2vec* (Mikolov et al., 2013) seem to have better performance than contextualised models (Nandakumar et al., 2019; King and Cook, 2018). These mixed results suggest that a controlled evaluation setup is needed to obtain comparable results across models and languages.

Therefore, we have carefully designed probing tasks to assess the representation of NCs in vector space models. As the same word can have different representations even in related paraphrased contexts (Shi et al., 2019), we adopt paraphrases with minimal modifications to compare the idiomatic and literal representations of a given NC.

3 Materials and Methods

3.1 Noun Compound Senses Dataset

The Noun Compound Senses (NCS) dataset is based on the NC Compositionality dataset, which contains NCs in English (Reddy et al., 2011), Portuguese and French (Cordeiro et al., 2019). Using the protocol by Reddy et al. (2011), human judgments were collected about the interpretation of each NC in 3 naturalistic corpus sentences. The task was to judge, for each NC, how literal the contributions of its component were for its meaning (e.g., "Is *climate change* truly/literally a *change* in *climate*?"). Each NC got a score, which was the average of the human judgments with a *Likert* scale from 0 (non-literal/idiomatic) to 5 (lit-

eral/compositional).[2]

For the NCS dataset, a set of probing sentences for the 280 NCs in English and the 180 NCs in Portuguese was added. For each NC, the sentences exemplify two conditions: (i) the naturalistic context provided by the original sentences (NAT), and (ii) a neutral context where the NCs appear in uninformative sentences (NEU). For the latter we use the pattern *This is a/an <NC>* (e.g., *This is an eager beaver*) and its Portuguese equivalent *Este/a é um(a) <NC>*. As some NCs may have both compositional and idiomatic meanings (e.g., *fish story* as either *an aquatic tale* or *a big lie*), these neutral contexts will be used to examine the representations that are generated for the NCs (and the sentences) in the absence of any contextual clues about the meaning of the NC. Moreover, they enable examining possible biases in the NC representation especially when compared to the representation generated for the NAT condition.

For each NC and condition, we created new sentence variants with lexical replacements, using synonyms of the NC as a whole or of each of its components. The synonyms of the NCs are the most frequent synonyms provided by the annotators of the original NC Compositionality dataset (e.g., *brain* for *grey matter*). The synonyms of each component were extracted from WordNet (Miller, 1995, for English) and from English and Portuguese dictionaries of synonyms (e.g., *alligator* for *crocodile* and *sobs* for *tears*). In cases of ambiguity (due to polysemy or homonymy), the most common meaning of each component was used. Experts (native or near-native speakers with linguistics background) reviewed these new utterances, keeping them as faithful as possible to the original ones, but with small modifications for preserving grammaticality after the substitution (e.g., modifications in determiners and adjectives related to gender, number and definiteness agreement).

NCS contains a total of 5,620 test items for English and 3,600 for Portuguese among neutral and naturalistic sentences, and it is freely available.[3]

[2]We averaged the Likert judgments for comparability with previous work, even though the median may reflect better the cases where there is more disagreement among the annotators. However, both mean and median are strongly correlated in our data: $\rho = 0.98$ (English) and $\rho = 0.96$ (Portuguese), $p < 0.001$.

[3]https://github.com/marcospln/noun_compound_senses

3.2 Probing Measures

This section presents the probing measures defined to assess how accurately idiomaticity is captured in vector space models. For these measures we consider comparisons between three types of embeddings: (i) the embedding for an NC out of context (i.e. the embedding calculated from the NC words alone), represented by ϵ_{NC}; (ii) the embedding for an NC in the context of a sentence S, represented by $\epsilon_{\text{NC} \subset \text{S}}$[4] (iii) finally, the sentence embedding that contains an NC, which is represented by $\epsilon_{\text{S} \supset \text{NC}}$. Here we use the standard output of some widely used models with no fine-tuning to avoid possible interference. However, in principle, these measures could apply to any embedding even after fine-tuning.

The similarities between embeddings are calculated in terms of cosine similarity: $\cos(\epsilon, \epsilon')$ where ϵ and ϵ' are embeddings from the same model with the same number of dimensions. In NAT cases, the similarity scores for each of the three available sentences for a given NC are averaged to generate a single score. We use Spearman ρ correlation between similarities and the NC idiomaticity scores (280 for English and 180 for Portuguese) to check for any effects of idiomaticity in the probing measures. We also calculate Spearman ρ correlation between different embedding models to determine how much the models agree, and between the NAT and NEU conditions to see how much the context affects the distribution of similarities. We also analyse the distribution of cosine similarities produced by different models for each of the probing measures. All probing measures are calculated for both NAT and NEU conditions.

P1: Probing the similarity between an NC and its synonym. If a contextualised model captures idiomaticity accurately, the embedding for a sentence containing an NC should be similar to the embedding for the same sentence containing a synonym of the NC (NC_{syn}, e.g., for *grey matter*, NC_{syn} = *brain*). Thus, $\text{sim}^{(\text{P1})}_{\text{Sent}} \simeq 1$, where $\text{sim}^{(\text{P1})}_{\text{Sent}} = \cos(\epsilon_{\text{S} \supset \text{NC}}, \epsilon_{\text{S} \supset \text{NC}_{syn}})$. This should occur regardless of how idiomatic the NC is, that is, similarity scores are not expected to correlate with NC idiomaticity scores ($\rho^{(\text{P1})}_{\text{Sent}} \simeq 0$). Moreover, this should also hold for the NC and NC_{syn} embeddings generated in the context of this sentence, which means that $\rho^{(\text{P1})}_{\text{NC}} \simeq 0$ and $\text{sim}^{(\text{P1})}_{\text{NC}} \simeq 1$

[4]For non-contextualised embeddings $\epsilon_{\text{NC} \subset \text{S}} = \epsilon_{\text{NC}}$.

Naturalistic sentence	NC	NC_{syn}	NC_{synW}
Field work *and practical archaeology are a particular focus.*	field work	research	area activity
The town centre is now deserted - it's almost like a ***ghost town****!*	ghost town	abandoned town	spectre city
How does it feel to experience a ***close call*** *only to come out alive and kicking?*	close call	scary situation	near claim
Eric was being an ***eager beaver*** *and left work late.*	eager beaver	hard worker	restless rodent
No wonder Tom couldn't work with him; he is a ***wet blanket****.*	wet blanket	loser	damp cloak

Table 1: Naturalistic examples with their NC_{syn} and NC_{synW} counterparts.

where $\text{sim}^{(\text{P1})}_{\text{NC}} = \cos(\epsilon_{\text{NC} \subset \text{S}}, \epsilon_{\text{NC}_{syn} \subset \text{S}})$. The baseline similarity scores can be approximated using the out-of-context embeddings for NC and NC_{syn}.

P2: Probing single component meaning preservation. As the meaning of a more compositional compound can be inferred from the meanings of its individual components, we evaluate to what extent an NC can be replaced by one of its component words and still be considered as representing a similar usage in a sentence. We measure $\text{sim}^{(\text{P2})}_{\text{Sent}} = \cos(\epsilon_{\text{S} \supset \text{NC}}, \epsilon_{\text{S} \supset \text{w}_i})$ and $\text{sim}^{(\text{P2})}_{\text{NC}} = \cos(\epsilon_{\text{NC} \subset \text{S}}, \epsilon_{\text{w}_i \subset \text{S}})$, where w_i is the component word (head or modifier) with the highest similarity, as for some NCs the main meaning may be represented by either its head or modifier. Similarity scores for idiomatic NCs should be low as they usually cannot be replaced by any of its components. In contrast, for more compositional NCs, the similarity is expected to be higher. For example, while for a more compositional NC like *white wine*, the head *wine* would provide a reasonable approximation as w_i, the same would not be the case for *grey matter*, a more idiomatic NC. Therefore, we expect significant correlations between the similarity values and the NC idiomaticity scores, that is $\rho^{(\text{P2})}_{\text{Sent}} > 0$ and $\rho^{(\text{P2})}_{\text{NC}} > 0$.

P3: Probing model sensitivity to disturbances caused by replacing individual component words by their synonyms. We examine whether vector representations are sensitive to the lack of individual substitutability of the component words displayed by idiomatic NCs (Farahmand and Henderson, 2016). To compare an NC with an expression made from synonyms of its component words (NC_{synW}, e.g., for *grey matter*, NC_{synW} = *silvery material*), we measure $\text{sim}^{(\text{P3})}_{\text{Sent}} = \cos(\epsilon_{\text{S} \supset \text{NC}}, \epsilon_{\text{S} \supset \text{NC}_{synW}})$ and $\text{sim}^{(\text{P3})}_{\text{NC}} = \cos(\epsilon_{\text{NC} \subset \text{S}}, \epsilon_{\text{NC}_{synW} \subset \text{S}})$. These substitutions should provide more similar variants for compositional than for idiomatic cases, and the similarity scores should correlate to the NC idiomaticity scores, that is $\rho^{(\text{P3})}_{\text{Sent}} > 0$ and $\rho^{(\text{P3})}_{\text{NC}} > 0$.

P4: Probing the similarity between the NC in the context of a sentence and out of context. To determine how much for a given model an NC in context differs from the same NC out of context we measure $\text{sim}^{(\text{P4})}_{\text{in-out}} = \cos(\epsilon_{\text{NC} \subset \text{S}}, \epsilon_{\text{NC}})$. We expect similarity scores to be higher in the NEU condition, given their semantically vague context, than for the NAT condition.

3.3 Calculating Embeddings

We use as a baseline the static non-contextualised GloVe model (Pennington et al., 2014) and, for contextualised embeddings, four widely adopted models: ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and two BERT variants, DistilBERT (DistilB) (Sanh et al., 2019) and Sentence-BERT (SBERT) (Reimers and Gurevych, 2019b). For all the contextualised models, we use their pre-trained weights publicly available through the Flair implementation[5]. For GloVe, the English and Portuguese models described in Pennington et al. (2014) and Hartmann et al. (2017). For ELMo, we use the small model provided by Peters et al. (2018), and for Portuguese we adopt the weights provided by Quinta de Castro et al. (2018). For all BERT-based models, we used the multilingual models for both English and Portuguese.[6]

To have a single embedding for the whole sentence or its parts, e.g., the NC representation, we use the standard procedure of averaging the vectors of the involved tokens.[7] In GloVe and ELMo, we average the output embeddings of each word, while in BERT-based models we obtain the final vector by averaging those of the sub-tokens (e.g., 'wet', 'blank' and '##et' for *wet blanket*).

Different combinations of the last five layers were probed in BERT-based models. However, they led to qualitatively similar results, and for reasons of presentation clarity, have been omitted

[5] https://github.com/flairNLP/flair

[6] We also investigated dedicated models for English, however, for allowing a more direct comparison between the languages, we report results only for the multilingual models.

[7] We discuss other operations in section 4.6.

from the discussion. We focus on embeddings calculated from a combination of the last four layers as they have been found to be representative of the other combinations. For ELMo, as it is intended to serve as a contextualised baseline, we represent the word embeddings using the concatenation of its three layers, albeit it is known that separate layers and weighting schemes generate better results in downstream tasks (Reimers and Gurevych, 2019a).

4 Results

This section discusses our results for each probing measure, using cosine similarities and Spearman ρ correlations. A qualitative analyses is also presented where we compare BERT and GloVe results of the five NCs in Table 1 (which shows the naturalistic sentences for each NC, together with their respective NC_{syn} and NC_{synW})[8]. We also discuss the average results of other NCs in both conditions and these results and other examples can be found in the Appendix.

4.1 Can contextualised models capture the similarity between an NC and its synonym?

If a contextualised model successfully captures idiomaticity, we would expect (i) high cosine similarity between a sentence containing an NC and its variant using a synonym of the NC (P1), and (ii) little or no correlation with the NC idiomaticity score. The results confirm high similarity values for all models, as shown in Figure 1a. However, this is not the case if we consider only the embeddings in context for NC and NC_{syn}, which display a larger spread of similarity values (see Figure 1b). Moreover, contrary to what was expected, a moderate correlation was found between most models and the idiomaticity scores (P1 in Table 2), indicating lower similarity scores for idiomatic than for compositional cases, for both NAT and NEU conditions.

Even though the high $\text{sim}^{(\text{P1})}_{\text{Sent}}$ values seem to suggest idiomaticity is captured, lower $\text{sim}^{(\text{P1})}_{\text{NC}}$ and moderate correlations with idiomaticity scores contradict it. Therefore a possible explanation for high similarities for Sent may be the effect of the overlap in words between a sentence and its variant (i.e., the context in Sent). This is also compatible with the larger similarities observed for NAT than for NEU condition since the average sentence length for the naturalistic sentences is 23.39 for English and 13.03 for Portuguese, while for the neutral it is five words for both languages. Moreover, a similar performance was also obtained with GloVe.[9] It is also worth noting that, in contrast to static embeddings, contextualised word representations are anisotropic, occupying a narrow cone in the vector space and therefore tending to produce higher cosine similarities (Ethayarajh, 2019).

The results with the first probing measure show that even though the similarities can be relatively high, they are consistently lower for idiomatic than for compositional cases, suggesting that idiomaticity may not be fully incorporated in the models.

Qualitative analysis: In Table 3, in P1, the similarity scores between NC in Table 1 and their respective NC_{syn} for BERT and GloVe models are shown. As expected, BERT shows higher scores than GloVe for all cases, and even if the values for P1 differ, both models follow the same tendency. There is a larger spread for GloVe (e.g., $\text{sim}^{(\text{P1})}_{\text{wet blanket}} = 0.21$ vs. $\text{sim}^{(\text{P1})}_{\text{ghost town}} = 0.80$) which could be explained by the choices of NC_{syn}. For *wet blanket* NC_{syn} = *loser*, which has probably a very dissimilar representation from both *wet* and *blanket*. On the other hand, *ghost town* with NC_{syn} = *abandoned town* not only shares a word with the original NC, but we can also argue that *ghost* and *abandoned* are likely to have similar embeddings. Finally, the average results of P1 show that BERT-based models tend to intensify lexical overlap, resulting in high cosine similarities when both the NC and NC_{syn} share (sub-)words. For instance, 47 (in English) and 49 (in Portuguese) out of the 50 compounds with highest $\text{sim}^{(\text{P1})}_{\text{NC-NAT}}$ share surface tokens, whether the NCs are more compositional (e.g., *music journalist* vs. *music reporter*) or more idiomatic (e.g., *ghost town* vs. *abandoned town*).

4.2 Can the lower semantic overlap between idiomatic NCs and their individual components be captured?

We would expect idiomatic NCs not to be similar to either of their individual components, which would be reflected by a larger spread of cosine similarity values for P2 than for P1. However, all models produced high similarities across the idiomaticity spectrum, see Figures 1c for Sent and 1d for NC.

[8] Neutral sentences are omitted since they all follow the same pattern *This is a/an* <NC>.

[9] GloVe-NC can be viewed as the baseline for the lack of contextual information.

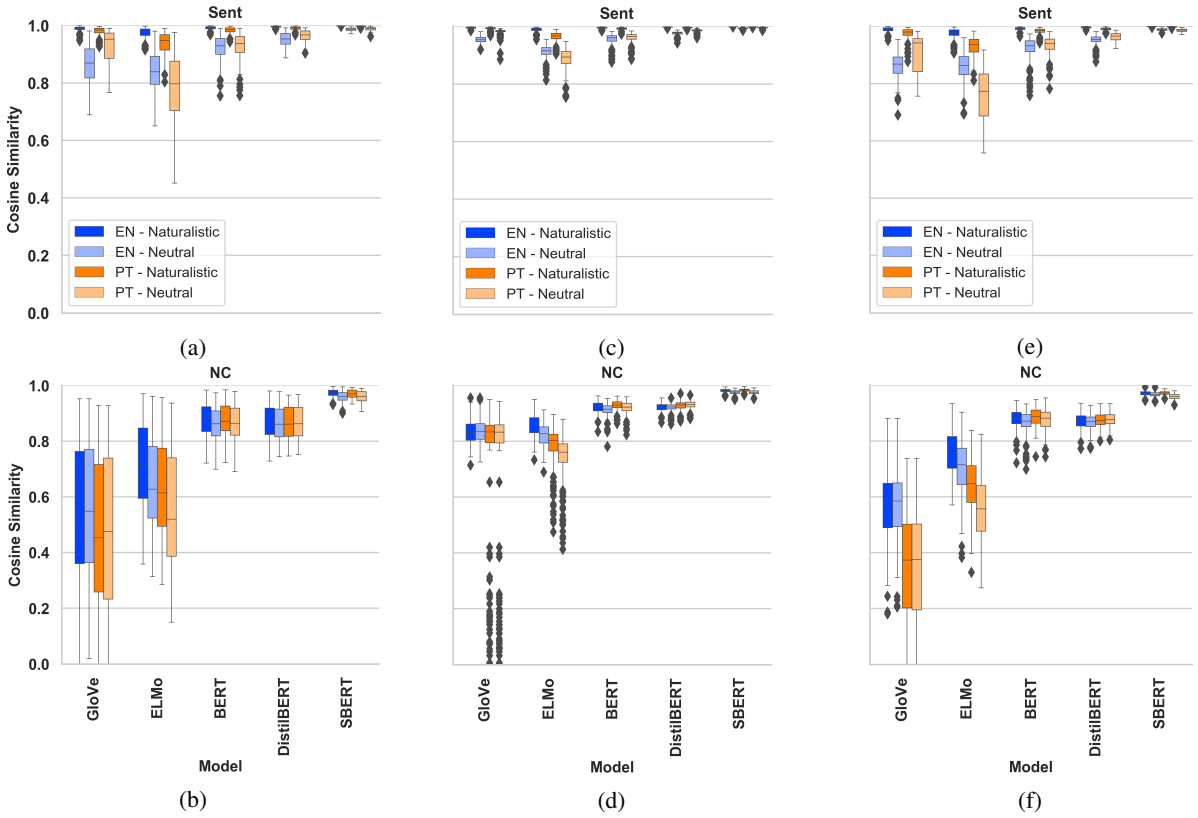

Figure 1: Cosine similarities in English (blue) and Portuguese (orange). First column for PI (a and b), second for P2 (c and d) and third for P3 (e and f). Sentence condition at the top and NC at the bottom.

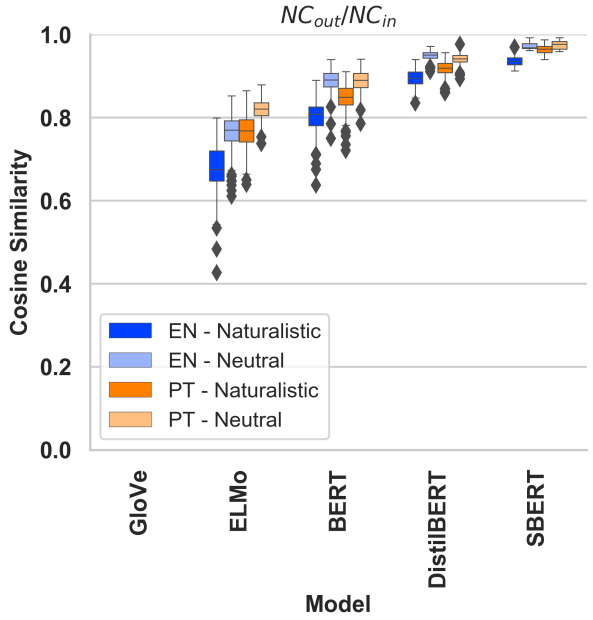

Figure 2: P4 ($\cos(\epsilon_{\text{NC}} \subset \text{s}, \epsilon_{\text{NC}})$).

The higher average similarities for P2 than for P1, compare Figures 1a and 1b with Figures 1c and 1d, reinforces the hypothesis that the models prioritise lexical overlap with one of the NC components rather than semantic overlap with a true NC synonym, even for idiomatic cases. Although there is some correlation with idiomaticity when it exists, it is lower than for P1, contrary to what would be expected (see P1 and P2 in Table 2). All of these indicate that these models cannot distinguish the partial semantic overlap between more compositional NCs and their components and the absence of overlap for idiomatic NCs.

Qualitative analysis: The P2 results in Table 3 show the highest similarity scores between each example in Table 1 and one of its components. These high similarity scores highlight the prioritisation of lexical over semantic overlap mentioned above. Furthermore, some idiomatic NCs also show strong similarities with their components, suggesting that the idiomatic meaning is not correctly represented. For instance, *poison pill* (meaning an emergency exit) has an average similarity of $\text{sim}^{(\text{P2})}_{\text{poison pill-NAT}} = 0.94$ with its head (*pill*).

4.3 Can they capture the lack of substitutability of individual components for idiomatic NCs?

We do not expect an idiomatic NC to keep the idiomatic meaning when each of its components is

	GloVe		ELMo		BERT		DistilB		SBERT		BERTRAM	
EN_{NAT}	ρSent	ρNC	ρ Sent	ρ NC	ρ Sent	ρ NC	ρ Sent	ρ NC	ρ Sent	ρ NC	ρ Sent	ρ NC
P1	0.31	0.62	0.43	0.60	0.51	0.67	0.38	0.58	0.30	0.43	0.14	0.30
P2	-	0.45	-	0.15	-	0.32	-	0.25	-	0.19	0.21	0.45
P3	-	0.18	-	-	-	0.21	-	0.15	-	0.20	0.18	0.39
EN_{NEU}												
P1	0.58	0.61	0.55	0.60	0.53	0.59	0.56	0.54	0.47	0.42	0.24	0.23
P2	0.29	0.44	-	0.22	-	-	-0.12	-	0.12	0.17	0.26	0.31
P3	-	0.18	-	-	-	-	-	-	0.17	0.19	0.32	0.26
PT_{NAT}												
P1	-	0.40	0.32	0.47	0.29	0.44	0.20	0.39	0.18	0.37	-	0.22
P2	-	0.20	-	0.28	-	-	-0.17	-	-	-	-	0.21
P3	-0.19	-	-	-	-	-	-	-	-	-	-	0.22
PT_{NEU}												
P1	0.22	0.41	0.37	0.47	0.30	0.35	0.31	0.37	0.30	0.36	-	0.18
P2	-	0.18	0.17	0.20	-	-	-	-	-	-	0.22	-
P3	-	-	-	-	-	-	-	-	-	-	0.22	0.18

Table 2: Spearman ρ correlation with human judgments, p$\leq$0.05. Non-significant results omitted from the table.

Noun Compound	P1			P2			P3		
	GloVe	BERT		GloVe	BERT		GloVe	BERT	
	NAT/NEU	NAT	NEU	NAT/NEU	NAT	NEU	NAT/NEU	NAT	NEU
field work	0.58	0.92	0.92	0.86(2)	0.94(2)	0.90(2)	0.54	0.90	0.88
ghost town	0.80	0.95	0.91	0.85(2)	0.93(2)	0.91(2)	0.66	0.90	0.84
close call	0.52	0.83	0.84	0.86(2)	0.94(2)	0.91(2)	0.61	0.86	0.84
eager beaver	0.43	0.82	0.83	0.84(2)	0.94(2)	0.92(2)	0.49	0.87	0.86
wet blanket	0.21	0.77	0.79	0.84(1)	0.94(2)	0.94(2)	0.69	0.91	0.90

Table 3: Similarities results from P1 to P3 at NC level of the examples in Table 1. In P2, number in parenthesis corresponds to the position of the w_i with highest similarity score in the NC.

individually replaced by synonyms, and this would be reflected in lower similarity values for P3 than for P1. However, high similarity values are found across the idiomaticity spectrum, and for all models and all conditions the average similarities are higher than those for P1 (see Figures 1e and 1f). Contrary to what would be expected, the correlations with idiomaticity scores are mostly nonexistent, and when they do exist they are much lower than for P1, (see P1 and P3 in Table 2).

The overall picture painted by P3 points towards contextualised models not being able to detect when a change in meaning takes place by the substitution of individual components by their synonyms.

Qualitative analysis: For P3, Table 3 shows the similarities scores at NC level between each NC and their NC_{synW} counterpart. Again, similarity scores for GloVe are considerably lower than for BERT. As expected for GloVe, $\text{sim}^{(P3)}_{\text{wet blanket}} = 0.69$ is noticeably higher than $\text{sim}^{(P1)}_{\text{wet blanket}} = 0.21$, since individually the words *damp* and *cloak* are closer in meaning to *wet* and *blanket*, respectively, than *loser* is. Another evidence that contextualised models are not modelling idiomaticity well is, for NAT cases, the considerably higher $\text{sim}^{(P3)}_{\text{wet blanket}} = 0.91$ in comparison to $\text{sim}^{(P1)}_{\text{wet blanket}} = 0.77$, for BERT. Although for the other NCs, $\text{sim}^{(P3)}_{\text{NC}}$ and $\text{sim}^{(P1)}_{\text{NC}}$ are comparable, the special case of the more idiomatic *wet blanket* highlights the issues of idiomaticity representation.

4.4 Is there a difference between an NC in and out of context?

For contextualised models, the greater the influence of the context, the lower we would expect the similarity to be between an NC in and out of context. However, especially for BERT models the results (Figure 2) show a high similarity between the NC in and out of context ($\text{sim}^{(P4)}_{\text{in-out}} > 0.8$). Moreover, a comparison with the similarities for the synonyms in P1 resulted in $\text{sim}^{(P4)}_{\text{in-out-NEU}} > \text{sim}^{(P1)}_{\text{NC-NEU}}$ and $\text{sim}^{(P4)}_{\text{in-out-NAT}} \simeq \text{sim}^{(P1)}_{\text{NC-NAT}}$, which indicates that these models consider the NC out of context to be a better approximation for the NC in context than its synonym. In addition, for BERT models $\text{sim}^{(P4)}_{\text{in-out}}$ is only weakly correlated with the idiomaticity score (Table 4), which suggests that the context may not play a bigger role for idiomatic than it does for more compositional NCs.

Qualitative analysis: The $\text{sim}^{(P4)}_{\text{in-out}}$ of the examples in Table 1 ranged from 0.78 (for *ghost town*) to 0.87 (*field work*) in the NAT condition, and from 0.84

	ELMo	BERT	DistilB	SBERT	BTRAM
EN_{NAT}	-	-	0.14	-0.16	0.14
EN_{NEU}	-	-	0.24	-0.24	-0.14
PT_{NAT}	0.25	0.17	0.18	-	0.21
PT_{NEU}	-	-	0.15	-	-

Table 4: Spearman ρ correlation with human judgments for P4, p$\leq$0.05. Non-significant results are omitted.

(also for *ghost town*) to 0.90 (*eager beaver* and *wet blanket*) in the neutral sentences for BERT.[10] Together with these examples, the general results of P4 show large differences not explained by the semantic compositionality of the NCs, as suggested by the weak correlation with the idiomaticity scores. In this respect, both the largest and smallest differences between $\text{sim}^{(P4)}_{\text{in-out}}$ in NAT and NEU conditions appear in compositional NCs (*engine room* with $\text{sim}^{(P4)}_{\text{in-out-NAT}} = 0.68$, $\text{sim}^{(P4)}_{\text{in-out-NEU}} = 0.89$, and *rice paper* with $\text{sim}^{(P4)}_{\text{in-out-NAT}} = 0.84$, $\text{sim}^{(P4)}_{\text{in-out-NEU}} = 0.86$).

Besides, we expected ambiguous compounds such as *bad apple* or *bad hat* to have large $\text{sim}^{(P4)}_{\text{in-out}}$ differences between both conditions, as they occur with an idiomatic meaning in the NAT sentences. However, the differences were of just 0.06 in both cases, while other less ambiguous idiomatic NCs showed higher variations (e.g., *melting pot*, with 0.16). In sum, the results of P4 suggest that contextualised models do not properly represent some NCs.

4.5 But how informative are the contexts?

As the neutral sentences do not provide informative contextual clues, if the NCs in NAT and NEU conditions are similar, this would provide an additional indication that for these models contexts are not playing an important role in distinguishing usages (in this case between a neutral and uninformative usage and a naturalistic one). Indeed, the two conditions follow the same trends in the two languages, see Figure 1. Furthermore, there are significant correlations between NAT and NEU conditions, and some are very strong correlations. For example, for SBERT the correlations between the NC in context in naturalistic and neutral conditions are $\rho^{(P1,P2,P3)}_{\text{NC(Nat/Neu)}} > 0.85$ for English and > 0.76 for Portuguese, for probes P1, P2 and P3. This indicates that to evaluate the effect of the variants in each of these probes, a neutral sentence is as good as a naturalistic one. This reinforces the possibility that these models do not adequately incorporate the context in a way that captures idiomaticity.

In terms of the similarity between a sentence and its variants, as we assumed that the representation of a sentence corresponds to the average of the individual components, sentence length may have a strong impact on cosine similarity. This would explain the high values obtained for sentence similarities throughout the probes, as they could be more the effect of the number of words in a sentence than of their semantic similarity. Indeed, the correlation between naturalistic sentence length and the cosine similarities for the first three probes is moderate to strong for all models (Table 5), and higher for some of the contextualised models than for the baseline (e.g., DistilB in English and P2).

EN	GloVe	ELMo	BERT	DistilB	SBERT
P1	0.71	0.47	0.52	0.66	0.67
P2	0.87	0.79	0.78	0.89	0.84
P3	0.88	0.71	0.80	0.87	0.77
PT					
P1	0.60	0.46	0.61	0.68	0.62
P2	0.80	0.68	0.72	0.84	0.75
P3	0.69	0.58	0.64	0.76	0.75

Table 5: Spearman ρ correlation between naturalistic sentence length and cosine similarity, p $\leq$ 0.001.

4.6 Other Operations

As referred in section 3.3 we have used vector averaging to obtain the NC embedding, as it is the standard procedure to represent not only MWEs but also out-of-vocabulary words, which are split into sub-tokens in contextualised models (Nandakumar et al., 2019; Wiedemann et al., 2019). However, we have also explored other methods to represent NCs in a single vector.

First, we have incorporated type-level vectors of the NCs into a BERT model, inspired by compositionality prediction methods (Baldwin et al., 2003; Cordeiro et al., 2019). To do so, we annotated the target NCs in large English and Portuguese corpora (Baroni et al., 2009; Wagner Filho et al., 2018) and used attentive mimicking with one-token-approximation (Schick and Schütze, 2019, 2020b) to learn up to 500 contexts for each NC. These new vectors encode each NC in a single representation, therefore avoiding possible biases produced by the compositional operations. Then, we used BERTRAM (Schick and Schütze, 2020a) to inject these type-level vectors in the BERT multilingual model. As expected, learning the vectors

[10] For Glove, $\text{sim}^{(P4)}_{\text{in-out}} = 1$.

of the NCs as single tokens improved the representation of idiomatic expressions (see BERTRAM in Tables 2 and 4), decreasing the correlation with idiomaticity in P1 (e.g., $\rho^{(P1)}_{\text{NC-NAT}} = 0.30$ in English), and increasing it in P2 ($\rho^{(P2)}_{\text{NC-NAT}} = 0.45$) and P3 ($\rho^{(P3)}_{\text{NC-NAT}} = 0.39 > \rho^{(P1)}_{\text{NC-NAT}}$). For P4, the correlation also increased in NAT contexts. In sum, these results were in general better and more statistically significant (at the expense of re-training a model).

Second, we compared the performance of averaging vs. concatenating the vectors of the NC subwords. In this case, we selected those utterances in English including NCs with the same number of sub-words of their synonyms (273 sentences), thus allowing for vector concatenation. Using this operation instead of average slightly improved the results of the BERT-based models (e.g., ≈0.06 higher correlations on average for P3 NAT) and obtained more significant values.

As the latter approach does not involve retraining a model, in further work we plan to probe other concatenation and pooling methods able to compare MWEs with different number of input vectors (e.g., *grey matter* vs. *brain*) which have achieved good results in sentence embeddings (Rücklé et al., 2018).

5 Conclusions

This paper presented probing tasks for assessing the ability of vector space models to retain the idiomatic meaning of NCs in the presence of lexical substitutions and different contexts. For these evaluations, we constructed the NCS dataset, with a total of 9,220 sentences in English and Portuguese, including variants with synonyms of the NC and of each of its components, in neutral and naturalistic sentences. The probing tasks revealed that contextualised models may not detect that idiomatic NCs have a lower degree of substitutability of the individual components when compared to more compositional NCs. This behaviour is similar in the controlled neutral and naturalistic conditions both in English and Portuguese.

The next steps are to extend the probing strategy with additional measures that go beyond similarities and correlations. Moreover, for ambiguous NCs, we intend to add probes for the different senses. Finally, we also plan to apply them to more languages, examining how multilingual information can be used to refine the representation of noun compounds and other MWEs.

Acknowledgments

Aline Villavicencio and Carolina Scarton are funded by the EPSRC project MIA: Modeling Idiomaticity in Human and Artificial Language Processing (EP/T02450X/1). Marcos Garcia is funded by the *Consellería de Cultura, Educación e Ordenación Universitaria* of the Galician Government (ERDF 2014-2020: Call ED431G 2019/04), and by a *Ramón y Cajal* grant (RYC2019-028473-I).

References

Laura Aina, Kristina Gulordava, and Gemma Boleda. 2019. Putting words in context: LSTM language models and lexical ambiguity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3348, Florence, Italy. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva, and Anderson da Silva Soares. 2018. Portuguese Named Entity Recognition Using LSTM-CRF. In *Proceedings of the 13th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2018)*, pages 83–92, Canela–RS, Brazil. Springer, Cham.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. 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 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* 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.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

David Alan Cruse. 1986. *Lexical semantics*. Cambridge university press.

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.

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.

Meghdad Farahmand and James Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 61–66, Berlin, Germany. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

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.

Rachel Giora. 1999. On the priority of salient meanings: Studies of literal and figurative language. *Journal of pragmatics*, 31(7):919–929.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv preprint arXiv:1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

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.

Milton King and Paul Cook. 2017. Supervised and unsupervised approaches to measuring usage similarity. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 47–52, Valencia, Spain. Association for Computational Linguistics.

Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of english verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 345–350. Association for Computational Linguistics.

Jean-Paul Laurent, Guy Denhières, Christine Passerieux, Galina Iakimova, and Marie-Christine Hardy-Baylé. 2006. On understanding idiomatic language: The salience hypothesis assessed by ERPs. *Brain Research*, 1068(1):151–160.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

James H. Neely, Dennis E. Keefe, and Kent L. Ross. 1989. Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(6):1003–1019.

Jeffrey Pennington, Richard Socher, and Christopher 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, Doha, Qatar. Association for Computational Linguistics.

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.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 210–218. The Association for Computer Linguistics.

Nils Reimers and Iryna Gurevych. 2019a. Alternative Weighting Schemes for ELMo Embeddings. *CoRR*, abs/1904.02954.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Joost Rommers, Ton Dijkstra, and Marcel Bastiaansen. 2013. Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, 25(5):762–776.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 1–15, Mexico City, Mexico. Springer, Berlin, Heidelberg.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019*, Vancouver, Canada.

Timo Schick and Hinrich Schütze. 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. 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 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020a. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020b. Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8766–8774.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Weijia Shi, Muhao Chen, Pei Zhou, and Kai-Wei Chang. 2019. Retrofitting contextualized word embeddings with paraphrases. 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 1198–1203, Hong Kong, China. Association for Computational Linguistics.

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.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)*, New Orleans, Louisiana.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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 the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Appendices

A Naturalistic examples in English

Table 6 includes naturalistic examples in English. We include the compositionality scores provided by the annotators and the BERT and GloVe results at NC level.

B Naturalistic examples in Portuguese

Table 7 includes naturalistic examples in Portuguese. We include the compositionality scores provided by the annotators and the BERT and GloVe results at NC level.

Table 6: Naturalistic examples in English, including human compositionality (HC). Results for BERT model at NC level ($\mathbf{sim}^{(P)}_{NC}$) together with GloVe results for the same measure. Average (Avg) values were calculated with BERT using the three NAT/NEU sentences for the same NC.

Probe	**Sentence (original compound in bold)**	**HC**	**Probe-specific results**		
			$\mathbf{NC}_{syn}$	$\mathbf{sim}^{(P1)}_{NC}$	**GloVe**
P1	He later became a **music journalist** covering the new wave and punk rock explosion [. . .]	4.54	music reporter	0.98	0.89
	The UN also held a **world conference** on Human Rights in Vienna.	3.96	global meeting	0.97	0.77
	The 3 month limit though is not a **brick wall**, if circumstances demand an extension of time [. . .]	3.79	obstacle	0.64	0.31
P2			$\mathbf{W}_i$	$\mathbf{sim}^{(P2)}_{NC}$	**GloVe**
	Allowing young people to opt out of the basic state pension is giving them a **poison pill**.	0.96	pill	0.92	0.85
	Arguably the king of comedy for the last ten years, Jim Carrey is **box office** gold.	0.88	box	0.81	0.81
P3			$\mathbf{NC}_{synW}$	$\mathbf{sim}^{(P3)}_{NC}$	**GloVe**
	It is not right that criminal enterprises try to use **dirty money** with a clean face.	2.21	smotty cash	0.93	0.63
	Formal evenings require a suit or dinner jacket for men and a **cocktail dress** for ladies.	3.04	appetizer costume	0.92	0.65
	If you burn coal without any kind of pollution control you get large amounts of ash and sulphur (and **radioactive waste** from natural Uranium decay in the coal).	4.58	dangerous rubbish	0.84	0.54
P4			$\mathbf{sim}^{(P4)}_{in\text{-}out}$	$\mathbf{Avg\ sim}^{(P4)}_{in\text{-}out\text{-}NAT}$	$\mathbf{Avg\ sim}^{(P4)}_{in\text{-}out\text{-}NEU}$
	The roll-on/roll-off nuclear cargo ferry Atlantic Osprey suffered an **engine room** fire on Monday.	4.93	0.66	0.68	0.89
	And we had to explain to her the difference between **rice paper** and ordinary paper.	4	0.83	0.84	0.86
	However, it will not work unless every single person does it, because one **bad apple** ruins the whole barrel.	1.13	0.82	0.83	0.89
	The jury heard the evidence presented, that he was general **bad hat**.	0.62	0.76	0.76	0.83
	Yet its heyday was down with the epochal **melting pot** of punk/funk/art/jazz/dub [. . .]	0.54	0.73	0.73	0.89

Table 7: Naturalistic examples in Portuguese, including human compositionality (HC). Results for BERT model at NC level ($\mathbf{sim}^{(P)}_{NC}$) together with GloVe results for the same measure. Average (Avg) values were calculated with BERT using the three NAT/NEU sentences for the same NC. English translations are in italic, together with the literal translation of the compounds where they are not word-to-word equivalents.

Probe	**Sentence (original compound in bold)**	**HC**	**Probe-specific results**		
			$\mathbf{NC}_{syn}$	$\mathbf{sim}^{(P1)}_{NC}$	**GloVe**
P1	Normalmente, os restaurantes encontram-se dentro de **centros comerciais**. *Restaurants are usually located inside **shopping malls** (lit. comercial centres).*	3.68	shoppings	0.94	0.45
	Foi um dia pesaroso, um **sexto sentido** me alertava que uma coisa ruim puxa outra. *It was a sorrowful day, a **sixth sense** alerted me that one bad thing pulls another.*	1.4	intuição *intuition*	0.79	0.11
	Existe mesmo no serviço secreto inglês um **agente secreto** com licença para matar! *There really is in the English secret service a **secret agent** licensed to kill!*	4.58	espião *spy*	0.81	0.56
P2			$\mathbf{W}_i$	$\mathbf{sim}^{(P2)}_{NC}$	**GloVe**
	Alguns dos estádios novos foram criticados por se tornarem "**elefantes brancos**" após a Copa. *Some of the new stadiums were criticized for becoming "**boondoggles**" (lit. white elephants) after the World Cup.*	0.16	elefantes *elephants*	0.96	0.81
	As espécies de **mar aberto** têm por princípio a natação contínua. *The **open sea** species have as a principle the continuous swimming.*	4.03	aberto *open*	0.9	0.79
P3			$\mathbf{NC}_{synW}$	$\mathbf{sim}^{(P3)}_{NC}$	**GloVe**
	Foices e facões são **armas brancas** de uso corriqueiro. *Scythes and machetes are commonplace **white weapons**.*	0.65	pistolas alvas *untanned guns*	0.92	0.50
	Não deu quase ninguém, só alguns **gatos-pingados**! *There's hardly anybody, just a **few people** (lit. dripping cats)!*	0	felinos chuviscados *drizzled felines*	0.92	0.01
P4			$\mathbf{sim}^{(P4)}_{in\text{-}out}$	$\mathbf{Avg\ sim}^{(P4)}_{in\text{-}out\text{-}NAT}$	$\mathbf{Avg\ sim}^{(P4)}_{in\text{-}out\text{-}NEU}$
	Troque o **leite integral** pelo desnatado e economize nas calorias. *Replace **whole milk** (lit. integral milk) with skimmed milk and save on calories.*	4.67	0.86	0.84	0.79
	Ganhou até uma fama de "**pé-frio**", por ter alguns rebaixamentos em seu currículo [...] *He even gained a reputation as an "**unlucky person**" (lit. cold foot), for having some downgrades in his resume [...]*	0.09	0.84	0.89	0.87
	Para muitos povos antigos, um novo mês era anunciado na passagem da **lua nova** para a lua crescente. *For many ancient peoples, a new month was announced as the **new moon** passed into the crescent moon.*	1.4	0.77	0.74	0.84
	Esse diagnostico é realizado através de **exame clínico** e radiográfico. *This diagnosis is made through **clinical** and radiographic **examination**.*	4.75	0.84	0.84	0.92