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.

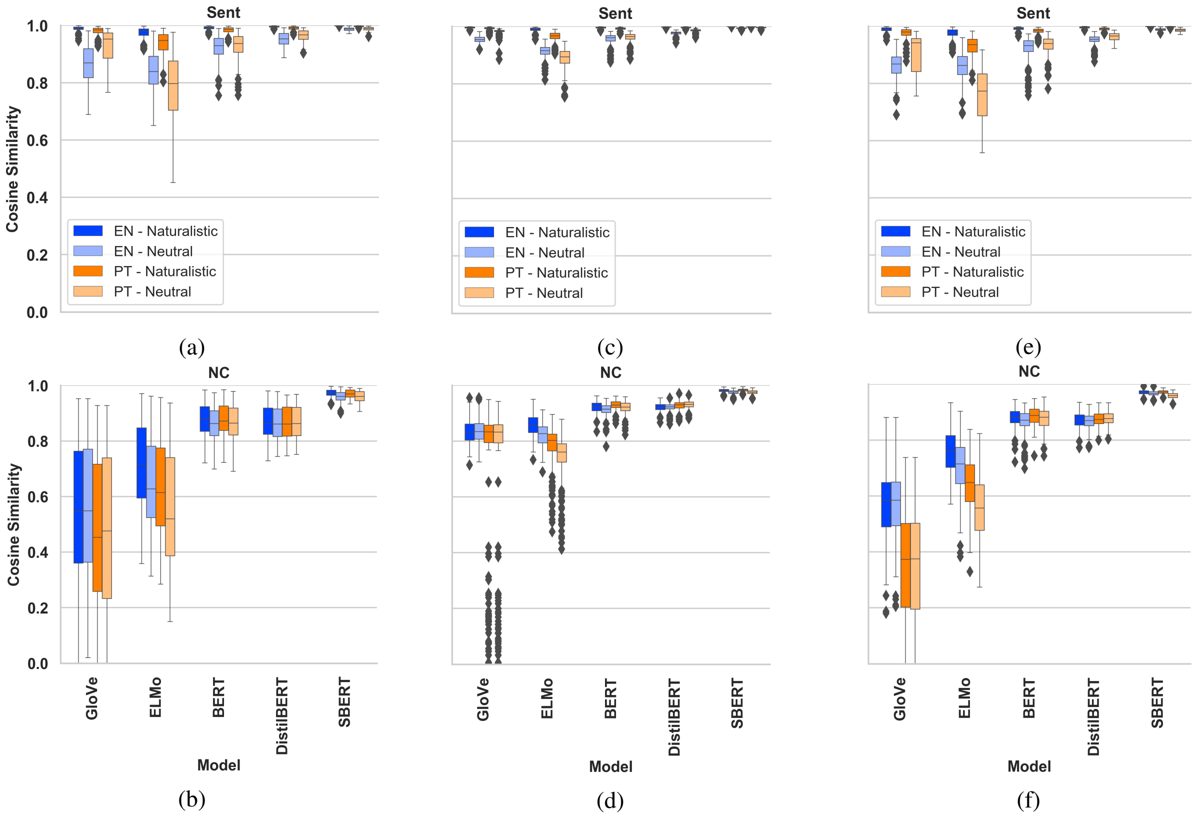
Qualitative analysis: In Table 3, in P1, the similarity scores between NC in Table 1 and their respective NC_{syn} for BERT and GloVe models are shown. As expected, BERT shows higher scores than GloVe for all cases, and even if the values for P1 differ, both models follow the same tendency. There is 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.



(a)

(c)

(e)

(b)

(d)

(f)

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.

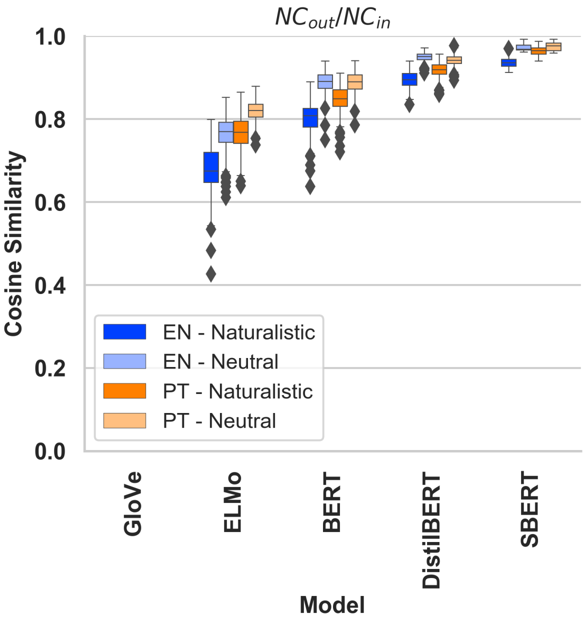


Figure 2: P4 ($\cos(\epsilon_{\text{NC} \subset \text{S}}, \epsilon_{\text{NC}})$).

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