	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