Assessing the Representations of Idiomaticity in Vector Models with a Noun Compound Dataset Labeled at Type and Token Levels

Marcos Garcia[*1], **Tiago Kramer Vieira**[*2],
Carolina Scarton[3], **Marco A. P. Idiart**[4], **Aline Villavicencio**[2,3]
[1] CiTIUS Research Centre, Universidade de Santiago de Compostela, Galiza (Spain)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3] Department of Computer Science, University of Sheffield (UK)
[4] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
marcos.garcia.gonzalez@udc.gal, tiagokv@hotmail.com,
{c.scarton, a.villavicencio}@sheffield.ac.uk,
marco.idiart@gmail.com

Abstract

Accurate assessment of the ability of embedding models to capture idiomaticity may require evaluation at token rather than type level, to account for degrees of idiomaticity and possible ambiguity between literal and idiomatic usages. However, most existing resources with annotation of idiomaticity include ratings only at type level. This paper presents the Noun Compound Type and Token Idiomaticity (NCTTI) dataset, with human annotations for 280 noun compounds in English and 180 in Portuguese at both type and token level. We compiled 8,725 and 5,091 token level annotations for English and Portuguese, respectively, which are strongly correlated with the corresponding scores obtained at type level. The NCTTI dataset is used to explore how vector space models reflect the variability of idiomaticity across sentences. Several experiments using state-of-the-art contextualised models suggest that their representations are not capturing the noun compounds idiomaticity as human annotators. This new multilingual resource also contains suggestions for paraphrases of the noun compounds both at type and token levels, with uses for lexical substitution or disambiguation in context.

1 Introduction

Multiword Expressions (MWEs) such as noun compounds (NCs), have been considered a challenge for NLP (Sag et al., 2002). This is partly due to the wide range of idiomaticity that they display, from more literal to idiomatic combinations (*olive oil* vs. *shrinking violet*). The task of identifying the degree of idiomaticity of MWEs has been investigated at type level, to determine the potential of an MWE to be idiomatic in general. Some of these approaches are based on the assumption that the distance between the representation of an MWE as a unit and the representation of the compositional combination of its components is an indication of the degree of idiomaticity: they are closer if the MWE is more compositional. Good performances are obtained even with non-contextualised word embeddings like *word2vec* (Mikolov et al., 2013), and vector operations like addition and multiplication (Mitchell and Lapata, 2010; Reddy et al., 2011; Cordeiro et al., 2019). Additionally, for some MWEs, there is a potential ambiguity between an idiomatic and a literal sense, like in the potentially idiomatic MWE *brass ring* which can be ambiguous between the more literal meaning a *ring made of brass* and the more idiomatic sense of a *prize*. Considering that these MWEs can have both idiomatic and literal senses, a related task of token-level identification evaluates whether in a particular context an MWE is idiomatic or not. For this task, models that incorporate the context in which an MWE occurs tend to be better equipped to distinguish idiomatic from literal occurrences (Sporleder and Li, 2009; King and Cook, 2018; Salton et al., 2016).

Contextualised embedding models, like BERT (Devlin et al., 2019), brought significant advances to a variety of downstream tasks (e.g. Zhu et al. (2020) for machine translation and Jiang and de Marneffe (2019) for natural language inference). They also seem to benefit tasks like idiomaticity and metaphor identification (Gao et al., 2018), since their interpretation is often dependent on contextual clues. Nonetheless, previous work found that non-contextualised models seem to still bring informative clues for these tasks (King and Cook, 2018), and their combination with contextualised models could improve results (e.g. for metaphor identification (Mao et al., 2019)). This complementarity between non-contextualised and contex-

[*] Equal contribution.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2730–2741
August 1–6, 2021. ©2021 Association for Computational Linguistics

tualised models may be an indication that enough core idiomatic information may already be available at type level. Moreover, type-based compositionality prediction measures that perform well with static embeddings may also perform well for token-based prediction with contextualised models.

To address these questions, in this paper, we present the Noun Compound Type and Token Idiomaticity (NCTTI) dataset, containing 280 NCs in English and 180 in Portuguese, annotated with the degree of idiomaticity perceived by human annotators, at type and token level.[1] NCTTI contains a total of 8,725 annotations in 840 different sentences in English, and 5,091 annotations in 540 sentences in Portuguese. Moreover, NCTTI has several paraphrases for each NC which are classified as either type level or token level equivalents. To control for the level of idiomaticity, the NCTTI dataset has a balanced amount of compositional, partly compositional and idiomatic items. As the importance of context to determine interpretation may be related to factors like the degree of idiomaticity, association strength or the frequency of an NC, we present an illustrative analysis of their impact for the performance of different models in capturing idiomaticity. We also examine how the performance obtained for human idiomaticity judgments per type differs from the performance obtained per token.

Our contributions can be summarised as: (1) building the NCTTI dataset with information about type and token idiomaticity for NCs in two languages, (2) evaluating to what extent models are able to detect idiomaticity at type and token level, analysing different levels of contextualisation and (3) proposing two new measures of idiomaticity. Moreover, the paraphrases provided for each NC at type and token level make NCTTI a useful resource for enhancing paraphrase datasets (e.g. PPDB (Ganitkevitch et al., 2013)), for tasks involving lexical substitution (McCarthy and Navigli, 2007; Mihalcea et al., 2010), or for improving the results of downstream tasks, such as text simplification (Paetzold, 2016; Alva-Manchego et al., 2020). Such paraphrases may also be useful for improving the task of machine translation, avoiding the need for parallel MWE corpora (Zaninello and Birch, 2020).

Section 2 gives an overview of existing idiomaticity datasets. Section 3 presents the NCTTI dataset and the annotations, and section 4 discusses the evaluation of the performance of different word embeddings in detecting idiomaticity.

2 Related Work

Datasets with type-level annotations are available for NCs in English (Farahmand et al., 2015; Reddy et al., 2011; Ramisch et al., 2016; Kruszewski and Baroni, 2014), German (Roller et al., 2013; Schulte im Walde et al., 2016), French (Cordeiro et al., 2019) and Portuguese (Cordeiro et al., 2019). However, datasets with idiomatic information at token level are scarce, e.g., the VNC-Tokens (Cook et al., 2008), containing almost 3k annotations for 53 Verb-Noun Combinations in English.

Regarding the use of contextualised embeddings to model idiomaticity, Nandakumar et al. (2019) compared different static and contextualised embeddings to predict the NCs compositionality, obtaining better results with static vectors learnt individually for each NC. Shwartz and Dagan (2019) train various classifiers initialised with static and contextualised embeddings for different compositional tasks, achieving the best results with BERT embeddings. Yu and Ettinger (2020), using partially idiomatic expressions of the BiRD dataset (Asaadi et al., 2019), show that contextualised embeddings from language models heavily rely on word content, missing additional information provided by compositional operations.

In this paper we take advantage of the NCTTI dataset to observe whether vector representations obtained with different strategies correlate with human annotations at both type and token levels.

3 The Noun Compound Type and Token Idiomaticity dataset

This section describes the procedure to create the NCTTI dataset and its main characteristics.[2]

3.1 Source data

We used as basis the English and Portuguese subsets of the NC Compositionality dataset (Cordeiro et al., 2019), which contain compositionality scores for 280 two-word NCs in English (90 of which came from Reddy et al. (2011)), and 180 in Portuguese, all of them labeled at type level: i.e., the annotators provided a compositionality value for a compound (from 0 –fully idiomatic– to 5, fully

[1] Type level annotations come from Cordeiro et al. (2019), the dataset used as source for the NCTTI.

[2] The NCCTI dataset can be downloaded from the following url: `https://github.com/marcospln/nctti`.

compositional) after reading various sentences with this NC.

To obtain more fine-grained compatible token-level annotations about the impact of different contexts in the interpretation of NCs, we used the same original sentences as in the source dataset (three sentences per compound with the same sense were selected from Reddy et al. (2011) dataset).[3]

Language experts classified each noun compound regarding their semantic compositionality as idiomatic (e.g., *gravy train*), partially idiomatic (e.g., *grandfather clock*), or compositional (e.g., *research project*). For English, this resulted in 103, 88, and 89 idiomatic, partially idiomatic, and compositional compounds. For Portuguese, each class has 60 compounds, as the selection had been balanced when the source dataset was created.

3.2 Annotation procedure

We used the same protocol as Reddy et al. (2011) and Cordeiro et al. (2019), asking each participant to give 0 to 5 scores for an NC and its components in a specific sentence (e.g., *glass ceiling* in "Women are continuing to slowly break through the *glass ceiling* of UK business [. . .]"). In particular, we asked participants for: (i) the contribution of the head to the meaning of the NC (e.g., is a *glass ceiling* literally a *ceiling*?); (ii) the contribution of the modifier to the meaning of the NC (e.g., is a *glass ceiling* literally of *glass*?); and (iii) the degree of compositionality of the compound (i.e., to what extent the meaning of the NC can be seen as a combination of its parts). Additionally, we asked for up to three synonyms of the NC in that particular sentence (e.g., synonyms at token level).

We used Amazon Mechanical Turk to obtain the annotations for English, and a dedicated online platform for the questionnaire in Portuguese,[4] as we could not find a suitable number of annotators for this language in AMT.[5] Taking this into account, the numbers of the Portuguese annotations are in general lower to those obtained for English.

For each language, we have included the three sentences of every compound in the dataset (840 sentences in English, and 540 in Portuguese), which were randomly submitted to the annotators. For English, we compiled at least 10 annotations per sentence, resulting in 8,725 annotations (10.4 annotations per sentence on average). A total of 412 annotators have taken part in the process, and on average, each participant labeled 21 instances. For Portuguese we set the threshold in 5 annotations per sentence: we got 5,091 annotations by 33 participants, so that each sentence has a mean of 9.4 annotations and each annotator labeled on average 154 sentences.

3.3 Results

Inter-annotator agreement: we computed the inter-annotator agreements for two and three annotators with the largest number of sentences in common (Table 1). For English, we obtained Krippendorff's α (Krippendorff, 2011) values of 0.30 for two annotators (199 sentences) and 0.22 for three annotators (76 sentences). The α values for Portuguese were of 0.52 for two annotators (131 sentences) and 0.44 for three annotators (60 sentences). Overall, and using the divisions proposed by Landis and Koch (1977), the agreement results can be classified as 'fair' (for English), and 'moderate' (for Portuguese).

Data	**English**		**Portuguese**	
	2	3	2	3
NC	0.30	0.22	0.52	0.44
Head	0.33	0.38	0.66	0.53
Modifier	0.45	0.42	0.56	0.48

Table 1: Krippendorff's α inter-annotator agreement for the NC, head, and modifiers for 2 and 3 annotators.

Data	**English**	**Portuguese**
All	0.92	0.90
Idiomatic	0.71	0.82
Partial	0.78	0.78
Compositional	0.66	0.91

Table 2: Spearman ρ correlations between the average compositionality values per compound of the NCTTI, and the original scores of the NC Compositionality dataset ($p < 0.01$ in all cases). *All* values were calculated with the all compounds for each language, while *Idiomatic*, *Partial*, and *Compositional* were computed on the three compositionality levels.

Correlation token vs. type scores: then, we calculated the correlations (Spearman ρ) between the average compositionality scores of the NCTTI

[3] Some contexts are spans of tokens instead of sentences, but usually enough to interpret the meaning of the NC.

[4] The platform was provided by Cordeiro et al. (2019).

[5] The annotation process was approved by the Ethics Committee of the University of Sheffield. This is a thorough evaluation process peer-reviewed by three ethical reviewers. The monetary compensation was deemed appropriate for the task.

Data	Noun Compound				Head				Modifier			
	English		**Portuguese**		**English**		**Portuguese**		**English**		**Portuguese**	
	Mean	*StD*	*Mean*	*StD*	*Mean*	*StD*	*Mean*	*StD*	*Mean*	*StD*	*Mean*	*StD*
Idiom.	0.95	0.58	1.52	0.81	1.53	1.37	1.83	1.07	1.69	1.29	2.02	1.18
Partial	2.34	1.01	2.46	0.91	3.34	1.41	3.65	1.03	2.75	1.26	2.67	1.15
Comp.	4.13	0.67	3.61	0.94	4.23	0.66	4.20	0.93	4.34	0.66	3.90	0.87

Table 3: Mean compositionality scores for each class in English and Portuguese (from 0, fully idiomatic, to 5, fully compositional), and standard deviations. Left columns contain the scores for the whole compound, while the values for the head and modifier are in the middle and right columns, respectively. The type averages for the NCs reported by Cordeiro et al. (2019) are 1.1, 2.4, and 4.2 for English and 1.3, 2.5, and 3.9 for Portuguese.

dataset and those of the original resource (NC Compositionality dataset). Table 2 contains the correlation results for each language and compositionality class. The strong to very strong significant correlations confirm the robustness between type-level and token-level human compositionality annotations for these two datasets.[6]

Idiomaticity values: with regards to the idiomaticity values of each class, Table 3 displays both the average scores and the standard deviation in both languages. As expected, for the whole compounds, partially idiomatic NCs are those with higher standard deviations, and their mean compositionality values are in the middle of the scale (2.34 and 2.46). In English, the results of both idiomatic and compositional compounds are more homogeneous, as they are clearly located on the margins of the scale (< 1 and > 4, respectively) with lower deviations. This is not the case in Portuguese, where the average values are > 1 and < 4 for idiomatic and compositional NCs, respectively, placing even the idiomatic cases closer towards the middle of the scale. With respect to the average values for the heads and modifiers, we can highlight the following observations: first, both head and modifier scores are consistently higher than the means for the whole compound in every scenario also suggesting at least a partial compositionality in their token occurrences. Second, for idiomatic NCs, the scores of the modifiers are higher than those of the heads, while for partially compositional NCs the results are the opposite.[7] Finally, regarding the compositional level, the modifier values are higher in English, while in Portuguese the heads seem to contribute more to the meaning of the NC.

Observing the variability across the annotations, we found some divergence in a few compounds (e.g., *brass ring* labeled as idiomatic for a compositional occurrence "Three drawers, each with a *brass ring* pull, provide plenty of storage whatever you use it for."), which hints at possible interference from a salient meaning (Giora, 1999). However, further investigation is needed.

Paraphrases: as mentioned, we asked the participants to provide synonyms or paraphrases for the noun compounds in each particular context. In this respect, it is worth noting that while some suggestions may be applicable across all the sentences for an NC (e.g. *spun sugar* for *cotton candy*, considered as a type level synonym), others are more dependent on context and differ for specific sentences (e.g. *flight recorder* and *unknown process*, for *black box*, which can be considered as token level paraphrases). We have classified the paraphrases as type or token level using the following procedure: to organise the large set of paraphrases provided by the annotators (see below), we performed an automatic classification as follows: we labeled as type level synonyms those paraphrases proposed for the three sentences of each compound, and those suggested for two sentences with a frequency $>= 3$; token level synonyms are those proposed only for one sentence with a frequency $>= 2$.

In English, 9,690 different paraphrases were proposed by the annotators (average 34.60 per NC), and 3,554 were suggested by at least 5 participants (average of 12.70 per NC). Out of them, 1,506 were classified as type level (5.4 synonyms per NC, on average), and 353 at token level (0.42 per sentence, 1.3 per NC). Overall, 118 NCs have token level synonyms for one sentence, 69 for two sentences, and 16 for the three sentences.

For Portuguese, the annotators suggested a total of 6,579 paraphrases (314 by at least 5 participants

[6] Removing annotators with low agreement (Spearman $\rho < 0.2$, and $\rho < 0.4$) resulted in almost identical correlations.

[7] The results for partially idiomatic compounds are expected to some extent as the head tends to bear more semantic load about the whole expression (e.g., as in collocations).

Sentence	*Mean*	*Paraphrase*
Keri enjoys music and has turned into a skilled *disc jockey*.	1.2	record player
Quality wedding *disc jockey* equipment comes at a cost.	2.5	broadcaster
Let one of our high energy *disc jockeys* entertain your next party.	1.7	announcer
Idiomaticity score at the type-level: 1.25. Most common (type-level) paraphrase: *DJ*.		

Table 4: Annotation example of the English NC *disc jockey*. Each row includes a sentence with the target NC together with the mean idiomaticity score and a token-level paraphrase. Bottom row shows the most common (type-level) paraphrase and the mean idiomaticity score from the original dataset (also at the type-level).

and 764 by $>= 3$, average of 4.2 per NC). 743 synonyms were proposed for the 180 compounds (an average of 4.1 per NC), being classified as type level. Concerning token level synonyms, we have collected 192 synonyms (1.1 per NC, on average). In this case the total number of annotations was lower, and the final resource contains 61 NCs with token level synonyms for one sentence, 38 for two sentences, and 6 compounds have token level synonyms for the three sentences.

The collection of paraphrases included in the NCTTI make this dataset a valuable resource for different evaluations, such as lexical substitution tasks and assessments of the performance of embedding models to correctly identify contextualised synonyms of NCs with different degrees of idiomaticity.

Table 4 shows an annotation example for the NC *disc jockey*, in English. It includes the three sentences together with the average idiomaticity score and both token-level and type-level paraphrases.

4 Experiments

This section displays some of the comparative analyses for the relevance of type and token annotation for idiomaticity detection. First, we adapt the type level compositionality prediction approaches used on static word vectors (Mitchell and Lapata, 2010) to contextualised models (Nandakumar et al., 2019), here computing the correlation also at token level. In particular, the assumption is that compositionality can be approximated as the distance between the representation for an NC and the representation for the compositional combination of its individual components. Then, we measure whether the vector representations reflect the variability of the human annotators, who capture different nuances of the NCs depending on the sentences in which they occur. Similarly, in a third experiment we use the standard deviations of the idiomaticity scores in the three contexts to observe how the interpretation of the NCs varies across sentences, and whether this correlates with the contextualised representations produced by various models. More specifically, we assume that, if models adequately incorporate contextual information, the standard deviations of the similarities between the NCs in different contexts should be correlated with those of the human annotators.

4.1 Models

We evaluate four contextualised models: three BERT variants, based on the Transformers architecture (Vaswani et al., 2017), and ELMo, which learns word vectors using bidirectional LSTMs (Peters et al., 2018). For English we used the ELMo small model provided by Peters et al. (2018), BERT-Large uncased (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), based on BERT-Base and distilled on SQuAD dataset, and Sentence-BERT (Reimers and Gurevych, 2019), trained on BERT-Large and both MultiNLI and SNLI.[8] For Portuguese we selected the ELMo pre-trained weights provided by Quinta de Castro et al. (2018) and the multilingual versions of the models used for English, namely mBERT (base cased), and both multilingual DistilBERT and Sentence-BERT (Reimers and Gurevych, 2020). As a static non-contextualised baseline we used GloVe (Pennington et al., 2014) (the English official models with 300 dimensions and trained on 840 billion tokens, and the equivalent Portuguese model released by Hartmann et al. (2017)). The vector representations were obtained with the *flairNLP* framework (Akbik et al., 2019) using the models provided by the *transformers* library (Wolf et al., 2020).

The representations of NCs (and their sentences) were obtained by averaging the word (or subword, if adopted by the model) embeddings. We used the concatenation of the three layers for ELMo and of

[8] https://www.nyu.edu/projects/bowman/multinli/
https://nlp.stanford.edu/projects/snli/

the last four hidden layers for the BERT models. In GloVe, words which are not in the vocabulary were skipped.

4.2 Experiment 1: Compositionality prediction

Unsupervised type idiomaticity identification with static non-contextualised word embeddings often assumes that the similarity between the NC embedding and the compositional embedding of the component words (e.g. *police_car* vs. *police* and *car*) is an indication of idiomaticity (Mitchell and Lapata, 2010): the more similar they are the more compositional the NC is. To approximate this with contextualised models, we calculate the cosine similarities between the contextualised vector of the NC in each sentence with two types of non-contextualised vectors. The first evaluates if even in the absence of an informative sentence context, each of the component words would be enough of a trigger to cue the NC meaning (e.g. *eager* for *eager beaver*). This is implemented as the vector for the NC out of context, obtained by feeding the model only with the compound, dubbed *NC_out*.[9] The second non-contextualised vector evaluates if the representations for the individual words have enough information to reconstruct the meaning of the NC in the absence of context and of the collocated component. It is implemented as the sum of the individual vectors of the NC components, where each NC component is fed individually to the model as a sentence, referred to as NC_out_{Comp}. On each case, we calculate two Spearman correlations with human judgments: at token level, using all the sentences for each language; and at type level, comparing the average cosine similarities of each NC with their compositionality scores at type level. We also compute correlations between the similarities and frequency-based data, namely the NC raw frequency, and the PPMI (Church and Hanks, 1990) between its component words, to verify whether they have any impact in these measures of idiomaticity. The frequency data were obtained from ukWaC, with 2.25B tokens in English (Baroni et al., 2009), and brWaC, containing 2.7B tokens in Portuguese (Wagner Filho et al., 2018).

The results by Cordeiro et al. (2019) suggested that if the two components of an NC are processed as a single token unit (for instance, by explicitly linking them with an underscore) the resulting static representation captures the NC idiomatic meaning. This is not surprising since by linking the two components we create a new word that would be treated by the model as completely independent of the preexisting component words. But such preprocessing may not be desirable or even feasible. In this sense the contextualised models would be a good promise, since we expected that by processing a sentence with an idiomatic NC, the context would be enough to lead the model into linking the component words and assigning the corresponding idiomatic meaning. Figuratively speaking, the contextualised models would put the underscore for us. Therefore, if contextualised models capture idiomaticity, the similarity between NC and NC_out_{Comp} (or *NC_out*) should have strong correlations with the idiomaticity scores of the NCs.

Table 5 shows the significant correlations in English (top rows) and Portuguese (bottom). These results indicate at best weak (NC_out_{Comp}) to moderate (*NC_out*) correlations between models' predictions and human judgments, both at type and token levels. Moreover, the correlations obtained are much smaller than those found by the static models used by Cordeiro et al. (2019). For English, the best correlations (0.37) were obtained by BERT, while ELMo and Sentence-BERT achieved the best performance in Portuguese (0.27 and 0.26, respectively). In both languages, the lower values were those of DistilBERT. It is worth noting that a direct comparison between the BERT models in both languages should not be done, as they are monolingual (for English) and multilingual (for Portuguese).

For PPMI, only weak positive correlations were found for ELMo and DistilBERT, indicating that for them higher cosine values weakly imply NCs with stronger association scores. Moreover, weak to moderate negative correlations with frequency were found for the BERT models, suggesting that cosine similarity is higher for less frequent NCs. The differences between *NC_out* and NC_out_{Comp} indicate the importance of some degree of contextualisation (also found by Yu and Ettinger (2020)), even if only as one component contextualising the other in *NC_out*, which may not be retrievable from the combination of the context-independent vectors of the components (NC_out_{Comp}). This is in line with the original strategy used with static embeddings, which learns the distribution of the NCs pre-identified as single tokens in corpora and that

[9] This representation equivalent to the *Avg Phrase* used by Yu and Ettinger (2020).

resulted in significantly better correlations per type than any of the contextualised models (Cordeiro et al., 2019).

To make a fairer comparison between both approaches, we injected into the BERT models single representations for the NCs, learnt from the referred ukWaC and brWaC corpora. We first annotated as single tokens in the corpus those NCs present in the dataset, and used attentive mimicking with one-token-approximation (Schick and Schütze, 2019, 2020b) to learn up to 500 contexts for each compound. After that, we injected these type level vectors into the BERT models using BERTRAM (Schick and Schütze, 2020a). For English, these new representations obtained lower results than the original BERT in *NC_out* (e.g., 0.37 vs. 0.28 at type level), but higher in NC_out_{Comp} (0.16 vs. 0.33 at type level). For Portuguese, including single representations for the NCs in BERT improved the correlations in three of the four scenarios (except for *NC_out* at token level), but the best results were almost identical to those of ELMo (see the full results in the bottom rows of Table 5).

Regarding the results reported by Nandakumar et al. (2019), for English, our experiments yielded higher correlations for BERT and lower for ELMo ($\approx$ 0.3 in both cases, depending on the setting), which may be due to differences in how the vectors are generated (e.g., the use of different input sentences, hidden layers or compositional operations).

In sum, the results of these evaluations suggest that the use of a straightforward adaptation of a compositionality prediction approach that led to good performance with static models was not as successful with contextualised models.

4.3 Experiment 2: Investigating idiomaticity with word embedding models

We analyse whether models are able to capture differences in idiomaticity perceived by human annotators across the sentences in which an NC occurs. That is, if an NC is found to be more idiomatic in one sentence than in others. For that, we created an annotator's vector for each sentence, combining the human scores to create a three dimensional vector representation, where the first dimension is the average NC compositionality, and the second and third are the average scores of the contributions of the head and of the modifier. For representing the sentence we obtain an embedding by averaging their (sub)words. We calculated the Euclidean distances between (i) the annotators' vectors and (ii) the cosine similarities between sentence embeddings of each of the possible combinations of the three sentences associated to each NC. Then, we measured the correlations between these values using Spearman ρ. We aim to assess if annotations and models indicate the same relative differences.[10] The results were averaged for the 280 (English) and 180 (Portuguese) NCs.

Table 6 shows the results for the whole datasets and divided by compositionality level. As we compare Euclidean distances with cosine similarities negative values are actually positive correlations and vice versa. The average ρ is close to 0 suggesting that the embedding models do not capture the nuances in idiomaticity perceived by the annotators between the different sentences per NC.

4.4 Experiment 3: NC idiomaticity across sentences

We also analysed the similarity among the annotations for each NC in the three sentences, computing the standard deviations of the average compositionality scores given by the annotators. In contrast to the previous experiment, here we represent the human annotations using only the idiomaticity scores of the whole NCs and the models' output as the contextualised embedding of the NCs in each sentence. At token level most compounds (85.7% in English and 91.1% in Portuguese) have mean idiomaticity scores with less than 0.6 of standard deviation. Very few NCs have deviations higher than 1: five in English and four in Portuguese. Looking at the contexts in which they occur, the variability seems to be due to the different topics to which the sentences refer. For instance, the annotators have identified two senses of *firing line*: one, more idiomatic, referring to a position in which someone is criticised (mean score of 1.25), and a second one (partially compositional, with an average of 2.7) referring to a specific position in an armed conflict. In Portuguese, *céu aberto* ('open-air', lit. 'open-sky') was interpreted as less compositional (1.2) when describing urban settings (e.g., open-air shopping centers) than when referring to wild places (e.g., *lobas que lutavam a céu aberto*, 'wolves fighting in the open'), with a mean idiomaticity score of 3.

[10] Spearman ρ is not used here as a statistical test but as a measure to evaluate if the sentence comparisons with two different metrics yield the same relative differences. As there are only three sentences to compare, ρ assumes only four values ± 0.5 or ± 1.

English												
Model	**NC_out**						**NC_out$_{\text{Comp}}$**					
	token level			**type level**			**token level**			**type level**		
	pred	PP	*freq*	*pred*	PP	*freq*	*pred*	PP	*freq*	*pred*	PP	*freq*
BERT	0.36	–	*-0.11*	0.37	–	–	0.20	–	-0.26	0.16	–	-0.34
DBERT	*0.07*	0.13	-0.26	–	*0.15*	-0.33	–	–	-0.27	–	–	-0.31
SBERT	0.20	–	-0.20	0.19	–	-0.22	–	–	-0.30	–	–	-0.33
ELMo	0.12	0.18	–	–	0.25	–	*0.07*	0.22	–	–	0.29	–
BERTRAM	0.16	–	0.15	0.28	–	0.23	0.20	–	–	0.33	–	–
Cordeiro et al. (2019) best prediction result at type-level (*word2vec* skip-gram): 0.73												

Portuguese												
Model	**NC_out**						**NC_out$_{\text{Comp}}$**					
	token level			**type level**			**token level**			**type level**		
	pred	PP	*freq*	*pred*	PP	*freq*	*pred*	PP	*freq*	*pred*	PP	*freq*
BERT	0.16	0.21	-0.12	*0.19*	0.24	–	–	0.23	*-0.11*	–	0.27	–
DBERT	0.16	0.19	0.12	*0.19*	0.24	*0.16*	0.19	0.46	-0.19	*0.17*	0.50	-0.20
SBERT	0.24	0.15	0.21	0.26	*0.16*	0.23	0.16	0.14	–	*0.19*	*0.15*	–
ELMo	0.26	0.17	0.15	0.27	0.22	*0.17*	0.27	0.17	-0.19	0.27	0.21	-0.12
BERTRAM	0.14	–	*0.09*	0.21	–	–	0.24	–	–	0.27	–	*0.17*
Cordeiro et al. (2019) best prediction result at type-level (PPMI model): 0.60												

Table 5: Spearman ρ correlations of contextualised models at token and type level (with the best type-level results from Cordeiro et al. (2019) for comparison). *NC_out* (left) refers to the results of the non-compositional approach, while *NC_out$_{Comp}$* are those of the compositional one (right). *Pred* are the results of the compositionality prediction measures proposed. PP and *freq* mean PPMI and frequency, respectively. Correlations have $p < 0.01$ except for values in italic ($p <= 0.05$). Non-significant results are omitted.

English								
Model	**Total**		**Idiomatic**		**Part. Comp.**		**Composit.**	
	ave. ρ	*StDev*	ave. ρ	*StDev*	ave. ρ	*StDev*	ave. ρ	*StDev*
BERT	-0.066	0.72	-0.058	0.71	-0.028	0.74	-0.111	0.70
DBERT	-0.032	0.71	0.047	0.71	-0.119	0.69	-0.036	0.74
SBERT	0.011	0.73	0.015	0.74	0.057	0.70	-0.038	0.74
ELMO	0.006	0.70	0.005	0.70	0.000	0.67	0.045	0.71
GLOVE	0.016	0.69	0.044	0.74	-0.063	0.66	0.030	0.71

Portuguese								
Model	**Total**		**Idiomatic**		**Part. Comp.**		**Composit.**	
	ave. ρ	*StDev*	ave. ρ	*StDev*	ave. ρ	*StDev*	ave. ρ	*StDev*
BERT	0.006	0.70	0.083	0.71	-0.050	0.71	-0.017	0.69
DBERT	0.031	0.72	0.050	0.75	0.083	0.71	-0.058	0.70
SBERT	0.001	0.72	-0.025	0.72	0.008	0.72	0.036	0.72
ELMO	-0.008	0.71	-0.017	0.75	0.042	0.72	-0.050	0.67
GLOVE	-0.006	0.72	-0.017	0.77	-0.058	0.66	0.058	0.73

Table 6: Average correlations (Spearman ρ) and standard deviations (*StDev*) on the whole dataset (Total) and in the three classes: *idiomatic*, *partially compositional*, and *compositional* noun compounds. Negative values are positive correlations and vice versa.

To observe whether language models capture these differences across sentences, we calculated the cosine similarities between the NCs in the three sentences and the standard deviation of these three values. We then computed the Spearman correlations between these deviations obtained from the models' representations and those of the human annotations: all correlations were very low and not significant, suggesting that the vector representations do not capture the variability perceived by the annotators. Finally, we have also selected two NCs in English with a combination of idiomatic and compositional meanings (*brick wall*, and *gold mine*). In these examples, we found that for BERT (our best model) the cosine similarities between the idiomatic meanings were higher (0.83 in both cases) than between idiomatic and compositional senses (0.68 and 0.7, respectively), suggesting that they are somehow identifying the different senses. However, since the highest standard deviations were achieved with NCs representing the same sense in all contexts (e.g., *big wig* and *grass root*), further analysis is needed.

As neither the cosine similarities obtained with BERT-based models nor the standard deviations between them were correlated with the variation in the human scores, these analyses suggest that state-of-the-art contextualised models still do not model semantic compositionality as human annotators do.

The experiments performed in this section have shown, on the one hand, some of the possibilities of a multilingual dataset labeled at type and token level; on the other hand, the results also suggest that capturing idiomaticity is a hard task for current language models, as only some of them show moderate correlations with human annotations in some scenarios.

5 Conclusions and Future Work

This paper presented the NCTTI, a dataset of NCs in English and Portuguese annotated at type and token level with human judgments about idiomaticity, and with suggestions of paraphrases. The very strong correlations found between type and token judgments confirm the robustness of the scores, while the paraphrases provide further validation of the interpretation of the NCs.

Moreover, evaluations involving embedding models with different levels of contextualisation suggest that they are still far from providing accurate estimates of NC idiomaticity, at least using the measures proposed and analysed in the paper. MWEs are still a pain in the neck for NLP, and datasets like the NCTTI can contribute towards finding better representations for them and better measures for idiomaticity identification.

Future work includes using these NCs as seeds in cross-lingual representations for enriching the dataset with NC equivalents in different languages. Besides, we also plan to enlarge the datasets including a subset of sentences with ambiguous NCs having idiomatic and compositional interpretations depending on the context.

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

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. 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 505–516, Minneapolis, Minnesota. 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.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

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

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.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Denver, Colorado. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating BERT for natural language inference: A case study on the CommitmentBank. 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 6086–6091, Hong Kong, China. 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 (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. Postprint version. Retrieved from http://repository.upenn.edu/asc_papers/43.

Germán Kruszewski and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 171–181, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala, Sweden. 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.

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

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.

Gustavo H. Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, The University of Sheffield, Sheffield, UK.

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.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Nils Reimers and Iryna Gurevych. 2019. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 32–41, Atlanta, Georgia, USA. Association for Computational Linguistics.

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.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv preprint arXiv:1706.03762.

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).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France. European Language Resources Association.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into Neural Machine Translation. In *Proceedings of the Eighth International Conference on Learning Representations*, Addis Ababa, Ethiopia.