Investigating Idiomaticity in Word Representations

Wei He[1], Tiago Kramer Vieira[2], Marcos Garcia[3], Carolina Scarton[1], Marco Idiart[4], and Aline Villavicencio[1,5,6]

[1]University of Sheffield, School of Computer Science
Wei.He1@sheffield.ac.uk, c.scarton@sheffield.ac.uk
[2]Federal University of Rio Grande do Sul, Institute of Informatics
tiagokv@hotmail.com
[3]University of Santiago de Compostela, CiTIUS Research Center
marcos.garcia.gonzalez@udc.gal
[4]Federal University of Rio Grande do Sul, Institute of Physics
marco.idiart@ufrgs.br
[5]University of Exeter, Institute for Data Science and Artificial Intelligence
a.villavicencio@exeter.ac.uk
[6]The Alan Turing Institute

*Idiomatic expressions are an integral part of human languages, often used to express complex ideas in compressed or conventional ways (e.g., **eager beaver** as a keen and enthusiastic person). However, their interpretations may not be straightforwardly linked to the meanings of their individual components in isolation and this may have an impact for compositional approaches. In this article, we investigate to what extent word representation models are able to go beyond compositional word combinations and capture multiword expression idiomaticity and some of the expected properties related to idiomatic meanings. We focus on noun compounds of varying levels of idiomaticity in two languages (English and Portuguese), presenting a dataset of minimal pairs containing human idiomaticity judgments for each noun compound at both type and token levels, their paraphrases and their occurrences in naturalistic and sense-neutral contexts, totalling 32,200 sentences. We propose this set of minimal pairs for evaluating how well a model captures idiomatic meanings, and define a set of fine-grained metrics of Affinity and Scaled Similarity, to determine how sensitive the models are to perturbations that may lead to changes in idiomaticity. Affinity is a comparative measure of the similarity between an experimental item, a target and a potential distractor, and Scaled Similarity incorporates a rescaling factor to magnify the meaningful similarities within the spaces defined by each specific model. The results obtained with a variety of representative and widely used models indicate that, despite superficial indications to the contrary in the form of high similarities, idiomaticity is not yet accurately represented in current models. Moreover, the performance of models with different levels of contextualization suggests that their ability to capture context is not yet able to go beyond*

Action Editor: Tal Linzen. Submission received: 22 April 2024; revised version received: 29 August 2024; accepted for publication: 22 October 2024.

https://doi.org/10.1162/coli_a_00546

more superficial lexical clues provided by the words and to actually incorporate the relevant semantic clues needed for idiomaticity. By proposing model-agnostic measures for assessing the ability of models to capture idiomaticity, this article contributes to determining limitations in the handling of non-compositional structures, which is one of the directions that needs to be considered for more natural, accurate, and robust language understanding. The source code and additional materials related to this paper are available at our GitHub repository.[1]

1. Introduction

The evolution of word representation models has resulted in models with seemingly remarkable language abilities. Not surprisingly, these models have been found to store a wealth of linguistic information (Henderson 2020; Manning et al. 2020; Vulić et al. 2020; Lenci et al. 2022), displaying high levels of performance on various tasks ranging from the abilities of even the static models of detecting semantic similarities between different words (Lin 1999; Mikolov et al. 2013; Baroni, Dinu, and Kruszewski 2014) to those of contextualized models of grouping representations in clusters that seem to be related to the various senses of the word (Schuster et al. 2019) and can be matched to specific sense definitions (Chang and Chen 2019). While substantial evaluation efforts have concentrated on word and subword units and on larger compositional combinations derived from them, there is less understanding about their ability for handling less compositional structures, such as those found on multiword expressions (MWEs), like noun compounds (NCs) (Garcia et al. 2021a), verb-noun combinations (King and Cook 2018; Hashempour and Villavicencio 2020), and idioms (Yu and Ettinger 2020; Dankers, Lucas, and Titov 2022). Indeed, MWEs include a variety of distinct phenomena and have been described as interpretations that cross word boundaries (Sag et al. 2002), whose meanings are not always straightforwardly derivable from the meanings of their individual components. Moreover, although they include, on the one hand, more transparent and compositional expressions (like *salt and pepper*) or expressions with implicit relations (like *olive oil* as oil *made from* olives), on the other hand they also include more idiomatic expressions (like *eager beaver* as a person who is willing to work very hard[2]), falling into a continuum of idiomaticity[3] (Sag et al. 2002; Fazly, Cook, and Stevenson 2009). This leads to potential problems for models if they follow the Principle of Compositionality (Frege 1956; Montague 1973), building the meaning of a larger unit (like a sentence or an expression) from a combination of the individual meanings of the words that are contained in it, as this would result in potentially incomplete or incorrect interpretation for more idiomatic cases (e.g., the idiomatic *eager beaver* interpreted literally as *impatient rodent*). Although understanding the meaning of an MWE may require knowledge that goes beyond that of the meanings of these individual words in isolation (Nunberg, Sag, and Wasow 1994), failure to take idiomaticity into account can affect the quality of downstream tasks (Sag et al. 2002; Constant et al. 2017; Cordeiro et al. 2019) such as reasoning and inference (Chakrabarty, Choi, and Shwartz 2022; Chakrabarty et al. 2022; Saakyan et al. 2022), information retrieval (Acosta, Villavicencio, and Moreira 2011), and machine translation (Dankers, Lucas, and Titov 2022). For machine translation, for example, the degree of idiomaticity and ambiguity of MWEs (literal vs. idiomatic usages) were found to have an impact on the quality of the results obtained (Dankers, Lucas,

1 `https://github.com/risehnhew/Finding-Idiomaticity-in-Word-Representations`.
2 Definition from the *Cambridge Dictionary*.
3 We understand idiomaticity as *semantic opacity*, and its continuum as different *degrees of opacity*.

and Titov 2022). Due to their non-compositional nature, idiomatic expressions result in lower-quality translations than literal expressions, as evidenced by lower BLEU scores for translations that are paraphrased rather than translated word-for-word. In this article, we investigate to what extent widely used word representation models are able to capture idiomaticity in MWEs. We focus, in particular, on their initial abilities for representing idiomaticity, looking at noun compounds of varying degrees of idiomaticity.[4] In addition to the complex interactions between MWEs, their component words, and their contexts (Sag et al. 2002), characteristics of languages and of word representation models may affect how accurately MWEs can be represented and processed, and we investigate the impact of some of these factors for compounds in two different languages (English and Portuguese).

One of the challenges is that uncovering how word representation models capture a specific type of knowledge is a non-trivial problem (Vulić et al. 2020), and may depend on factors like the particular model and the way it encodes different types of linguistic information (Yu and Ettinger 2020). For instance, whereas in Transformer-based models, the initial layers seem to represent more lexical level knowledge and the final layers seem to capture more semantic and pragmatic information (Rogers, Kovaleva, and Rumshisky 2020), determining where phenomena that sit at the interface of various levels are encoded, like multiword expressions (Sag et al. 2002), is challenging because they could potentially involve information distributed across different layers. Moreover, the possible findings from an investigation about where in the architecture of a given model idiomaticity is encoded, or about the role of particular components in representing it, may not generalize to other models and architectures. In this article we propose instead a set of model-agnostic idiomatic probes for assessing the representation of idiomaticity. These probes contain NCs of different levels of idiomaticity, ranging from idiomatic to compositional cases, which form the basis for minimal pairs. In these pairs one of them contains an NC and the other contains a semantically related item (such as a synonym) or a distractor. The hypothesis is that if a model is able to accurately represent an NC, higher similarities will be observed for minimal pairs involving NCs and their synonyms (e.g., for the idiomatic *eager beaver* and *hardworking person*). Conversely, for minimal pairs with variants that may incorporate changes in meaning, such as those containing NCs and synonyms of their individual component words (e.g., the idiomatic *eager beaver* and *impatient rodent*) or other distractors, lower similarities should be observed.

As word representation models may form spaces that are anisotropic (Ethayarajh and Jurafsky 2021) with representations concentrating on parts of the space, or may have rogue dimensions that dominate similarity measures (Timkey and van Schijndel 2021), these could lead to high similarities overall (Liu et al. 2020), affecting the ability to distinguish meaningful similarities from spurious ones arising from specific characteristics of a given space. In this article, we propose two new measures to assess idiomaticity within a model while taking into account its potential for high similarities. The first, *Assessment of Feature Familiarity and Idiomatic Nuance by Interpreting Target Yielding* (**Affinity**), takes two representations of different levels of relatedness to a given target, and can be used to determine if a model accurately reflects their degree of

4 We use the off-the-shelf publicly available pre-trained versions of widely adopted word representation models, standard operations, and common similarity measures. Even in scenarios in which adopting additional optimizations, more complex operations or fine-tuning could lead to improvements in performance, this may depend on the availability of comprehensive training data for the target model, domain, and language. Measuring the initial idiomatic abilities of models can help understand the potential loss of idiomatic meaning that could be propagated to the downstream tasks that use them off the shelf.

similarity to the target. Focusing on idiomaticity, we use Affinity to assess if greater similarities are observed for NCs and related words (in this case their synonyms), than for NCs and other potentially less related alternatives including distractors. The second measure, **Scaled Similarity**, determines a new lowerbound for a given space in terms of similarities for unrelated representations, rescaling the space to help distinguish them from the meaningful similarities for related representations. For idiomaticity, we analyze the similarities between the NCs and their synonyms adopting the similarities between the NCs and random items as a new lowerbound. These measures of Affinity and Scaled Similarity do not directly address the problem of rogue dimensions, and we discuss this further in the Conclusions section.

Using these metrics and minimal pairs for evaluation, this article presents a fine-grained analysis of the ability of a model to capture idiomaticity, looking at the following questions:

Q1 **To what extent is idiomaticity captured by word representation models?** We assess this by comparing the predictions of models for NCs and their synonyms against human judgments about idiomaticity in the same sentences, analyzing how sensitive these models are to potential changes in meaning resulting from the lexical variations in the minimal pairs.

Q2 **Is this ability affected by the degree of idiomaticity of the NCs, the informativeness of the contexts, or the languages involved?** To determine if more idiomatic expressions are more challenging for models, we present an analysis of the impact of the level of idiomaticity of the NCs. We also analyze more informative contexts provided by naturalistic sentences against uninformative neutral contexts to determine their impact on idiomaticity representation. These evaluations include two languages, to measure the potential language dependence of these results.

Q3 **Do contextualized models (from Transformer-based models) perform better compared to static models in idiomaticity representation?** In addressing this question, we conduct a comparative analysis across different static and contextualized models, focusing on their ability to capture idiomatic expressions. This involves examining how each model represents idiomatic NCs of varying levels of idiomaticity, in sentences that contain more (or less) informative contexts, and the accuracy with which they reflect the nuanced meanings that idiomaticity often entails. The analyses consider various linguistic scenarios that can change the idiomatic meaning to comprehensively assess the accuracy of contextualized models over their static counterparts.

The main contributions of this work include:

- The Noun Compound Idiomaticity Minimal Pairs (NCIMP) Dataset, a dataset of minimal pair sentences containing NCs of varying levels of idiomaticity, along with human judgments about the degree of NC idiomaticity and gold standard paraphrases, at both type and token level.

In total, the dataset contains 32,200 sentences for two languages (19,600 in English and 12,600 in Portuguese).[5]

- A comparative measure of Affinity to help determine how accurately idiomaticity is incorporated in these representations contrasting similarities for semantically related and unrelated representations.
- A novel model-agnostic measure of Scaled Similarity, which rescales a space in relation to a new lowerbound taking into account expected similarities among random items to magnify meaningful similarities among semantically related representations.
- In-depth analyses of the representation of idiomaticity in widely used word representation models, examining their ability to display sensitivity to changes in idiomaticity.

The remainder of this article is organized as follows: Section 2 presents related work. Section 3 presents the NCIMP dataset (Section 3.1), the models (Section 3.2), and the proposed idiomatic probes and measures (Section 3.3). Finally, in Section 4 we discuss the results of our experiments and draw conclusions in Section 5.

2. Representing Multiword Expressions and Idiomaticity

2.1 Static and Contextualized Models for Representing MWEs

A variety of vector models have been used to investigate the representation of MWEs, ranging from static to contextualized representations, each with its own set of challenges (Contreras Kallens and Christiansen 2022; Garcia et al. 2021a; Liu and Neubig 2022). The former include models like Word2Vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), and fastText (Bojanowski et al. 2017), which represent words at type-level, producing a single vector for each word that conflates all its senses. At this level, MWEs are often represented based on their overall syntactic and semantic properties as they are generally understood, without taking into account the variability of contexts. For example, both the literal and the idiomatic meaning of *gold mine*[6] would be represented jointly in a single vector regardless of its use in any specific sentence. At the other end of the scale are the contextualized models, from ELMo (Peters et al. 2018), BERT (Devlin et al. 2019) and GPT-3 (Brown et al. 2020) to LlaMA (Touvron et al. 2023) and other large language models, which produce token-level dynamic representations dedicated to capturing specific usages of a word in a particular context, resulting in several vectors for each word (Lenci et al. 2022; Apidianaki 2022). Token-level representations focus on the specific occurrences of words or subwords within contexts, and how their meaning or function may vary or be influenced by the surrounding text. Therefore, they have the potential for accurately representing MWEs, capturing the interdependence of the idiomatic meaning on a particular configuration of words, while also anchoring the MWEs in relation to their immediate linguistic environment. The primary challenge at token-level is accurately determining the presence, meaning, and role of MWEs in specific

5 This work extends the idiomatic probes proposed by Garcia et al. (2021b) and the type and token annotations by Garcia et al. (2021a), also introducing new measures, additional tests, and substantially expanding the analyses with new baselines and results from a larger set of models.

6 "Opportunity for making a lot of money" (definition from the *Cambridge Dictionary*).

contexts, especially when they have possibly multiple literal and idiomatic readings or when they are part of complex syntactic structures (Zeng and Bhat 2021).

Evaluation of successive generations of word representation models, ranging from static (Landauer and Dumais 1997; Lin 1999; Baroni and Lenci 2010; Mikolov et al. 2013; Bojanowski et al. 2017) to contextualized models (Peters et al. 2018; Devlin et al. 2019; Brown et al. 2020; Touvron et al. 2023), has devoted considerable attention to their linguistic abilities (Mandera, Keuleers, and Brysbaert 2017; Wang et al. 2018; Henderson 2020; Rogers, Kovaleva, and Rumshisky 2020; Lenci et al. 2022). On lexical semantics, the representations extracted from contextualized models seem to be able to reflect word senses in clusters of vectors (e.g., Wiedemann et al. [2019] for BERT) including in cross-lingual alignments involving polysemous words (e.g., Schuster et al. [2019] for ELMo). However, controlled uniform evaluations of different generations of word representation models settings have also reported strong performances from static models, which were able to outperform contextualized models in most tasks (Lenci et al. 2022).

2.2 Vector Model Evaluation on Idiomaticity

Regarding idiomaticity, uniform assessment of the performance of different models on the processing of MWEs are particularly important, as independent evaluations have reported mixed results (King and Cook 2018; Nandakumar, Baldwin, and Salehi 2019; Cordeiro et al. 2019; Hashempour and Villavicencio 2020; Garcia et al. 2021b; Klubička, Nedumpozhimana, and Kelleher 2023). For instance, for the task of identifying the degree of idiomaticity of MWEs at type level (i.e., the potential of an MWE to be idiomatic in general), good performances have been obtained with static word embeddings (Mitchell and Lapata 2010; Reddy, McCarthy, and Manandhar 2011; Cordeiro et al. 2019), and they have even been reported as obtaining better performance than contextualized models for capturing idiomaticity in MWEs in some evaluations (King and Cook 2018; Nandakumar, Baldwin, and Salehi 2019). Likewise, BERT-based models obtained similar results to those of static vector representations for predicting the degree of compositionality of a given NC (Miletic and Schulte im Walde 2023).[7]

However, a potential limitation of static models is that in representing different word senses in the same vector, the literal usage of an expression may differ considerably from its idiomatic usage (e.g., a *brass ring* as an idiomatic prize or as a literal ring made of brass), and complex operations may be required to deal with semantic phenomena like polysemy (Erk 2012). In this sense, contextualized models may provide the means for distinguishing literal from idiomatic usages, along with fine-grained sense distinctions. In this respect, Garcia et al. (2021a) proposed probing metrics to investigate and understand the linguistic information encoded in the models' representations. Similarly, using a method of probing with noise and a repurposed idiomatic usage probing task revealed better performance by BERT in encoding idiomaticity compared to GloVe (Klubička, Nedumpozhimana, and Kelleher 2023). These types of intrinsic evaluations have also been framed as shared tasks, like SemEval-2022 task 2B (Tayyar Madabushi et al. 2022), which proposed the assessment of idiomaticity representation in multilingual texts (English, Portuguese, and Galician) while also requiring models to predict the semantic text similarity scores between sentence pairs, regardless of whether or not either sentence contains an idiomatic expression.

7 See Miletić and Walde (2024) for a recent survey on the representation of MWEs in Transformer-based models.

Extrinsic evaluations have measured how well the representation of idiomaticity in a model impacts downstream tasks, for example, sentence generation (Zhou, Gong, and Bhat 2021), or conversational systems (Adewumi, Liwicki, and Liwicki 2022). For instance, evaluations of different classifiers initialized with static and contextualized embeddings in five tasks related to lexical composition (including the literality of NCs) found that contextualized models led to better performance across all tasks (Shwartz and Dagan 2019), and supervised methods that used contextualized models also outperformed alternatives on the classification of potentially idiomatic expressions in both monolingual and cross-lingual (English and Russian) scenarios (Kurfalı and Östling 2020; Fakharian and Cook 2021). Alternatively, both types of representations can be combined, as, for example, in a supervised neural architecture to identify and classify potentially idiomatic expressions combining contextualized and static embeddings in an attention flow (Zeng and Bhat 2021). Regarding machine translation, a recent evaluation of compositional generalization in Transformer models found that they tend to perform too compositional translations even for idiomatic expressions (Dankers, Lucas, and Titov 2022). Furthermore, an analysis of GPT-3 (Brown et al. 2020) reported 50.7% accuracy in idiom comprehension (Zeng and Bhat 2022), suggesting that the models' ability to deal with idiomaticity is not yet adequate.

2.3 Vector Operations and Idiomatic Knowledge Induction

In addition to the level of contextualization, the performance of vector space models may also be affected by the way the target words of an expression are composed, with functions like sum, concatenation, and multiplication used for combining the words of static models (Cordeiro et al. 2019; Mitchell and Lapata 2010; Reddy, McCarthy, and Manandhar 2011) or the subwords of contextualized models (Garcia et al. 2021b). For the embeddings extracted from language models, other potential sources of variation include which input is given to the model (e.g., one vs. several sentences including the target MWE in evaluations at the type level), or the number of layers that will be taken into account to obtain the vector representation (Miletić and Walde 2024). In this regard, the intermediate and last layers seem to encode more semantic information at the token level (Tenney, Das, and Pavlick 2019; Garcia 2021), whereas other evaluations at the type level found that averaging the initial layers of the target expressions achieved the best results (e.g., Miletic and Schulte im Walde [2023] for NCs and Vulić et al. [2020] for single word semantic tasks). With respect to semantic composition, Yu and Ettinger (2020) explored the type level representation of two word phrases (which in many cases correspond to NCs as the ones used in our study) in various contextualized models, showing that phrase representations miss compositionality effects as they heavily rely on word content. Similar conclusions, for neural machine translation, can be inferred from Dankers, Lucas, and Titov (2022). While some of these evaluations rely on substitutivity and the changes to a larger phrase representation caused by substitutions to its constituents (Garcia et al. 2021b; Yu and Ettinger 2020), alternatively, the notion of localism has also been analyzed (Liu and Neubig 2022) focusing on whether the operations of a model are local (Hupkes et al. 2021), that is, the extent to which the representation of a phrase is derivable from its local structure.

Crucially, a substantial amount of the discussed studies evaluate idiomaticity at the type-level, that is, they obtain the embedding of a given MWE by averaging its representation in several sentences that have been previously extracted in an automatic way. A more detailed controlled comparison of type-level and token-level idiomaticity reported

compatible results for both levels, with type-level being a close approximation for token-level (Garcia et al. 2021a) in sentences where the NC occurs with the same sense. Further analysis of the occurrences of these NCs in fine-grained sense annotations of literal and idiomatic usages (Tayyar Madabushi et al. 2021) provided additional confirmation that the ability of contextualized models to capture idiomaticity during pre-training was limited, with approaches for building single token representations (Phelps et al. 2022) and for fine-tuning leading to more accurate representations (Tayyar Madabushi et al. 2022). Recent alternatives for representing idiomatic expressions also include adding a new adapter module which has been developed and trained to recognize idioms (Zeng and Bhat 2022). This module functions as a language expert for idioms, augmenting the learning process of BART (Lewis et al. 2019) with additional information, and this approach effectively improves the representation of idiomatic expressions in off-the-shelf pre-trained language models, equipping them with greater ability to navigate the intricacies of natural language. Zeng and Bhat (2023) also proposed PIER+, a language model improvement for handling both literal and figurative language. This is achieved by combining a base model with an additional curriculum learning framework that gradually introduces more complex potentially idiomatic expressions. Compared with other models, PIER+ demonstrates better performance at identifying, understanding, and maintaining proficiency in both types of expressions. Finally, Zeng et al. (2023) introduce a knowledge graph designed to enhance the understanding of idiomatic expressions, which integrates commonsense knowledge to aid in deciphering the non-literal meanings of idioms. This work demonstrates how to inject MWE-related knowledge into pre-trained language models effectively. However, it is still unclear to what extent the context and its representation in contextualized models are contributing to a more accurate representation of MWEs according to their idiomaticity level (Nedumpozhimana and Kelleher 2021; Miletic and Schulte im Walde 2023).

2.4 Towards a More Controlled Assessment of Idiomaticity in Vector Space Models

Shedding some light on these questions requires a more controlled evaluation setup and measures that can abstract away from the particularities of these word representation spaces. In this effort, we take inspiration from psycholinguistic methodologies, which have been traditionally used to examine how humans process language in controlled experimental setups, to allow the removal of obvious biases and potentially confounding factors from evaluations (Linzen, Dupoux, and Goldberg 2016; Gulordava et al. 2018). They also enable comparative analyses of performance in artificially constructed but controlled sentences and in naturally occurring sentences.

Setups like these have been used, for instance, to investigate how models represent syntax, if they understand negation (van Schijndel and Linzen 2018; Prasad, van Schijndel, and Linzen 2019; Ettinger 2020; Kassner and Schütze 2020), and if they are aware of which properties are relevant for which concepts (Misra, Rayz, and Ettinger 2023). Adopting evaluation protocols that use minimal pair sentences (e.g., Warstadt et al. 2020; Misra, Rayz, and Ettinger 2023) allows for a controlled comparison of the target item against carefully selected distractors that may share linguistic properties with them. For instance, a dataset of Conceptual Minimal Pair Sentences was used to compare the performance of 22 large language models including both masked language models (like BERT) and autoregressive language models (like GPT-2), where the models have to validate which of two concepts a given property belongs to (e.g., *stripes* for *zebras* vs. *oaks*). Although the models seem to obtain relatively high accuracies for attributing properties to concepts, when semantically related concepts are involved or distractors

are included, performance drops substantially, and goes below chance even for models like GPT-3 (Misra, Rayz, and Ettinger 2023). Similarly, in targeted syntactic evaluation (Marvin and Linzen 2018), models are assessed using minimal pairs datasets focused on specific syntactic phenomena, such as those included in the BLiMP dataset for English (Warstadt et al. 2020). Analyses like these highlight the importance of adding controls to the experimental setup to distinguish seemingly sophisticated behavior with high performance that gives the illusion of knowledge from robust understanding with access to meaning (Misra, Rayz, and Ettinger 2023; de Dios-Flores, Garcia Amboage, and Garcia 2023). With this is mind, we follow Garcia et al. (2021b) and use minimal pairs to propose a set of intrinsic evaluations including probes and affinity measures aimed at gaining a better understanding of how vector space models represent MWEs with different degrees of semantic compositionality in context.

2.5 Datasets for Exploring Idiomaticity in Computational Models

Concerning experimental data, the first datasets to evaluate computational models were composed of different types of multiword expressions annotated at the type-level (McCarthy, Keller, and Carroll 2003; Venkatapathy and Joshi 2005). Further studies released annotations of MWEs in context, such as the VNC-tokens dataset (Cook, Fazly, and Stevenson 2008), which includes 60 English verb-noun combinations occurring in almost 3,000 sentences annotated as idiomatic or literal, or the IDIX corpus (Sporleder et al. 2010), with almost 6,000 labeled sentences of 78 expressions extracted from the British National Corpus. Using a crowdsourcing platform, Reddy, McCarthy, and Manandhar (2011) released a dataset with numerical ratings of the compositionality degree of 90 noun compounds in English, which also includes the contribution of each component to the meaning of the MWEs. Similar efforts were carried out for other languages, such as the GhoSt-NN dataset for German (Schulte im Walde et al. 2016), or the NC Compositionality (NCC) dataset (Cordeiro et al. 2019), which expanded the resource provided by Reddy, McCarthy, and Manandhar (2011) with additional NCs for English, and new data for Portuguese and French. Semi-automatic techniques combined with crowdsourced annotations were used to compile MAGPIE (Haagsma, Bos, and Nissim 2020), a large resource of more than 50,000 sentences with binary annotations at the token level of potentially idiomatic expressions. Similarly, the AStitchInLanguageModels dataset (Tayyar Madabushi et al. 2021), used in SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al. 2022), also contains potentially idiomatic expressions annotated in naturalistic sentences.

Recently, Garcia et al. (2021a; 2021b) enriched the English and Portuguese data of the NCC dataset with crowdsourced annotations of the compositionality degree of NCs and their components at the token level, paraphrases of the NCs in context, and different types of controlled replacements. These variants compose a large set of minimal pairs that allow for the systematic exploration of the representation of idiomaticity in vector space models.[8]

In this article, we adopt the minimal pairs paradigm as one of the bases for the evaluation and present the Noun Compound Idiomaticity Minimal Pairs dataset, which contains a set of idiomatic probes to explore to what extent idiomaticity is captured in word representation models. To do so, we rely on the datasets for English and Portuguese

8 We refer to Ramisch (2023) for a recent review on MWEs processing, including datasets, and to Schulte im Walde (2023) for a comprehensive overview on compositionality ratings for MWEs.

by Garcia et al. (2021a; 2021b) and extend them with new semantically related variants and distractors and sets of minimal pairs as discussed in the next section to conduct in-depth intrinsic evaluations.

3. Materials and Methods

3.1 Noun Compound Idiomaticity Minimal Pairs Dataset

The Noun Compound Idiomaticity Minimal Pairs (NCIMP) dataset contains 32,200 sentences targeting two-word NCs in two languages, 280 in English (EN) and 180 in Portuguese (PT), with idiomatic (e.g., *gravy train*[9]), partly compositional (e.g., *grandfather clock*[10]), and compositional (e.g., *research project*) NCs.[11] For each NC, the dataset contains minimal pairs formed by a first sentence with the target NC and a second sentence where the NC was replaced by an experimental item. These experimental items were selected on the basis of MWE properties, like more limited substitutability (or greater lexical fixedness), and can be used to determine if models are sensitive to perturbations to these properties, and if this is affected by how idiomatic the NCs are. For example, depending on the degree of lexical fixedness of an NC, the variants generated may not fully retain its original meaning (e.g., *panda car*[12] and *bear automobile*). In particular, we analyze the following:

- NC_{Syn}: the minimal pairs are formed by the NC being replaced by one of the **gold standard synonyms** provided holistically for the NC by the annotators (e.g., *brain* for *grey matter*). In this case, we adopted the synonyms provided by the Noun Compound Senses (NCS) dataset (Garcia et al. 2021b), which were selected on the basis of the most frequent paraphrases given by native speaker annotators. These pairs are used to assess if the models provide similar representations for NCs and their synonyms, even if they involve lexically diverse surface forms.
- $NC_{WordsSyn}$: minimal pairs where each component word of the NC is replaced individually by a synonym generating new **two-word compositional replacements** (e.g., forming *alligator sobs* for the NC *crocodile tears* by replacing *alligator* for *crocodile* and *sobs* for *tears*). The synonyms were manually selected from WordNet (Miller 1995) for English, and OpenWordNet (Rademaker et al. 2014) for Portuguese, and from online dictionaries of synonyms where additional coverage was required. In case of ambiguity (due to polysemy or homonymy), the most common meaning of each component was selected. For each NC, 5 compositional replacements were generated. These pairs are used to evaluate how sensitive a model is to the conventionality and lexical fixedness of these NCs, especially the more idiomatic ones, and if it can detect when the (idiomatic) meaning changes with the replacements.

9 Referring to an easy way of making money without doing much work (*Cambridge Dictionary*).
10 A type of tall free-standing clock.
11 The NCIMP dataset is based on the Noun Compound Senses (Garcia et al. 2021b), the Noun Compound Type and Token Idiomaticity (Garcia et al. 2021a), and the NC Compositionality (Cordeiro et al. 2019) datasets, significantly extending them with new data.
12 Referring to a police car.

- NC_{Comp}: the minimal pairs are formed by replacing the NC by only one of its **component words**, namely, replacing the NC by its head in one minimal pair, and by the modifier in the other pair (e.g., *crocodile* for *crocodile tears* and *tears* for *crocodile tears*). These pairs are used to explore if the models can detect when the meaning of an NC is related to the meaning of a component (in more compositional cases) from when it is not (in more idiomatic cases).
- NC_{Rand}: the **random replacement controlled by frequency** is a two word expression in which the words are chosen to match the frequencies of the components of the target NC. The frequency values were extracted from corpora (in this case ukWaC and brWaC) as follows: We averaged the frequency of each NC and of its components ($f_{avg} = (f_{NC} + f_{w1} + f_{w2})/3$), and extracted the compound with the closest average value (e.g., *police car* and *supermarket city*). For each NC, 5 random replacements were used for each sentence. These pairs are used as controls to determine the lowerbound similarities for the target NCs, avoiding the potential impact of any differences in frequency.

The NCs were pre-selected by experts trying to maintain a balance between the 3 classes (idiomatic, partial, and compositional),[13] and they appear in the context of three **naturalistic sentences** (*Nat*) from corpora that exemplify the same compound sense (Garcia et al. 2021a). Using Amazon Mechanical Turk (for English) and a dedicated custom built online platform (for Portuguese), compositionality scores for each NC and its components were obtained following the procedure of Reddy, McCarthy, and Manandhar (2011) and Cordeiro et al. (2019). A Likert scale from 0 (idiomatic) to 5 (compositional) was used for the human judgments, and the resulting scores were aggregated from the average of the different annotators (Garcia et al. 2021a).[14] The annotators also provided synonyms or paraphrases for the NCs in these sentences, which were used by language experts to manually generate the NC_{Syn} variants (Garcia et al. 2021a). These annotations, including the synonyms, were collected at two levels of granularity: a more fine-grained token level, where annotations for each sentence are collected individually, and a more rough-grained type level, where a single annotation for each NC is collected considering all three sentences at once (Garcia et al. 2021b). This allows for analyses of the impact of each individual context in the interpretation of the NC. A total of 8,725 annotations was obtained for English (421 annotators, each labelling an average of 21 sentences, resulting in 10.4 annotations per sentence). In Portuguese, 5,091 annotations were provided by 33 annotators (with an average of 154 annotated sentences per annotator, and 9.4 annotations per sentence).

In addition, NCIMP also contains **sense-neutral sentences** (*Neut*) in which the NCs appear in uninformative contexts containing only 5 words and following the pattern *This is a/an <NC>* for English (e.g., "This is an *eager beaver*") and the Portuguese equivalent

13 The two-word compounds were selected to be representative cases of compositional NCs (meaning related to the two words), partly idiomatic (meaning related to one of the words), and idiomatic (meaning unrelated to either of the two words), as our aim is to investigate to what extent the degree of idiomaticity affects the ability of models to generate an accurate representation. For English, the dataset contains 103, 88, and 89 idiomatic, partial, and compositional expressions, respectively, while for Portuguese it has 60 NCs per class.

14 On average, the compositionality scores were of 0.95/2.34/4.13 for English, and of 1.52/2.46/3.61 for Portuguese (idiomatic/partial/compositional).

Table 1
Naturalistic sentence containing the NC *front man* (in row 1) forming minimal pairs with sentences in rows 2–4, and with control random baselines in row 5.

#	NC	Sentence
1	Original	John Paul II was an effective *front man* for the catholic church.
2	NC_{Syn}	John Paul II was an effective *representative* for the catholic church.
3	$NC_{WordsSyn}$	John Paul II was an effective *forepart woman* for the catholic church.
4	NC_{Comp}	John Paul II was an effective *man* for the catholic church.
		John Paul II was an effective *front* for the catholic church.
5	NC_{Rand}	John Paul II was an effective *battlefront serviceman* for the catholic church.

Este/a é um(a) <NC>.[15] These neutral contexts can be used to examine how much contextual information is added to a representation in the more informative naturalistic contexts. Moreover, as some NCs may have more than one meaning (e.g., *fish story* as either the literal *aquatic tale* or the idiomatic *big lie*), they can also be used to determine the default usage elicited for the NC in the absence of any informative contextual clues, in particular, whether it leans towards an idiomatic or a literal sense, potentially serving as an indication of the predominant sense sampled during training.

Experts (native or near-native speakers with background in linguistics) reviewed both the naturalistic and the sense-neutral sentences in the minimal pairs, editing them if needed for preserving grammaticality after substitution (e.g., revising gender, number, and definiteness agreement with determiners and adjectives). However, some of the variants generated may be semantically nonsensical, especially those involving random replacements. Table 1 displays an example with the original sentence in the first row and the relevant sentences for each of the minimal pairs in the other rows.

Finally, each NC was also annotated with frequency, Pointwise Mutual Information (Church and Hanks 1989) and Positive Pointwise Mutual Information values, calculated from the ukWaC (2.25B tokens; Baroni et al. 2009) and brWaC corpora (2.7B tokens; Wagner Filho et al. 2018), which can serve as approximations for their familiarity and conventionality.

3.2 Word Representation Models

We evaluate representative static and contextualized models. For the former, we compare GloVe and Word2Vec, using the official models for English, and the 300 dimensions vectors for Portuguese (Hartmann et al. 2017).

For the latter, we evaluate a large set of models, including the Bi-LSTM-based ELMo (Peters et al. 2018), and several Transformer-based language models: BERT (Devlin et al. 2019) and some of its variants, such as multilingual BERT (mBERT[16]) (Pires, Schlinger, and Garrette 2019), multilingual DistilBERT (mDistilB[17]) (Sanh et al. 2019),

15 NCIMP also contains a second longer pattern of uninformative neutral sentences (10 words in English and 9 in Portuguese) following the patterns *This is what a/an <NC> is supposed to be* and the Portuguese equivalent *Isto é o que um/uma <NC> deveria ser*, to measure the potential impact of the length of the neutral context and of the position of the NC in the sentence. As the two types of neutral sentences elicit similar results, in the article we only present the results for the short neutral sentences.

16 `https://huggingface.co/google-bert/bert-base-multilingual-cased`.

17 `https://huggingface.co/distilbert/distilbert-base-multilingual-cased`.

and multilingual Sentence-BERT (mSBERT[18]) (Reimers and Gurevych 2019b). The recent flagship model Llama2 (Touvron et al. 2023) is also included in our experiments. OpenAI text embeddings (Neelakantan et al. 2022) are included in the evaluations at sentence-level as they can only be accessed by the API[19] rather than by direct inspection of the whole model, which would be required for analyses at the NC-level. Therefore, the latter are not conducted for OpenAI text embeddings. For ELMo, we use the small model provided by Peters et al. (2018), and for Portuguese we adopt the weights released by Quinta de Castro, Félix Felipe da Silva, and da Silva Soares (2018). For Llama2 and OpenAI's embeddings, we use the *13B* version and *text-embedding-ada-002* version, respectively. For all other contextualized models, we use the pre-trained weights publicly available through Flair[20] (Akbik et al. 2019) and HuggingFace[21] (Wolf et al. 2020). For BERT-based models (and for DistilB in English), we report the results obtained both by the multilingual uncased (ML) and by monolingual models for English (large, uncased) and Portuguese (large, cased), all available through HuggingFace.

3.2.1 Sentence and NC Embeddings. Embeddings for the whole sentence as well as for the NCs are generated by averaging the (sub)word embeddings[22] of the relevant tokens involved, according to the model:

- for static models, the word embeddings are derived directly from the vocabulary, with missing out-of-vocabulary words being ignored;
- for ELMo the output word embeddings are averaged, and the concatenation of its three layers is adopted;
- for Transformer-based models, the word embeddings are generated by averaging the representations of the sub-tokens and we report results using the last four layers.[23]

In general we adopt standard widely used configurations to determine what the landscape of results is before any task optimization, even if alternative tokenization approaches (Gow-Smith et al. 2022), dedicated representations for MWEs as single-tokens (Cordeiro et al. 2019; Phelps et al. 2022), and different combinations of layers and weighting schemes (Reimers and Gurevych 2019a; Vulić, Korhonen, and Glavaš 2020; Rogers, Kovaleva, and Rumshisky 2020) may generate better results in downstream tasks. Additional configurations were also extensively analyzed and as they produced qualitatively similar results, they are not included in the article.

18 `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased`.
19 `https://platform.openai.com/docs/guides/embeddings/what-are-embeddings`.
20 `https://github.com/flairNLP/flair`.
21 `https://github.com/huggingface/transformers`.
22 In our preliminary experiments, we tested various pooling strategies, including max pooling, min pooling, the CLS token from BERT, concatenation, and mean pooling. The performance was similar across these methods, but to maintain simplicity and avoid complications from variable vector lengths, we chose mean pooling for the reported experiments.
23 Extensive evaluation of the individual layers and their combination were performed, but as the results follow the trend of those reported here, they are not included in the article.

3.3 Measuring Idiomatic Meaning

The general premises of this work, shared by many similar investigations, are the following:

1. **Vector embeddings approximate meaning.** We assume that the vector embeddings produced by the models are representations of usages in a semantic space that can approximate meaning. Because there is no absolute reference frame for meaning in that space, the meaning of a word/sentence is always relative and it is evaluated in terms of its similarity to other relevant words/sentences in the same semantic space.
2. **Word/multiword/sentence representations are the combinations of the (sub)word representations.** We adopt as the meaning of a word, multiword expression, or of a sentence, the compositional combination of its components. In this article we focus on the additive combination, summing/averaging the vector embeddings of each token in the word, expression, or sentence, and summing/averaging the vector embeddings of the relevant layers, when more than one layer is used.
3. **Similarity of meanings can be approximated by similarity of vectors.** Similarity is a measure of the proximity between two vector embeddings. Throughout this article we adopt cosine similarity as the similarity metric.[24] As contextualized models provide different vector representations for the same linguistic expression in different contexts, its vector representations would be found among different clusters of meaning as it transitions between its meanings in different sentences, and this would be reflected by the similarity measures.

3.3.1 The Probing Strategies. To evaluate how word representation models deal with idiomaticity, we propose a probing strategy where a target item in a sentence, in this case an NC, is systematically replaced by a set of different paraphrases or probes (P), forming the minimal pairs discussed in Section 3.1. We then use similarity measures to compare the representation for the sentence before and after replacing NC by P. Given the focus on idiomaticity we select a set of probes specifically for the expected changes in meaning they would induce in a sentence, and we refer to these potential changes in meaning as Linguistic Predictions (LPs). If the representations generated by a model reflect these predictions, passing the probing tests, then we consider that particular model as capturing to some extent the idiomatic meaning in NCs. The idiomatic probes are defined as follows, where *Comp* is the average human annotation compositionality score:

- P_{Syn} - The true synonym. The replacement is a single word or a two word compositional noun compound that represents closely the meaning of the target NC, forming the minimal pair NC_{Syn}. Linguistic Prediction: After the replacement, the resulting sentence should be a near perfect

24 Other compositional operations and measures of distance were also used during these analyses, but with qualitatively similar results, and have been omitted from the article.

paraphrase of the original sentence. Therefore high similarities are expected for all minimal pairs independently of the degree of compositionality of the target NC, from the more idiomatic *grey matter* (and *brain*) to the more literal *economic aid* (and *financial assistance*), with no correlation expected with *Comp*.

- P_{Comp} - The partial expression. The replacement is one of the component words of the target compound, and in particular we consider the one that preserves most of the meaning, forming the minimal pair NC_{Comp}. Linguistic Prediction: The resulting sentence may preserve some of the original meaning for more compositional cases, but not for idiomatic cases. Therefore, high similarities are only expected between minimal pairs involving compositional and partly compositional cases (e.g., *economic aid* and *aid*, *crocodile tears* and *tears*, but not for *wet blanket* and *blanket* or *wet*), with some correlation expected with *Comp*.
- $P_{WordsSyn}$ - The literal synonyms of the individual NC components. The replacement is a two-word expression formed from frequent out-of-context synonyms for each of the component words of an NC when considered independently, forming the minimal pair $NC_{WordsSyn}$. Linguistic Prediction: After replacement, the resulting sentence may not preserve the meaning of the original sentence, especially for more idiomatic cases. Therefore, higher similarities are only expected for minimal pairs involving more compositional NCs (e.g., *wedding day* and *marriage date* but not *eager beaver* and *restless rodent*), with a high correlation expected with *Comp*.
- P_{Rand} - The random replacement controlled by frequency. The replacement is a two word expression where the words are chosen to match the frequencies of the components of the target NC, forming the NC_{Rand} minimal pair. Linguistic Prediction: After replacement, the resulting sentence should not preserve the meaning of the original sentence, independently of the level of idiomaticity of the original NC (e.g., for *police car* and *supermarket city*), with no correlation expected with *Comp*.

For a more in-depth analysis of expected changes in meaning, we follow Garcia et al. (2021b), comparing representations both at a macro sentence level and also at a micro NC level, analyzing the representations of NC (and its variants P) extracted from the context of the sentence. Although any differences in meaning should be reflected both at sentence and at NC representation levels (only magnified in the latter), this comparison aims to highlight the impact of the level of granularity used when analyzing idiomaticity.[25]

3.4 Metrics

3.4.1 The Human Compositionality Score (Comp). Assuming a list of N NCs, chosen to provide balanced test scenarios of different levels of idiomaticity, we denote NC_α with $\alpha = 1, \ldots, N$, the different NCs to be evaluated. The meaning of these NCs is exemplified by a set of $N \times M$ sentences $Sent_{\alpha\beta}$ with $\alpha = 1, \ldots, N$ and $\beta = 1, \ldots, M$ the sentence

25 Our prior work reveals that only looking at similarities at sentence level when comparing the representations of the original and the resulting sentences may not accurately reflect their differences (Garcia et al. 2021b).

index. The dataset contains $M = 3$ naturalistic sentences to exemplify the use of each NC (see section 3.1), with each sentence annotated by human judges according to the compositionality of the target NC in the sentence. The resulting scores are denoted $\text{Comp}_{\alpha\beta j}$, with $\alpha = 1, \ldots, N$, $\beta = 1, \ldots, M$, and $j = 1, \ldots, A_{\alpha\beta}$ where $A_{\alpha\beta}$ is the number of annotators for sentence $\text{Sent}_{\alpha\beta}$. $\text{Comp}_{\alpha\beta j}$ are integer values derived from a Likert scale and range from 0 (totally idiomatic) to 5 (totally compositional). We define the compositionality score for a specific NC_α as the average of the annotations for sentences $\text{Sent}_{\alpha\beta}$,

$$\text{Comp}(\text{NC}_\alpha) = \left\langle \left\langle \text{Comp}_{\alpha\beta j} \right\rangle_{Annot} \right\rangle_{Sent} \tag{1}$$

where $\langle \cdots \rangle_{Sent}$ are averages on sentences and $\langle \cdots \rangle_{Annot}$ averages on annotations. These average values are the gold standard in this work.

3.4.2 The Similarity Score (Sim). Probing the meaning of a compound NC_α in a sentence $\text{Sent}_{\alpha\beta}$ requires the generation of a new set of modified sentences $\text{Sent}_{Pi\,\beta\gamma}$ where NC_α is replaced by a probe P*i* (discussed in Section 3.3.1). We measure the effect of the probe substitution directly from the similarity between the representation of the original expression, X, and the representation of the new expression after substitution, Y, adopting, throughout this article, cosine similarity as a measure of the similarity of meaning between two vector embeddings.

$$\text{cossim}(X, Y) = \frac{\epsilon_X \cdot \epsilon_Y}{||\epsilon_X||\ ||\epsilon_Y||} \tag{2}$$

where ϵ_X and ϵ_Y are vector embeddings of D components, $\epsilon_X \cdot \epsilon_Y$ their inner products, and $||\epsilon_X||$, $||\epsilon_Y||$ are their L2 norms. Therefore the average similarity between the original expression and the probe-modified expression for a given NC can be defined as

$$\text{Sim}(\text{P}i, \textit{Target}) = \langle\ \text{cossim}(\textit{expr}(\text{P}i), \textit{expr}(\text{NC}))\ \rangle_{Pi} \tag{3}$$

where *expr*(NC) is the target NC expression, and *expr*(P*i*) is the expression where NC is replaced by a probe of the type P*i*, and $\langle \cdots \rangle_{Pi}$ means the average over possible substitutions of this type. We use more than one substitution only for random probes (P_{Rand}); for all other probes a single substitution is reported.

3.4.3 The Affinity Score (Aff). Cosine similarity measures are not sensitive enough to capture subtle meaning differences, especially in anisotropic representation spaces (Ethayarajh and Jurafsky 2021). Additionally, there may be a "horizon of interest," beyond which word connections lose meaningful inference (Karlgren and Kanerva 2021), which may be a challenge for representing idiomatic expressions, as the necessary context may lie within this critical boundary. Investigating measures that account for anisotropic spaces and for a horizon of interest are interesting avenues for future research for improving idiomaticity detection. In this article, we propose a comparative measure that we refer to as Affinity (Assessment of Feature Familiarity and Idiomatic Nuance by Interpreting Target Yielding), that identifies which between two representations is the closest to a given target representation.

Given a target representation *Target* and two possible probes P*i* and P*j*, the Affinity is defined as:

$$\text{Aff}(\text{P}i, \text{P}j \mid \textit{Target}) = \text{Sim}\left(\text{P}i, \textit{Target}\right) - \text{Sim}\left(\text{P}j, \textit{Target}\right) \tag{4}$$

Affinities closer to 1 or larger indicate a greater similarity between the target and the first probe P*i*, values closer to -1 or lower indicate the opposite situation where the target is more similar to the second probe P*j*, and values near zero indicate no preference. Given the focus of this article on detecting idiomaticity in representations, we measure the Affinities involving the minimal pairs defined in Section 3.3.1, analyzing if, as expected, the target NCs have higher similarities with probes with substitutions that maintain the original meaning as P*i* than with probes that involve potential changes in meaning as P*j*. In particular:

- Affinity $A_{Syn|WordsSyn} = \text{Aff}(\text{P}_{Syn}, \text{P}_{WordsSyn}|\text{NC})$ measures if the target NCs have greater similarities with their gold synonyms than with synonyms of the individual components (e.g., *eager beaver* with *hardworking person* than with *restless rodent*).
- Affinity $A_{Syn|Rand} = \text{Aff}(\text{P}_{Syn}, \text{P}_{Rand}|\text{NC})$ compares if the target NCs display greater similarities to their gold synonyms than to random substitutions.

Our Affinity measure extends traditional forced-choice evaluations (Warstadt et al. 2020) by quantifying the degree of similarity preference between two options. Unlike binary choices, Affinity provides a continuous measure of relative similarity, offering a more detailed assessment of how well models capture idiomatic meanings. This nuanced analysis reveals subtle differences in model performance, providing deeper insights into the representation of idiomatic expressions.

3.4.4 The Scaled Similarity Score (Sim_R). Even though Affinity is an advance over the simple similarity measure, additional measures may still need to be adopted for models if the average similarity between two random embeddings is larger than zero, as Affinities will tend to have small values even for very dissimilar probes (see discussion). To address this issue, we propose a scaled version of the similarity:

$$\text{Sim}_R(\text{P}i|\textit{Target}) = \left\langle \frac{\text{Sim}(\text{P}i, \textit{Target}) - \text{Sim}(\text{P}_{Rand}, \textit{Target})}{1 - \text{Sim}(\text{P}_{Rand}, \textit{Target})} \right\rangle_{Sent} \tag{5}$$

where $\langle \cdots \rangle_{Sent}$ denotes the average over the M sentences that illustrate the meaning of a particular NC and P_{Rand} is a random substitution. The scaled similarity is defined such that if replacing the target with a probe P*i* results in cosine similarities close to one ($\text{Sim}(\text{P}i, \textit{Target}) \approx 1$), the scaled similarity is also close to one, $\text{Sim}_R \approx 1$. Conversely, if the replacement is similar to a random replacement ($\text{Sim}(\text{P}i, \text{Target}) \approx \text{Sim}(\text{P}_{Rand}, \textit{Target})$),

then $\text{Sim}_R \approx 0$. This approach is equivalent to a max-min normalization[26] in the anisotropic space of a model.

In particular, given the focus on idiomaticity, we focus as before on two similarities:

- $\text{Sim}_{R|Syn} = \text{Sim}_R(\text{P}_{Syn}|\text{NC})$, where the NCs are replaced by gold synonyms and no changes in meaning are expected, therefore $\text{Sim}_{R|Syn}$ should be close to 1.
- $\text{Sim}_{R|WordsSyn} = \text{Sim}_R(\text{P}_{WordsSyn}|\text{NC})$, where the NCs are replaced by synonyms of the individual components and greater changes in meaning, and therefore small values (~ 0) of Sim_R, are expected for more idiomatic cases.

3.4.5 The Correlation Measure (ρ). Finally, to assess the impact of idiomaticity for the probe substitutions we use Spearman correlation between the different measurements and the gold standard human annotations of compositionality (Comp) given by Equation (1).

4. Probing for Idiomaticity

4.1 Are the Representations of the NCs and Their Synonyms Similar?

A first indication of the successful modeling of idiomaticity is if a model assigns similar representations for the target NCs and for their synonyms, regardless of their level of compositionality. We measure this using the minimal pairs of probe P_{Syn} and compare it with less appropriate substitutions represented by the other probes Pj. The distribution of similarities obtained for each of the probes is shown in Figure 1, along with the correlations of these similarities with the human compositionality scores for the NCs at sentence (ρ_{Sent}) and NC (ρ_{NC}) levels, in Tables 2 and 3. Considered in isolation, the high similarity scores for P_{Syn} at sentence level (close to 1 for naturalistic sentences, and mostly above 0.75 for neutral sentences, Figure 1 (P_{Syn})) seem to suggest that these models are able to capture idiomaticity. However, when compared against the scores for the minimal pairs of the other probes a different story emerges.

When the components of a target NC are replaced with one of their component words (Figure 1 (P_{Comp})) or with the synonyms of their component words (Figure 1 ($\text{P}_{WordsSyn}$)), lower similarities should be observed between the minimal pairs since, although these substitutions could preserve some of the meaning of the more compositional cases, they would not do so for the more idiomatic cases. Moreover, random substitutions should lead to even lower similarities for all NCs (Figure 1 (P_{Rand})), since they could result in nonsensical sentences. This expected staggered pattern of similarities, highest for P_{Syn}, moderate for P_{Comp} and $\text{P}_{WordsSyn}$, and lower for P_{Rand}, illustrated in Figure 1 (Ideal Values), does not seem to be reflected by a visible reduction of the similarities at sentence

26 Given a value x in a dataset, the max-min normalization of x is calculated as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

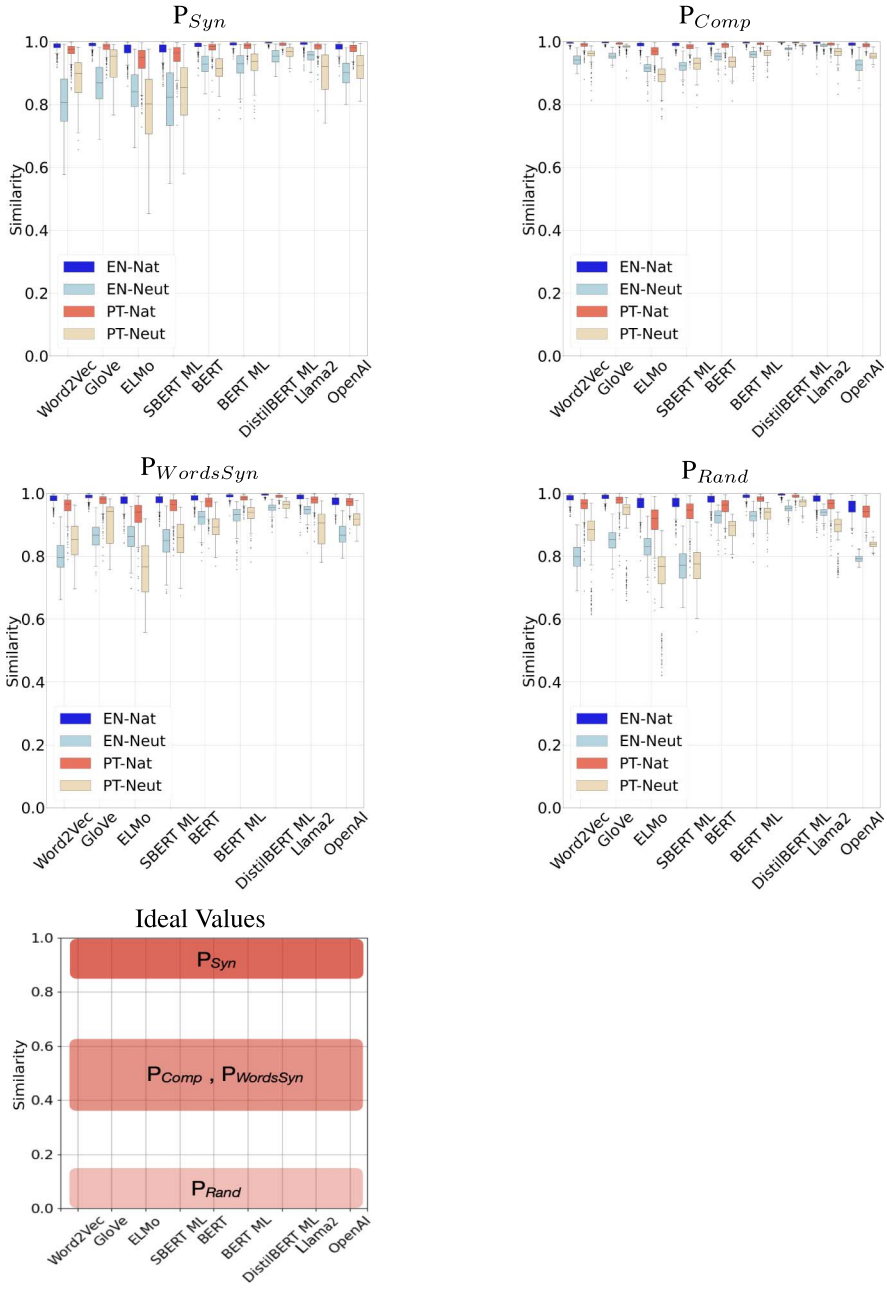

Figure 1
Distribution of cosine similarities between the minimal pairs at sentence level, with the original NC and the probe-modified substitution for English (EN, in blue) and Portuguese (PT, in orange), with naturalistic (Nat) sentences in darker shade and neutral (Neut) in lighter. The lower panel (Ideal Values) is an illustration of similarity values ideally expected for the different probes. The means and standard deviations are in Table 10 in the Appendix.

Table 2
Spearman ρ correlation between cosine similarities and human compositionality judgments (Comp) at sentence level. Only significant results ($p \leq 0.05$) are displayed, for P_{Syn}, P_{Comp}, $P_{WordsSyn}$, and P_{Rand}, for English (EN) and Portuguese (PT), naturalistic (Nat) and neutral (Neut) sentences.

ρ_{Sent}	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilB ML	Llama2	OpenAI
P_{Syn}									
EN-Nat	0.30	0.31	0.43	0.47	0.39	0.51	0.38	0.15	0.41
EN-Neut	0.60	0.58	0.55	0.60	0.51	0.53	0.56	0.37	0.54
PT-Nat	0.18	0.13	0.33	0.31	0.32	0.29	0.20	0.27	0.46
PT-Neut	0.31	0.22	0.37	0.46	0.35	0.30	0.31	0.31	0.51
P_{Comp}									
EN-Nat	–	–	–	–	0.17	–	–	–	0.37
EN-Neut	0.19	0.29	–	–	–	–	−0.12	–	0.51
PT-Nat	–	−0.12	0.12	–	0.16	–	−0.15	–	0.21
PT-Neut	0.13	–	0.17	–	–	−0.14	–	–	0.27
$P_{WordsSyn}$									
EN-Nat	–	–	–	–	–	–	–	–	0.21
EN-Neut	0.19	–	–	−0.13	−0.15	–	–	0.20	0.13
PT-Nat	−0.12	−0.19	–		–	–	−0.14	–	0.11
PT-Neut	–	−0.13	–	–	–	–	–	–	0.17
P_{Rand}									
EN-Nat	–	−0.11	−0.13	−0.16	−0.27	−0.11	−0.18	−0.11	–
EN-Neut	0.11	–	−0.31	−0.36	−0.29	–	−0.13	–	–
PT-Nat	−0.17	−0.20	−0.13	−0.11	−0.14	−0.12	–	−0.18	–
PT-Neut	0.13	−0.17	−0.14	−0.11	−0.22	−0.11	–	–	–

level, in Figure 1. In fact, even random substitutions seem to result in high sentence similarities, even if they are not as high as the other substitutions.

Another important point relates to the correlation of these similarities for the different NCs with human judgments for compositionality. It is expected that there would be almost no correlation for the similarities derived from P_{Syn} and P_{Rand}, and a moderate correlation for P_{Comp} and $P_{WordsSyn}$, as they may be more acceptable for compositional NCs than for idiomatic ones. However, this expected pattern is not observed in the results presented in Table 2. For most models, ρ_{Sent} (P_{Syn}) shows moderate correlation, while ρ_{Sent} (P_{Comp}) and ρ_{Sent} ($P_{WordsSyn}$) are either weak or non-significant.

Because in these minimal pairs only the target NCs and their substitutions change, the high similarities found may be an effect of the lexical overlap between the sentences of a minimal pair. Indeed, comparing the output of the models in relation to sentence lengths for naturalistic sentences, there is a significant moderate to strong positive correlation between the lexical overlap and the cosine similarity of a pair, for both English and Portuguese (Table 4), where the greater the overlap between the sentences, the higher their similarity. This can also explain the higher similarities observed for naturalistic than for neutral sentences, since the former are longer than the latter with a higher lexical overlap proportional to the length of the sentence: average sentence length for naturalistic sentences is 23.4 words for English (lexical overlap $> 91\%$) and 13.0 words for Portuguese (overlap $> 84\%$), while for the neutral sentences it is five words (overlap $> 60\%$) for

both languages.[27] It could be argued that the influence of lexical overlap is expected, given that a compositional representation is used for sentences, where the embeddings for each token are added. However, although this holds true for static models, it may not necessarily apply to contextualized models. In contextualized models, it is expected that each token/word would interact with others via attention heads, and if the model accurately captures semantics, all tokens/words will adjust to the context of the sentence as a whole. Ideally, even with the simple compositional representation of the sentence, we would anticipate that a correct sentence would exhibit low similarity with the mostly nonsensical sentences produced by random probes. Even though similarities coming from contextualized models seem to present lower correlations with sentence size, still lexical overlap appears to dominate across all types of models.

To minimize the effect of the lexical overlap in the similarities, we now focus our analyses only on the similarities among the tokens representing the NCs and their substitutions in the context of the target sentences. In this case, lower similarities were obtained for all probes and all models compared with those at sentence level (Figure 2 vs. Figure 1). This is even the case for similarities for the NCs and their synonyms (P_{Syn}), which are centered around the same values as those for the NCs and synonyms of the individual components ($P_{WordsSyn}$). Those for the random replacements (P_{Rand}) also follow this trend. They are all lower than those for the NCs and only one NC component (P_{Comp}). In fact, similarities for the gold standard synonyms are lower than for many of the other probes, regardless of the extent to which the original NC meaning is changed, as probes P_{Comp} to P_{Rand} involve some change in meaning whereas P_{Syn} does not. Finally, there is more variation displayed among the models, as there are lower similarities for static than for most contextualized models. Overall, the resulting similarities at NC level do not follow the expected patterns for representing idiomaticity, illustrated in Figure 2 (Ideal Values). The same holds true for their correlations with human judgments. In line with what occurs at the sentence level, the similarities at the NC level exhibit correlations that contradict linguistic expectations. In particular, it is expected that true synonymous substitutions work well across the idiomatic-compositionality spectrum. Therefore, no correlation should be expected for P_{Syn}, while for P_{Comp} and $P_{WordsSyn}$, a moderate correlation is expected and no correlation for P_{Rand}. However, Table 3 indicates that for most models, $\rho_{NC}(P_{Syn}) > \rho_{NC}$ (P_{Comp} or $P_{WordsSyn}$) with the latter being either weak or not significant.

In the next section we analyze if, at least at a detailed level, the similarities between NCs and their synonyms are mostly higher than of other alternatives.

4.2 Are the Representations of the NCs and Their Synonyms Relatively More Similar When Compared to Other Alternatives?

If a model accurately represents idiomaticity, the representation of a given NC should be more similar to its synonym than to other alternatives, including distractors and random representations. Using the proposed comparative measures of Affinity (introduced in Section 3.4.3), we now assess whether the models we are evaluating are able to reliably distinguish between a substitution that preserves meaning (P_{Syn}) from those that do not ($P_{WordsSyn}$ for more idiomatic NCs, P_{Rand} for all NCs). The results from the previous section demonstrated that, on average, the models do not seem to represent idiomaticity correctly.

27 We also compared longer neutral contexts with 10 words for English (> 80%), and 9 words for Portuguese (> 77%), and found similar results.

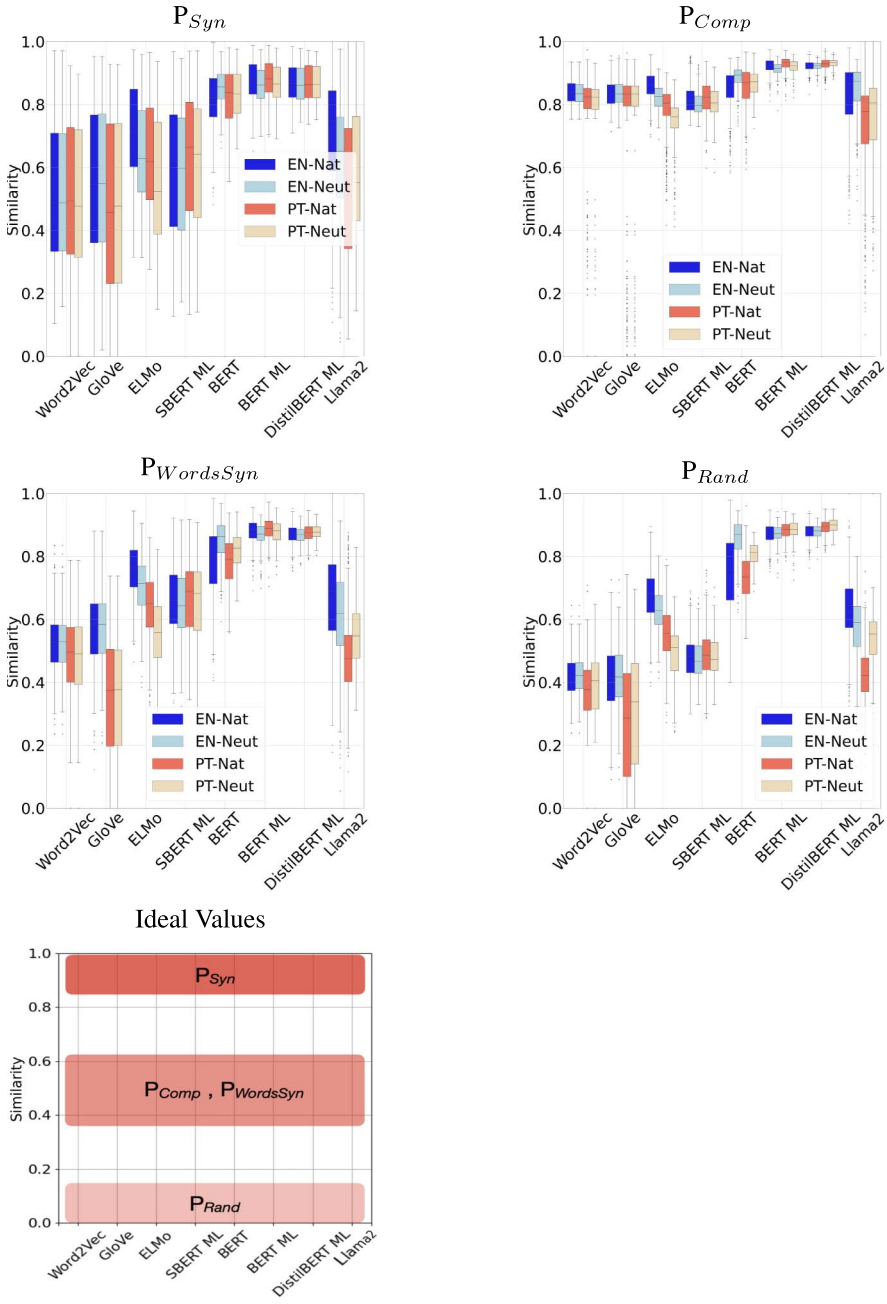

Figure 2
Distribution of cosine similarities between the minimal pairs at NC level, with the original NC and the probe-modified substitution for English (blue) and Portuguese (orange), with naturalistic sentences in darker shade and neutral in lighter. The lower panel (Ideal Values) is an illustration of similarity values ideally expected for the different probes. The means and standard deviations are in Table 10 in the Appendix.

Table 3
Spearman ρ correlation between cosine similarities and human compositionality judgments (Comp) at NC level. Only significant results ($p \leq 0.05$) are displayed, for P_{Syn}, P_{Comp}, $P_{WordsSyn}$, and P_{Rand}, for English (EN) and Portuguese (PT), naturalistic (Nat) and neutral (Neut) sentences.

ρ_{NC}	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilB ML	Llama2
P_{Syn}								
EN-Nat	0.62	0.62	0.60	0.66	0.39	0.67	0.58	0.36
EN-Neut	0.62	0.61	0.60	0.65	0.34	0.58	0.54	0.37
PT-Nat	0.45	0.40	0.47	0.48	0.57	0.44	0.39	0.37
PT-Neut	0.43	0.41	0.47	0.48	0.48	0.35	0.37	0.31
P_{Comp}								
EN-Nat	0.20	0.45	0.17	0.34	0.17	0.35	0.26	0.15
EN-Neut	0.20	0.44	0.23	0.28	−0.31	–	–	0.12
PT-Nat	0.30	0.20	0.29	0.11	0.43	0.16	–	0.22
PT-Neut	0.27	0.18	0.21	–	0.24	–	–	0.13
$P_{WordsSyn}$								
EN-Nat	–	0.18	–	–	−0.40	0.21	0.15	0.29
EN-Neut	0.11	0.18	–	–	−0.40	–	–	0.22
PT-Nat	–	–	0.13	–	0.17	0.11	–	–
PT-Neut	–	–	–	–	0.14	–	–	–
P_{Rand}								
EN-Nat	0.11	0.18	−0.18	−0.23	−0.58	−0.22	−0.29	–
EN-Neut	0.12	0.18	−0.21	−0.20	−0.49	–	−0.24	0.13
PT-Nat	–	–	–	–	–	–	–	–
PT-Neut	–	–	–	–	–	−0.11	–	–

Table 4
Spearman ρ correlation between naturalistic sentence length and cosine similarity, $p \leq 0.05$, for P_{Syn}, P_{Comp}, $P_{WordsSyn}$, and P_{Rand}.

	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilB ML	Llama2	OpenAI
P_{Syn}									
EN-Nat	0.71	0.71	0.49	0.49	0.58	0.53	0.67	0.46	0.44
PT-Nat	0.66	0.59	0.42	0.46	0.48	0.57	0.65	0.51	0.26
P_{Comp}									
EN-Nat	0.82	0.86	0.74	0.80	0.75	0.78	0.89	0.55	0.52
PT-Nat	0.72	0.80	0.67	0.58	0.63	0.72	0.83	0.59	0.50
$P_{WordsSyn}$									
EN-Nat	0.86	0.87	0.70	0.74	0.75	0.81	0.87	0.54	0.60
PT-Nat	0.74	0.70	0.58	0.62	0.69	0.67	0.78	0.60	0.46
P_{Rand}									
EN-Nat	0.87	0.88	0.77	0.85	0.82	0.85	0.87	0.62	0.80
PT-Nat	0.77	0.79	0.73	0.74	0.87	0.76	0.81	0.62	0.68

For instance, Figure 2 shows that probe P_{Comp} yields larger average similarities than probe P_{Syn}, and that P_{Syn} and $P_{WordsSyn}$ have similar averages, but with P_{Syn} exhibiting more variance. Both results are incompatible with a good idiomatic representation. Affinity will allow us to verify this on a per-NC basis. In particular, we compare the Affinities for NCs

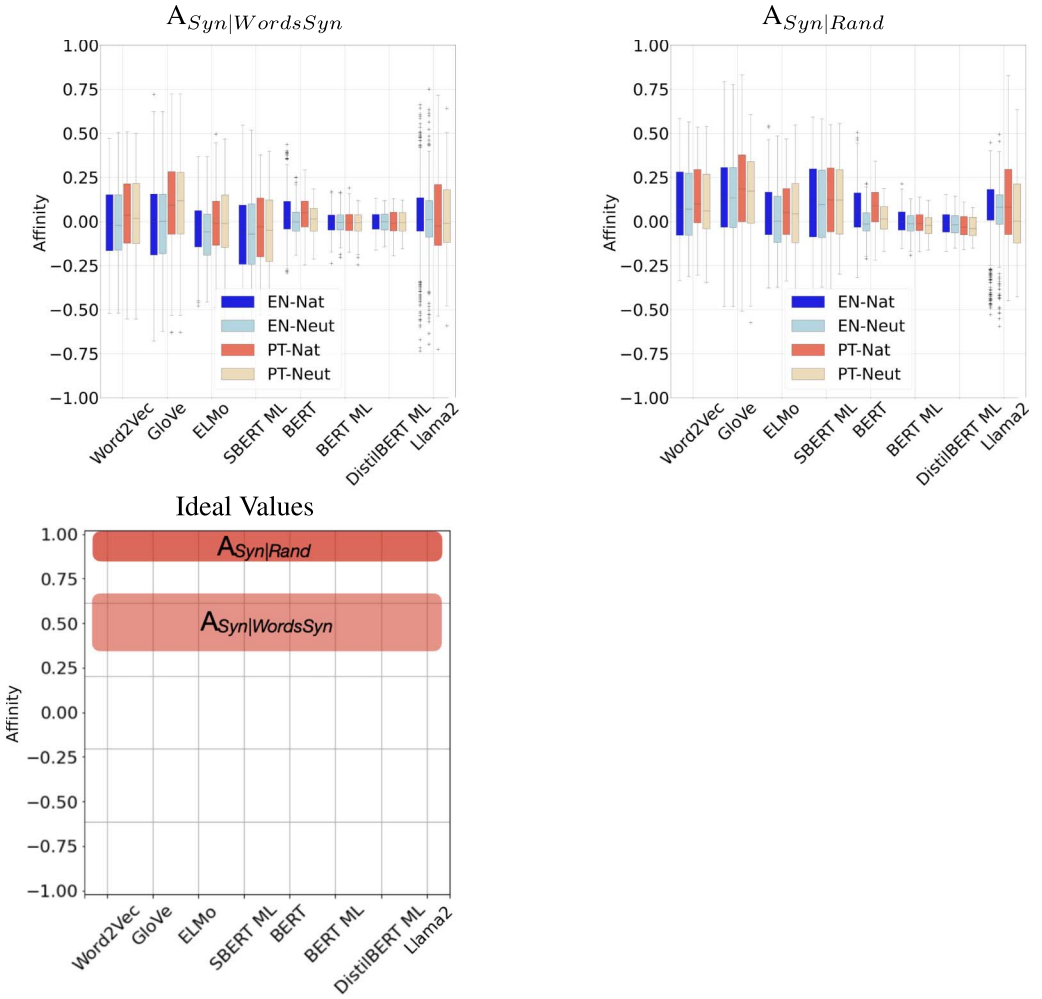

Figure 3
Affinity at the NC level for English (blue) and Portuguese (orange), with naturalistic sentences in darker shade and neutral in lighter. The lower panel (Ideal Values) is an illustration of values ideally expected for the different Affinities. The means and standard deviations are in Table 12 in the Appendix.

and their synonyms against the synonyms of their individual components ($A_{Syn|WordsSyn}$), and against random substitutions ($A_{Syn|Rand}$), with the expected affinity ranges shown in Figure 3 (Ideal Values).

First of all, comparing against synonyms of the individual components (Figure 3 ($A_{Syn|WordsSyn}$)) on the whole the models display comparable abilities in term of averages, around 0 for all models, but differ to some extent in their variances. As the Affinities obtained are mostly neutral (around 0) the models do not display the higher similarities between the NCs and their gold synonyms to the extent that would be expected. Moreover, this holds even for random replacements (Figure 3 ($A_{Syn|Rand}$)), where some models display small positive averages, but are far from the expected ideal (Figure 3 (Ideal Values)).

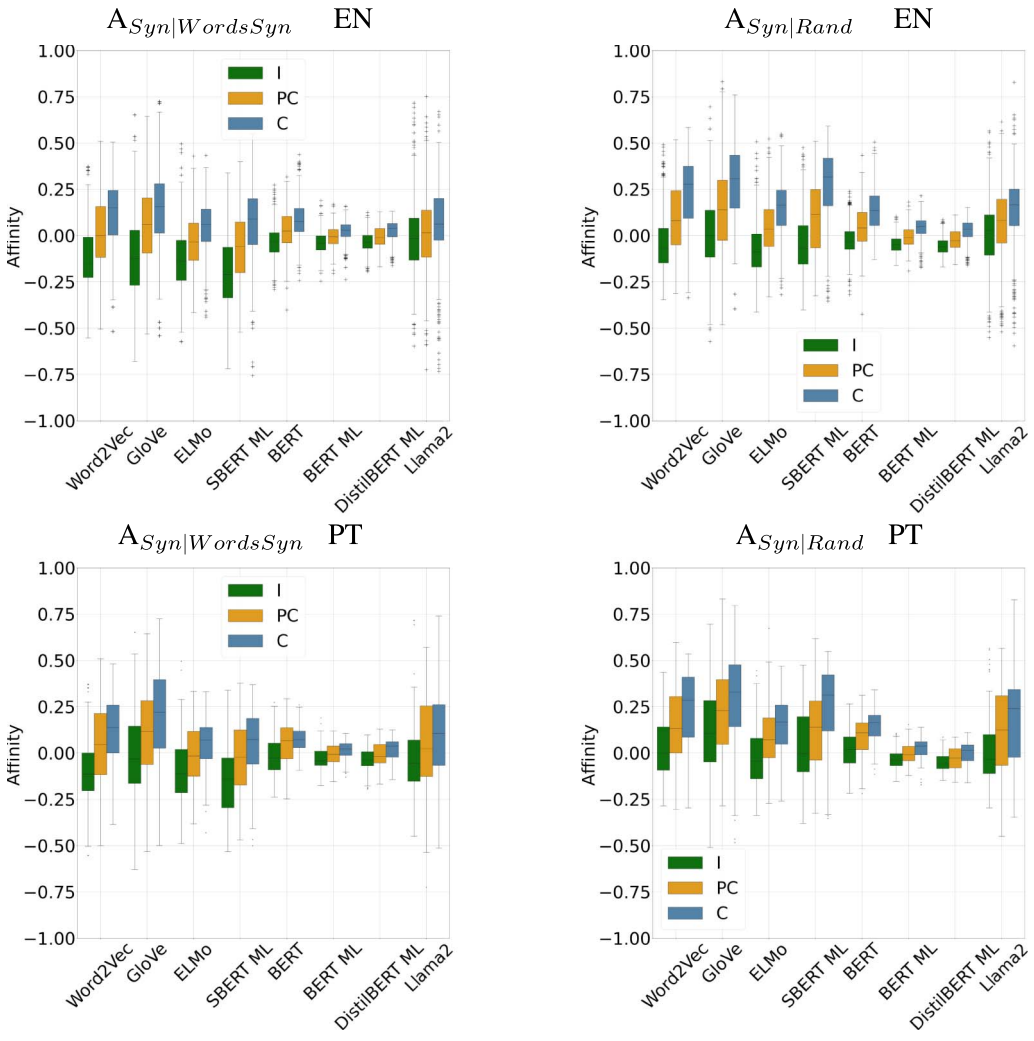

Figure 4
Affinity by idiomaticity Class at NC level for English (EN) and Portuguese (PT) naturalistic sentences. Idiomatic (I) in green, partly compositional (PC), in yellow, and compositional NCs (C) in blue.

The relatively important variances in Figure 3 call for an analysis of the Affinities according to idiomaticity level. This is displayed in Figure 4 for English naturalistic sentences where the classification of NCs as compositional (C), partly compositional (PC), and idiomatic (I) from Garcia et al. (2021a) is adopted. A striking pattern emerges in all the figures. For each model, its distribution of Affinities splits into three distinct distributions with similar variances but different averages ordered according to compositionality. Compositional NCs exhibit higher Affinities than partly compositional NCs, and the latter show higher Affinities than idiomatic NCs. This is confirmed by the correlation analysis in Table 5, with most models displaying significant weak to moderate correlations between Affinities and human compositionality judgments, for all Affinities types, including

Table 5
Spearman ρ correlation between the *Affinity* and human judgments for English and Portuguese for naturalistic (Nat) and neutral (Neut) sentences. Non-significant ($p > 0.05$) results omitted from the table. Although these values shown are for the correlations at the NC level, the correlations at the sentence level are comparable.

	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
$A_{Syn\|WordsSyn}$								
EN-Nat	0.58	0.52	0.55	0.58	0.57	0.55	0.51	0.17
EN-Neut	0.58	0.52	0.54	0.58	0.53	0.49	0.50	0.23
PT-Nat	0.43	0.37	0.39	0.38	0.35	0.34	0.35	0.27
PT-Neut	0.44	0.37	0.41	0.40	0.34	0.32	0.36	0.31
$A_{Syn\|Rand}$								
EN-Nat	0.60	0.54	0.63	0.66	0.69	0.68	0.61	0.39
EN-Neut	0.59	0.53	0.63	0.64	0.59	0.55	0.57	0.36
PT-Nat	0.48	0.41	0.54	0.49	0.51	0.48	0.42	0.36
PT-Neut	0.44	0.41	0.50	0.47	0.46	0.46	0.41	0.33

neutral sentences and Portuguese data.[28] This contradicts what was generally expected: Affinity $A_{Syn|WordsSyn}$ values should exhibit a negative correlation with compositionality, while Affinity $A_{Syn|Rand}$ should show no correlation at all.

These results suggest that representations of idiomatic NCs may not be accurately incorporating their meanings, since NCs are not closer to their synonyms than to other alternatives, even if they are random. Moreover, the more idiomatic NCs seem to be more similar to synonyms of their individual components, which suggests that the surface clues about their individual components may be playing a greater role in driving these similarities, even in contextualized models. This result remains valid even after removing compounds from the dataset that have lexical overlaps with the NC_{Syn} produced by the annotators (see Table 14 in the Appendix).

4.3 Can a More Meaningful Similarity Measure be Found to Unveil NC Meaning?

If random substitutions that should result in Affinities around 1 ($A_{Syn|Rand}$ in Figure 3 (Ideal Values)) result instead in values mostly below 0.5, the latter may represent the de facto upperbound for Affinity for these models. In this case, a rescaling factor may need to be adopted that could magnify meaningful similarity values. To implement this, we propose the Scaled Similarity (Equation (5)), which takes into account the threshold defined by random replacements when calculating the cosine similarities between the target representation and a given probe. In this section we explore the behavior of $Sim_{R|Syn}$ and $Sim_{R|WordsSyn}$ defined in Section 3.4.

The Scaled Similarity values (Figure 5) reveal, even more than the Affinities, the equivalences displayed by the behavior of these models, with Sim_R being able to abstract away from the particularities of the spaces defined by each of these models. Interestingly, comparing different levels of contextualization (e.g., static models on the left and contextualized on the right half of Figure 5) the Scaled Similarities produced by static models

28 We omitted the equivalent of Figure 4 for neutral sentences and Portuguese data due to their visual similarity to the English naturalistic version.

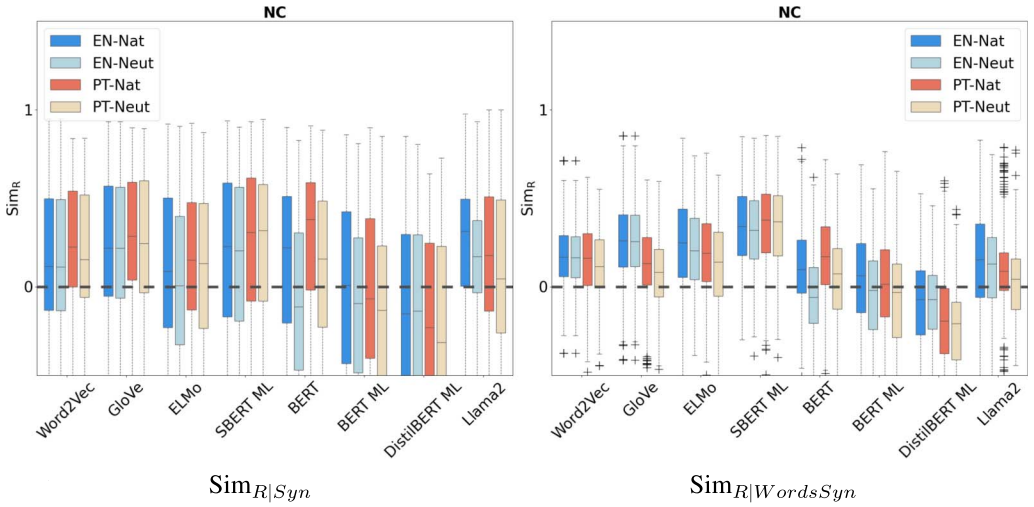

Figure 5
Average Scaled Similarity when the original NCs are replaced by gold synonyms ($Sim_{R|Syn}$) or by the synonyms of component words ($Sim_{R|WordsSyn}$), in relation to random substitutions. English (blue) and Portuguese (orange), with naturalistic sentences in darker shades and for neutral in lighter. The means and standard deviations are in Table 13 in the Appendix.

like Word2Vec and GloVe are comparable to those by a contextualized large language model like Llama2. These results seem to hold independently of how informative the context is (naturalistic vs. neutral sentences), with NC representations from naturalistic sentences displaying no real advantage over those from neutral sentences. Overall, these results suggest that the informative contexts provided by the naturalistic sentences may not yet be adequately incorporated even by the larger contextualized models.

Inspecting the $Sim_{R|Syn}$ values according to idiomaticity level (Figure 6), the models display lower Scaled Similarities for the more idiomatic than for the more compositional NCs, confirming what was already indicated by the Affinities that the models are less able to capture the idiomatic meanings and as a consequence the expected high similarities with their gold standard synonyms are not observed. This is further confirmed by analyzing the values obtained for the synonyms of the individual components (Figures 5 and 7) with the distributions of $Sim_{R|WordsSyn}$ values having similar averages but considerably lower variances when compared to $Sim_{R|Syn}$, whereas the expected result would be the opposite: lower averages and variances for $Sim_{R|Syn}$. In fact the average and standard deviation for the ratio $Sim_{R|Syn}/Sim_{R|WordsSyn}$ (Figure 8) show that the ratio oscillates around 1, which indicates that as a whole the models respond similarly to P_{Syn} and $P_{WordsSyn}$ substitutions. In addition, the average values and variances for $Sim_{R|WordsSyn}$ do not depend on the degree of compositionality of the target NC (Figure 7 for $Sim_{R|WordsSyn}$ and Figure 9 for the average and standard deviation for the ratio $Sim_{R|Syn}/Sim_{R|WordsSyn}$, according to idiomaticity level). The whole picture indicates that for all models (contextualized or not) replacing the NC by literal synonyms of the component words is more effective (produces higher similarities) than using their gold synonyms. In particular, for idiomatic NCs, we observe that $Sim_{R|Syn} < Sim_{R|WordsSyn}$, which indicates that the lexical similarity (as opposed to the similarity of meaning) is still a dominant factor in the representations even for the contextualized models, and provides additional confirmation for the possibility that the component words of an

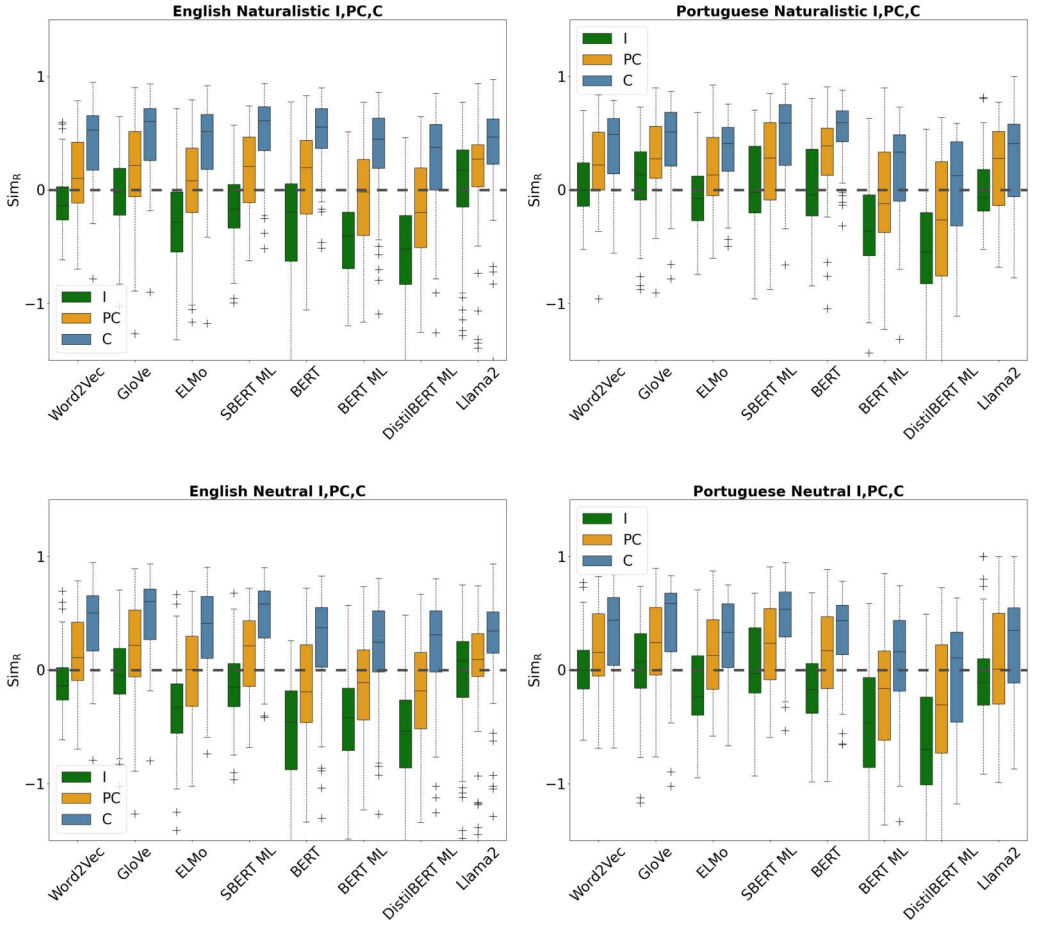

Figure 6
Sim_{RP1} per compositionality class: green for idiomatic (I), yellow for partly compositional (PC) and blue for compositional (C), in English (EN) and Portuguese (PT), in naturalistic and neutral sentences.

idiomatic NC may be represented individually and combined compositionally by these models.

Table 6 summarizes these results in terms of the Spearman correlations between Sim_R values and the human judgments for compositionality. It shows that, considering the different models, $Sim_{R|Syn}$ is almost always moderately correlated with the compositionality score: The higher the compositionality score, the higher the value $Sim_{R|Syn}$, and consequently the more the meaning is preserved with a P_{Syn} substitution. $Sim_{R|WordsSyn}$, in contrast, rarely displays significant correlation with compositionality score. As discussed above, this is a demonstration that the idiomatic meaning is not captured by these models, not even by those that are contextualized. As with Affinities, this discrepancy in the behavior of Scaled Similarities persists even after removing compounds from the dataset that have lexical overlaps with the NC_{Syn} produced by the annotators (see Table 15 in the Appendix).

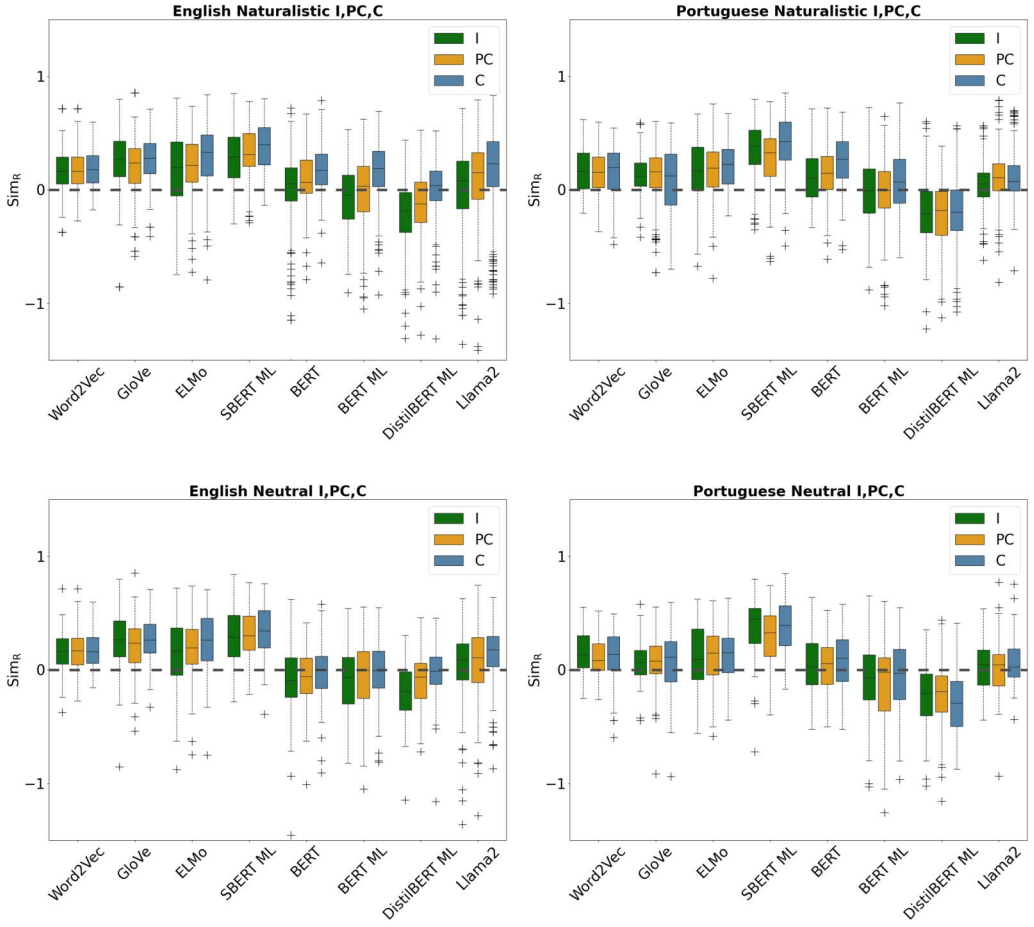

Figure 7
Sim_{RP3} per compositionality class: green for idiomatic (I), yellow for partly compositional (PC), and blue for compositional (C), in English (EN) and Portuguese (PT), in naturalistic and neutral sentences.

4.4 How Are the Results Across Models and Languages?

We have evaluated several vector models from different architectures in two languages, ranging from static to contextual representations as well as monolingual and multilingual models. Although the results are generally far from being satisfactory, in this section we highlight some differences and similarities between models and languages.

Across models, the similarities are in general higher for Transformer-based models than for static representations. In this respect, it is worth noting that the results of ELMo and mSBERT are as similar to those of Word2Vec and GloVe than to the other BERT variants (for instance in Figure 2). Although further research would be needed to determine the precise factors, for ELMo this behavior could be due either to the different vector space constructed by LSTMs or to the smaller number of hidden layers when compared with the other models (2 vs. 6 and 12 layers), which may imply lower contextualization effects across the network (Ethayarajh 2019).

For the Transformer-based models, there are clear differences between the similarities produced by the BERT-based models and those of the autoregressive models, which

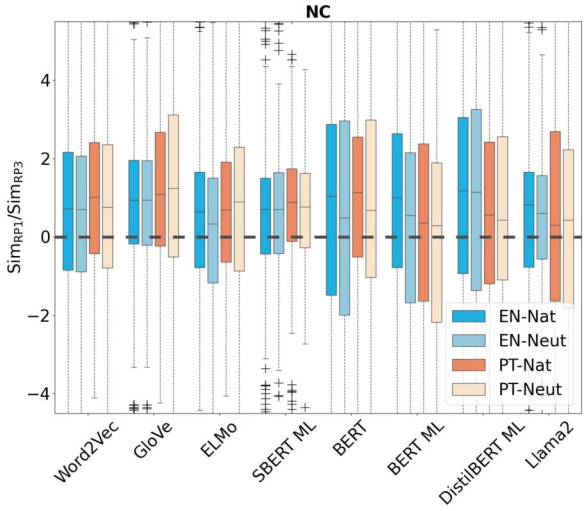

Figure 8
Ratio of average of Scaled Similarity ($Sim_{R|Syn}/Sim_{R|WordsSyn}$). English data are in blue, Portuguese data in orange, values for naturalistic sentences in darker shade and for neutral in lighter.

are lower and with a wider range, especially for neutral sentences. When comparing monolingual and multilingual models, namely, BERT and BERT ML, similar tendencies are found both in similarities and in correlations with the human judgments. In general, multilingual models seem to place the vector representations in a more restricted space, implying higher degrees of similarity and lower ranges of variation. Similar tendencies are found for DistilBERT-ML.

The proposed measures also suggest that the representations of the large autoregressive models are more similar to those of the static embeddings than to the other Transformer-based encoder models.

Although the results of the different models across languages follow very similar trends, they also display two main differences. The first one is that when comparing the minimal pairs of the naturalistic data, the representations in English seem to be closer and occupy less space than those in Portuguese, in both monolingual and multilingual models of all types. The second is that for neutral sentences, there are larger differences than for naturalistic sentences, especially at the sentence level in both languages, and similar results at the NC level, except for ELMo and BERT embeddings in $P_{WordsSyn}$ and P_{Rand} (Figure 2). The trends are even more aligned when considering Affinities and Scaled Similarities for most models in both languages.

Indeed, high correlations were found among all models, reflected by the correlogram in Figure 10. Correlations are particularly high for the expected congruent variants involving NC_{Syn}, as reflected by the darker red shades: P_{Syn}, $A_{Syn|WordsSyn}$, and $Sim_{R|Syn}$. They are also higher for Affinities and Scaled Similarities, indicating that taking into account the relative preferences and random similarities within each model reveals how comparable they are in their ability to represent idiomaticity. That is, regardless of any superiority of specific models for other tasks, and in spite of their seemingly different individual performances in terms of cosines similarities in terms of idiomaticity representation, this sample of models has not revealed one that is clearly better than the others. Moreover, high correlations with the static models also suggests that the relevant

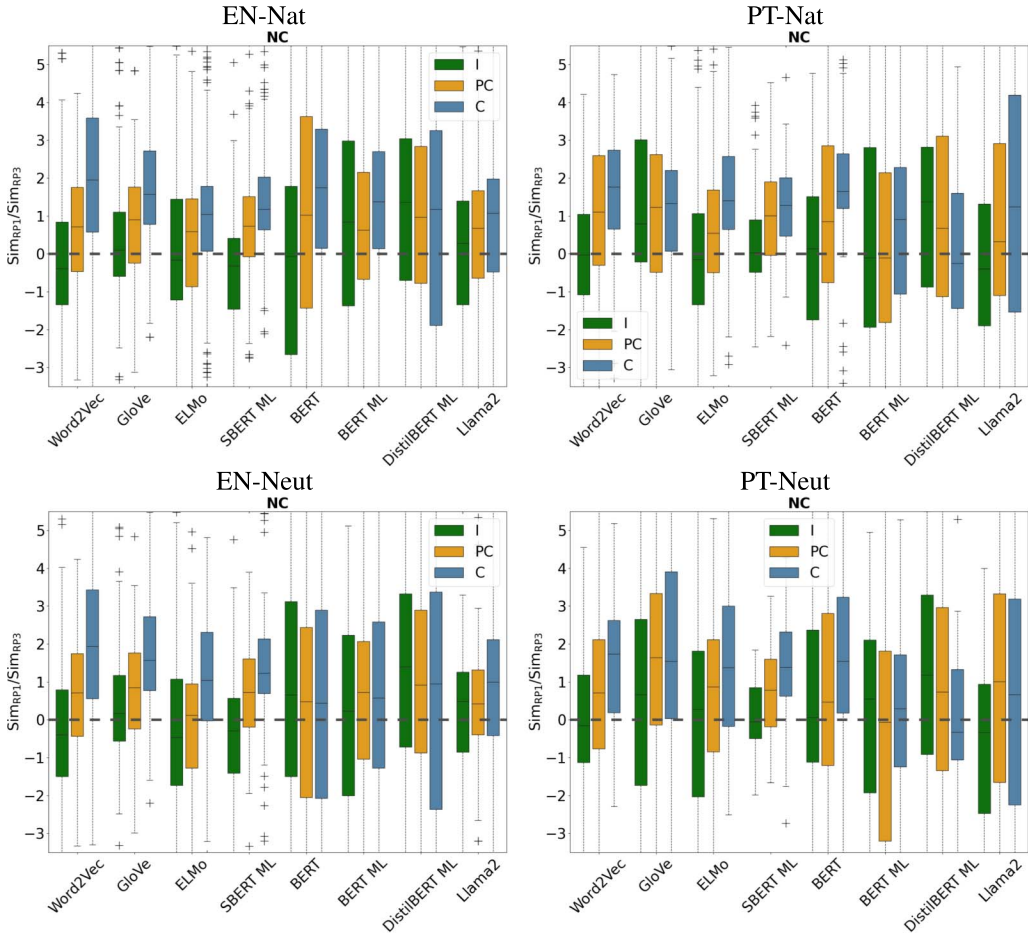

Figure 9
Ratio of average of Scaled Similarity ($Sim_{R|Syn}/Sim_{R|WordsSyn}$), per compositionality class: green for idiomatic (I), orange for partly compositional (PC), and blue for compositional (C), in English (EN) and Portuguese (PT), in naturalistic (Nat) and neutral (Neut) sentences.

contextual clues for idiomaticity representation are not yet adequately incorporated by the contextualized models.

In sum, our results indicate that the different models evaluated in general follow the same tendencies when representing idiomaticity in context, suggesting that they are not yet able to adequately capture the semantics of the MWEs. More investigation is needed to determine how to effectively achieve this with these architectures and training regimes, or whether a change in paradigm is required. We will now discuss some representative cases, to give a flavor of how these models handle a spectrum of idiomaticity.

4.5 Analyzing Example Cases

For a more concrete qualitative overview of the ability of models in representing different levels of idiomaticity, we now look at some representative English NCs evenly distributed among the three levels of compositionality (compositional, partly compositional, and

Table 6
Spearman ρ correlation between the Scale Similarities and human judgments, for $Sim_{R|Syn}$ and $Sim_{R|WordsSyn}$ in both English and Portuguese. Non-significant ($p > 0.05$) results were omitted from the table. Although these values shown are for the correlations at the NC level, the correlations at the sentence level are comparable.

$Sim_{R\|Syn}$	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
EN-Nat	0.61	0.57	0.61	0.66	0.63	0.67	0.61	0.38
EN-Neut	0.60	0.56	0.62	0.64	0.60	0.54	0.57	0.37
PT-Nat	0.46	0.39	0.47	0.48	0.51	0.45	0.37	0.40
PT-Neut	0.44	0.41	0.45	0.48	0.46	0.41	0.38	0.32
$Sim_{R\|WordsSyn}$	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
EN-Nat	–	–	0.17	0.18	0.19	0.36	0.33	0.25
EN-Neut	–	–	0.14	0.12	–	–	0.29	0.17
PT-Nat	–	–	0.10	0.14	0.25	0.16	–	–
PT-Neut	–	–	–	–	0.12	–	–	–

idiomatic) in three naturalistic sentences (Table 7). We start with the probes for 6 English NCs and then look at the highest and lowest values for the $A_{Syn|WordsSyn}$ Affinity focusing on the relation between a given NC and its NC_{Syn} and $NC_{WordSyn}$ variants.

Probes. Considering the probing measures in terms of the average scores of all sentences for each of the 6 NCs (Table 8), we focus on the cosine similarities for the probes and whether they differ from the expected behavior compatible with capturing idiomatic meaning.

First of all, for P_{Syn} the similarities should be close to 1. Indeed, at sentence level all similarities for all models are above 0.9, and tend to be higher for compositional NCs (0.98) than for partly compositional (0.95) and than for idiomatic NCs (0.90). However, at NC level, they display considerable variation, and while the similarities are high for all models for compositional NCs, for idiomatic NCs, in particular, the similarities are the lowest and vary considerably per model (from 0.27 for SBERT ML and Word2Vec to 0.81 for Llama2 for *grey matter*). For partly compositional NCs, although some of the models assign the expected high similarities for some NCs (0.94 for BERT for *Dutch courage*), other NCs have lower similarities (0.43 for Word2Vec for *eternal rest*).

For P_{Comp}, lower similarities are expected for idiomatic NCs, as the idiomatic meaning may be lost when one of the component words is missing (e.g., *grey matter* vs. *grey* or vs. *matter*). However, at sentence level they are higher than 0.93 for all models. At NC level, although these idiomatic NCs have lower similarities they are still high (from 0.77 for SBERT ML for *grey matter* to 0.94 for BERT ML for *eager beaver*). For partly compositional and compositional NCs they are mostly high for all models, except for Llama 2 for *Dutch courage* (0.67).

Although lower $P_{WordsSyn}$ were also expected for more idiomatic NCs, at sentence level all idiomatic, partly compositional, and compositional NCs display similarities above 0.95, even though the *NCWordSyn* in $P_{WordsSyn}$ does not preserve the idiomatic meaning (e.g., *grey matter* vs. *silvery material*). At the NC level, even if lower values were found for idiomatic NCs with static models (Word2Vec and GloVe), high similarities were still found (e.g., 0.91 for BERT for *grey matter*).

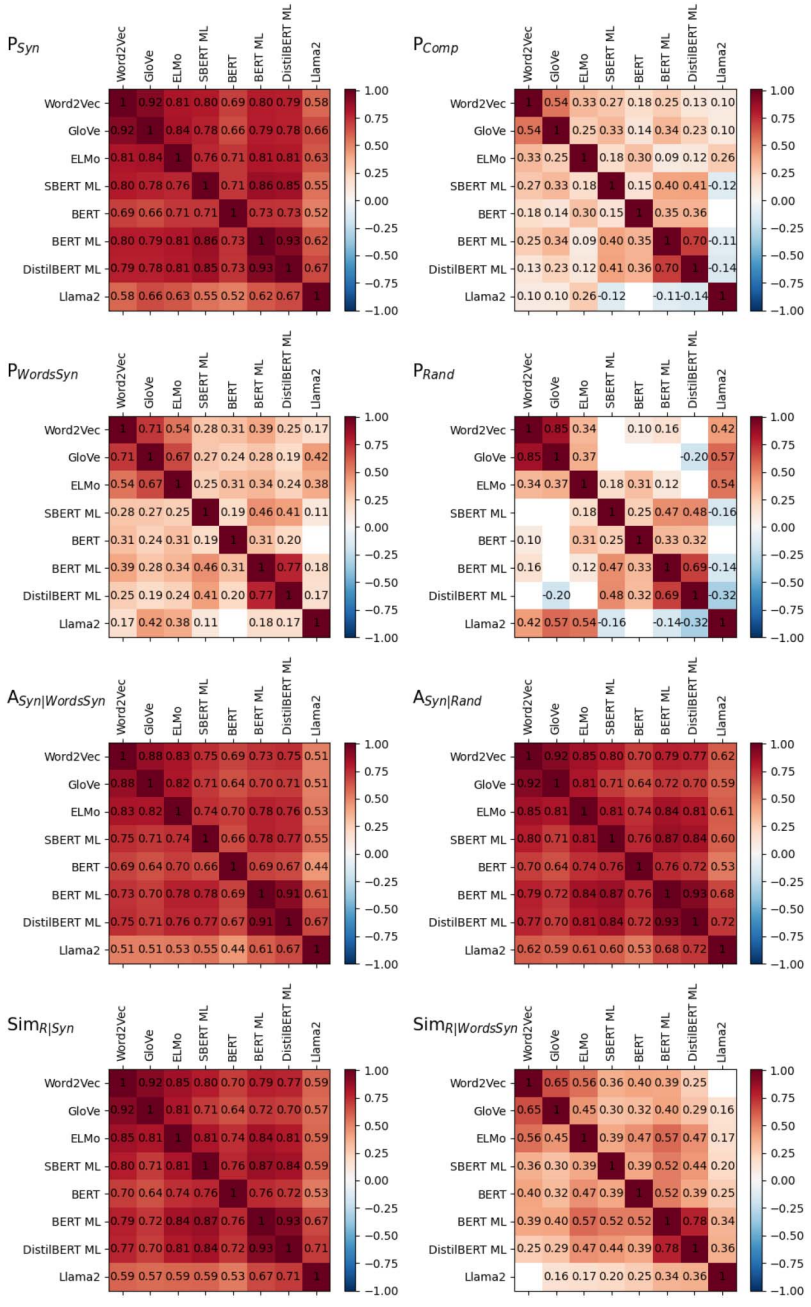

Figure 10
Correlograms for all models for all sentences and languages, with only significant values displayed ($p < 0.05$). Red indicates positive correlation and blue negative; darker shades are for higher values, lighter for lower values.

Finally, for P_{Rand}, there should be low similarities for all NCs and randomly generated substitutions. However, most of the similarities are still high, regardless of the level of idiomaticity (e.g., 0.92 for BERT for *grey matter* and for BERT ML for *Dutch courage*).

Table 7
Compositional NCs, NC_{Syn}, $NC_{WordSyn}$ and three sentences for qualitative analyses.

	NC	NC_{Syn}	$NC_{WordsSyn}$	Examples
Idiomatic	grey matter	brain	silvery material	S1: *Give your **grey matter** the workout that it needs to stay sharp and focused.* S2: *More ideas will follow when I get the **grey matter** functioning.* S3: *These youngsters can be encouraged to use their **grey matter** if the presentation is right.*
	eager beaver	hard worker	restless rodent	S1: *Eric was being an **eager beaver** and left work late.* S2: ***Eager beavers** willing to learn your job for less pay are almost always waiting in the wings.* S3: *If you are a really **eager beaver** you can pre-order the DVD now from either of the below retailers.*
Partly Compositional	Dutch courage	alcoholic courage	Hollander bravery	S1: *We had to go down to the pub to get some **Dutch courage**!* S2: *We suggest you try the following cocktail to work up a bit of **Dutch courage** to get you through the match!* S2: *After some **Dutch courage** (a few vodkas) in the nightclub, and a nerve-racking conversation, we kissed!*
	eternal rest	death	permanent break	S1: *They have been called home to their **eternal rest** and we are left behind.* S2: *These tolls announce the death of a nun and call for prayers for her **eternal rest**.* S3: *The passengers, with early morning porridge complexions, don't look far from **eternal rest**.*
Compositional	economic aid	financial assistance	budgetary assistance	S1: *We have already extended to Greece certain types of relief and **economic aid** but these are inadequate.* S2: *The USSR was soon giving Cuba **economic aid**, technical support and military 'advisers' from the USSR.* S3: *A government's success in reducing population movement should be a key factor in allocating **economic aid**.*
	research lab	research facility	investigation workplace	S1: *The fourth year is spent doing a research project in a 'real' **research lab**.* S2: *Being part of a **research lab** provides at times very exciting fieldwork experiences for individual students.* S3: *Bath operates several undergraduate degree programmes that include a professional placement year in industry or a **research lab**.*

Overall, the expected high similarities for P_{Syn} are not displayed by these models at the NC level, and for the other probes the perturbations to the idiomatic meaning are not reflected by lower similarities.

Affinities. For the Affinity measures, considering the examples with the highest and lowest values for $A_{Syn|WordsSyn}$ as a proxy for how a particular model represents an NC compared to its synonym and to a word-by-word replacement ($NC_{Syn|WordsSyn}$), we focus on the results for BERT in the naturalistic sentences in English. As discussed in Section 4.1,

Table 8
Similarity, Affinity, and Scaled Similarity values for the NCs selected in Table 7. Values in parentheses represent the standard deviations among the three sentences. The static models are independent of context, and for them, the variance is omitted, except in the case of *eager beaver*, where there is a sentence where the compound appears in plural form.

grey matter	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
P_{Syn}	0.27	0.37	0.45 (0.03)	0.27 (0.01)	0.68 (0.14)	0.78 (0.05)	0.77 (0.01)	0.81 (0.04)
P_{Comp}	0.86	0.84	0.80 (0.01)	0.77 (0.00)	0.89 (0.01)	0.90 (0.02)	0.92 (0.01)	0.89 (0.01)
$P_{WordsSyn}$	0.58	0.59	0.66 (0.01)	0.59 (0.02)	0.91 (0.02)	0.88 (0.03)	0.85 (0.02)	0.69 (0.00)
P_{Rand}	0.47	0.52	0.61 (0.02)	0.49 (0.00)	0.92 (0.02)	0.87 (0.02)	0.87 (0.02)	0.68 (0.02)
$A_{Syn\|WordsSyn}$	−0.31	−0.22	−0.21 (0.02)	−0.32 (0.02)	−0.23 (0.15)	−0.10 (0.02)	−0.08 (0.01)	0.12 (0.04)
$A_{Syn\|Rand}$	−0.20	−0.15	−0.16 (0.02)	−0.22 (0.01)	−0.25 (0.16)	−0.09 (0.04)	−0.09 (0.01)	0.14 (0.04)
$Sim_{R\|Syn}$	0.11	−0.13	−0.37	−0.39	−1.54	−1.64	−1.72	0.42
$Sim_{R\|WordsSyn}$	−0.02	0.25	0.15	0.15	0.04	0.02	−0.08	0.04

eager beaver	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
P_{Syn}	0.34 (0.02)	0.41 (0.04)	0.68 (0.05)	0.40 (0.02)	0.78 (0.06)	0.82 (0.02)	0.83 (0.01)	0.66 (0.05)
P_{Comp}	0.87 (0.01)	0.85 (0.01)	0.79 (0.05)	0.82 (0.02)	0.89 (0.00)	0.94 (0.01)	0.93 (0.01)	0.78 (0.16)
$P_{WordsSyn}$	0.45 (0.03)	0.49 (0.00)	0.84 (0.03)	0.58 (0.04)	0.86 (0.04)	0.87 (0.03)	0.85 (0.01)	0.53 (0.03)
P_{Rand}	0.41 (0.01)	0.33 (0.04)	0.72 (0.14)	0.51 (0.02)	0.92 (0.03)	0.86 (0.00)	0.89 (0.01)	0.60 (0.10)
$A_{Syn\|WordsSyn}$	−0.10 (0.00)	−0.08 (0.04)	−0.16 (0.04)	−0.18 (0.02)	−0.07 (0.04)	−0.05 (0.00)	−0.02 (0.00)	0.13 (0.04)
$A_{Syn\|Rand}$	−0.06 (0.03)	0.08 (0.00)	−0.04 (0.15)	−0.11 (0.01)	−0.13 (0.08)	−0.04 (0.02)	−0.07 (0.00)	0.06 (0.05)
$Sim_{R\|Syn}$	0.09	0.08	−0.10	−0.14	−0.79	−0.78	−0.90	0.15
$Sim_{R\|WordsSyn}$	−0.13	0.15	0.23	0.26	−0.14	−0.24	−0.42	−0.20

Dutch courage	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
P_{Syn}	0.63	0.64	0.88 (0.02)	0.75 (0.01)	0.94 (0.03)	0.93 (0.01)	0.90 (0.00)	0.77 (0.00)
P_{Comp}	0.77	0.76	0.91 (0.01)	0.82 (0.00)	0.88 (0.04)	0.90 (0.01)	0.90 (0.01)	0.67 (0.14)
$P_{WordsSyn}$	0.57	0.53	0.82 (0.01)	0.69 (0.02)	0.88 (0.03)	0.87 (0.01)	0.89 (0.01)	0.54 (0.03)
P_{Rand}	0.41	0.35	0.79 (0.01)	0.51 (0.00)	0.90 (0.04)	0.92 (0.01)	0.90 (0.00)	0.59 (0.00)
$A_{Syn\|WordsSyn}$	0.07	0.11	0.06 (0.02)	0.06 (0.02)	0.06 (0.04)	0.05 (0.02)	0.00 (0.00)	0.23 (0.03)
$A_{Syn\|Rand}$	0.22	0.28	0.09 (0.02)	0.24 (0.01)	0.04 (0.06)	0.01 (0.01)	−0.01 (0.00)	0.18 (0.00)
$Sim_{R\|Syn}$	0.32	0.37	0.42	0.46	0.38	0.28	0.09	0.44
$Sim_{R\|WordsSyn}$	0.21	0.32	0.22	0.26	0.06	−0.17	−0.33	−0.12

eternal rest	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
P_{Syn}	0.43	0.53	0.48 (0.01)	0.53 (0.02)	0.74 (0.04)	0.86 (0.01)	0.76 (0.02)	0.71 (0.04)
P_{Comp}	0.92	0.89	0.80 (0.02)	0.82 (0.01)	0.87 (0.03)	0.93 (0.01)	0.09 (0.00)	0.75 (0.06)
$P_{WordsSyn}$	0.43	0.53	0.60 (0.02)	0.71 (0.02)	0.83 (0.05)	0.83 (0.05)	0.82 (0.01)	0.66 (0.03)
P_{Rand}	0.44	0.38	0.64 (0.01)	0.43 (0.01)	0.80 (0.08)	0.87 (0.03)	0.89 (0.00)	0.62 (0.04)
$A_{Syn\|WordsSyn}$	−0.00	−0.01	−0.12 (0.02)	−0.18 (0.01)	−0.09 (0.08)	0.03 (0.06)	−0.06 (0.01)	0.05 (0.01)
$A_{Syn\|Rand}$	−0.00	0.14	−0.15 (0.00)	0.10 (0.02)	−0.06 (0.08)	−0.01 (0.03)	−0.13 (0.02)	0.09 (0.02)
$Sim_{R\|Syn}$	0.15	0.14	−0.06	−0.00	−0.25	−0.15	−0.62	0.23
$Sim_{R\|WordsSyn}$	−0.34	0.16	0.05	0.22	0.17	0.10	−0.29	0.11

economic aid	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
P_{Syn}	0.65	0.80	0.90 (0.01)	0.77 (0.01)	0.89 (0.02)	0.96 (0.01)	0.94 (0.01)	0.95 (0.02)
P_{Comp}	0.80	0.88	0.89 (0.00)	0.78 (0.00)	0.90 (0.04)	0.93 (0.03)	0.92 (0.00)	0.90 (0.02)
$P_{WordsSyn}$	0.64	0.74	0.92 (0.00)	0.78 (0.01)	0.89 (0.00)	0.95 (0.01)	0.92 (0.01)	0.90 (0.02)
P_{Rand}	0.65	0.73	0.70 (0.07)	0.58 (0.01)	0.76 (0.05)	0.90 (0.02)	0.90 (0.01)	0.70 (0.03)
$A_{Syn\|WordsSyn}$	0.02	0.06	−0.01 (0.01)	−0.00 (0.00)	−0.00 (0.02)	0.01 (0.01)	0.02 (0.00)	0.05 (0.01)
$A_{Syn\|Rand}$	0.01	0.07	0.20 (0.06)	0.20 (0.01)	0.14 (0.05)	0.06 (0.01)	0.04 (0.01)	0.25 (0.02)
$Sim_{R\|Syn}$	0.27	0.22	0.32	0.47	0.56	0.53	0.52	0.84
$Sim_{R\|WordsSyn}$	0.42	0.28	0.24	0.41	0.58	0.49	0.42	0.66

research lab	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
P_{Syn}	0.71	0.82	0.91 (0.03)	0.78 (0.02)	0.93 (0.01)	0.97 (0.00)	0.95 (0.00)	0.96 (0.01)
P_{Comp}	0.86	0.88	0.90 (0.00)	0.88 (0.01)	0.89 (0.04)	0.94 (0.01)	0.94 (0.01)	0.87 (0.05)
$P_{WordsSyn}$	0.47	0.51	0.68 (0.02)	0.72 (0.02)	0.67 (0.09)	0.90 (0.01)	0.90 (0.01)	0.78 (0.11)
P_{Rand}	0.39	0.40	0.72 (0.04)	0.40 (0.01)	0.73 (0.07)	0.88 (0.01)	0.90 (0.01)	0.68 (0.08)
$A_{Syn\|WordsSyn}$	0.23	0.30	0.23 (0.01)	0.06 (0.02)	0.26 (0.08)	0.07 (0.01)	0.05 (0.01)	0.18 (0.11)
$A_{Syn\|Rand}$	0.32	0.41	0.19 (0.01)	0.38 (0.03)	0.20 (0.06)	0.09 (0.01)	0.06 (0.01)	0.28 (0.07)
$Sim_{R\|Syn}$	0.52	0.69	0.68	0.63	0.73	0.72	0.55	0.87
$Sim_{R\|WordsSyn}$	0.14	0.18	−0.17	0.53	−0.20	0.17	0.06	0.35

we expect higher $A_{Syn|WordsSyn}$ values for idiomatic NCs, since the model should display a stronger preference for a semantically related synonym than to a potentially unrelated substitution, representing the former as closely as possible from the NC in a vector space. In contrast, for more compositional cases, both substitutions may be possible and close to one another (reflected by $A_{Syn|WordsSyn}$ values around 0). However, the NCs with the highest $A_{Syn|WordsSyn}$ values were mostly compositional (starting with *video game* with $A_{Syn|WordsSyn} = 0.44$, and *parking lot* with $A_{Syn|WordsSyn} = 0.40$), with the first partly compositional NC appearing at position 16 (*sparkling water* with $A_{Syn|WordsSyn} = 0.32$). The idiomatic NC with the highest $A_{Syn|WordsSyn}$ value is at position 53 (*box office*, referring to the popularity of a movie with $A_{Syn|WordsSyn} = 0.24$).

At the other end of the ranking, we find mostly idiomatic cases. Among the top 10 examples with the lowest values we find 7 idiomatic (e.g., *agony aunt* with $A_{Syn|WordsSyn} = -0.29$ and the NC with the lowest value, *grey matter*, $A_{Syn|WordsSyn} = -0.40$), with 2 partly compositional and only one compositional NC in position 10 (*cooking stove* with $A_{Syn|WordsSyn} = -0.24$).

In sum, these confirm that the models do not display the expected preference for representing NCs closer to their synonyms than to distractors, even when these involve idiomatic NCs and/or random items.

5. Conclusions

This article presented an evaluation of the ability of widely available word representation models to capture idiomatic meaning, focusing on noun compounds in two languages, English and Portuguese. For evaluation we introduced the NCIMP dataset, containing NCs in English and Portuguese in naturalistic and neutral sentences forming minimal pairs with idiomatic probes using their component words, synonyms, and other variant replacements, resulting in a dataset containing 29,900 items, extending the datasets by Garcia et al. (2021a) and Garcia et al. (2021b). These pairs can be used to measure the ability of models to detect the loss of the idiomatic meaning in the presence of lexical substitutions and different contexts. We also propose two types of measure for quantifying this ability: Affinities and Scaled Similarities. Affinity is a relative measure of the proximity of the NC to two alternative probes, determining which of them is the closest to the NC. Focusing on idiomaticity, we analyzed if the models were able to generate a representation for a given NC that was more similar to a semantically related paraphrase given by the gold standard synonym than to an alternative possibly semantically unrelated representation. The proposed measures of scaled similarities, Sim_R, take sample random similarities into account for rescaling the space of a given model, to magnify high similarities and distinguish them from those that are artifacts of the characteristics of the landscape of that model. As a consequence, Sim_R also seems to abstract away from the particularities of the semantic space of each model and provides a more direct way of comparing idiomaticity representation across models. The results obtained indicate that models are not able to accurately capture idiomaticity, as they fail to reflect actual similarities between NCs and their gold synonyms, especially for idiomatic cases, while at the same time not displaying enough awareness of perturbations that lead to changes in meaning, such as those involving the synonyms of the component words, and even random words. It seems that the lexical clues provided by the component words are prioritized when representing an NC over a more holistic combination of the relevant semantic clues needed for representing its idiomatic meaning. Moreover, although the contexts could provide relevant information about the idiomatic meanings, they do not seem to be adequately incorporated in these widely adopted models, regardless of their

degree of contextualization. They also seem to fail to incorporate the relevant context for idiomaticity, seeing as static and contextualized models show comparable performances.

In this article we evaluated the proposed measures focusing on idiomaticity, but they may be applied to other tasks, and serve as a basis to detect unwanted biases towards non-target meanings more generally. Moreover, they may be informative when fine-tuning models to assess if the changes are going towards the intended target representations.

5.1 Future Work

In this article, we inspected the similarities produced by a number of models to determine how accurately they represent idiomatic expressions. The results obtained are that not even large models like Llama2 seem to display the expected patterns that would confirm idiomatic understanding.

It is important to note that some of the difficulties in extracting information from cosine similarity measures may be attributed to the presence of rogue dimensions (Timkey and van Schijndel 2021) rather than anisotropy in semantic space. Measures like Affinity and Scaled Similarity may not fully address this issue. We conducted a preliminary analysis using Timkey and van Schijndel (2021) method to identify and standardize the top three rogue dimensions per model/layer. After standardization, we conducted an analysis focusing on P_{Syn} measures and found correlations mostly above 0.85, except for BERT-PT-Neut (0.79) and Llama2-EN-Neut (0.65) (see Table 16 in the Appendix). Further investigation is needed to assess the impact of standardizing these dimensions and different approaches for standardization, but given the high correlations with our original results, we will leave this for future work.

Although our proposed assessment protocol and measures are model-independent, they rely on access to the models and to their representations for subwords, words, and multiwords. Therefore, probing large generative AI chatbots for their understanding of idiomaticity, especially closed-source models, presents additional challenges potentially requiring adaptation in the application of the protocol, due to the restricted access to their base models and of the potential variation in their answers. These warrant further investigations that are outside the scope of this paper. However, one possible alternative would be to perform probing using question-answering, following Zeng and Bhat (2022). We illustrate this question-answering approach with recent AI chatbots: GPT-3, Gemini Pro[29] (Team et al. 2023), and ChatGPT4.[30] For testing these models, simple questions containing idiomatic expressions are used, after having instructed each model to provide the shortest answer to each question before asking them. The assumption is that the questions could only be answered correctly if the model understood the meaning of the idioms in context. The questions and answers are included in Table 9.

The responses from different systems vary in terms of correctly interpreting the idioms. The responses from GPT-3 often miss the mark, while the responses from Gemini Pro and ChatGPT 4 are mixed, with some answers suggesting correct interpretations and some incorrect. For instance, for "Every trick in the book", GPT-3 responds with "A magician", which is a literal interpretation, while ChatGPT 4 correctly identifies the figurative meaning with "Determined student". The Gemini Pro response to the question is "Cheater". The idiom "every trick in the book" generally means to use all available means or strategies to achieve one's goal, often implying ingenuity or resourcefulness

29 https://gemini.google.com/app.
30 https://chat.openai.com/.

Table 9
Questions used to probe the understanding of idioms and the answers provided by recent generative models. The leftmost column lists idiomatic expressions, the second column presents hypothetical questions using these expressions, and the following columns show the responses from different models, including GPT-3, Gemini Pro, and ChatGPT 4. Answers by GPT-3 are from Zeng and Bhat (2022).

Idiom	Question	GPT-3	Gemini Pro	ChatGPT 4
Never say die	If I have a never say die attitude, would I run the marathon injured or forfeit?	I would never run a marathon injured.	Run.	Run the marathon.
All at sea	If I am all at sea with my math assignment, am I making progress or am I lost?	You are making progress.	Lost.	Lost.
Every trick in the book	If I use every trick in the book to guarantee my grade, am I a magician or a determined student?	A magician.	Cheater.	Determined student.
Kill two birds in one stone	If I wanted to kill two birds in one stone, what kind of a workplace should I work in?	A slaughterhouse.	Multiskilled.	A multitasking environment.
Ahead of the game	If I want to be ahead of the game, would I study early or procrastinate?	Procrastinate.	Early.	Study early.

rather than dishonesty. The response from Gemini Pro could either be due to "trick" or it could be seen as a misinterpretation. This table could also be seen as indicative of the evolution of AI language models over time, with newer models potentially being trained to better handle idiomatic expressions and context, as seen in the generally more accurate responses from ChatGPT 4 compared to GPT 3. Although the questions in the table are indeed useful for exemplifying the comprehension of idiomatic expressions by these models they only cover a very limited and focused sample. In this paper, we propose the use of minimal pairs containing synonyms and other distractors for a more in-depth assessment of idiomatic understanding. Although their adaptation for a question-answering setting is left for future work, our results for open models is in line with comparative analyses of the ability of some of these models for idiomatic and figurative language (Phelps et al. 2024).

Moreover, as idiomatic expressions can be extremely diverse and nuanced, a comprehensive evaluation of the ability of a model to understand them requires a controlled but extensive set of idiomatic expressions and their variations. Therefore, we plan to extend the test items to contain additional types of multiword expressions, including verb-noun combinations and phrasal verbs. In addition for a larger crosslingual examination of idiomaticity, and in particular of whether multilingual models capture language-specific realizations of idiomatic expressions, we plan to extend the dataset with additional languages. These would also allow the investigation of factors relevant to specific tasks, such as machine translation, for which the translatability of MWEs from source into target languages may also affect performance when processing MWEs (Dankers, Lucas, and Titov 2022).

Possible next steps also include extending the probing strategy with additional measures that go beyond similarities and correlations. Moreover, for ambiguous NCs in particular, we intend to add sense-specific probes that could be used to measure and address training biases towards particular senses. Finally, this article has focused the

evaluation on off-the-shelf pre-trained models to provide an analysis of their ability to capture idiomaticity, and left the investigation of fine-tuned models for future work. In particular, although fine-tuning can improve model performance (Tayyar Madabushi et al. 2022), it is unclear to what extent the models are able to generalize beyond the specific items seen to other unseen idiomatic expressions, or if each new expression would have to be individually learned by the model. But these points are left for future investigation.

Appendix A. Measures for English and Portuguese

In this section we present the mean and standard deviation for the NCs in English and Portuguese in naturalistic and neutral sentences, for the different probes at the sentence level (Table 10), for the different probes at the NC level (Table 11), for Affinities (Table 12), and for Scaled Similarities (Table 13).

Table 10
Mean and standard deviation (std) at Sentence level for P_{Syn}, P_{Comp}, $P_{WordsSyn}$, and P_{Rand}, for English (EN) and Portuguese (PT) for naturalistic (Nat) and neutral (Neut) sentences.

Model Name	EN-Nat		EN-Neut		PT-Nat		PT-Neut	
	mean	std	mean	std	mean	std	mean	std
P_{Syn}								
Word2Vec	0.985	0.012	0.811	0.083	0.968	0.025	0.883	0.062
GloVe	0.990	0.008	0.868	0.063	0.980	0.018	0.931	0.054
ELMo	0.974	0.022	0.841	0.070	0.938	0.045	0.782	0.116
SBERT ML	0.974	0.022	0.810	0.101	0.955	0.035	0.833	0.096
BERT	0.988	0.011	0.927	0.035	0.980	0.017	0.915	0.041
BERT ML	0.992	0.007	0.924	0.040	0.984	0.012	0.929	0.044
DistilBERT ML	0.996	0.003	0.952	0.023	0.991	0.007	0.966	0.018
Llama2	0.992	0.010	0.955	0.020	0.981	0.018	0.903	0.065
P_{Comp}								
Word2Vec	0.996	0.004	0.941	0.018	0.987	0.011	0.957	0.026
GloVe	0.996	0.003	0.955	0.011	0.993	0.006	0.982	0.012
ELMo	0.989	0.009	0.914	0.019	0.966	0.020	0.890	0.035
SBERT ML	0.990	0.007	0.922	0.021	0.982	0.013	0.929	0.029
BERT	0.992	0.007	0.951	0.016	0.986	0.013	0.933	0.025
BERT ML	0.996	0.003	0.957	0.016	0.993	0.005	0.962	0.016
DistilBERT ML	0.998	0.001	0.977	0.006	0.996	0.002	0.987	0.005
Llama2	0.995	0.008	0.986	0.007	0.991	0.008	0.964	0.020
$P_{WordsSyn}$								
Word2Vec	0.983	0.013	0.797	0.049	0.958	0.031	0.845	0.060
GloVe	0.989	0.009	0.863	0.041	0.974	0.025	0.904	0.062
ELMo	0.975	0.020	0.861	0.048	0.930	0.042	0.760	0.088
SBERT ML	0.977	0.017	0.844	0.057	0.956	0.033	0.855	0.060
BERT	0.983	0.014	0.919	0.032	0.967	0.025	0.891	0.038
BERT ML	0.991	0.006	0.925	0.036	0.983	0.012	0.934	0.032
DistilBERT ML	0.995	0.003	0.952	0.016	0.990	0.006	0.963	0.014
Llama2	0.986	0.014	0.945	0.021	0.977	0.017	0.891	0.052
P_{Rand}								
Word2Vec	0.984	0.012	0.799	0.043	0.960	0.033	0.851	0.099
GloVe	0.988	0.009	0.849	0.038	0.974	0.026	0.911	0.095
ELMo	0.966	0.025	0.829	0.040	0.912	0.048	0.725	0.115
SBERT ML	0.968	0.023	0.769	0.053	0.935	0.043	0.768	0.063
BERT	0.979	0.018	0.924	0.027	0.956	0.028	0.886	0.033
BERT ML	0.990	0.008	0.925	0.024	0.980	0.013	0.933	0.030
DistilBERT ML	0.995	0.004	0.951	0.012	0.990	0.007	0.967	0.016
Llama2	0.980	0.019	0.937	0.015	0.962	0.026	0.879	0.058

Table 11
Mean and standard deviation (std) at NC level for P_{Syn}, P_{Comp}, $P_{WordsSyn}$, and P_{Rand}, for English (EN) and Portuguese (PT) for naturalistic (Nat) and neutral (Neut) sentences.

Model Name	EN-Nat		EN-Neut		PT-Nat		PT-Neut	
	mean	std	mean	std	mean	std	mean	std
P_{Syn}								
Word2Vec	0.517	0.209	0.517	0.207	0.498	0.251	0.488	0.258
GloVe	0.551	0.227	0.555	0.222	0.465	0.278	0.473	0.275
ELMo	0.714	0.147	0.646	0.155	0.629	0.166	0.551	0.192
SBERT ML	0.591	0.208	0.577	0.203	0.632	0.199	0.612	0.198
BERT	0.816	0.086	0.854	0.060	0.824	0.090	0.831	0.079
BERT ML	0.876	0.061	0.861	0.059	0.880	0.056	0.866	0.063
DistilBERT ML	0.867	0.058	0.864	0.057	0.868	0.059	0.870	0.056
Llama2	0.702	0.189	0.612	0.200	0.533	0.216	0.589	0.205
P_{Comp}								
Word2Vec	0.840	0.039	0.838	0.039	0.714	0.269	0.703	0.280
GloVe	0.835	0.041	0.837	0.040	0.715	0.276	0.710	0.282
ELMo	0.859	0.042	0.823	0.040	0.781	0.080	0.733	0.093
SBERT ML	0.815	0.042	0.805	0.038	0.823	0.050	0.808	0.052
BERT	0.849	0.060	0.886	0.037	0.855	0.066	0.864	0.041
BERT ML	0.923	0.022	0.913	0.020	0.930	0.021	0.921	0.023
DistilBERT ML	0.922	0.015	0.922	0.013	0.929	0.018	0.931	0.014
Llama2	0.828	0.102	0.844	0.086	0.741	0.174	0.749	0.174
$P_{WordsSyn}$								
Word2Vec	0.524	0.098	0.524	0.097	0.459	0.185	0.450	0.189
GloVe	0.569	0.119	0.572	0.116	0.356	0.196	0.357	0.198
ELMo	0.759	0.083	0.707	0.091	0.644	0.100	0.557	0.110
SBERT ML	0.659	0.112	0.645	0.112	0.670	0.119	0.662	0.122
BERT	0.780	0.105	0.850	0.064	0.783	0.077	0.820	0.054
BERT ML	0.881	0.035	0.867	0.040	0.887	0.035	0.877	0.039
DistilBERT ML	0.870	0.029	0.868	0.027	0.875	0.027	0.877	0.026
Llama2	0.668	0.148	0.601	0.151	0.490	0.137	0.560	0.118
P_{Rand}								
Word2Vec	0.419	0.064	0.423	0.065	0.460	0.185	0.371	0.151
GloVe	0.413	0.108	0.419	0.108	0.356	0.196	0.293	0.219
ELMo	0.674	0.082	0.628	0.069	0.644	0.100	0.482	0.097
SBERT ML	0.479	0.067	0.473	0.067	0.670	0.119	0.479	0.072
BERT	0.746	0.117	0.855	0.061	0.783	0.077	0.808	0.037
BERT ML	0.872	0.031	0.872	0.028	0.887	0.035	0.883	0.032
DistilBERT ML	0.879	0.024	0.879	0.021	0.875	0.027	0.898	0.021
Llama2	0.631	0.100	0.568	0.105	0.490	0.137	0.544	0.102

Table 12
Mean and standard deviation (std) at NC level for $A_{Syn|WordsSyn}$ and $A_{Syn|Rand}$, for English (EN) and Portuguese (PT) for naturalistic (Nat) and neutral (Neut) sentences.

Model Name	EN-Nat		EN-Neut		PT-Nat		PT-Neut	
	mean	std	mean	std	mean	std	mean	std
$A_{Syn\|WordsSyn}$								
Word2Vec	−0.002	0.149	0.004	0.156	0.025	0.152	0.038	0.160
GloVe	−0.009	0.166	−0.006	0.170	0.058	0.193	0.072	0.193
ELMo	−0.023	0.108	−0.040	0.134	−0.003	0.124	0.008	0.182
SBERT ML	−0.036	0.160	−0.051	0.178	−0.019	0.154	−0.036	0.176
BERT	0.021	0.090	0.006	0.067	0.027	0.077	0.017	0.073
BERT ML	−0.002	0.044	−0.003	0.056	−0.003	0.044	−0.008	0.059
DistilBERT ML	−0.001	0.041	−0.003	0.045	−0.003	0.047	−0.002	0.049
Llama2	0.020	0.137	0.011	0.154	0.024	0.169	0.021	0.166
$A_{Syn\|Rand}$								
Word2Vec	0.049	0.156	0.054	0.162	0.076	0.165	0.074	0.182
GloVe	0.070	0.177	0.077	0.179	0.110	0.213	0.100	0.219
ELMo	0.024	0.116	0.015	0.135	0.051	0.127	0.062	0.198
SBERT ML	0.059	0.173	0.072	0.186	0.081	0.165	0.099	0.171
BERT	0.040	0.103	0.001	0.065	0.057	0.085	0.026	0.070
BERT ML	0.003	0.048	−0.006	0.052	0.000	0.042	−0.011	0.054
DistilBERT ML	−0.005	0.047	−0.007	0.049	−0.012	0.044	−0.015	0.046
Llama2	0.042	0.124	0.031	0.133	0.064	0.165	0.034	0.162

Table 13
Mean and standard deviation (std) at NC level for $Sim_{R|Syn}$ and $Sim_{R|WordsSyn}$, for English (EN) and Portuguese (PT) for naturalistic (Nat) and neutral (Neut) sentences.

Model Name	EN-Nat		EN-Neut		PT-Nat		PT-Neut	
	mean	std	mean	std	mean	std	mean	std
$Sim_{R\|Syn}$								
Word2Vec	0.164	0.365	0.159	0.362	0.221	0.356	0.183	0.373
GloVe	0.221	0.407	0.220	0.406	0.264	0.391	0.225	0.424
ELMo	0.076	0.512	0.012	0.470	0.154	0.384	0.104	0.412
SBERT ML	0.190	0.441	0.172	0.429	0.259	0.419	0.244	0.395
BERT	0.075	0.735	−0.166	0.659	0.289	0.486	0.098	0.437
BERT ML	−0.024	0.533	−0.128	0.525	−0.057	0.510	−0.194	0.566
DistilBERT ML	−0.147	0.566	−0.166	0.544	−0.257	0.589	−0.320	0.618
Llama2	0.194	0.506	0.095	0.466	0.129	0.389	0.056	0.448
$Sim_{R\|WordsSyn}$								
Word2Vec	0.173	0.182	0.167	0.181	0.165	0.187	0.124	0.204
GloVe	0.245	0.236	0.243	0.233	0.113	0.237	0.061	0.246
ELMo	0.231	0.276	0.193	0.272	0.185	0.234	0.118	0.262
SBERT ML	0.336	0.231	0.315	0.230	0.339	0.258	0.340	0.256
BERT	0.092	0.307	−0.058	0.272	0.169	0.275	0.057	0.238
BERT ML	0.034	0.294	−0.068	0.357	0.007	0.296	−0.098	0.366
DistilBERT ML	−0.105	0.284	−0.104	0.246	−0.196	0.300	−0.244	0.304
Llama2	0.094	0.380	0.058	0.368	0.099	0.240	0.039	0.255

Appendix B. Results After Removing Examples with Synonym Lexical Overlaps

As the NC_{Syn} were selected from the synonyms proposed by the human annotators, and chosen according to frequency, this led to cases of lexical overlap. Removing the NCs with lexical overlap with their NC_{Syn} and analyzing the correlations for Affinities and Scaled Similarities, the results are as shown in Tables 14 and 15. The results are compatible with those of Tables 5 and 6 for the complete set of NCs. As expected the correlations are smaller and less significant than those obtained for the full set; as with the removal of the NCs with lexical overlap a smaller set was used to calculate correlations. The ultimate test will be to redo the analysis with the full list of NCs but only using NC_{Syn} without lexical overlap, but this requires additional human annotation and is left for future work.

Table 14
Spearman ρ correlation between the Affinity and human judgments for English and Portuguese for naturalistic (Nat) and neutral (Neut) sentences after removing NCs with lexical overlap between NC and NC_{Syn}. Non-significant ($p > 0.05$) results omitted from the table.

	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
$A_{Syn\|WordsSyn}$								
EN-Nat	0.44	0.35	0.44	0.48	0.42	0.44	0.32	–
EN-Neut	0.44	0.35	0.42	0.48	0.34	0.36	0.29	–
PT-Nat	0.26	0.19	0.23	0.16	0.32	0.12	–	–
PT-Neut	0.28	–	0.26	–	0.21	–	–	–
$A_{Syn\|Rand}$								
EN-Nat	0.49	0.37	0.57	0.53	0.57	0.56	0.46	0.16
EN-Neut	0.47	0.36	0.54	0.51	0.41	0.39	0.37	–
PT-Nat	0.25	0.16	0.25	0.25	0.40	0.20	–	–
PT-Neut	–	–	0.28	0.22	0.31	–	–	–

Table 15
Spearman ρ correlation between the Scaled Similarities and human judgments, for $Sim_{R|Syn}$ and $Sim_{R|WordsSyn}$ in both English and Portuguese after removing NCs with lexical overlap between NC and NC_{Syn}. Non-significant ($p > 0.05$) results were omitted from the table.

	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
$Sim_{R|Syn}$								
EN-Nat	0.49	0.37	0.57	0.53	0.57	0.56	0.46	0.15
EN-Neut	0.47	0.36	0.54	0.51	0.41	0.39	0.37	–
PT-Nat	0.29	0.21	0.37	0.24	0.32	0.24	–	–
PT-Neut	–	–	0.40	–	0.29	–	–	–
$Sim_{R|WordsSyn}$								
EN-Nat	–	–	0.14	0.11	0.38	0.26	0.28	0.20
EN-Neut	–	–	–	–	–	–	0.22	–
PT-Nat	–	–	–	–	–	–	–	0.17
PT-Neut	–	–	–	–	–	–	–	–

Appendix C. The Impact of Rogue Dimensions

C.1 Standardization Process

To mitigate the impact of rogue dimensions, a standardization process using z-scores[31] was applied as proposed by Timkey and van Schijndel (2021). The mean vector μ was calculated across the NC sentences and subtracted from each embedding vector to center the data. Each dimension of the embedding was divided by its standard deviation σ.

C.2 Spearman Correlation Analysis

To assess the impact of standardization, Spearman correlation was calculated between the P_{Syn} cosine similarities before and after standardization:

- **Pre-standardization:** Cosine similarities calculated using the original representations.
- **Post-standardization:** Cosine similarities recalculated after standardization.

The results are reported in Table 16.

Table 16
Spearman ρ correlation for P_{Syn} cosine similarities before and after standardization (results significant for $p < 0.05$.)

Sent	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
EN-Nat				0.974	0.964	0.964	0.954	0.960
EN-Neut				0.965	0.888	0.876	0.908	0.650
PT-Nat				0.976	0.860	0.955	0.955	0.960
PT-Neut				0.952	0.874	0.874	0.911	0.927
NC	Word2Vec	GloVe	ELMo	SBERT ML	BERT	BERT ML	DistilBERT ML	Llama2
EN-Nat				0.991	0.967	0.951	0.953	0.937
EN-Neut				0.984	0.940	0.916	0.947	0.875
PT-Nat				0.987	0.852	0.939	0.939	0.939
PT-Neut				0.975	0.795	0.903	0.910	0.925

31 $z = (x - \mu)/\sigma$.

Acknowledgments

This work was partly supported by UKRI EPSRC EP/T02450X/1 and NAF/R2/202209 (UK), by CNPq 311497/2021-7 and CAPES/PRINT 88887.583995/2020-00 (Brazil), by MCIN/AEI/10.13039 /501100011033 (grants PID2021-128811OA-I00 and TED2021-130295B-C33, the latter also funded by "European Union Next Generation EU/PRTR"), by the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, ED431F 2021/01, and ED431F 2021/01), and by a Ramón y Cajal grant (RYC2019-028473-I), and by COST-Action UniDive.

References

Acosta, Otavio, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 101–109.

Adewumi, Tosin, Foteini Liwicki, and Marcus Liwicki. 2022. Vector representations of idioms in conversational systems. *Sci*, 4(4):37. https://doi.org/10.3390/sci4040037

Akbik, Alan, 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.

Apidianaki, Marianna. 2022. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523.

Baroni, Marco, 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. https://doi.org/10.1007/s10579-009-9081-4

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. https://doi.org/10.3115/v1/P14-1023

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721. https://doi.org/10.1162/coli_a_00016

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Quinta de Castro, Pedro Vitor, 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. https://doi.org/10.1007/978-3-319-99722-3_9

Chakrabarty, Tuhin, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606. https://doi.org/10.1162/tacl_a_00478

Chakrabarty, Tuhin, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159. https://doi.org/10.18653/v1/2022.emnlp-main.481

Chang, Ting Yun 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. https://doi.org/10.18653/v1/D19-1627

Church, Kenneth Ward and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83. https://doi.org/10.3115/981623.981633

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos

Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892. https://doi.org/10.1162/COLI_a_00302

Contreras Kallens, Pablo and Morten H. Christiansen. 2022. Models of language and multiword expressions. *Frontiers in Artificial Intelligence*, 5:781962. https://doi.org/10.3389/frai.2022.781962, PubMed: 35252848

Cook, Paul, 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.

Cordeiro, Silvio, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57. https://doi.org/10.1162/coli_a_00341

Dankers, Verna, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? Analyzing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626. https://doi.org/10.18653/v1/2022.acl-long.252

Devlin, Jacob, 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.

de Dios-Flores, Iria, Juan Garcia Amboage, and Marcos Garcia. 2023. Dependency resolution at the syntax-semantics interface: Psycholinguistic and computational insights on control dependencies. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 203–222. https://doi.org/10.18653/v1/2023.acl-long.12

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653. https://doi.org/10.1002/lnco.362

Ethayarajh, Kawin. 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. https://doi.org/10.18653/v1/D19-1006

Ethayarajh, Kawin and Dan Jurafsky. 2021. Attention flows are Shapley Value explanations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 49–54. https://doi.org/10.18653/v1/2021.acl-short.8

Ettinger, Allyson. 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. https://doi.org/10.1162/tacl_a_00298

Fakharian, Samin and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32. https://doi.org/10.18653/v1/2021.mwe-1.4

Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103. https://doi.org/10.1162/coli.08-010-R1-07-048

Frege, Gottlob. 1956. The thought: A logical inquiry. *Mind*, 65(259):289–311. https://doi.org/10.1093/mind/65.1.289

Garcia, Marcos. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640. https://doi.org/10.18653/v1/2021.acl-long.281

Garcia, Marcos, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741. https://doi.org/10.18653/v1/2021.acl-long.212

Garcia, Marcos, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564. https://doi.org/10.18653/v1/2021.eacl-main.310

Gow-Smith, Edward, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Improving tokenization by alternative treatment of spaces. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11430–11443. https://doi.org/10.18653/v1/2022.emnlp-main.786

Gulordava, Kristina, 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. https://doi.org/10.18653/v1/N18-1108

Haagsma, Hessel, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287.

Hartmann, Nathan, 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.

Hashempour, Reyhaneh and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80.

Henderson, James. 2020. The unstoppable rise of computational linguistics in deep learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6294–6306. https://doi.org/10.18653/v1/2020.acl-main.561

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2021. Compositionality decomposed: How do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages IJCAI'20. https://doi.org/10.24963/ijcai.2020/708

Karlgren, Jussi and Pentti Kanerva. 2021. Semantics in high-dimensional space. *Frontiers in Artificial Intelligence*, 4:698809. https://doi.org/10.3389/frai.2021.698809, PubMed: 34532704

Kassner, Nora and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818. https://doi.org/10.18653/v1/2020.acl-main.698

King, Milton 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, Volume 2: Short Papers*, pages 345–350. https://doi.org/10.18653/v1/P18-2055

Klubička, Filip, Vasudevan Nedumpozhimana, and John Kelleher. 2023. Idioms, probing and dangerous things: Towards structural probing for idiomaticity in vector space. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 45–57. https://doi.org/10.18653/v1/2023.mwe-1.8

Kurfalı, Murathan and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94.

Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211. https://doi.org/10.1037//0033-295X.104.2.211

Lenci, Alessandro, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, 56(4):1269–1313. https://doi.org/10.1007/s10579-021-09575-z

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman

Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324. https://doi.org/10.3115/1034678.1034730

Linzen, Tal, 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. https://doi.org/10.1162/tacl_a_00115

Liu, Chunxi, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52.

Liu, Emmy and Graham Neubig. 2022. Are representations built from the ground up? An empirical examination of local composition in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9053–9073. https://doi.org/10.18653/v1/2022.emnlp-main.617

Mandera, Paweł, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78. https://doi.org/10.1016/j.jml.2016.04.001

Manning, Christopher D., Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, U.S.A.*, 117(48):30046–30054. https://doi.org/10.1073/pnas.1907367117, PubMed: 32493748

Marvin, Rebecca and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. https://doi.org/10.18653/v1/D18-1151

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80. https://doi.org/10.3115/1119282.1119292

Mikolov, Tomas, 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*, pages 3111–3119.

Miletic, Filip and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512. https://doi.org/10.18653/v1/2023.eacl-main.110

Miletić, Filip and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612. https://doi.org/10.1162/tacl_a_00657

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41. https://doi.org/10.1145/219717.219748

Misra, Kanishka, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949. https://doi.org/10.18653/v1/2023.eacl-main.213

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x, PubMed: 21564253

Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 221–242. https://doi.org/10.1007/978-94-010-2506-5_10

Nandakumar, Navnita, 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. https://doi.org/10.18653/v1/W19-2004

Nedumpozhimana, Vasudevan and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62. https://doi.org/10.18653/v1/2021.mwe-1.7

Neelakantan, Arvind, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538. https://doi.org/10.1353/lan.1994.0007

Pennington, Jeffrey, 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. https://doi.org/10.3115/v1/D14-1162

Peters, Matthew E., 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. https://doi.org/10.18653/v1/N18-1202

Phelps, Dylan, Xuan-Rui Fan, Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Sample efficient approaches for idiomaticity detection. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 105–111.

Phelps, Dylan, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. https://doi.org/10.18653/v1/P19-1493

Prasad, Grusha, 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. https://doi.org/10.18653/v1/K19-1007

Rademaker, Alexandre, Valeria de Paiva, Gerard de Melo, Livy Real, and Maira Gatti. 2014. OpenWordNet-PT: A project report. In *Proceedings of the Seventh Global Wordnet Conference*, pages 383–390.

Ramisch, Carlos. 2023. *Multiword Expressions in Computational Linguistics. Down the Rabbit Hole and Through the Looking Glass*. Dissertation, Aix Marseille University. Available https://tel.archives-ouvertes.fr/tel-0421622

Reddy, Siva, 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.

Reimers, Nils and Iryna Gurevych. 2019a. Alternative weighting schemes for ELMo embeddings. *CoRR*, abs/1904.02954.

Reimers, Nils 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. https://doi.org/10.18653/v1/D19-1410

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Saakyan, Arkadiy, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183. https://doi.org/10.18653/v1/2022.flp-1.26

Sag, Ivan A., 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. https://doi.org/10.1007/3-540-45715-1_1

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

Schuster, Tal, 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. https://doi.org/10.18653/v1/N19-1162

Shwartz, Vered 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. https://doi.org/10.1162/tacl_a_00277

Sporleder, Caroline, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Tayyar Madabushi, Harish, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121.

Tayyar Madabushi, Harish, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477. https://doi.org/10.18653/v1/2022.semeval-1.13

Team, Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://doi.org/10.18653/v1/2021.findings-emnlp.294

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546. https://doi.org/10.18653/v1/2021.emnlp-main.372

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LlaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

van Schijndel, Marten 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. https://doi.org/10.18653/v1/D18-1499

Venkatapathy, Sriram and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-n) collocations by integrating features. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 899–906. https://doi.org/10.3115/1220575.1220688

Vulić, Ivan, Anna Korhonen, and Goran Glavaš. 2020. Improving bilingual lexicon induction with unsupervised post-processing of monolingual word vector spaces. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 45–54. https://doi.org/10.18653/v1/2020.repl4nlp-1.7

Vulić, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240. https://doi.org/10.18653/v1/2020.emnlp-main.586

Wagner Filho, Jorge A., 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).

Schulte im Walde, Sabine. 2023. Collecting and investigating features of compositionality ratings. *Multiword Expressions in Lexical Resources. Linguistic, Lexicographic and Computational Perspectives, Phraseology and Multiword Expressions*. Language Science Press.

Schulte im Walde, Sabine, 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. European Language Resources Association (ELRA), Portorož, Slovenia.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. https://doi.org/10.18653/v1/W18-5446

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392. https://doi.org/10.1162/tacl_a_00321

Wiedemann, Gregor, 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.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yu, Lang 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)*. https://doi.org/10.18653/v1/2020.emnlp-main.397

Zeng, Ziheng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562. https://doi.org/10.1162/tacl_a_00442

Zeng, Ziheng and Suma Bhat. 2022. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137. https://doi.org/10.1162/tacl_a_00510

Zeng, Ziheng and Suma Bhat. 2023. Unified representation for non-compositional and compositional expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11696–11710. https://doi.org/10.18653/v1/2023.findings-emnlp.783

Zeng, Ziheng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A commonsense knowledge graph for idiomatic expressions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264. https://doi.org/10.18653/v1/2023.emnlp-main.881

Zhou, Jianing, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48. https://doi.org/10.18653/v1/2021.mwe-1.5