UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

TIAGO KRAMER VIEIRA

Finding Idiomaticity in Word Representations

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Advisor: Prof.Dr. Claudio Rosito Jung

Porto Alegre
February 2023

CIP — CATALOGING-IN-PUBLICATION

Vieira, Tiago Kramer

Finding Idiomaticity in Word Representations / Tiago Kramer Vieira. – Porto Alegre: PPGC da UFRGS, 2023.

70 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2023. Advisor: Claudio Rosito Jung.

1. Multi-Word Expressions. 2. MWE. 3. Deep Learning. I. Rosito Jung, Claudio. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos André Bulhões
Vice-Reitora: Prof[a]. Patricia Pranke
Pró-Reitor de Pós-Graduação: Prof. Celso Giannetti Loureiro Chaves
Diretora do Instituto de Informática: Prof[a]. Carla Maria Dal Sasso Freitas
Coordenador do PPGC: Prof. Claudio Rosito Jung
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

ABSTRACT

Contextualised word representation models have been successfully used for capturing different word usages, and they may be an attractive alternative for representing idiomaticity in language. However, it is not clear how these models represent idiomaticity or to what extent they capture it. In this work, probing tasks are proposed to assess if some of the expected linguistic properties of noun compounds, especially those related to idiomatic meanings, and their dependence on context and sensitivity to lexical choice, are readily available in some standard and widely used representations. To evaluate that, the Noun Compound Idiomaticity (NCI) dataset was constructed, which contains annotations for noun compounds and their paraphrases, in neutral and informative naturalistic sentences, in two languages: English and Portuguese. The dataset, composed of 27,600 items, also contains human idiomaticity judgements for each noun compound at both type and token levels. For evaluation, four types of probing measures are proposed to assess how well the models distinguish idiomatic and literal meanings and is also defined as a set of metrics, that it is called affinity measures, to determine how much of these senses the compound representation captures. Results obtained with models like ELMo, BERT, and some of its variants, indicate that idiomaticity is not yet accurately represented by contextualised models. This work is a by-product of the two published papers in top-level conferences.
Keywords: Multi-Word Expressions. MWE. Deep Learning.

Encontrando Idiomaticidade na Representação das Palavras

RESUMO

Modelos que representam palavras com seu contexto vem sendo utilizados para capturar diferentes uso de palavras, e podem ser uma alternativa atrativa para representar idiomaticidade na linguagem. Entretanto, não é claro como esses modelos representam a idiomaticidade ou em qual extensão conseguem capturá-la. Nesse trabalho, são propostas medidas para avaliar se algumas das propriedades linguísticas esperadas em compostos substantivos, especialmente aqueles relacionados a significados idiomáticos, suas dependências com o contexto ao redor e as suas sensibilidades a escolhas lexicais, estão disponíveis em algumas das representações amplamente utilizadas na área. Para avaliar esses pontos, foi construído o conjunto de dados *Noun Compound Idiomaticity (NCI)*, que contém anotações para compostos substantivos e suas paráfrases, em contexto neutro e informativo, em dois idiomas: Inglês e Português. O conjunto, composto por 27.600 sentenças, também contém avaliações idiomáticas humanas para cada composto substantivo em âmbito de tipo (isolado) e contextualizado. Para avaliação, é proposto quatro tipos de medidas que avaliam quão bem os modelos distinguem significados idiomáticos e literais, e também é definido medidas um conjunto de medidas, chamadas de afinidades, que determinam o quanto desses sentidos são capturados na representação do composto. Resultados obtidos com modelos como ELMo, BERT e algumas de suas variantes, indicam que idiomaticidade ainda não é representada com precisão por modelos contextualizados. Esse trabalho é um resultado de dois artigos já publicados em conferências de alto nível.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LSTM	Long-Short Term Memory
NC	Noun Compound
NCI	Noun Compound Idiomaticity
NLP	Natural Language Processing
MWE	Multi-word expressions
PLM	Probabilistic Language Models
PMI	Pointwise Mutual Information
PPMI	Positive Pointwise Mutual Information

SUMÁRIO

1 INTRODUCTION

In natural language processing (NLP), word embeddings have been the standard mathematical tool to map words to vector spaces. The traditional methods of representing these spaces such as *word2vec* (MIKOLOV et al., 2013) and GloVe (PENNINGTON; SOCHER; MANNING, 2014), are limited in that they used a single representation to capture the different meanings of words. This shortcoming was addressed by the introduction of *contextual* word representations such as ELMo (PETERS et al., 2018a), which provided different representations of the same word depending on the context it appeared in. These contextual embeddings were further enhanced by the introduction of Transformer based pre-trained language models such as BERT (DEVLIN et al., 2019), which not only capture the different senses of polysemous words (SCHUSTER et al., 2019; CHANG; CHEN, 2019), but also seem to capture a variety of information (ROGERS; KOVALEVA; RUMSHISKY, 2020) such as about parts of speech, entity types (TENNEY et al., 2019), syntactic trees (VILARES et al., 2020) and world knowledge (PETRONI et al., 2019). However, multi-word expressions (MWEs) still represent a challenge to both model classes as their meanings may not be directly related to the meanings of their individual words (e.g., *graduate student* vs. *eager beaver* as a hardworking person). This disambiguation is only possible if the MWE is contained in a proper informative context. However, work like Dankers, Lucas and Titov (2022) show that even though models capable of taking the context into account tend to translate idiomatic expressions literally, which can be seen by the example of Figure 1.1, and the literature shows mixed results when comparing both static and non-static models (SHWARTZ; DAGAN, 2019; KING; COOK, 2018; NANDAKUMAR; BALDWIN; SALEHI, 2019; FAKHARIAN; COOK, 2021). Therefore, one open question is whether word representation models store information about MWE accurately.

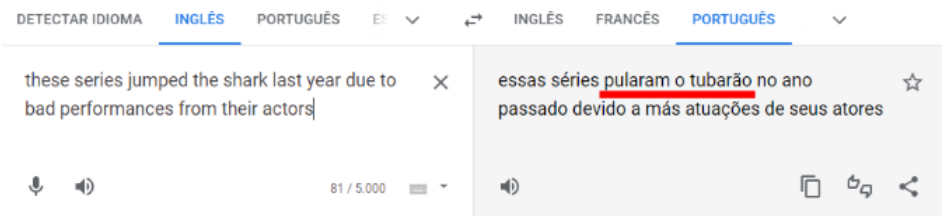


Figura 1.1 – Translation of a sentence from English to Portuguese using Google Translator. It contains the MWE *jump the shark*, which indicates that a series is declining in quality, and highlights the literal translation in Portuguese.

Hence, the goal of this work is to investigate to what extent idiomaticity in noun com-

pounds (NC) is captured by representations generated by different models, being static or contextualised.

In order to support the research, this investigation is divided into the following questions:

(i) are models sensitive to changes in meaning and idiomaticity resulting from replacing NCs by variants?

(ii) if this sensitivity exists, is it affected by the informativeness of the contexts?

(iii) is it affected by the characteristics, like its compositionality, of the NCs or the language?

The work proposes a set of probing tasks to examine how accurately idiomaticity in NCs is captured in word representation models which are widely used and freely available. In those tasks, the word embeddings from these NCs or a sentence containing them will be compared to other handcrafted examples. The aim of the work is that these tasks can serve as an indicator of the potential success (or lack of success) of a model before any subsequent additional training, e.g. for classification, or optimisation, such as fine-tuning. To answer the question i) the NC will be compared to variants like its synonyms, artificial noun compounds that are composed by the synonym of its components and another NC with a similar frequency in a specific corpus. For answering ii) the word embeddings will be extracted from two different kinds of sentences, three with naturalistic context and another one manually crafted to contain a neutral context. Last but not least, for question iii) the probing tasks will be executed in two different languages, English and Portuguese, also analysing the NCs with respect to their idiomaticity level.

All the data required for the analysis are part of a dataset called Noun Compound Idiomaticity (NCI, which is introduced in the dissertation. In total, it contains 27,600 sentences, 16,800 in English and 10,800 in Portuguese, along with human judgements from native speakers about the degree of NC idiomaticity in the naturalistic sentences both at type, meaning that is classified regardless of its context, and token level, by sentence its located in.

Therefore, the combination of both probing tasks and NCI can be used to evaluate the performance of widely adopted models in predicting idiomaticity in terms of their agreement with human judgements about idiomaticity.

Hence, the main contributions of this work include:

- a new large dataset of NCs in two languages, the NCI dataset, that can be used to evaluate to what extent models are able to detect idiomaticity at type and token level,

- in-depth analysis of the representation of idiomaticity in widely used word representation models, examining their ability to display sensitivity to perturbations linked to idiomaticity.

- the design of novel affinity measures that use relative distances between representations to determine idiomaticity in models.

The results obtained using these measures suggest that the standard and widely adopted models and composition operations display a limited ability to reflect behaviour linked to NC idiomaticity. This work is a by-product of the idiomatic probes proposed in Garcia et al. (2021b) introducing new probes, novel affinity measures, and substantially expanding the analyses with new baselines and results from a larger set of models. Moreover, it combined and considerably extended the Noun Compound Senses and the Noun Compound Type and Token Idiomaticity datasets in the new NCI dataset (GARCIA et al., 2021b; GARCIA et al., 2021a).

The remainder of this work is organised as follows: Chapter 2 dives into the necessary background literature to understand basic concepts from NLP used throughout the thesis. Chapter 3 presents related works. Then, Chapter 4 describes how the NCI dataset is built and the evaluated models. The proposed idiomatic probes and affinity measures are presented in Chapters 5 and 6, respectively. Finally, the conclusions of the work and possibly new follow-up are described in Chapter 7.

2 BACKGROUND

This chapter explores the NLP background required to understand how the work will achieve its objectives. Section 2.1 explains one of the core parts of natural language processing, how the words are mapped to vector spaces and which are the tools required for it. As the focus of this work is to explore models' capabilities with respect to idiomatic expressions, section 2.2 discusses more in-depth their complexity and why exploring them is interesting. Last but not least, section 2.3 dives into which options are available to extract what linguistic information is stored by the models.

2.1 Word Embedding

The distributional hypothesis, which predicts that "difference of meaning correlates with a difference of distribution" (HARRIS, 1954), has been empirically implemented in computational models which represent a given linguistic unit (e.g., a word type) as an n-dimensional vector (LUND; BURGESS, 1996; LANDAUER; DUMAIS, 1997; SCHÜTZE, 1998), which are also referred to as *word embedding* (BENGIO et al., 2003). These vectors, learnt in an unsupervised manner from the contexts of each word type in large corpora, have been used in different research fields such as linguistics, psychology, and cognitive sciences (MILLER, 1971; MCDONALD; RAMSCAR, 2001; SAHLGREN, 2008). Recently, the emergence of neural network architectures in NLP (COLLOBERT; WESTON, 2008) fostered the use of vector representations, now exploiting the distributional hypothesis by predicting (instead of previous count-based approaches) the surrounding contexts of the target words to generate word embeddings, a method popularised by the introduction of *word2vec* (MIKOLOV et al., 2013), a more efficient method of generating neural embeddings. Despite their success, a potential drawback of these language models, both using count-based and predicted-based strategies, is that they represent the different senses of a word in the same (static) vector, so that complex operations are needed to deal with semantic phenomena such as homonyms or polysemy (ERK, 2012).

The use of probabilistic neural language models (BENGIO et al., 2003) has been proposed to overcome this issue, as they generate different word representations in each context (thus being contextualised embeddings). In this respect, models such as ELMo (PETERS et al., 2018a), which uses LSTM networks (HOCHREITER; SCHMIDHUBER, 1997), or BERT (DEVLIN et al., 2019), trained with a Transformer architecture (VASWANI et al.,

2017), have become ubiquitous in the NLP field as they have considerably advanced the state-of-the-art of downstream tasks. For Henderson (2020), the Transformer-based models succeed in several NLP tasks due to their ability to learn not only vector representations of words but also to induce linguistic structures by means of their multi-head attention mechanism. The good performance of these neural language models has given rise to a number of studies exploring their linguistic capabilities, namely from a syntactic viewpoint (LINZEN; DUPOUX; GOLDBERG, 2016; LINZEN; BARONI, 2021).
After the word embeddings are generated, often is interesting to compare how close two vectors are in a specific n-dimensional space. For this task, one of the most used operations in literature is cosine similarity (SCHONE; JURAFSKY, 2001; PENNINGTON; SOCHER; MANNING, 2014; REDDY; MCCARTHY; MANANDHAR, 2011; SALEHI et al., 2015).

2.2 Context Importance for Words and Expressions

There are several examples of different languages that the same word representing different senses. For example, in English, the word *bank* in the following two sentences *The **bank** was robbed yesterday* and *The house is located near the river **bank*** have different meaning, the first by representing the building and the latter the margin of the river. One of the reasons we are capable of disambiguating which sense is being used is by the context of the word in the sentence, which itself is an NLP research area (FINLAYSON; KULKARNI, 2011; SCHNEIDER et al., 2016).
Context also plays a very important role in multi-word expressions, which are composed of two or more words that can be treated by a unit having one or more senses, which can or cannot be inferred by the meaning of its components, which is often referred as semantic idiomaticity (MASINI, 2019). In Portuguese, the expression *bater as botas*, can be either recognised by its literal meaning of someone stomping their boots on the floor (to remove some dirt or muck) or by the idiomatic meaning of dying. Three characteristics from those expressions can be highlighted:

- **Idiomaticity or non-compositionality**, that is, the degree with which the meaning of an MWE cannot be inferred from a composition of the meanings of its parts (*big fish* as an important person). It is also understood as *semantic opacity* and its continuum as different *degrees of opacity* (CRUSE, 1986).

- **Polysemy and ambiguity**, or to what extent the MWE can appear both in a literal

and an idiomatic meaning and the impact of context in determining the specific sense (e.g., *bad apple* as either a rotten fruit or a troublemaker) in (*This/He is a **bad apple***).

- **Non-substitutability** of individual MWE components, or the degree with which a specific lexical item cannot be replaced for a noun-compound to still retain its meaning (e.g. *police car* and *police vehicle* vs. *panda car* and **bear automobile*),

An interesting subset of multi-word expressions, which serves as the focus of the work, is the nominal compounds which are noun phrases containing two or more content words, whose head is a noun, but is accompanied by an adjective or another noun. Depending on the language they occur in different combinations, like noun-noun (e.g. *field work* in English and *acampamento militar* in Portuguese), adjective-noun (e.g. *bad apple* or *black hole* in English), or noun-adjective (e.g. *algodão doce* or *buraco negro* in Portuguese).

2.3 Probing Models for Linguistic Information

Although the language models seem to contain considerable linguistic information, uncovering how they capture a specific type of knowledge is a non-trivial problem, and it depends on factors like the particular model and the way words are encoded (YU; ETTINGER, 2020; VULIĆ et al., 2020).

Therefore there are three primary methods of *probing* or answering the question of whether or not a model has access to specific linguistic (or similar) information: i) is by use of a 'probing classifier' to find a linear transform between the sentence representation output by a model and the property being investigated, thus using the performance of the classifier as a proxy to evaluate the representations (VELDHOEN; HUPKES; ZUIDEMA, 2016). ii) usage of an information theoretic perspective by observing that any regularity in the labels can be used to compress them and translates probing to the task of transmitting label information in as few bits as possible (VOITA; TITOV, 2020). iii) Finally, known as 'inoculation by fine-tuning' is the fine-tuning of the model using a specific (and much smaller) training set (RICHARDSON et al., 2020; LIU; SCHWARTZ; SMITH, 2019). If a model can be fine-tuned to attain an acceptable performance this indicates that somehow the necessary linguistic information is present.

The methods like the ones proposed by ii) are interesting because do not require any further training, be it of classifiers (even if linear) or of fine-tuning because that introduces new variables and analysis for the operations themselves. For example, there is debate

surrounding the amount of training data used during inoculation and the complexity of the probing classifier so as to distinguish between probing for a phenomenon and training on a task (ROGERS; KOVALEVA; RUMSHISKY, 2020).

3 RELATED WORKS

This chapter explores the related works which the dissertation uses as references to construct the proposed probing tasks and the NCI dataset, as well as a new relative similarity measure called affinities, additionally, it also mentions the works that serve as inspiration to question whether the contextualised models can represent successfully linguistic information with respect to idiomaticity.

In establishing the extent to which PLMs capture idiomaticity, there have been two primary directions of exploration: The first has been in terms of their ability to detect if a sentence contains an idiom and the second, which we focus on in this work, has been in establishing their ability to correctly encode the idiomatic meaning. Studies have shown that PLMs are surprisingly good at identifying idiomaticity (SHWARTZ; DAGAN, 2019; TAN; JIANG, 2021) and that they seem to encode this information mostly on the representation of the idiomatic expressions itself, while the sentential context also plays an important role (NEDUMPOZHIMANA; KELLEHER, 2021). In fact, Zeng and Bhat (2021) have recently presented a supervised neural architecture —by combining contextualised and static embeddings— which exploits the notion of *semantic compatibility*, i.e., whether the MWE is semantically compatible with its context, to identify and classify potentially idiomatic expressions. As we show in this work, PLMs are rather poor at representing idiomaticity.

With respect to lexical semantics, contextualised models seem to represent words more accurately than static word embeddings. On the one hand, static word embeddings can still be obtained from contextualised representations, achieving better performance on lexical semantic tasks at type level (VULIĆ et al., 2020). On the other hand, and more relevant to our work, diverse studies have shown that representations of a word in several contexts can be grouped in different clusters which seem to be related to the various senses of the word (COENEN et al., 2019; SCHUSTER et al., 2019; WIEDEMANN et al., 2019). It is worth noting, however, that the results of Haber and Poesio (2020) suggest that contextualised word embeddings produced by BERT represent contextual information rather than word senses.

The comparison of static and contextualised models for representing MWEs has been reported with mixed results. Evaluating different classifiers initialised with contextualised and non-contextualised embeddings in tasks related to lexical composition (including the literality of NCs), Shwartz and Dagan (2019) found that contextualised models, especially

BERT, obtained better performance across tasks. However, for capturing idiomaticity in MWEs, static models like *word2vec* seem to have better performance than contextualised models (KING; COOK, 2018; NANDAKUMAR; BALDWIN; SALEHI, 2019). Nevertheless, the supervised method of Fakharian and Cook (2021), using Transformer models, obtained better results than previous approaches on the classification of potentially idiomatic expressions in both monolingual and cross-lingual scenarios (in English and Russian).

With respect to semantic composition, Yu and Ettinger (2020) explored the representation of two-word phrases (which in many cases correspond to NCs as the ones used in our study) in various Transformer models, showing that phrase representations miss compositionality effects as they heavily rely on word content. Crucially, most of these experiments evaluate idiomaticity at the type level, i.e., they obtain the embedding of a given MWE by averaging its representation in several sentences that have been previously extracted in an automatic way.

The boundaries between metaphorical and idiomatic expressions are not always clear, as both of them convey non-literal meanings. In fact, conventionalised metaphors can be classified as idiomatic MWEs, and therefore similar approaches can be used to their identification (DINH; EGER; GUREVYCH, 2018). Even though the field of automatic identification of metaphors has its own tradition, some datasets could be used in both areas (e.g., the adjective-noun pairs of Tsvetkov et al. (2014)), and recent papers also explore the representation of metaphorical expressions on current neural language models (PEDINOTTI et al., 2021; AGHAZADEH; FAYYAZ; YAGHOOBZADEH, 2022), as mentioned for MWEs. More generally, studies on figurative knowledge have shown that the PLMs, while good at identifying figurative language struggle with correctly representing it, but can be improved through the incorporation of external knowledge (CHAKRABARTY; CHOI; SHWARTZ, 2022)

Inspired by psycholinguistic experiments, which investigate how the psychological processes involved in the use of language (KENNISON; MESSER, 2014), Ettinger (2020) proposes a diagnostics in order to understand the language models' linguistic knowledge. In the paper, the usage of handcraft tasks inspired by real human studies shows that, in general, BERT distinguishes good from bad completions, but it is easily distracted by negation, which is also confirmed by Kassner and Schütze (2020). This method is interesting because it does not involve any fine-tuning or training of any linear model on it to extract a prediction.

4 METHODOLOGY

In this work, we rely on probing tasks on models, both static and non-static, without any fine-tuning or additional classifiers. Figure 4.1 depicts how the evaluation is done in general. Sentences with linguistic relationships are input into a model, generating the word embedding representation, and then we compare those vector spaces using a similarity-base metric, evaluating them for the characteristic we are interested in.

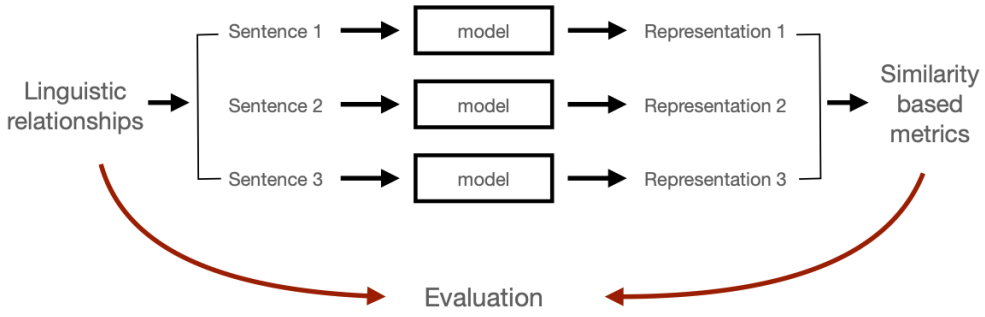


Figura 4.1 – Diagram illustrating the process of probing a model.

This chapter presents how the sentences for the probing tasks are formed and how they build the Noun Compound Idiomaticity (NCI) dataset, which contains NCs with different levels of idiomaticity in two languages, English (EN) and Portuguese (PT), as well as how the sentences and their variants. The last section explores which models are probed together with their publicly available pre-trained weights.

4.1 The Noun Compound Idiomaticity Dataset

The NCI dataset is based on the Noun Compound Senses and on the Noun Compound Type and Token Idiomaticity datasets (GARCIA et al., 2021b; GARCIA et al., 2021a), significantly extended with new annotations and data. The English and Portuguese subsets of the NC Compositionality (NCC) dataset (CORDEIRO et al., 2019), with 280 (EN) and 180 (PT) two-word NCs, where each compound is annotated with human judgements about idiomaticity at the type level, in the context of three naturalistic corpus sentences that exemplify the same compound sense serve as a basis for our annotations. For each NC there is a single idiomaticity score, which is the average of the human judgments using a *Likert* scale from 0 (idiomatic) to 5 (compositional), along with the synonyms provided by the annotators for the NC in the target sentences. However, for NCI, the human judgements

about idiomaticity were collected at the token level, with each NC having an idiomaticity score for each sentence in which it occurs. This allows for a fine-grained investigation of the impact of each context in the interpretation of the NC, compared with the type-level judgements of the original dataset. Then, for each NC, paraphrases were collected both at the type level and at the token level, with different paraphrases for these different contexts. Finally, several subsets of naturalistic and neutral sentences are created to define probing tasks aimed at assessing the perception of variants of NCs at different levels of idiomaticity, together with baseline sets with automatically generated compounds.
Each noun compound was classified by language experts regarding their semantic compositionality as idiomatic (e.g., *gravy train*), partly compositional (e.g., *grandfather clock*), or compositional (e.g., *research project*). For English, this resulted in 103, 88, and 89 idiomatic, partly idiomatic, and compositional compounds. For Portuguese, each class has 60 compounds, as the selection had been balanced when the source dataset was created.

4.1.1 Token-level annotations

The idiomaticity annotations at the token level were obtained using the same original sentences as in the source NCC dataset, asking each annotator to give a 0 (idiomatic) to 5 (compositional) score for an NC in a specific context (e.g., *glass ceiling* in "Women are continuing to slowly break through the *glass ceiling* of UK business [...]"). The same protocol as in Reddy, McCarthy and Manandhar (2011) is used and Cordeiro et al. (2019), asking participants for: (i) the contribution of the head to the meaning of the NC (e.g., is a *glass ceiling* literally a *ceiling*?); (ii) the contribution of the modifier to the meaning of the compound (e.g., is a *glass ceiling* literally of *glass*?); and, (iii) the degree of compositionality of the compound (i.e., to what extent the meaning of the NC can be seen as a combination of its parts). In addition, it was asked the participants to provide up to three synonyms of each NC in each particular sentence. To obtain the annotations, it was used Amazon Mechanical Turk for English, and an online platform for Portuguese,[1] as it was not found an adequate number of native speakers of Portuguese in AMT, and therefore there is, on average, fewer annotations per compound in Portuguese than in English.
Before submitting the data to the annotators, it was randomised the three sentences per compound of the original dataset (840 and 540 sentences in English and Portuguese, respectively). Then, it was compiled at least 10 annotations per sentence in English,

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

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

(a) Krippendorff's α inter-annotator agreement for the NC, head, and modifiers for 2 and 3 annotators.

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

(b) Spearman ρ correlations between the idiomaticity scores at type (from the NCC dataset) and token level.

Tabela 4.1 – Inter-annotator agreement for the NCs and their components (Table 4.1a, at left), and Spearman ρ correlations ($p < 0.01$ in all cases) for all the compounds (*All*) and for each of the three compositionality classes (Table 4.1b, at right).

and 5 in Portuguese. In English, it was obtained 8,725 annotations, an average of 10.4 annotations per sentence. A total of 412 participants labelled on average 21 instances. In Portuguese, 5,091 annotations (9.4 annotations per sentence) by 22 annotators, so each participant annotated an average of 154 sentences.

Inter-annotator agreement To evaluate the labelling process, the inter-annotator agreement is computed for the two and three annotators with the largest number of sentences in common. For English, Krippendorff's α (KRIPPENDORFF, 2011) of 0.30 for two annotators (with 199 sentences in common) and 0.22 for three participants (with 76 sentences) is calculated. For Portuguese, the α values were of 0.52 and 0.44 for three and two annotators (with 131 and 60 sentences, respectively). The full results are shown in Table 4.1a. In general, and using the divisions proposed by Landis and Koch (1977), the agreement results can be classified as 'fair' and 'moderate' for English and Portuguese, respectively.

Correlation token vs. type scores It is then analysed the compatibility between the original scores at type-level from the NCC dataset and the new token-level annotations, using the micro-average compositionality values of each NC (Table 4.1b). Strong to very strong correlations are obtained, which confirm the robustness between the human annotations of idiomaticity at both type-level and token-level: overall Spearman $\rho = 0.92$ for English and $\rho = 0.90$ for Portuguese.[2]

Idiomaticity values Regarding the annotation of idiomaticity, Table 4.2 contains the mean compositionality scores and standard deviations for each class in the two languages. The results of idiomatic and compositional compounds are more homogeneous in English, as they are clearly located on the margins of the scale (< 1 and > 4, respectively), also with lower deviations. However, in Portuguese the average values are > 1 and < 4 for idiomatic and compositional NCs, respectively, even ranking the idiomatic cases closer to the middle

[2]It is worth noting that removing annotators with the low agreement (Spearman $\rho < 0.2$, and $\rho < 0.4$) produced almost identical results.

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

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

of the spectrum.

Looking at the average values for heads and modifiers, the following observations can be remarked: First, both head and modifier scores are consistently higher than the means for the whole compound in every scenario also suggesting at least a partial compositionality in their token occurrences. Second, for idiomatic NCs, the scores of the modifiers are higher than those of the heads, while for partially compositional NCs the results are the opposite.[3] Finally, regarding the compositional level, the modifier values are higher in English, while in Portuguese the heads seem to contribute more to the meaning of the NC.

Paraphrases of NCs The original NCC dataset includes paraphrases for each NC at type-level, provided by the annotators after reading three sentences for a given compound. Here, this resource is enriched with new synonyms which can be classified at type or token levels. To do so, it was asked the participants to provide synonyms or paraphrases for the noun compounds in each particular context. It is worth noting that while some suggestions may be applicable across all the sentences for an NC (e.g. *spun sugar* for *cotton candy*, considered as a type-level synonym), others are more dependent on context and differ for specific sentences (e.g. *flight recorder* and *unknown process*, for *black box*, which can be considered as token-level paraphrases). The paraphrases are classified as type or token level using the following procedure: type-level synonyms are those paraphrases proposed for the three sentences of each compound, and those suggested for two sentences with a frequency $>= 3$; token-level synonyms are those proposed only for one sentence with a frequency $>= 2$.

In English, 9,690 different paraphrases were proposed by the annotators (average 34.60 per NC), and 3,554 were suggested by at least 5 participants (average of 12.70 per NC). Out of them, 1,506 were classified as type-level (5.4 synonyms per NC, on average), and 353 at

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

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

Tabela 4.3 – Annotation example of the English NC *disc jockey*. Each row includes a sentence with the target NC together with the mean idiomaticity score and a token-level paraphrase. Bottom row shows the most common (type-level) paraphrase and the mean idiomaticity score from the original dataset (also at the type-level).

token-level (0.42 per sentence, 1.3 per NC). Overall, 118 NCs have token-level synonyms for one sentence, 69 for two sentences, and 16 for three sentences.

For Portuguese, the annotators suggested a total of 6,579 paraphrases (314 by at least 5 participants and 764 by $>= 3$, an average of 4.2 per NC). 743 synonyms were proposed for the 180 compounds (an average of 4.1 per NC), being classified as type-level. Concerning token-level synonyms, it was collected 192 equivalents (1.1 per NC, on average). In this case, the total number of annotations was lower, and the final resource contains 61 NCs with token-level synonyms for one sentence, 38 for two sentences, and 6 compounds have token-level synonyms for three sentences.

Table 4.3 shows an annotation example for the NC *disc jockey*, in English. It includes the three sentences together with the average idiomaticity scores and both token-level and type-level paraphrases. The collection of paraphrases included in the NCI dataset makes it a valuable resource for different evaluations, such as lexical substitution tasks and assessments of the performance of embedding models to correctly identify contextualised synonyms of NCs with different degrees of idiomaticity.

Finally, and aimed at observing the effect of statistical data on our experiments, each NC is annotated with frequency, PMI and PPMI (CHURCH; HANKS, 1989) values, calculated from the ukWaC (2.25B tokens, Baroni et al. (2009)) and brWaC corpora (2.7B tokens, Filho et al. (2018)).

4.1.2 Probing sentences

Out of these annotations, are created several subsets of probing sentences for the English and Portuguese NCs. For each compound, the sentences exemplify two conditions: (i) the naturalistic context provided by the original sentences (Nat), and (ii) a neutral context

where the NCs appear in uninformative sentences ($Neutral$), with only 5 words[4] following the pattern *This is a/an <NC>* (e.g. "This is an *eager beaver*") and the Portuguese equivalent *Este/a é um(a) <NC>*. As some NCs may have both compositional and idiomatic meanings (e.g. *fish story* as either *an aquatic tale* or *a big lie*), these neutral contexts will be used to examine the representations that are generated for the NCs (and the sentences) in the absence of any contextual clues about the meaning of the NC. Moreover, they enable the examination of possible biases in the NC representation towards an idiomatic or literal sense, especially when compared to the representation generated for the Nat condition.

For each NC and condition, three sentence variants are created with lexical replacements of the target compound:

- NC$_{Syn}$: Selected synonyms of NC as a whole (referred to throughout the work as the true or holistic synonym), using the most frequent type synonyms provided by the annotators of the original NCC dataset (e.g. *brain* for *grey matter*). In a few cases, where the same synonym did not fit the three sentences, paraphrases classified as type-level in the process described above are selected. A total of two synonyms for each NC in both languages.

- NC$_{WordsSyn}$: New NCs using synonyms of each component (e.g. *alligator* for *crocodile* and *sobs* for *tears*, in the NC *crocodile tears*), which were manually extracted from WordNet (MILLER, 1995; RADEMAKER et al., 2014, for English and Portuguese, respectively) and from dictionaries of synonyms. In cases of ambiguity (due to polysemy or homonymy), the most common meaning of each component was used.

- NC$_{comp}$: Two additional sentence variants for each compound, including only one component of the NC, i.e., replacing the NC by its head in one sentence, and by the modifier in the other one. These examples will be used to explore the contribution of each component to the representation of the whole compound.

Experts (native or near-native speakers with a background in Linguistics) reviewed all the newly generated sentences, keeping them as faithful as possible to the original ones, but with small modifications for preserving grammatic after the substitution (e.g. modifications in determiners and adjectives related to gender, number and definiteness agreement). The

[4]Additionally, a longer uninformative sentence (with 10 words in EN and 9 in PT) was also probed and the results are further discussed in the appendix.

NC	Sentence
Original	John Paul II was an effective *front man* for the catholic church.
NC_{Syn}	John Paul II was an effective *representative|leader* for the catholic church.
$NC_{WordsSyn}$	John Paul II was an effective *forepart woman* for the catholic church.
NC_{comp}	John Paul II was an effective *man* for the catholic church.
NC_{Rand2}	John Paul II was an effective *long beach* for the catholic church.
W_{Rand}	John Paul II was an effective *battlefront serviceman|homo* for the catholic church.

Tabela 4.4 – Example of the lexical replacements in a naturalistic sentence for the original NC *front man*. The top rows include the probing sentences, while the bottom rows contain the baseline substitutions.

top rows of Table 4.4 display the sentence variants generated by the lexical replacements in a Nat example in English.

4.2 Baseline sentences

Three additional subsets of sentences were also created to provide baselines for probes, replacing the original NCs with:

- NC_{Random}: Other compounds with similar frequency values extracted from large corpora (in this case ukWaC and brWaC) as follows: the frequency of each NC and of its components ($avg = \frac{f_{NC}+f_{w1}+f_{w2}}{3}$) are averaged, and extracted the compound with the closest average value and the same morphosyntactic pattern. For each NC, 5 random replacements were used for each sentence. Once again, although the morphosyntactic agreement of the resulting sentences was manually reviewed, the sentences may still be semantically incongruous.

- W_{Random}: Artificial compounds built by means of automatically extracted semantically related words. For each NC, WordNet (for English) and OpenWordNet.PT (for Portuguese) are used to compile each component synonyms (or hypernyms and cohyponyms in cases with less than 2 synonyms) and generate 4 combinations.

The NCI dataset contains a total of 16,800 test items for English and 10,800 for Portuguese among neutral and naturalistic sentences.

4.3 Models

We evaluate current contextualised representations obtained from neural language models together with static non-contextualised embeddings as baselines. For the latter, the com-

parison is among GloVe, *word2vec* and *fastText* (GRAVE et al., 2018), using the official models for English, and the 300 dimensions vectors described by Hartmann et al. (2017) for Portuguese.
Regarding contextualised representations, the following models were probed:

- ELMo (PETERS et al., 2018b): Composed by two Long-short Term Memory (LSTM) (HOCHREITER; SCHMIDHUBER, 1997) lanes, one in the forward and another in the backwards direction.

- BERT (DEVLIN et al., 2019): Transformer-based model (VASWANI et al., 2017) which generates an embedding for each token (word or sub-word) of a fixed-length sentence. Differently from ELMo, it does not recur to recursion.

- DistilBERT (DistilB) (SANH et al., 2019): This is a model derived from BERT, which was generated by using the knowledge distillation technique, meaning that it's intended to have the same power as BERT but with a much smaller and faster model.

- Sentence-BERT (SBERT) (REIMERS; GUREVYCH, 2019b): BERT fine-tuned by approximating sentences with similar meaning and moving away those that don't share, possibly yielding semantically richer embedding.

With respect to pre-trained weights probed, for ELMo, the work explores the small model provided by Peters et al. (2018a), and for Portuguese is adopted the weights provided by Castro, Silva and Soares (2018). For all the other contextualised models, their pre-trained weights are publicly available through their Flair[5] (AKBIK et al., 2019) and HuggingFace[6] (WOLF et al., 2020) implementations. For BERT-based models (and for DistilB in English), the results reported are both by using the multilingual uncased (ML) and by monolingual models for English (large, uncased) and Portuguese (large, cased).
Embeddings for the whole sentence as well as for the NCs are generated by averaging the word (sub-word) embeddings of the relevant tokens involved. Each model requires a different approach for generating these representations: for the static models, the word embeddings are derived directly from its vocabulary, with missing out-of-vocabulary words being ignored; for the Transformer based models, the word embeddings are generated by averaging the representations of the sub-tokens; and for ELMo, although its processing is a character-based input, it outputs an embedding by word, which is also averaged.

[5] <https://github.com/flairNLP/flair>
[6] <https://github.com/huggingface/transformers>

For the Transformer models, different combinations of layers were used in the analyses. However, as they led to qualitatively similar results, for reasons of presentation clarity, the discussion will mostly focus on the performance obtained by adopting the last four layers, as this is a widely adopted configuration. A discussion of alternative layer combinations can be investigated later as a follow-up work. For ELMo, as it is intended to serve as a contextualised baseline, it is represented the word embeddings using the concatenation of its three layers, albeit it is known that separate layers and weighting schemes generate better results in downstream tasks (REIMERS; GUREVYCH, 2019a).

5 PROBING FOR IDIOMATICITY

This chapter presents a set of idiomatic probes for analysing how sensitive vector space models are to some of the properties of NCs, including their potential for non-compositionality (*big fish* as an important person), non-substitutability (*police/panda/*bear car*) and ambiguity (*bad apple* as either a rotten fruit or a troublemaker). In particular, the following aspects are analysed:

1. What are the idiomatic properties captured by the NC representations given a specific sentence? The **embedding of an NC in context** is represented from the contextualised embeddings of its components in that particular sentence, as $\epsilon_{\text{NC} \subset \text{S}}$.

2. What are the idiomatic properties of the NC representations in the absence of any context, that is, the embedding calculated using only the NC as input sentence? The **embedding for an NC out of context** is analysed are derived from the model output, denoted as ϵ_{NC}.

3. What are the idiomatic properties captured by the representations of sentences containing NCs? For this, the **embedding of a sentence that contains an NC** is investigated, which is denoted as $\epsilon_{\text{S} \supset \text{NC}}$.

Similarities between embeddings of words and sentences are calculated using cosine similarity, $\text{cossim}(\epsilon, \epsilon')$, where ϵ and ϵ' are embeddings from the same model with the same number of dimensions. These embeddings are calculated according to the specificity of each of the models as described in section 4.3. Moreover, for static embeddings $\epsilon_{\text{NC} \subset \text{S}} = \epsilon_{\text{NC}}$.

For naturalistic sentences, the similarity scores for each NC are calculated at an individual sentence level[1] The Spearman ρ correlation is applied between the cosine similarities and the NC idiomaticity scores to check for any effects of idiomaticity in the probes. It is also calculated Spearman ρ correlation between cosine similarities coming from different embedding models to determine how much the models agree, and between similarities coming from the same model but in a different context conditions (*Nat* and *Neutral*) to see how the context affects the similarities. These results are complemented by analyses of the distribution of cosine similarities produced by different models. The results also include a qualitative analysis where the results for the five English NCs in Table 5.1 are compared,

[1]As the NCC dataset also provides the score independently of the sentence, an evaluation is done also by averaging all sentences, which is further detailed in the appendix.

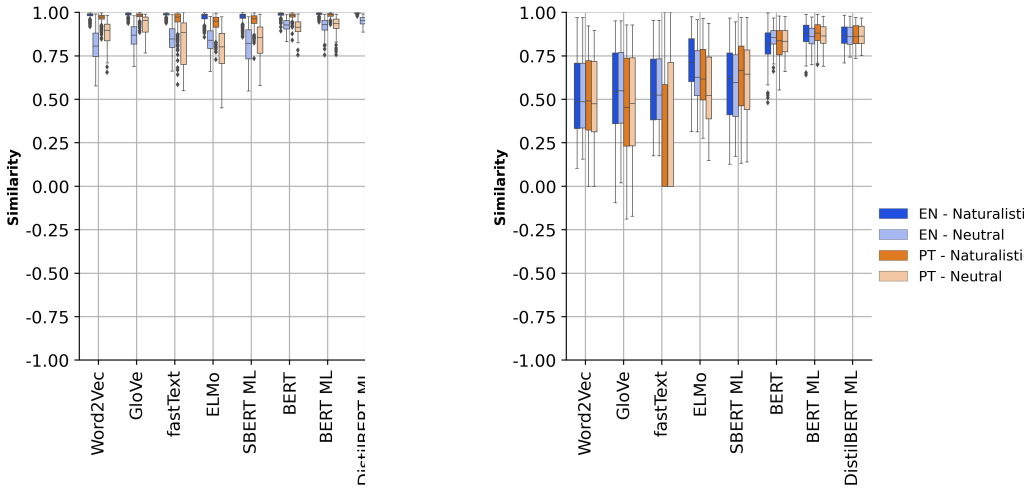
Naturalistic sentence	**NC**	**NC_{Syn}**	**$NC_{WordsSyn}$**
Field work *and practical archaeology are a particular focus.*	field work	research	area activity
The town centre is now deserted - it's almost like a ***ghost town****!*	ghost town	abandoned town	spectre city
How does it feel to experience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

Tabela 5.1 – Naturalistic examples with their NC_{Syn} and $NC_{WordsSyn}$ counterparts.

which shows a naturalistic sentence for each NC, together with their respective holistic synonyms (NC_{Syn}) and a compositional synonym derived by replacing each component word individually by a synonym ($NC_{WordsSyn}$).[2] Some sets of naturalistic sentences and their results and other examples can be found in the Appendix.

5.1 P1: Can vector space models capture the similarity between an NC and its synonym?

This first probe assesses the expected proximity between an NC (e.g. *grey matter*) and one of its (holistic) synonyms (e.g. *brain*). The semantic similarity between an NC and its synonym (NC_{Syn}) should be reflected in their embeddings, both in context (i.e. $\text{sim}^{(P1)}_{NC \subset S} \simeq 1$ where $\text{sim}^{(P1)}_{NC \subset S} = \text{cossim}(\epsilon_{NC \subset S}, \epsilon_{NC_{Syn} \subset S})$) and out of context (i.e. $\text{sim}^{(P1)}_{NC} = \text{cossim}(\epsilon_{NC}, \epsilon_{NC_{Syn}}) \simeq 1$), with the embeddings of their sentences also being similar ($\text{sim}^{(P1)}_{S \supset NC} \simeq 1$, where $\text{sim}^{(P1)}_{S \supset NC}$ = $\text{cossim}(\epsilon_{S \supset NC}, \epsilon_{S \supset NC_{Syn}})$). Moreover, an NC should be close to its synonym, regardless of how idiomatic it is, so that the similarities obtained for P1 (between NCs and synonyms) are not expected to correlate with NC idiomaticity scores (i.e., $\rho^{(P1)}_{S \supset NC} \simeq 0$, $\rho^{(P1)}_{NC \subset S} \simeq 0$ and $\rho^{(P1)}_{NC} \simeq 0$).



(a) Sentence (b) NC

Figura 5.1 – P1 Cosine similarities for English (blue) and Portuguese (orange), for the Sentence (on the left) and for NC representations (on the right). Values for naturalistic sentences are shown in darker shade and for neutral long and short in lighter shades.

5.1.1 Results

The results confirm the expected high cosine similarities between a sentence with NC ($\epsilon_{S \supset NC}$) and its variant with NC_{Syn} for all models (Figure 5.1, left). However, all similarities are saturated and decrease when the analyses on the similarities between the embeddings of NC and NC_{Syn} are focused (Figure 5.1, right), which display a much larger spread of similarities.[3] This suggests that a high cosine for sentence embeddings provides a misleading indication of how similar an NC is to its synonym. The similarities are even lower for NCs in neutral contexts, especially for short contexts (*Neutral*). Contrary to what was expected, moderate to fair correlations were found between most models and the idiomaticity scores of human annotators (see Table 5.6 for static, and Table 5.7 for contextualised models), indicating lower similarities for idiomatic than for compositional cases.[4]

If, at first sight, the high similarity scores for sentences ($sim^{(P1)}_{S \supset NC}$) seem to suggest that these models are able to capture idiomaticity, when combined with lower scores for the representations of the NCs in context ($sim^{(P1)}_{NC \subset S}$), especially in neutral sentences, along with the correlations with idiomaticity scores, a different story seems to emerge. The high cosine similarities may be partly due to contextualised models being anisotropic, occupying a narrow cone in the vector space and therefore tending to produce higher

[2]Neutral sentences are omitted, since they all follow the same patterns.

[3]The same is observed for out-of-context compound nouns, shown in the Appendix.

[4]For the static models, the slight differences in Spearman are due to differences in morphological inflexion in the naturalistic sentences.

cosine similarities (ETHAYARAJH, 2019) in contrast to static embeddings. In addition, the high similarities found for the sentences may also be an effect of the overlap in words between a sentence and its variants (only the NC and the synonym change). This would also explain the larger similarities observed for naturalistic than for neutral sentences since the average sentence length for naturalistic is 23.4 and $\sigma \approx 8.10$ for English (lexical overlap $> 91\%$) and 13.0 and $\sigma \approx 4.11$ for Portuguese ($> 84\%$), while for the neutral it is 5 words ($> 60\%$) for both languages. The correlation between the models' outputs and the naturalistic sentence length presented in the table 5.8 shows a moderate positive correlation having both ELMo and BERT ML with lowest $\rho \approx 0.50$ and the highest with DistilBERT $\rho \approx 0.68$ for both English and Portuguese.

Another effect that could influence the similarity in the NC comparison towards 1 is the overlap between the NC and its synonym. In the NCI dataset, out of the 280 compounds in English, 102 have an overlap of at least one component (e.g. *ancient history* with *history*) and for Portuguese, from a total of 180, there are 74 (e.g. *alta-costura* with *alta costura*). Analysing this probe without the components that have overlap, the Tables 5.2 and 5.3 refer to in-context and 5.4 to out-of-context correlations.

	word2vec		GloVe		*fastText*	
	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	-	0.51	-	0.47	-	0.5
$EN_{Neutral}$	0.47	0.52	0.44	0.47	0.47	0.5
PT_{Nat}	-0.13	0.4	-	0.33	-	-
$PT_{Neutral}$	-	0.36	-	0.32	-	-

Tabela 5.2 – Spearman ρ correlation between the static models output and human judgments only for the components witch do not have their synonym with an overlapping component, p≤0.05, for P1. Non-significant results omitted from the table.

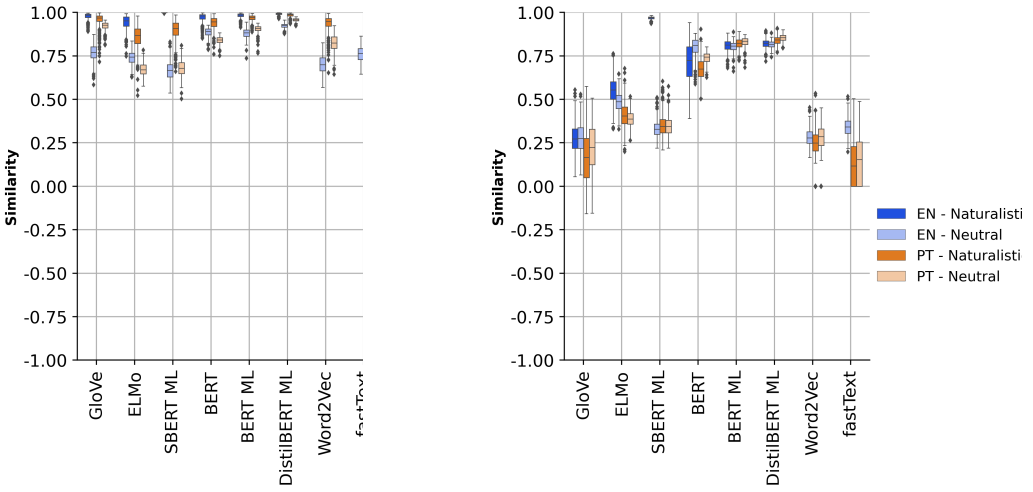
	ELMo		BERT		BERT ML		DistilB ML		SBERT ML	
	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	0.1	0.44	-	0.09	0.16	0.55	-	0.39	0.13	0.53
$EN_{Neutral}$	0.45	0.49	0.34	-	0.39	0.47	0.42	0.37	0.53	0.56
PT_{Nat}	-	0.3	-	0.49	-	0.18	-0.15	-	-	0.3
$PT_{Neutral}$	-	0.35	-	0.4	-	-	-	-	-	-

Tabela 5.3 – Spearman ρ correlation between the non-static models output and human judgments only for the components witch do not have their synonym with an overlapping component, p≤0.05, for P1. Non-significant results omitted from the table.

	word2vec	GloVe	*fastText*	ELMo	BERT	BERT ML	DistilB ML	SBERT ML
EN	0.52	0.47	0.5	0.5	0.21	0.3	0.38	0.56
PT	0.36	0.32	-	0.3	0.28	-	-	0.33

Tabela 5.4 – Spearman ρ correlation between the all models output and human judgments, p≤0.05, for P1 in out-of-context probe measure for both English and Portuguese, and only for the components without an overlapping component. Non-significant results omitted from the table.

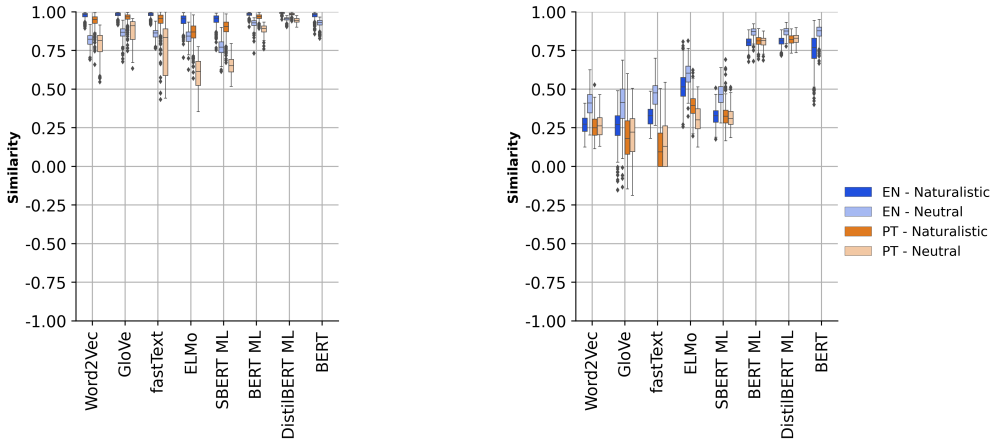
In Figure 5.2, it is seen a general probing baseline, which is calculated by averaging the cosine similarity between the noun-compound and five different frequency compatible two-words combination.



(a) Sentence (b) NC

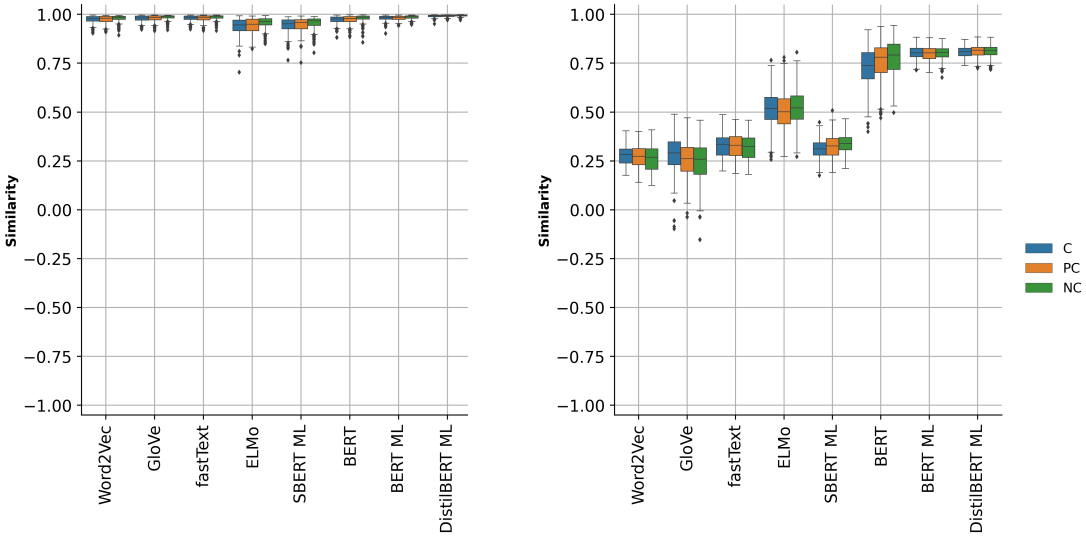
Figura 5.2 – Average of cosine similarities among for five different frequency compatible two-word combination words in English (blue) and Portuguese (orange), for the Sentence (on the left) and for NC representations (on the right). Values for naturalistic sentences are shown in darker shade and for neutral long and short in lighter shades.

In Figures 5.3 and 5.4, dedicated baselines for probe 1 are shown, the boxplots represent an average cosine similarity between the NC true synonym and the chosen frequency compatible two-words combination. The latter shows only the English naturalistics sentences clustered by compositionality.



(a) Sentence (b) NC

Figura 5.3 – Average of cosine similarities between the NC true synonym and five different frequency compatible two-word combination words in English (blue) and Portuguese (orange), for the Sentence (on the left) and for NC representations (on the right). Values for naturalistic sentences are shown in darker shade and for neutral long and short in lighter shades.



(a) Sentence (b) NC

Figura 5.4 – Average of cosine similarities between the NC true synonym and five different frequency compatible two-word combination words in English clustered by compositionality, for the Sentence (on the left) and for NC representations (on the right) in naturalistics sentences.

	word2vec	GloVe	*fastText*	ELMo	BERT	BERT ML	DistilB ML	SBERT ML
EN	0.59	0.58	0.62	0.4	0.38	0.4	0.5	0.62
PT	0.46	0.44	0.27	0.48	0.43	0.36	0.41	0.53

Tabela 5.5 – Spearman ρ correlation between the all models output and human judgments, p$\leq$0.05, for P1 in out-of-context probe measure for both English and Portuguese. Non-significant results omitted from the table.

Qualitative analysis: Table 5.9 shows the similarity scores between each NC in Table 5.1 and their respective NC$_{syn}$ for three representative models, BERT, ELMo and GloVe. As expected, BERT and ELMo show higher scores than GloVe for all cases, and even if the values for P1 differ, all models display the same tendency. Some of these high similarities may come from the overlap in words between the NC and NC$_{syn}$, as in *ghost town* and NC$_{syn}$ = *abandoned town*, which not only share a word, but where *ghost* and *abandoned* are likely to have similar embeddings. This is even more prominent in ELMo, where *ghost town* has a significantly higher score than other idiomatic NCs. Indeed, 47 (in English) and 49 (in Portuguese) out of the 50 compounds with the highest $\text{sim}^{(\text{P1})}_{\text{NC} \subset \text{S-Nat}}$ share surface tokens, including more compositional NCs (e.g., *music journalist* vs. *music reporter*) but also partly compositional cases (e.g., *ghost town* vs. *abandoned town*).

In sum, the results with the first idiomatic probe show that even though the similarities can be relatively high, they are consistently lower for idiomatic than for compositional cases.

	word2vec		GloVe		*fastText*	
	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}						
Baseline	-	-	-0.16	0.25	-	-
$P1_{Baseline}$	-0.15	-	-0.19	0.11	-0.17	0.08
P1	0.14	0.57	0.14	0.57	0.12	0.57
P2	-0.11	0.2	-0.14	0.46	-0.12	0.26
P3	-0.12	0.11	-0.2	0.16	-0.14	0.07
$P3_{ALT}$	-0.14	0.14	-0.18	0.23	-0.15	0.11
$EN_{Neutral}$						
Baseline	0.21	0.2	0.16	0.25	0.22	0.2
$P1_{Baseline}$	-	-	-	-	-	-
P1	0.58	0.59	0.56	0.58	0.58	0.58
P2	0.2	0.22	0.32	0.47	0.23	0.27
P3	0.19	0.13	-	0.16	0.17	-
$P3_{ALT}$	0.18	0.14	0.16	0.25	0.18	0.12
PT_{Nat}						
Baseline	-0.18	-	-0.2	-	-0.11	-
$P1_{Baseline}$	-0.17	-	-0.17	0.11	-0.18	-
P1	0.1	0.47	-	0.42	-	0.21
P2	-	0.33	-0.12	0.22	0.1	0.12
P3	-0.15	0.09	-0.23	-	-	-0.23
$P3_{ALT}$	-0.13	0.2	-0.2	0.11	-	-
$PT_{Neutral}$						
Baseline	-	0.15	-0.2	-	-	-
$P1_{Baseline}$	-	-	-	-	-	-
P1	0.3	0.46	0.22	0.44	0.21	0.27
P2	0.18	0.34	-	0.23	0.22	0.18
P3	-	-	-0.16	-	-	-0.18
$P3_{ALT}$	-	0.21	-0.15	-	-	-

Tabela 5.6 – Spearman ρ correlation between the static models output and human judgments, p$\leq$0.05, for P1, P2 and P3 in both English and Portuguese. Non-significant results omitted from the table.

5.2 P2: Can these models detect the semantic overlap between more compositional NCs and their individual components?

This idiomatic probe examines the potential overlap in meaning between NCs and their components, evaluating to what extent an NC can be replaced by one of its component words and still be considered as representing a similar usage in a sentence. For example, for a more compositional NC like *white wine*, the head *wine* would provide a reasonable approximation, but the same would not be the case for *matter* for *grey matter*, a more idiomatic NC. Both $\text{sim}^{(\text{P2})}_{\text{S}\supset\text{NC}} = \max_i \text{cossim}(\epsilon_{\text{S}\supset\text{NC}}, \epsilon_{\text{S}\supset\text{w}_i})$ and $\text{sim}^{(\text{P2})}_{\text{NC}\subset\text{S}} = \max_i \text{cossim}(\epsilon_{\text{NC}\subset\text{S}}, \epsilon_{\text{w}_i\subset\text{S}})$, where w_i is one of the component words (head or modifier), are measured. For more compositional NCs, the similarities are expected to be higher, while for idiomatic NCs they should be lower as they may not be replaceable by any of their components. Therefore, significant correlations between the similarity values and the NC idiomaticity scores are

	ELMo		BERT		BERT ML		DistilB ML		SBERT ML	
	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}										
Baseline	-0.17	-0.1	-0.28	-0.51	-0.19	-	-0.23	-0.16	-0.11	-
$P1_{Baseline}$	-0.2	-	-0.2	-0.21	-0.17	-	-0.21	-0.1	-0.16	-0.21
P1	0.26	0.53	0.2	0.31	0.32	0.61	0.2	0.53	0.31	0.61
P2	-0.16	0.15	-	0.12	-	0.34	-0.16	0.26	-0.1	0.33
P3	-0.12	-	-0.18	-0.36	-0.07	0.22	-0.18	0.16	-0.15	-
$P3_{ALT}$	-0.19	-	-0.25	-0.42	-0.15	0.1	-0.23	-	-0.19	-
$EN_{Neutral}$										
Baseline	-0.23	-0.2	-0.23	-0.46	-	-	-	-0.23	-0.37	-0.19
$P1_{Baseline}$	-0.16	-0.12	-	-0.2	-	-	-	-0.15	-0.2	-0.21
P1	0.54	0.58	0.48	0.31	0.51	0.57	0.54	0.51	0.58	0.63
P2	-	0.21	-	-0.31	-	-	-0.14	-	-	0.29
P3	-	-	-0.17	-0.4	-	-	-	-	-0.15	-
$P3_{ALT}$	-0.18	-	-0.25	-0.45	-	-	-	-	-0.18	-
PT_{Nat}										
Baseline	-0.11	-	-0.17	-	-0.14	-	-0.15	-	-0.12	-
$P1_{Baseline}$	-0.14	-	-	-	-0.13	-	-0.18	-	-0.17	-0.21
P1	0.27	0.46	0.24	0.55	0.18	0.42	0.11	0.36	0.26	0.51
P2	0.09	0.32	0.13	0.45	-0.09	0.16	-0.18	0.09	-	0.18
P3	-	0.14	-	0.21	-	0.11	-0.18	-	-	-
$P3_{ALT}$	-	0.21	-	0.16	-	0.11	-0.15	-	-	0.15
$PT_{Neutral}$										
Baseline	-	0.16	-0.2	-	-	-	-	-	-	-
$P1_{Baseline}$	-	-	-	-	-	0.16	-	-	-	-
P1	0.37	0.5	0.32	0.47	0.32	0.39	0.29	0.37	0.48	0.52
P2	0.22	0.3	-	0.28	-	-	-	-	-	0.15
P3	-	-	-	0.16	-	-	-	-	-	-
$P3_{ALT}$	-	0.22	-	-	-	-	-	-	-	-

Tabela 5.7 – Spearman ρ correlation between the contextual models output and human judgments, $p \leq 0.05$, for P1, P2 and P3 in both English and Portuguese. Non-significant results omitted from the table.

	word2vec	GloVe	fastText	ELMo	BERT	BERT ML	DistilB ML	SBERT ML
EN								
P1	0.76	0.75	0.78	0.51	0.57	0.53	0.69	0.52
P2	0.84	0.87	0.86	0.71	0.69	0.74	0.88	0.76
P3	0.88	0.88	0.90	0.68	0.72	0.75	0.87	0.73
PT								
P1	0.72	0.67	0.48	0.50	0.51	0.59	0.68	0.49
P2	0.78	0.86	0.16	0.70	0.65	0.75	0.86	0.65
P3	0.81	0.77	0.50	0.65	0.71	0.69	0.81	0.67

Tabela 5.8 – Spearman ρ correlation between naturalistic sentence length and cosine similarity, $p \leq 0.05$.

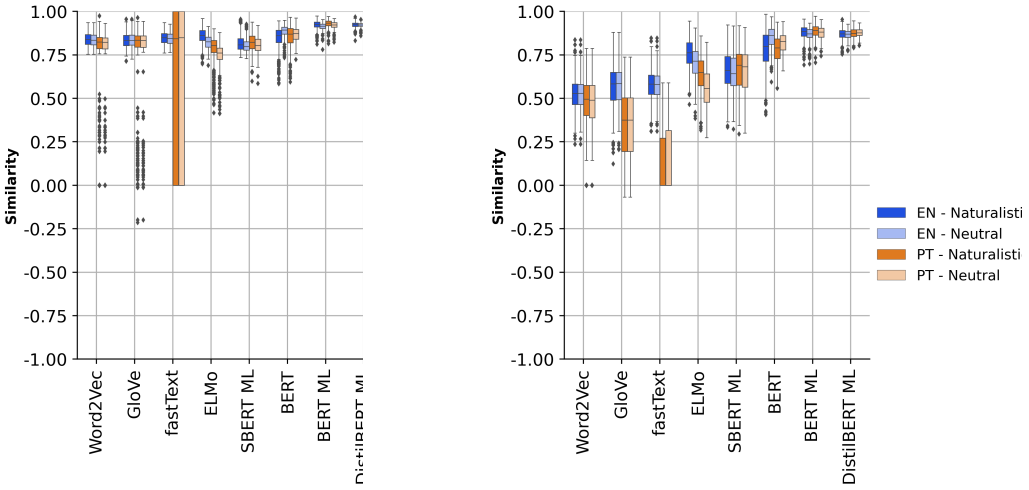
	GloVe	ELMo		BERT	
		Nat	Neu	Nat	Neu
civil marriage	0.88	0.90	0.86	0.93	0.95
close call	0.52	0.57	0.54	0.78	0.85
eager beaver	0.43	0.68	0.58	0.78	0.85
field work	0.58	0.67	0.54	0.80	0.89
ghost town	0.80	0.85	0.79	0.80	0.88
wet blanket	0.21	0.50	0.45	0.70	0.81

Tabela 5.9 – Similarities results from P1 at NC level of the examples in Table 5.1.

expected, that is $\rho^{(P2)}_{S \supset NC} > 0$ and $\rho^{(P2)}_{NC \subset S} > 0$.

5.2.1 Results

Contrary to what was expected, all models produced high similarities across the idiomaticity spectrum (Figure 5.5a), and correlations with idiomaticity that are either non existent or that are lower than those obtained for P1 (Tables 5.6 and 5.7). In fact, although a larger spread of cosine similarity values was expected for P2 than for P1, the results showed higher average similarities for P2 than for P1. These reinforce the hypothesis that the models may be more sensitive to the lexical overlap with the NC components rather than to the semantic overlap with a true NC synonym, even for idiomatic cases.



(a) P2 Cosine similarities (b) P3 Cosine similarities.

Figura 5.5 – Cosine similarities for in English (blue) and Portuguese (orange). Values for naturalistic sentences in darker shade and for neutral in lighter. As the similarity scores for sentences are all saturated at the top of the scale, it is only shown the more informative results for NC representations.

	GloVe	ELMo		BERT	
		Nat	Neu	Nat	Neu
civil marriage	0.87(2)	0.88	0.81(2)	0.89	0.91(1)
close call	0.86(2)	0.79	0.73(2)	0.77	0.87(2)
eager beaver	0.85(2)	0.79	0.75(2)	0.89	0.94(1)
field work	0.86(1)	0.84	0.86(2)	0.84	0.85(2)
ghost town	0.85(2)	0.87	0.80(1)	0.80	0.83(1)
wet blanket	0.84(1)	0.85	0.83(2)	0.78	0.93(1)

Tabela 5.10 – Similarities results from P2 at NC level of the examples in Table 5.1. The number in parenthesis corresponds to the position of the w_i with highest similarity score in the NC.

Qualitative analysis: The P2 results in Table 5.10 show the highest similarity scores between each NC in Table 5.1 and one of its components. These include some idiomatic

NCs, like *poison pill* (meaning an emergency exit), for which strong similarities were found with their components, (average similarity of $\text{sim}^{(\text{P2})}_{\text{poison pill} \subset \text{S-Nat}} = 0.94$ with its head *pill*), exemplifying the priority of lexical over semantic overlap.
All of these indicate that these models cannot distinguish the partial semantic overlap between more compositional NCs and their components and the absence of overlap for idiomatic NCs. Overall, probe P2 also suggests that the idiomatic meaning is not correctly represented by current language models.

5.3 P3: Are models sensitive to perturbations to idiomaticity caused by lexical variations?

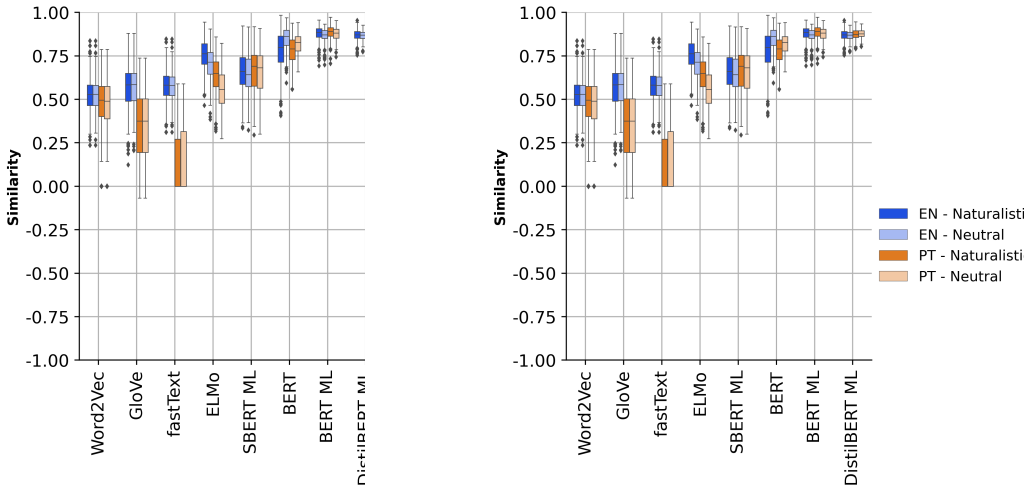
P3 examines model sensitivity to potential perturbations in the meaning of an NC caused by replacing each of its component words individually by their synonyms, $\text{NC}_{\text{WordsSyn}}$, so that for an NC like *grey matter*, *grey* is replaced by *silvery* and *matter* by *material*. Both $\text{sim}^{(\text{P3})}_{\text{S} \supset \text{NC}} = \text{cossim}(\epsilon_{\text{S} \supset \text{NC}}, \epsilon_{\text{S} \supset \text{NC}_{\text{WordsSyn}}})$ and $\text{sim}^{(\text{P3})}_{\text{NC} \subset \text{S}} = \text{cossim}(\epsilon_{\text{NC} \subset \text{S}}, \epsilon_{\text{NC}_{\text{WordsSyn}} \subset \text{S}})$ are measured. Since idiomatic cases display a lower degree of substitutability of their individual components (SAG et al., 2002; FARAHMAND; HENDERSON, 2016) potentially loosing the idiomatic meaning or forming anti-collocations (PEARCE, 2001), similarity values would be expected to correlate to the NC idiomaticity scores, that is $\rho^{(\text{P3})}_{\text{S} \supset \text{NC}} > 0$ and $\rho^{(\text{P3})}_{\text{NC} \subset \text{S}} > 0$.

	GloVe	ELMo		BERT	
		Nat	Neu	Nat	Neu
civil marriage	0.60	0.77	0.72	0.78	0.83
close call	0.61	0.67	0.59	0.75	0.81
eager beaver	0.49	0.84	0.81	0.86	0.79
field work	0.54	0.72	0.72	0.76	0.91
ghost town	0.66	0.79	0.67	0.68	0.75
wet blanket	0.69	0.89	0.85	0.87	0.94

Tabela 5.11 – Similarities results from P3 at NC level of the examples in Table 5.1.

5.3.1 Results

Contrary to what would be expected, high similarity values were found across the idiomaticity spectrum (Figure 5.5b), comparable to those for P1, with correlations with idiomaticity scores being mostly nonexistent and when they do exist being much lower than for P1.
Qualitative analysis: Table 5.11 shows the similarities scores at NC level between each



(a) P3 by Idiomaticity Class

(b) P3 Using Additional False Synonyms

Figura 5.6 – Cosine similarities for Probe 3 for both English and Portuguese in naturalistics sentences. Left side shows the results only with the components synonym from the original dataset (CORDEIRO et al., 2019) and in the right side with the additional synonyms collected from other sources (W_{Random} detailed in section 4).

NC and a compositional synonym, $NC_{WordsSyn}$. For some idiomatic cases the similarity with the compositional synonym (e.g. for GloVe $sim^{(P3)}_{\text{wet blanket} \subset S} = 0.692$) is noticeably higher than that with the holistic synonym ($sim^{(P1)}_{\text{wet blanket} \subset S} = 0.213$), suggesting that individually the words *damp* and *cloak* are considered to be closer in meaning to *wet* and *blanket*, respectively, than *loser* is. It worth noticing that ELMo seems particularly bad for idiomatic cases like *wet blanket* and *close call*, where the values in Table 5.11 are significantly higher than the values in Table 5.9, showing results worse than BERT ML.

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

5.4 P4: Just how informative contexts are?

Contextual information may be helpful in determining the particular meaning of an NC in a sentence (e.g. the literal *gold mine* vs. the idiomatic *source of valuable information*). As contextualised models allow dedicated representations for different usages of a word, this probe examines to what extent the representation of an NC in context actually differs from the representation of the same NC out of context, for different degrees of informativeness of the context ($sim^{(P4)}_{\text{Nat | Out}} = \text{cossim}(\epsilon_{\text{NC} \subset \text{S-Nat}}, \epsilon_{\text{NC}})$ and $sim^{(P4)}_{\text{Neutral | Out}} = \text{cossim}(\epsilon_{\text{NC} \subset \text{S-Neutral}}, \epsilon_{\text{NC}})$). The more they differ, the more information is incorporated from the context in the represen-

		ELMo	BERT	BERT$_{ML}$	DistilB$_{ML}$	SBERT$_{ML}$
In/Out	EN$_{\text{Nat}}$	0.13	0.35	-	0.19	-
	EN$_{\text{Neutral}}$	-	0.37	-0.12	0.25	0.20
	PT$_{\text{Nat}}$	0.26	0.34	0.16	0.16	0.24
	PT$_{\text{Neutral}}$	-	0.15	-	-	-
Nat/Neutral	EN	-	0.13	-	-	0.22
	PT	0.15	0.16	-	0.11	0.21

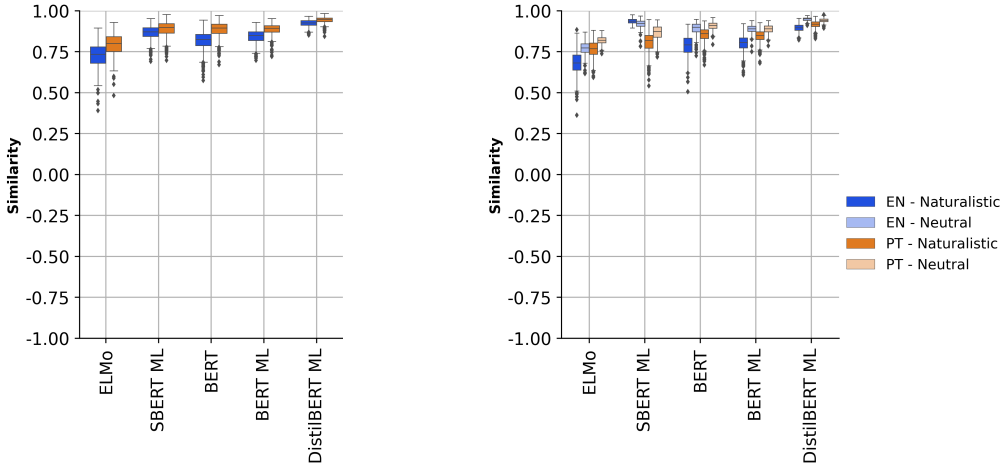
Tabela 5.12 – Spearman ρ correlation with human judgments for P4, p$\leq$0.05. Non-significant results omitted from the table.

tation. It is also compared how similar the representation of the NC in naturalistic context is to the representation in neutral context ($\text{sim}^{(\text{P4})}_{\text{Nat | Neutral}} = \text{cossim}(\epsilon_{\text{NC} \subset \text{S-Nat}}, \epsilon_{\text{NC} \subset \text{S-Neutral}})$).

5.4.1 Results

The results (Figure 5.7b) show high similarity values between the NC in and out of context (values of $\text{sim}^{(\text{P4})}_{\text{* | Out}}$ are mostly higher than 0.75). As some of these values are similar or even higher than those obtained for synonyms in P1 ($\text{sim}^{(\text{P4})}_{\text{Nat | Out}} \simeq \text{sim}^{(\text{P1})}_{\text{NC} \subset \text{S-Nat}}$ and $\text{sim}^{(\text{P4})}_{\text{Neutral | Out}} > \text{sim}^{(\text{P1})}_{\text{NC} \subset \text{S-Neutral}}$) they indicate that these models consider NCs out of context as good as (or better) approximations for NCs in context than their synonyms. In addition, for most models $\text{sim}^{(\text{P4})}_{\text{* | Out}}$ is only weakly correlated with the idiomaticity score (Table 5.12), which suggests that for these models the role of the context for idiomatic NCs does not appear to be bigger than for more compositional NCs. When analysing the informativeness of contexts, the results of comparing the NCs in naturalistic and in neutral contexts (Figure 5.7a) show a high similarity between them. In fact, the naturalistic and the neutral condition follow the same trends in the two languages and are significantly correlated, with some very strong correlations for some models, which is further analyzed in the appendix tables B.1 and B.2. For example, for SBERT the correlations between the NC in context in naturalistic and neutral conditions are $\rho^{(\text{P1,P2,P3})}_{\text{NC} \subset \text{(Nat | Neutral)}} > 0.85$ for English and > 0.76 for Portuguese, for probes P1, P2 and P3. These results suggest that NCs in and out of context are similar, with limited information being incorporated from a specific context, even if these contexts are informative.

Although unlikely, one possible explanation for these results is that the NC meaning exemplified in these sentences is predominant in the training corpus, which would lead to the representations in and out of context being similar.



(a) (b)

Figura 5.7 – Cosine similarities for P4: (a) comparing NC representation in naturalistic and neutral contexts. For static models the representations in and out of context are the same; (b) comparing NC representations in and out of context (cossim($\epsilon_{NC \subset S}, \epsilon_{NC}$))

Qualitative analysis: The $\text{sim}^{(P4)}_{\text{Nat} \mid \text{Out}}$ of the examples in Table 5.1 ranged from 0.78 (for *ghost town*) to 0.87 (*field work*), while $\text{sim}^{(P4)}_{\text{NeuLong} \mid \text{Out}}$ ranged from 0.82 (also for *ghost town*) to 0.87 (*wet blanket*) for BERT_{ML}. For ELMo, the $\text{sim}^{(P4)}_{\text{Nat} \mid \text{Out}}$ ranges from 0.63 (*close call*) to 0.77 (*field work*) and $\text{sim}^{(P4)}_{\text{NeuLong} \mid \text{Out}}$ ranges from 0.60 (*close call*) to 0.77 (*ghost town*). Interestingly, both the largest and smallest differences between $\text{sim}^{(P4)}_{\text{NeuLong} \mid \text{Out}}$ and $\text{sim}^{(P4)}_{\text{Nat} \mid \text{Out}}$ are found for compositional NCs (*engine room* with 0.21 , and *rice paper* with 0.02). Besides, ambiguous compounds such as *bad apple* or *bad hat* are expected to have large $\text{sim}^{(P4)}_{* \mid \text{Out}}$ differences between both conditions, as they occur with an idiomatic meaning in the naturalistic sentences. However, the differences were of just 0.06 in both cases, while other less ambiguous idiomatic NCs showed higher variations (e.g., *melting pot*, with 0.16).

According to the results for P4, the high similarities displayed by NCs in and out of context suggest that the context is not bringing much additional information to that already captured by the representation for the NC out of context.

	ELMo		BERT	
	Nat	Neu	Nat	Neu
civil marriage	0.74	0.79	0.86	0.90
close call	0.63	0.76	0.76	0.92
eager beaver	0.71	0.73	0.73	0.76
field work	0.69	0.71	0.85	0.88
ghost town	0.69	0.79	0.78	0.89
wet blanket	0.64	0.78	0.84	0.93

Tabela 5.13 – Similarities results from P4 at In/Out level of the examples in Table 5.1.

	ELMo	BERT
civil marriage	0.72	0.85
close call	0.74	0.80
eager beaver	0.77	0.84
field work	0.57	0.78
ghost town	0.84	0.84
wet blanket	0.77	0.87

Tabela 5.14 – Similarities results from P4 at Neutral/Naturalistics comparison of the examples in Table 5.1.

5.5 Idiomatic Probes

The results obtained with the idiomatic probes are of high similarities for NCs with their holistic synonyms across models and languages. However, these results also revealed that the high similarities obtained may be due to overlap in words between a sentence containing an NC and its variants. The correlations found between similarities and sentence length, shown in Table 5.8, confirm this, with strong correlations for most probes and models for English and moderate for Portuguese. Even if focusing the analyses on NC representations, the overall high similarities obtained with all these probes regardless of how semantically incongruous the NC variants are, how idiomatic the NCs are or how uninformative the contexts are, all suggest a possible saturation in the semantic space. More concerning, downstream tasks that rely on these high similarities are vulnerable to using representations that do not capture idiomaticity accurately, and may equate an idiomatic *eager beaver* with a literal *restless rodent*.

6 THE AFFINITY MEASURES

In the preceding chapter, similarity measures are used to evaluate the effect of replacing the target NC by a series of possible probing substitutes. Since the linguistic behaviour in these substitutions depends on the idiomaticity of the NC, a model that captures well idiomaticity will reproduce this behaviour. For instance, if an NC is very idiomatic and rigid, replacing it in a sentence with its literal synonym will produce a sentence with a completely different meaning. If the contextual model is linguistically faithful then the similarity between the original sentence representation and the new sentence representation should be low. However, there is no good way of saying what is a low similarity. Another obstacle is that the representation of a sentence in contextual models is still a matter of debate since there is no clear rule of how to combine layers and tokens in a way that allows all sentences to be compared in the same semantic space. In addition, the number of tokens in common between the sentences is expected to play an important role in its similarity value when measured by cosine. Even if a word is wrongly substituted in a long sentence, the original sentence and its new version will still be very similar since they share most of the words.

One way to avoid this scaling problem is to use relative measures that compare similarities in two different substitutions. Consider that there is an original sentence $Sent$ containing an NC (or more generally an MWE) and two possible paraphrases for this sentence, $Para_1$ and $Para_2$, where only the target NC (or MWE) is replaced. The relative similarities or affinities are defined as the expression:

$$\text{aff}(Para_1, Para_2 \,|Sent) = \text{cossim}\,(\epsilon_{Sent}, \epsilon_{Para_1}) - \text{cossim}\,(\epsilon_{Sent}, \epsilon_{Para_2}) \tag{6.1}$$

where ϵ_{Sent}, ϵ_{Para_1} and ϵ_{Para_2} are sentences (or NC) embeddings.

In all the measures a positive value indicates that paraphrase $Para_1$ is closer to the original meaning $Sent$ than paraphrase $Para_2$, and a value near zero indicates no preference for a paraphrase. In the following sections, affinity measures that examine the expected behaviour of NCs in relation to some key anchors are proposed.

This chapter introduces two affinity measures: the first one detailed in Section 6.1 which is a relative comparison that captures both Probe P1 and Probe P3, mentioned in Sections 5.1 and 5.3 respectively, using sentence variants containing the true NC synonym (NC_{Syn}) and the built variant ($\text{NC}_{WordsSyn}$) with the components' synonyms. Section 6.2 presents the second affinity which verifies whether the contextualised models prefer a component from

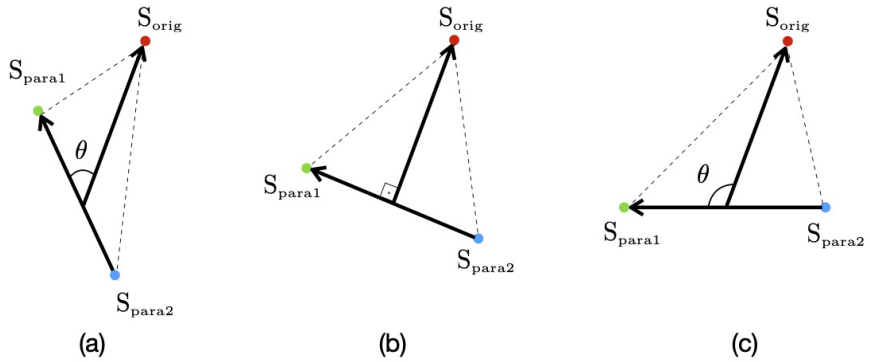


Figura 6.1 – Geometrical interpretation of the affinities. (a) aff$(para1, para2|orig) > 0$. (b) aff$(para1, para2|orig) = 0$ and (c) aff$(para1, para2|orig) < 0$

the NC or the true synonym. Last but not least, Section 6.2.1 details the results found in both experiments.

6.1 A1: Idiomatic NCs have Greater Affinity for the Holistic NC Synonym than for the Synonyms of Individual Components

The affinity A1 is modelled to indicate whether a model is sensitive to the semantic perturbations of considering synonyms of the individual parts rather than the NC synonym, which can be considered as a comparison between Probes 5.1 and 5.3 but in one shot. The definition come as $\text{aff}^{(\text{A1})}_{\text{NC} \subset \text{S}}(\epsilon_{\text{NC}_{\text{Syn}}} \subset \text{S}, \epsilon_{\text{NC}_{\text{WordsSyn}}} \subset \text{S} | \epsilon_{\text{NC}} \subset \text{S})$ for comparing the NC embeddings and $\text{aff}^{(\text{A1})}_{\text{S} \supset \text{NC}}(\epsilon_{\text{S}} \supset \text{NC}_{\text{Syn}}, \epsilon_{\text{S}} \supset \text{NC}_{\text{WordsSyn}} | \epsilon_{\text{S}} \supset \text{NC})$ for the whole sentence embedding comparison. This affinity reflects whether a model displays a marked preference for the NC synonym than for the synonyms of the individual components from the NC, with a higher positive affinity value indicating that *eager beaver* is considered to be more similar to *hardworking person* than to *restless rodent* which is mainly expected for idiomatic noun compounds, therefore, as in 5.3, it's also expected that affinity values would be negatively correlated to the NC idiomaticity scores ($\rho^{(\text{A1})}_{\text{S} \supset \text{NC}} < 0$, $\rho^{(\text{A1})}_{\text{NC} \subset \text{S}} < 0$)[1].

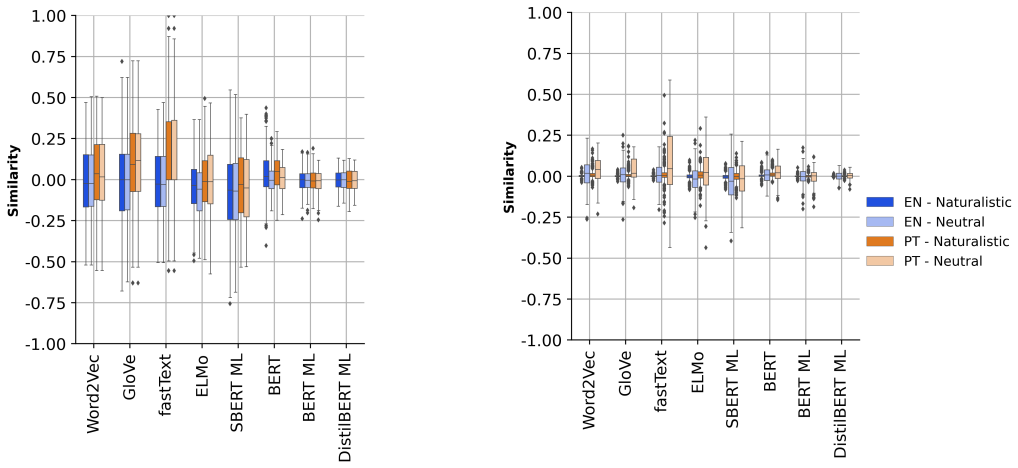
6.1.1 Results

Most models display affinity values of around 0 (Figure 6.2) with no preference for either of the synonyms. This is confirmed in the more in-depth examination of the affinity per

[1] It's negatively correlated because while the affinity 1 increases with the proximity with the synonym, the idiomaticity scores reduce for idiomatic NCs

level of idiomaticity (Figure 6.3) which also reveals lower affinities for idiomatic cases than for compositional cases. These results suggest that these models are somewhat insensitive to the semantic perturbations of not preferring holistic synonyms for idiomatic cases. Moreover, these affinity patterns are displayed both at the NC level (Figure 6.2(a)) and at the sentence level (Figure 6.2(b)), indicating the robustness of this measure to the granularity of the representation. Finally, the same trends were found for the naturalistic and neutral conditions, confirming that even a simple neutral context can be used to measure this affinity in models.

From a correlation point-of-view, the tables 6.1 and 6.2 show negligible or very low positive Spearman correlation when analysing the NCs separated by their compositionality independently from the sentence type (neutral or naturalistic). When looking at the table 6.5 that contains the correlation without separation, even some moderate positive correlations (e.g. 0.50 for dedicated BERT in English, naturalistic sentence). Therefore, it can also be inferred that the contextualised models still struggle to match the idiomatic NC to the correct synonym.



(a) NC (b) Sentence

Figura 6.2 – Affinity A1. English in blue and Portuguese in orange. Naturalistic conditions in a darker shade and neutral conditions in a lighter.

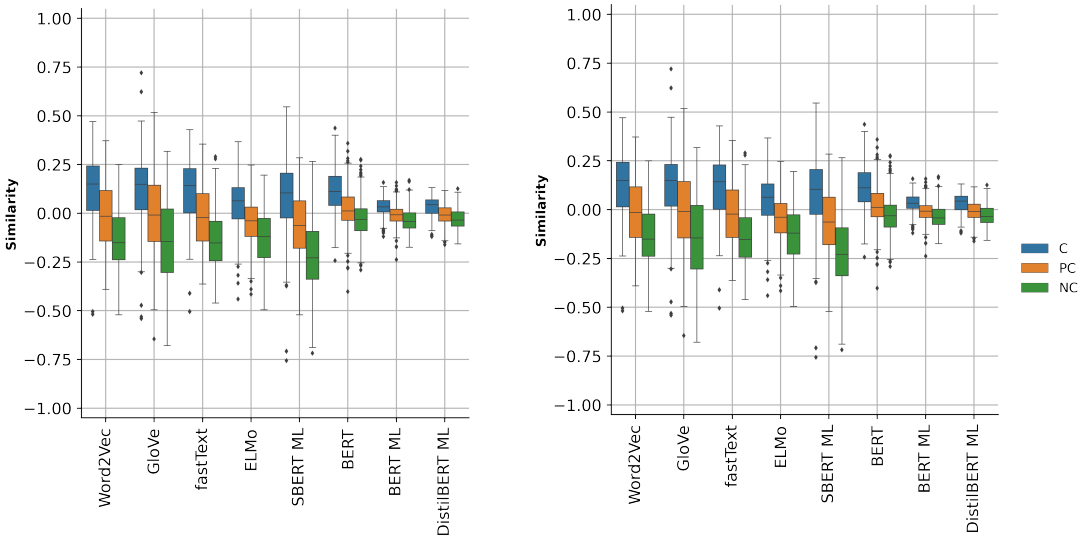
As this affinity also can suffer from the same impact in Probe 5.1 of overlapping components between its NC and the true synonym, this same analysis is also taken into account for a1. The Tables 6.3 and 6.4 show the correlation between the model's output and the idiomaticity score from the annotators, for both static and contextualised models respectively, but excluding those NCs that overlap with their true synonym. There it can be seen that indeed some models have abrupt changes like BERT for EN_{Nat} which went from 0.50 to 0.09, but in overall low correlation is still found like 0.18 for BERT ML in PT_{Nat} to 0.53 in

		word2vec		GloVe		*fastText*	
		Sent	NC	Sent	NC	Sent	NC
C	EN_{Nat}	0.14	0.19	0.15	0.20	-	0.18
	$EN_{Neutral}$	-	-	0.26	0.26	-	-
	PT_{Nat}	0.18	0.17	-	0.22	-	-
	$PT_{Neutral}$	-	-	-	-	-	-
PC	EN_{Nat}	-	-	-	-	-	-
	$EN_{Neutral}$	-	-	-	-	-	-
	PT_{Nat}	0.22	0.18	0.33	0.16	0.17	0.15
	$PT_{Neutral}$	0.30	0.22	0.30	-	0.28	-
I	EN_{Nat}	0.22	0.33	0.29	0.33	0.30	0.34
	$EN_{Neutral}$	0.32	0.40	0.32	0.41	0.36	0.41
	PT_{Nat}	-	-	-	-	-	-
	$PT_{Neutral}$	-	-	-	-	-	-

Tabela 6.1 – Spearman ρ correlation between static model prediction and human judgements, for Compositional (C). Partly Compositional (PC) and idiomatic (I) NCs. $p \leq 0.05$. Non-significant results omitted from the table.

		ELMo		BERT		BERT ML		DistilBERT ML		SBERT ML	
		Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
C	EN_{Nat}	-	-	-	-	-	-	0.14	-	0.16	-
	$EN_{Neutral}$	-	-	0.28	-	-	-	-	-	-	-
	PT_{Nat}	-	-	-	-	-0.16	-	-	-	-	-
	$PT_{Neutral}$	-	-	-	-	-	-	-	-	-	-
PC	EN_{Nat}	-	-	-	-	-	-	-	-	0.17	-
	$EN_{Neutral}$	-	-	-	-	-	-	-	-	-	-
	PT_{Nat}	0.14	0.16	-	-	-	-	0.21	-	-	-
	$PT_{Neutral}$	0.26	0.26	-	-	-	-	-	-	-	-
I	EN_{Nat}	0.27	0.35	0.28	0.27	0.34	0.32	0.39	0.29	0.35	0.27
	$EN_{Neutral}$	0.26	0.35	0.27	-	0.30	0.32	0.37	0.34	0.42	0.35
	PT_{Nat}	-	-	-	-	-	-	-	-	-	-
	$PT_{Neutral}$	-	-	-	-	-	-	-	-	-	-

Tabela 6.2 – Spearman ρ correlation between contextualised model prediction and human judgments, for Compositional (C). Partly Compositional (PC) and idiomatic (I) NCs. $p \leq 0.05$. Non-significant results omitted from the table.



(a) Naturalistic (b) Neutral

Figura 6.3 – Affinity A1 for English. Compositional NCs in blue, partly compositional in orange and idiomatic in green. Left: Naturalistic condition. Right: Neutral condition.

SBERT ML in EN_{Nat}, which still goes against the negative correlation hypothesis.

	word2vec		*GloVe*		fastText	
	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	0.23	0.43	0.32	0.36	0.24	0.43
$EN_{Neutral}$	0.35	0.44	0.33	0.37	0.37	0.44
PT_{Nat}	-	0.22	0.14	0.14	-0.16	0.17
$PT_{Neutral}$	-	0.2	-	-	-	-

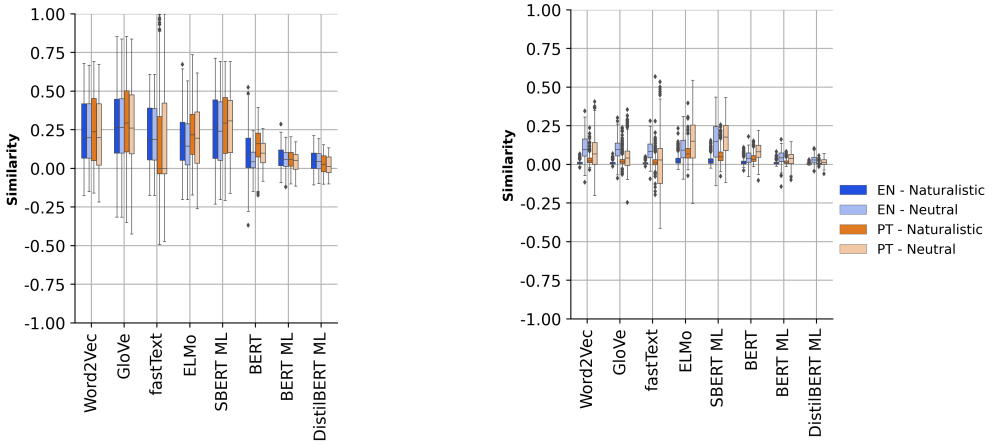
Tabela 6.3 – Spearman ρ correlation between the static models' output and human judgements only for the components which do not have their synonym with an overlapping component, $p \leq 0.05$, for A1. Non-significant results were omitted from the table.

	ELMo		*BERT*		BERT ML		DistilB ML		SBERT ML	
	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	0.1	0.44	-	0.09	0.16	0.55	-	0.39	0.13	0.53
$EN_{Neutral}$	0.45	0.49	0.34	-	0.39	0.47	0.42	0.37	0.53	0.56
PT_{Nat}	-	0.3	-	0.49	-	0.18	-0.15	-	-	0.3
$PT_{Neutral}$	0.45	0.49	0.34	-	0.39	0.47	0.42	0.37	0.53	0.56

Tabela 6.4 – Spearman ρ correlation between the non-static models' output and human judgements only for the components which do not have their synonym with an overlapping component, $p \leq 0.05$, for A1. Non-significant results were omitted from the table.

As in Chapter 5, the affinity using the baseline is also accounted for comparison. In other words, instead of considering the built synonym with the synonym of the components, the variant NC_{Random} is used. This means that for all NCs, regardless of idiomaticity level, an affinity higher than zero is expected, as the models should prefer synonyms rather than words which share just statistical characteristics (e.g. frequency, in this case). The affinity comparison can be seen in Figure 6.4 for English and Portuguese in both informative and non-informative setup, still not showing a marked preference (positive value) for the

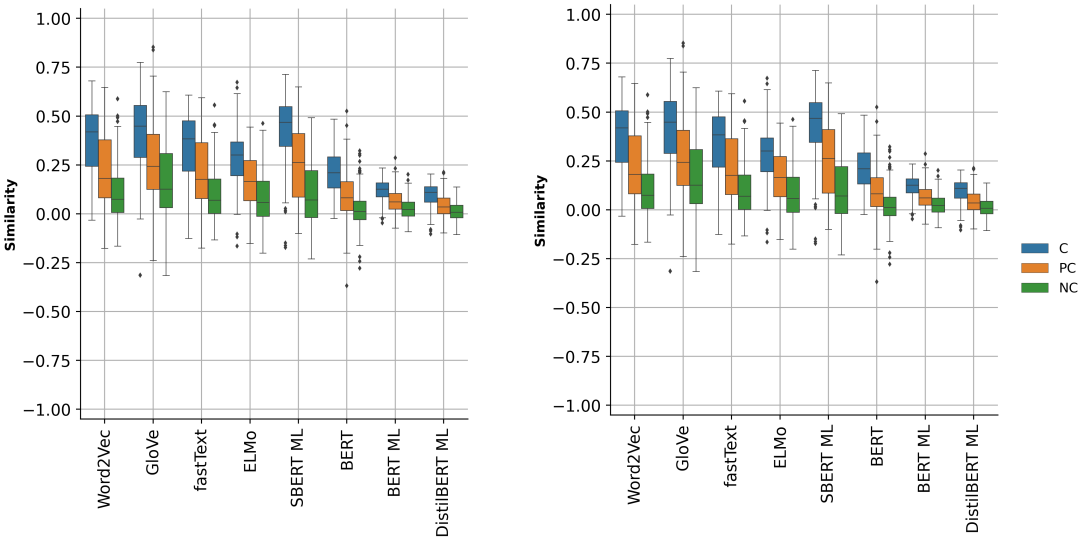
synonyms, but when comparing to similarities shown in A1, show slightly more positive values with an average of 0.105 and standard deviation of 0.133 for BERT in the baseline and with A1's average of 0.04 and standard deviation of 0.125.



(a) NC (b) Sentence

Figura 6.4 – Baseline for affinity A1. English in blue and Portuguese in orange. Naturalistic conditions in a darker shade and neutral conditions in a lighter.

Under the same baseline condition, a more dissected analysis by idiomaticity classes in English naturalistic sentences is shown in Figure 6.5, which shows affinity values a bit higher than A1 (expected as aforementioned) but also the same pattern of compositional NC showing higher affinity values than the partially-compositional and the idiomatic, which can show that for those classes the model does refer to statistical information.



(a) Naturalistic (b) Neutral

Figura 6.5 – Baseline affinity A1 for English. Compositional NCs in blue, partly compositional in orange and idiomatic in green. Left: Naturalistic condition. Right: Neutral condition.

Qualitative analysis: The affinities for the NCs in Table 5.1 confirm these findings (Table

	word2vec		GloVe		*fastText*	
	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}						
A1	0.39	0.52	0.47	0.47	0.42	0.53
A2	0.25	0.54	0.28	0.51	0.24	0.52
$EN_{Neutral}$						
A1	0.49	0.54	0.45	0.49	0.51	0.55
A2	0.55	0.55	0.52	0.53	0.55	0.54
PT_{Nat}						
A1	0.33	0.41	0.36	0.36	-	0.31
A2	-	0.21	0.12	0.23	0.09	0.12
$PT_{Neutral}$						
A1	0.29	0.43	0.29	0.37	0.23	0.35
A2	0.18	0.21	0.17	0.22	0.15	-

Tabela 6.5 – Spearman ρ correlation with human judgements, $p \leq 0.05$. Non-significant results were omitted from the table.

	ELMo		BERT		BERT ML		DistilB ML		SBERT ML	
	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}										
A1	0.42	0.49	0.38	0.50	0.44	0.49	0.47	0.45	0.49	0.54
A2	0.40	0.50	0.25	0.22	0.42	0.52	0.36	0.46	0.41	0.56
$EN_{Neutral}$										
A1	0.49	0.51	0.47	0.50	0.38	0.48	0.46	0.47	0.59	0.57
A2	0.54	0.53	0.46	0.44	0.50	0.52	0.52	0.49	0.58	0.58
PT_{Nat}										
A1	0.29	0.38	0.19	0.32	0.25	0.33	0.32	0.34	0.34	0.43
A2	0.23	0.31	0.20	0.25	0.28	0.38	0.22	0.36	0.35	0.48
$PT_{Neutral}$										
A1	0.32	0.41	0.20	0.32	0.27	0.35	0.31	0.37	0.43	0.47
A2	0.26	0.34	0.29	0.30	0.33	0.39	0.30	0.40	0.46	0.48

Tabela 6.6 – Spearman ρ correlation with human judgements, $p \leq 0.05$. Non-significant results were omitted from the table.

6.7): the more compositional NC *civil marriage* has higher affinity value compared to idiomatic cases like *eager beaver* and *wet blanket*, which is not the expected if the models were capable of assimilate idiomatic information.

	GloVe	ELMo		BERT	
	NAT/NEU	NAT	NEU	NAT	NEU
civil marriage	0.28	0.12	0.13	0.15	0.12
close call	-0.09	-0.08	-0.05	0.03	0.04
eager beaver	-0.08	-0.16	-0.22	-0.07	0.06
field work	0.04	-0.05	-0.18	0.04	-0.02
ghost town	0.13	0.05	0.14	0.12	0.13
wet blanket	-0.48	-0.39	-0.41	-0.17	-0.13

Tabela 6.7 – A1 results at NC level for the NCs in Table 5.1.

6.2 A2: More Compositional NCs also Have Affinity for a Component Word

The second measure assesses if there's a higher affinity of compositional NCs with at least one of its individual component words compared to idiomatic ones that would prefer the holistic synonyms as they cannot be represented by just one of the components. Therefore, the affinity 2 is defined as $\text{aff}^{(\text{A2})}_{\text{NC}\subset\text{S}}(\epsilon_{\text{NC}_{\text{Syn}}\subset\text{S}}, \epsilon_{\text{NC}_{w_i}\subset\text{S}}|\epsilon_{\text{NC}\subset\text{S}})$ for measuring the proximity for NC embeddings and $\text{aff}^{(\text{A2})}_{\text{S}\supset\text{NC}}(\epsilon_{\text{S}\supset\text{NC}_{\text{Syn}}}, \epsilon_{\text{S}\supset\text{NC}_{w_i}}|\epsilon_{\text{S}\supset\text{NC}})$ for calculating the affinity for sentence embeddings, where NC_{w_i} is the most similar component word, either the head or the modifier, defined by $\text{NC}_{w_i} = \text{argmax}(\text{cossim}(\epsilon_{\text{NC}\subset\text{S}}, \epsilon_{\text{NC}_{\text{head}}\subset\text{S}}), \text{cossim}(\epsilon_{\text{NC}\subset\text{S}}, \epsilon_{\text{NC}_{\text{modifier}}\subset\text{S}}))$. In this case, an idiomatic NC would display a strong affinity only to its holistic synonym i.e. high $\text{aff}^{(\text{A2})}$ values compared to both compositional and partly compositional NCs as their components are essential for the noun-compound meaning. Hence, it's expected that there is a negative correlation between the affinity and the compositionality scores ($\rho^{(\text{A2})}_{\text{S}\supset\text{NC}} < 0$, $\rho^{(\text{A1})}_{\text{NC}\subset\text{S}} < 0$).

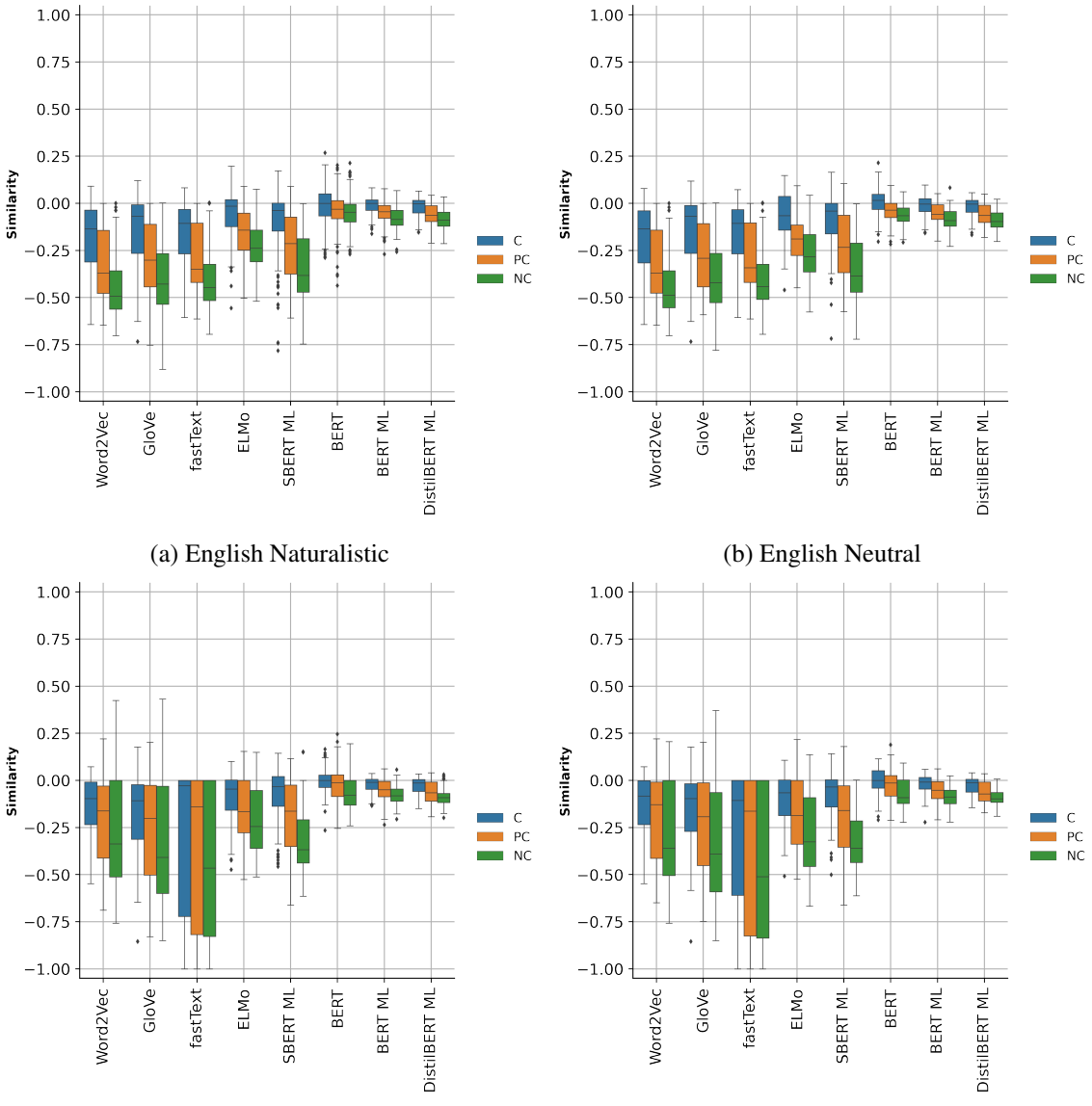
6.2.1 Results

Figure 6.6 shows that most models display a stronger preference between an NC and an individual component than between their true synonyms, especially for idiomatic NCs (in green), due to their values being towards -1, regardless of the model, which is very similar to the one found in A1.

What can be seen in Tables 6.5 and 6.6 for the correlation between each model's output with human judgements, for static and contextualised models respectively, is that no negative

correlation can be seen, but quite the opposite, positive low to moderate correlation can be found ranging from 0.52 for $\rho^{(A2)}_{NC \subset S}$ for BERT ML and 0.22 to BERT.

The correlation is also not expected even when separating the correlations per idiomaticity, which could indicate that the models perform better under some level of compositionality. Both Tables 6.8 and 6.9 show low to moderate positive correlation indicating that the affinity has a higher value when the idiomaticity is higher.



(a) English Naturalistic

(b) English Neutral

(c) Portuguese Naturalistic

(d) Portuguese Neutral

Figura 6.6 – NC representation: Affinity A2. Results for target NCs compared with holistic synonyms and to the most similar individual component word. Compositional NCs in blue, partly compositional in orange and idiomatic in green.

Qualitative analysis: Table 6.10 confirms these findings for the target NCs, with higher affinity scores for compositional NCs (e.g. *civil marriage*) than for idiomatic NCs (e.g. *wet blanket* with the lowest A2 scores). There is also virtually no difference between NAT

		word2vec		GloVe		*fastText*	
		Sent	NC	Sent	NC	Sent	NC
C	EN_{Nat}	0.23	0.21	0.32	0.29	0.25	0.24
	$EN_{Neutral}$	-	-	-	0.29	-	-
	PT_{Nat}	0.22	-	-	-	0.25	0.16
	$PT_{Neutral}$	-	-	-	-	-	-
PC	EN_{Nat}	0.43	0.49	0.45	0.40	0.44	0.46
	$EN_{Neutral}$	0.37	0.43	0.37	0.32	0.37	0.42
	PT_{Nat}	0.26	0.29	0.28	0.25	0.23	0.30
	$PT_{Neutral}$	0.23	0.30	-	0.24	-	0.26
I	EN_{Nat}	0.52	0.40	0.49	0.36	0.53	0.38
	$EN_{Neutral}$	0.34	0.41	0.32	0.36	0.35	0.40
	PT_{Nat}	-	0.18	-	0.30	-	-
	$PT_{Neutral}$	-	-	-	0.30	0.29	-

Tabela 6.8 – Spearman ρ correlation between static model prediction and human judgments, for Compositional (C). Partly Compositional (PC) and idiomatic (I) NCs, $p \leq 0.05$, in affinity A2. Non-significant results omitted from the table.

		ELMo		BERT		BERT ML		DistilBERT ML		SBERT ML	
		Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
C	EN_{Nat}	0.20	-	-	-	0.17	-	0.22	0.15	0.27	0.21
	$EN_{Neutral}$	-	-	-	-	0.38	0.27	-	-	-	0.25
	PT_{Nat}	0.23	-	0.28	-	0.18	-	-	-	0.26	0.23
	$PT_{Neutral}$	-	-	-	-	-	-	-	-	0.33	-
PC	EN_{Nat}	0.31	0.28	0.25	0.25	0.43	0.44	0.46	0.37	0.45	0.31
	$EN_{Neutral}$	-	-	-	-	0.31	0.35	0.40	0.38	0.29	0.29
	PT_{Nat}	0.16	0.19	-	-	0.14	0.15	0.19	0.18	-	0.14
	$PT_{Neutral}$	-	-	-	-	-	-	-	-	-	-
I	EN_{Nat}	0.52	0.45	0.35	0.28	0.53	0.43	0.53	0.41	0.61	0.44
	$EN_{Neutral}$	-	0.28	0.43	0.41	-	0.33	0.35	0.41	0.45	0.40
	PT_{Nat}	0.25	0.27	-	0.26	0.20	0.30	-	0.24	-	0.30
	$PT_{Neutral}$	-	-	-	0.37	-	0.31	-	-	-	0.36

Tabela 6.9 – Spearman ρ correlation between contextualised model prediction and human judgements, for Compositional (C). Partly Compositional (PC) and idiomatic (I) NCs. $p \leq 0.05$, in affinity A2. Non-significant results were omitted from the table.

and NEU cases.

	GloVe	ELMo		BERT	
	NAT/NEU	NAT	NEU	NAT	NEU
civil marriage	0.01	0.02	0.04	0.04	0.04
close call	-0.34	-0.21	-0.19	0.00	-0.02
eager beaver	-0.44	-0.12	-0.17	-0.11	-0.09
field work	-0.28	-0.18	-0.31	-0.05	0.04
ghost town	-0.06	-0.02	0.00	0.00	0.05
wet blanket	-0.63	-0.36	-0.38	-0.08	-0.12

Tabela 6.10 – A2 results at NC level of the examples in Table 5.1.

6.3 Idiomatic Affinities

Affinity tasks show advantages when compared to probing tasks shown in Chapter 5, as we are able to compare models' preferences and have better-scaled values of similarities which is easy to analyse as the values are not saturated towards one.
The affinities can be confirmed as a valuable metric to display preferences of the models with a sentence between two paraphrases, as consistent results can be analysed in both A1 and A2.
For affinity A1, where the anchor is the NC and the references are both the true synonym and the synonym of the parts, the contextualised models are not preferring the first variant if the anchor is an idiomatic noun compound, regardless of how informative the context is, as the expected is that models that place idiomatic NCs much closer to their holistic synonyms than to the synonyms for the individual components would display greater compatibility with idiomaticity awareness.
What is interesting is that the baseline analysed shows that models are capable of preferring a bit more the synonyms, but not relevantly enough to not show correlations with human judgements, which can possibly indicate that they hold linguistic information, but are still severely impacted by statistical one.
The affinity A2 results further suggest that the similarities that emerge between these representations reflect the lexical overlap between NCs and their component words, rather than between the NCs and lexically different representations that share their meaning (true synonym in this case).

7 CONCLUSIONS

This work presented a large-scale evaluation of the ability of contextualised models to retain (or not) the idiomatic meaning of NCs in the presence of lexical substitutions and different contexts, which is a by-product of the two published papers Garcia et al. (2021a) and Garcia et al. (2021b).

For these evaluations the Noun Compound Idiomaticity (NCI) dataset is constructed, with a total of 27,600 sentences in English and Portuguese, with annotations at the token level on idiomaticity level ranging from 0 (idiomatic) to 5 (compositional) and including variants i) with synonyms of the NC, ii) synonyms of each of its components, iii) each component of the NC and iv) noun compounds with similar frequency in a specific corpus, in neutral and naturalistic probing sentences. These phrases were input in both static and contextualised most representative models of the literature, with their publicly available pre-trained weights.

Four different probes were analysed with the variants aforementioned, by extracting the embeddings generated by the models and comparing them using cosine similarity, which was later correlated with the annotations collected from experts. The results from those probes indicate that contextualised models do not capture idiomaticity accurately, since, for instance, they do not seem to detect the lower degree of individual component substitutability for idiomatic than for more compositional NCs, and consider them as good as synonyms for the NC as a whole. This behaviour is similar in the controlled neutral and with the real naturalistic conditions and regardless of the language. It also suggests that noun compounds may be represented using a mixture of senses, which could be reflecting their distribution in the training corpora.

Although the idiomatic probes reveal the general tendencies in performance for these models, for more in-depth analyses, this work introduces new measures of idiomatic affinities for even more sensitive ways of analysing idiomaticity. There it is possible to compare, using relative similarity, an anchor, e.g. the sentence containing the original NC, with two other paraphrases. Using the same variants from NCI, two different experiments were done 1) by analysing which preference the model has when entering NCs in their sentences and comparing both the true synonym and the fake compound with the synonyms and 2) by comparing the NC in the sentence with both the true synonym and with the sentence containing either the head or the modifier, whichever is closer in vector space (higher similarity).

The results from both affinities reinforce the results yielded from probes, that the contextualised models are highly impacted by statistical information and also struggle to detect the lower degree of individual component substitutability especially for idiomatic and that is independent of how informative the context is and the language analysed.
In summary, the contributions of this work are:

- The new Noun Compound Idiomaticity (NCI) dataset contains 27,600 sentences in EN and PT, containing sentences, in both naturalistic and neutral contexts, with noun compounds and different variants of the multi-word expression. Those are annotated by experts on the token level with respect to their idiomaticity, hence being able to correlate with other measures.

- Four different probing tasks along with their baselines using NCI as input, each one is crafted to analyse the static and non-static models' sensibility to the idiomatic property of the noun compounds.

- Two novel affinity tasks using relative distances between representations to analyse how accurate idiomaticity is represented in the word embedding generated by contextualised models. They are also supported by the variants in NCI and do not need any fine-tuning or additional post-processing on the pre-trained weights.

- As a by-product of the dissertation, two published papers were published, both Garcia et al. (2021a) and Garcia et al. (2021b), in top-level conferences from NLP area.

In future work, affinity measures can also be used to address unwanted biases towards the non-target meaning, directing the fine-tuning of the model in the direction of the relevant meaning and reducing the impact of statistical information. Also, different similarity measure functions than cosine similarity can be used, as Zhou et al. (2022) points that it captures more training data frequency, and other pooling functions different than averaging of the MWE embeddings, which Arora, Liang and Ma (2017) shows that can yield better results.
For ambiguous NCs, adding probes for the different senses is also an opportunity for future work as could be used to measure any training bias towards one of the senses. Additionally, applying the probing and affinity tasks to more languages, and examining how multilingual information can be used to refine the representation of noun compounds and other MWEs is also an interesting path to follow.

REFERÊNCIAS

AGHAZADEH, E.; FAYYAZ, M.; YAGHOOBZADEH, Y. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In: **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 2037–2050. Available from Internet: <https://aclanthology.org/2022.acl-long.144>.

AKBIK, A. et al. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 54–59. Available from Internet: <https://aclanthology.org/N19-4010>.

ARORA, S.; LIANG, Y.; MA, T. A simple but tough-to-beat baseline for sentence embeddings. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2017.

BARONI, M. et al. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. **Language resources and evaluation**, Springer, v. 43, n. 3, p. 209–226, 2009.

BENGIO, Y. et al. A neural probabilistic language model. **Journal of machine learning research**, v. 3, n. Feb, p. 1137–1155, 2003.

CASTRO, P. V. Quinta de; SILVA, N. Félix Felipe da; SOARES, A. da S. Portuguese Named Entity Recognition Using LSTM-CRF. In: VILLAVICENCIO, A. et al. (Ed.). **Proceedings of the 13th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2018)**. Canela–RS, Brazil: Springer, Cham, 2018. p. 83–92. ISBN 978-3-319-99722-3. Available from Internet: <https://link.springer.com/chapter/10.1007/978-3-319-99722-3_9>.

CHAKRABARTY, T.; CHOI, Y.; SHWARTZ, V. It's not Rocket Science: Interpreting Figurative Language in Narratives. **Transactions of the Association for Computational Linguistics**, v. 10, p. 589–606, 05 2022. ISSN 2307-387X. Available from Internet: <https://doi.org/10.1162/tacl_a_00478>.

CHANG, T.-Y.; CHEN, Y.-N. 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)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 6064–6070. Available from Internet: <https://aclanthology.org/D19-1627>.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. In: **27th Annual Meeting of the Association for Computational Linguistics**. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 1989. p. 76–83. Available from Internet: <https://aclanthology.org/P89-1010>.

COENEN, A. et al. Visualizing and Measuring the Geometry of BERT. In: WALLACH, H. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2019. v. 32, p. 8594–8603. Available

from Internet: <https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf>.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: **Proceedings of the 25th international conference on Machine learning**. [S.l.: s.n.], 2008. p. 160–167.

CORDEIRO, S. et al. Unsupervised compositionality prediction of nominal compounds. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 45, n. 1, p. 1–57, mar. 2019. Available from Internet: <https://aclanthology.org/J19-1001>.

CRUSE, D. A. **Lexical semantics**. [S.l.]: Cambridge university press, 1986.

DANKERS, V.; LUCAS, C.; TITOV, I. Can transformer be too compositional? analysing idiom processing in neural machine translation. In: **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 3608–3626. Available from Internet: <https://aclanthology.org/2022.acl-long.252>.

DEVLIN, J. et al. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Available from Internet: <https://aclanthology.org/N19-1423>.

DINH, E.-L. D.; EGER, S.; GUREVYCH, I. Killing four birds with two stones: Multi-task learning for non-literal language detection. In: **Proceedings of the 27th International Conference on Computational Linguistics**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 1558–1569. Available from Internet: <https://aclanthology.org/C18-1132>.

ERK, K. Vector space models of word meaning and phrase meaning: A survey. **Language and Linguistics Compass**, Wiley Online Library, v. 6, n. 10, p. 635–653, 2012.

ETHAYARAJH, K. 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)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 55–65. Available from Internet: <https://aclanthology.org/D19-1006>.

ETTINGER, A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 8, p. 34–48, 2020. Available from Internet: <https://aclanthology.org/2020.tacl-1.3>.

FAKHARIAN, S.; COOK, P. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In: **Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)**. Online: Association for Computational Linguistics, 2021. p. 23–32. Available from Internet: <https://aclanthology.org/2021.mwe-1.4>.

FARAHMAND, M.; HENDERSON, J. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In: **Proceedings of the 12th Workshop on Multiword Expressions**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 61–66. Available from Internet: <https://www.aclweb.org/anthology/W16-1809>.

FILHO, J. A. W. et al. The brWaC corpus: A new open resource for Brazilian Portuguese. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Available from Internet: <https://aclanthology.org/L18-1686>.

FINLAYSON, M.; KULKARNI, N. Detecting multi-word expressions improves word sense disambiguation. In: **Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World**. Portland, Oregon, USA: Association for Computational Linguistics, 2011. p. 20–24. Available from Internet: <https://aclanthology.org/W11-0805>.

GARCIA, M. et al. 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)**. Online: Association for Computational Linguistics, 2021. p. 2730–2741. Available from Internet: <https://aclanthology.org/2021.acl-long.212>.

GARCIA, M. et al. 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**. Online: Association for Computational Linguistics, 2021. p. 3551–3564. Available from Internet: <https://aclanthology.org/2021.eacl-main.310>.

GRAVE, E. et al. Learning word vectors for 157 languages. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Available from Internet: <https://aclanthology.org/L18-1550>.

HABER, J.; POESIO, M. Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In: **Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics**. Barcelona, Spain (Online): Association for Computational Linguistics, 2020. p. 114–124. Available from Internet: <https://aclanthology.org/2020.starsem-1.12>.

HARRIS, Z. S. Distributional structure. **Word**, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954.

HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. Uberlândia, Brazil: Sociedade Brasileira de Computação, 2017. p. 122–131. Available from Internet: <https://aclanthology.org/W17-6615>.

HENDERSON, J. The unstoppable rise of computational linguistics in deep learning. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 6294–6306. Available from Internet: <https://aclanthology.org/2020.acl-main.561>.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

KASSNER, N.; SCHÜTZE, H. 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**. Online: Association for Computational Linguistics, 2020. p. 7811–7818. Available from Internet: <https://aclanthology.org/2020.acl-main.698>.

KENNISON, S. M.; MESSER, R. H. **Psycholinguistics**. Oxford University Press (OUP), 2014. Available from Internet: <https://doi.org/10.1093/obo/9780199828340-0153>.

KING, M.; COOK, P. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of english verb-noun combinations. In: GUREVYCH, I.; MIYAO, Y. (Ed.). **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers**. Association for Computational Linguistics, 2018. p. 345–350. Available from Internet: <https://www.aclweb.org/anthology/P18-2055/>.

KRIPPENDORFF, K. **Computing Krippendorff's Alpha-Reliability**. 2011. Postprint version. Retrieved from <http://repository.upenn.edu/asc_papers/43>.

LANDAUER, T. K.; DUMAIS, S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. **Psychological review**, American Psychological Association, v. 104, n. 2, p. 211, 1997.

LANDIS, J. R.; KOCH, G. G. The Measurement of Observer Agreement for Categorical Data. **Biometrics**, JSTOR, v. 33, p. 159–174, 1977.

LINZEN, T.; BARONI, M. Syntactic structure from deep learning. **Annual Review of Linguistics**, v. 7, n. 1, p. 195–212, 2021. Available from Internet: <https://doi.org/10.1146/annurev-linguistics-032020-051035>.

LINZEN, T.; DUPOUX, E.; GOLDBERG, Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 4, p. 521–535, 2016. Available from Internet: <https://aclanthology.org/Q16-1037>.

LIU, N. F.; SCHWARTZ, R.; SMITH, N. A. Inoculation by fine-tuning: A method for analyzing challenge datasets. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 2171–2179. Available from Internet: <https://aclanthology.org/N19-1225>.

LUND, K.; BURGESS, C. Producing high-dimensional semantic spaces from lexical co-occurrence. **Behavior research methods, instruments, & computers**, Springer, v. 28, n. 2, p. 203–208, 1996.

MASINI, F. **Multi-Word Expressions and Morphology**. Oxford University Press, 2019. Available from Internet: <https://doi.org/10.1093/acrefore/9780199384655.013.611>.

MCDONALD, S.; RAMSCAR, M. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In: **Proceedings of the Annual Meeting of the Cognitive Science Society**. [S.l.: s.n.], 2001. v. 23.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2**. USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119. Available from Internet: <http://dl.acm.org/citation.cfm?id=2999792.2999959>.

MILLER, G. A. Empirical methods in the study of semantics. **Semantics, an interdisciplinary reader in philosophy, linguistics, and psychology**, p. 569–585, 1971.

MILLER, G. A. WordNet: a lexical database for English. **Communications of the ACM**, ACM New York, NY, USA, v. 38, n. 11, p. 39–41, 1995. Available from Internet: <https://dl.acm.org/doi/10.1145/219717.21974>.

NANDAKUMAR, N.; BALDWIN, T.; SALEHI, B. 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**. Minneapolis, USA: Association for Computational Linguistics, 2019. p. 27–34. Available from Internet: <https://www.aclweb.org/anthology/W19-2004>.

NEDUMPOZHIMANA, V.; KELLEHER, J. Finding BERT's idiomatic key. In: **Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)**. Online: Association for Computational Linguistics, 2021. p. 57–62. Available from Internet: <https://aclanthology.org/2021.mwe-1.7>.

PEARCE, D. Using conceptual similarity for collocation extraction. In: **Proceedings of the Fourth annual CLUK colloquium**. [S.l.: s.n.], 2001.

PEDINOTTI, P. et al. A howling success or a working sea? testing what BERT knows about metaphors. In: **Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP**. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 192–204. Available from Internet: <https://aclanthology.org/2021.blackboxnlp-1.13>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Available from Internet: <https://aclanthology.org/D14-1162>.

PETERS, M. E. et al. 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)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 2227–2237. Available from Internet: <https://aclanthology.org/N18-1202>.

PETERS, M. E. et al. Deep contextualized word representations. In: **Proc. of NAACL**. [S.l.: s.n.], 2018.

PETRONI, F. et al. Language models as knowledge bases? 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)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 2463–2473. Available from Internet: <https://aclanthology.org/D19-1250>.

RADEMAKER, A. et al. OpenWordNet-PT: A project report. In: **Proceedings of the Seventh Global Wordnet Conference**. Tartu, Estonia: University of Tartu Press, 2014. p. 383–390. Available from Internet: <https://aclanthology.org/W14-0153>.

REDDY, S.; MCCARTHY, D.; MANANDHAR, S. An empirical study on compositionality in compound nouns. In: **Proceedings of 5th International Joint Conference on Natural Language Processing**. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 2011. p. 210–218. Available from Internet: <https://aclanthology.org/I11-1024>.

REIMERS, N.; GUREVYCH, I. Alternative Weighting Schemes for ELMo Embeddings. **CoRR**, abs/1904.02954, 2019. Available from Internet: <http://arxiv.org/abs/1904.02954>.

REIMERS, N.; GUREVYCH, I. 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)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3982–3992. Available from Internet: <https://aclanthology.org/D19-1410>.

RICHARDSON, K. et al. Probing natural language inference models through semantic fragments. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 05, p. 8713–8721, Apr. 2020. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/6397>.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in BERTology: What we know about how BERT works. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 8, p. 842–866, 2020. Available from Internet: <https://aclanthology.org/2020.tacl-1.54>.

SAG, I. A. et al. 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)**. Mexico City, Mexico: Springer, Berlin, Heidelberg, 2002. p. 1–15. Available from Internet: <https://link.springer.com/chapter/10.1007/3-540-45715-1_1>.

SAHLGREN, M. The Distributional Hypothesis. **Rivista di Linguistica (Italian Journal of Linguistics)**, v. 20, n. 1, p. 33–53, 2008.

SALEHI, B. et al. The impact of multiword expression compositionality on machine translation evaluation. In: **Proceedings of the 11th Workshop on Multiword Expressions**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 54–59. Available from Internet: <https://aclanthology.org/W15-0909>.

SANH, V. et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, 2019.

SCHNEIDER, N. et al. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In: **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**. San Diego, California: Association for Computational Linguistics, 2016. p. 546–559. Available from Internet: <https://aclanthology.org/S16-1084>.

SCHONE, P.; JURAFSKY, D. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In: **Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing**. [s.n.], 2001. Available from Internet: <https://aclanthology.org/W01-0513>.

SCHUSTER, T. et al. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 1599–1613. Available from Internet: <https://aclanthology.org/N19-1162>.

SCHÜTZE, H. Automatic word sense discrimination. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 24, n. 1, p. 97–123, 1998. Available from Internet: <https://aclanthology.org/J98-1004>.

SHWARTZ, V.; DAGAN, I. Still a pain in the neck: Evaluating text representations on lexical composition. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 7, p. 403–419, 2019. Available from Internet: <https://aclanthology.org/Q19-1027>.

TAN, M.; JIANG, J. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)**. Held Online: INCOMA Ltd., 2021. p. 1397–1407. Available from Internet: <https://aclanthology.org/2021.ranlp-main.156>.

TENNEY, I. et al. What do you learn from context? Probing for sentence structure in contextualized word representations. In: **Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019)**. New Orleans, Louisiana: [s.n.], 2019. Available from Internet: <https://arxiv.org/pdf/1905.06316.pdf>.

TSVETKOV, Y. et al. Metaphor detection with cross-lingual model transfer. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**

(Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 248–258. Available from Internet: <https://aclanthology.org/P14-1024>.

VASWANI, A. et al. **Attention Is All You Need**. 2017. ArXiv preprint arXiv:1706.03762.

VELDHOEN, S.; HUPKES, D.; ZUIDEMA, W. H. Diagnostic classifiers revealing how neural networks process hierarchical structure. In: **CoCo@NIPS**. [s.n.], 2016. Available from Internet: <http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper6.pdf>.

VILARES, D. et al. Parsing as pretraining. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 05, p. 9114–9121, Apr. 2020. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/6446>.

VOITA, E.; TITOV, I. Information-theoretic probing with minimum description length. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 183–196. Available from Internet: <https://aclanthology.org/2020.emnlp-main.14>.

VULIĆ, I. et al. Probing pretrained language models for lexical semantics. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 7222–7240. Available from Internet: <https://aclanthology.org/2020.emnlp-main.586>.

WIEDEMANN, G. et al. 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**. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, 2019. p. 161–170. Available from Internet: <https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_43.pdf>.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**. Online: Association for Computational Linguistics, 2020. p. 38–45. Available from Internet: <https://aclanthology.org/2020.emnlp-demos.6>.

YU, L.; ETTINGER, A. Assessing phrasal representation and composition in transformers. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 4896–4907. Available from Internet: <https://aclanthology.org/2020.emnlp-main.397>.

ZENG, Z.; BHAT, S. Idiomatic expression identification using semantic compatibility. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 9, p. 1546–1562, 2021. Available from Internet: <https://aclanthology.org/2021.tacl-1.92>.

ZHOU, K. et al. Problems with cosine as a measure of embedding similarity for high frequency words. In: **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Dublin, Ireland: Association

for Computational Linguistics, 2022. p. 401–423. Available from Internet: <https://aclanthology.org/2022.acl-short.45>.

APÊNDICE A — RESUMO EXPANDIDO

Modelos que representam palavras com seu contexto vem sendo utilizados para capturar diferentes uso de palavras, e podem ser uma alternativa atrativa para representar idiomaticidade na linguagem. Entretanto, não é claro como esses modelos representam a idiomaticidade ou em qual extensão conseguem capturá-la. Nesse trabalho, são propostas medidas para avaliar se algumas das propriedades linguísticas esperadas em compostos substantivos, especialmente aqueles relacionados a significados idiomáticos, suas dependências com o contexto ao redor e as suas sensibilidades a escolhas lexicais, estão disponíveis em algumas das representações amplamente utilizadas na área.

Trabalhos Relacionados Modelos computacionais que representam uma unidade linguística, por exemplo uma palavra, como vetores n-dimensionais (LUND; BURGESS, 1996; LANDAUER; DUMAIS, 1997; SCHÜTZE, 1998) vem sendo amplamente utilizados em diversos campos de pesquisa. Esses são geralmente aprendidos de maneira não-supervisionada explorando a sua hipótese distribuicional prevendo o contexto de uma palavra alvo, assim gerando os chamados *word embeddings*, método popularmente introduzido em *word2vec* (MIKOLOV et al., 2013). Apesar do sucesso dos modelos baseados em contagem de palavras ou predição, eles representam os diferentes sentidos de uma palavra em um único vetor estático, assim operações complexas sobre os mesmos são necessárias para tratar com polisemia (ERK, 2012).

Com o nascimento dos modelos baseados em redes neurais, foram possíveis gerar representação de palavras baseados em diferentes contextos. Modelos como ELMo (PETERS et al., 2018a), que usa redes LSTM (HOCHREITER; SCHMIDHUBER, 1997), or BERT (DEVLIN et al., 2019), que são treinados com a arquitetura Transformer (VASWANI et al., 2017), se tornaram unanimidade no campo de PLN por causa das suas performances de estado-da-arte em tarefas finais.

Existem duas principais direções quando se fala na avaliação dos modelos linguísticos probabilísticos: a primeira se dá entorno da capacidade do modelo detectar que uma sentação contém uma expressão idiomática (SHWARTZ; DAGAN, 2019; TAN; JIANG, 2021) e a outra, que é o foco do trabalho, se da na detecção da habilidade de codificar corretamente o significado idiomático.

A comparação dos modelos estáticos e contextualizados em relação a representação de expressões multi-palavra são misturadas. Shwartz and Dagan (2019) encontrou resultados melhores para o BERT em tarefas de composição lexical (incluindo a literalidade de

compostos substantivos), entretando para a capturação de idiomaticidade King and Cook (2018) e Nandakumar, Baldwin and Salehi (2019) mostram que modelos estáticos como *word2vec* obtiveram melhor performance.

Um ponto de destaque é que a maioria dos experimentos que avaliam idiomaticidade são a um nível de típo, ou seja, eles obtém a representação de uma expresão com mais de uma palavra fazendo a média de sua representação em diferentes sentenças que são extraídas de maneira automática.

Neste trabalho é apresentado um conjunto de medidas que explorão a representação da idiomaticidade nos modelos que representam vetorialmente as palavras. Ele é uma extensão de dois artigos Garcia et al. (2021a) e Garcia et al. (2021b) já publicados em conferências internacionais.

Materiais Para avaliar esses pontos, foi construído o conjunto de dados *Noun Compound Idiomaticity (NCI)*, que contém anotações de anotadores humanos para compostos substantivos e suas paráfrases, em contexto neutro e informativo, em dois idiomas: Inglês e Português. O conjunto, composto por 27.600 sentenças, também contém avaliações idiomáticas humanas para cada composto substantivo em âmbito de tipo (isolado) e contextualizado.

Métodos Para avaliação, é proposto quatro tipos de medidas que avaliam quão bem os modelos distinguem significados idiomáticos e literais. A primeira gira em torno da capacidade dos modelos em capturar a similaridade entre um composto substantivo e seu sinônimo em uma mesma sentença; a segunda avalia se os modelos são capazes de detectar uma sobreposição semântica entre compostos substantivos composicionais e seus componentes individuas; a terceira se os modelos são capazes de detectar perturbações idiomáticas causas por variações léxicas; e a quarta avalia o quanto de contexto as representações vetoriais dos modelos representam.

Também é definido medidas um conjunto de medidas, chamadas de afinidades, que determinam o quanto desses sentidos são capturados na representação do composto. Isso é feito de duas formas, a primeira comparando o sinônimo verdadeiro com composto formado pelos sinônimos das componentes e segundo determinando se um composto substantivo composicional tem afinidade por uma das componentes dele.

Resultados e Conclusão Resultados obtidos com modelos como ELMo, BERT e algumas de suas variantes, indicam que idiomaticidade ainda não é representada com precisão por modelos contextualizados.

APÊNDICE B — SANITY CHECKS

B.1 Correlation between Naturalistic and Neutral Sentence Variants

As described in 4 the intention of neutral sentences is to create a sentence with very little context information and understand if it can represent/or be analyzed as a naturalistic sentences. In order to comprehend this, it is calculated the spearman correlation between both naturalistic and neutral cosine similarities, which can be found in both B.1 and B.2.

	word2vec		*GloVe*		fastText	
	Sent	NC	Sent	NC	Sent	NC
EN						
P1	0,51	0,99	0,49	0,99	0,46	0,99
A1	0,87	0,99	0,85	0,99	0,89	0,99
PT						
P1	0,60	0,98	0,61	0,98	0,59	0,94
A1	0,79	0,97	0,78	0,98	0,55	0,93

Tabela B.1 – Spearman ρ correlation between naturalistic and neutral sentence variants for both English and Portuguese, only static models, P1 and A1. $p \leq 0.05$. Non-significant results were omitted from the table.

	ELMo		BERT		BERT ML		DistilB ML		SBERT ML	
	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN										
P1	0,54	0,92	0,48	0,76	0,51	0,85	0,51	0,95	0,66	0,99
A1	0,75	0,90	0,63	0,80	0,59	0,81	0,73	0,94	0,90	0,99
PT										
P1	0,75	0,96	0,59	0,89	0,61	0,83	0,68	0,94	0,79	0,98
A1	0,80	0,97	0,61	0,80	0,60	0,82	0,81	0,95	0,91	0,98

Tabela B.2 – Spearman ρ correlation between naturalistic and neutral sentence variants for both English and Portuguese, only non-static models, P1 and A1. $p \leq 0.05$. Non-significant results were omitted from the table.

B.2 Does it depend on the granularity of the judgment?

The results obtained for these probes for each NC (at type level) are also mirrored by those found for the probes for individual sentences for each NC (at token level). For the latter it is considered similarity at token level, and derived similarity at type level by averaging them for the 3 sentences.

The performance reported for some of these models for idiomatic probes displayed at

		word2vec		GloVe		fastText	
	Granularity	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	Type	0,3	0,62	0,3	0,61	0,28	0,61
EN_{Nat}	Token	0,14	0,57	0,14	0,57	0,12	0,57
$EN_{NeuShort}$	Type	0,6	0,62	0,58	0,61	0,6	0,61
$EN_{NeuShort}$	Token	0,58	0,59	0,56	0,58	0,58	0,58
PT_{Nat}	Type	0,15	0,45	0,1	0,39	0,13	0,19
PT_{Nat}	Token	0,1	0,47	-	0,42	-	0,21
$PT_{NeuShort}$	Type	0,31	0,43	0,22	0,41	0,2	0,24
$PT_{NeuShort}$	Token	0,3	0,46	0,22	0,44	0,21	0,27

Tabela B.3 – Comparison between the Spearman ρ correlation for P1 experiment and for both type and token granularity, only static models. p$\leq$0.05. Non-significant results omitted from the table.

		ELMo		BERT		BERT ML		DistilB ML		SBERT ML	
	Granularity	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	Type	0,39	0,58	0,36	0,36	0,47	0,64	0,36	0,57	0,45	0,65
EN_{Nat}	Token	0,26	0,53	0,2	0,31	0,32	0,61	0,2	0,53	0,31	0,61
$EN_{NeuShort}$	Type	0,55	0,6	0,51	0,34	0,53	0,58	0,56	0,54	0,6	0,65
$EN_{NeuShort}$	Token	0,54	0,58	0,48	0,31	0,51	0,57	0,54	0,51	0,58	0,63
PT_{Nat}	Type	0,28	0,45	0,28	0,54	0,25	0,42	0,17	0,38	0,28	0,47
PT_{Nat}	Token	0,27	0,46	0,24	0,55	0,18	0,42	0,11	0,36	0,26	0,51
$PT_{NeuShort}$	Type	0,37	0,47	0,34	0,48	0,3	0,35	0,31	0,37	0,46	0,48
$PT_{NeuShort}$	Token	0,37	0,5	0,32	0,47	0,32	0,39	0,29	0,37	0,48	0,52

Tabela B.4 – Comparison between the Spearman ρ correlation for P1 experiment and for both type and token granularity, only non-static models. p$\leq$0.05. Non-significant results omitted from the table.

most weak correlation to human judgments (GARCIA et al., 2021b). However, the human judgments used to evaluate the models were collected at type level (at most one judgment per NC per participant) and the comparison was done against the average of all human judgements per NC. To determine the impact in the performance of these models of using finer-grained human judgments of NC idiomaticity at a token level (at most 1 judgment per NC per participant and per sentence), it is also compared performance at token and type level. It is expected a high agreement if the different sentences selected per NC are predominantly displaying the same sense.

The B.3 and B.4compares the Spearman correlation for both granularities for probe P1 and B.5 and B.6 for affinity A1.

		word2vec		GloVe		fastText	
	Granularity	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	Type	0,38	0,58	0,48	0,52	0,42	0,58
EN_{Nat}	Token	0,39	0,52	0,47	0,47	0,42	0,53
$EN_{NeuShort}$	Type	0,52	0,58	0,47	0,52	0,53	0,58
$EN_{NeuShort}$	Token	0,49	0,54	0,45	0,49	0,51	0,55
PT_{Nat}	Type	0,33	0,43	0,34	0,36	-	0,28
PT_{Nat}	Token	0,33	0,41	0,36	0,36	-	0,31
$PT_{NeuShort}$	Type	0,29	0,44	0,26	0,37	0,21	0,31
$PT_{NeuShort}$	Token	0,29	0,43	0,29	0,37	0,23	0,35

Tabela B.5 – Comparison between the Spearman ρ correlation for A1 experiment and for both type and token granularity, only static models. $p \leq 0.05$. Non-significant results omitted from the table.

		ELMo		BERT		BERT ML		DistilB ML		SBERT ML	
	Granularity	Sent	NC	Sent	NC	Sent	NC	Sent	NC	Sent	NC
EN_{Nat}	Type	0,45	0,54	0,37	0,54	0,45	0,54	0,47	0,5	0,54	0,58
EN_{Nat}	Token	0,42	0,49	0,38	0,5	0,44	0,49	0,47	0,45	0,49	0,54
$EN_{NeuShort}$	Type	0,52	0,54	0,48	0,53	0,38	0,49	0,48	0,5	0,6	0,58
$EN_{NeuShort}$	Token	0,49	0,51	0,47	0,5	0,38	0,48	0,46	0,47	0,59	0,57
PT_{Nat}	Type	0,3	0,38	0,19	0,33	0,25	0,33	0,32	0,35	0,31	0,38
PT_{Nat}	Token	0,29	0,38	0,19	0,32	0,25	0,33	0,32	0,34	0,34	0,43
$PT_{NeuShort}$	Type	0,33	0,41	0,2	0,34	0,25	0,32	0,3	0,36	0,39	0,4
$PT_{NeuShort}$	Token	0,32	0,41	0,2	0,32	0,27	0,35	0,31	0,37	0,43	0,47

Tabela B.6 – Comparison between the Spearman ρ correlation for A1 experiment and for both type and token granularity, only non-static models. $p \leq 0.05$. Non-significant results omitted from the table.