From 'It's All Greek to Me' to 'Nur Bahnhof verstehen': An Investigation of mBERT's Cross-Linguistic Capabilities

Aria Rastegar[1,*], Pegah Ramezani[1]

[1]FAU Erlangen-Nuremberg, Germany

Abstract

This study investigates the impact of cross-linguistic similarities on idiom representations in mBERT, focusing on English and German idioms categorized by different degrees of similarity. We aim to determine whether different degrees of cross-linguistic similarities significantly affect mBERT's representations and to observe how these representations change across its 12 layers. Contrary to our initial hypothesis, cross-linguistic similarity did not uniformly impact idiom representations across all layers. While early and middle layers showed no significant differences among idiom categories, higher layers (from Layer 8 onwards) revealed more nuanced processing. Specifically, significant differences between the control category and idioms with similar meaning (SM), as well as between idioms with similar lexical items (SL) and those with similar semantics (SM) were observed. Our analysis revealed that early layers provided general representations, while higher layers showed increased differentiation between literal and figurative meanings. This was evidenced by a general decrease in cosine similarities from Layer 5 onwards, with Layer 8 demonstrating the lowest cosine similarities across all categories. Interestingly, a trend suggests that mBERT performs slightly better with more literal hints. The order of cosine similarity for the categorizations was: idioms with a degree of formal similarity, control idioms, idioms with both formal and semantic similarity, and finally idioms with only semantic similarity. These findings indicate that mBERT's processing of idioms evolves significantly across its layers, with cross-linguistic might affect more significantly in higher layers where more abstract semantic processing likely occurs.

Keywords

mBERT, Multi-word Expressions, Idioms, Bertology, computationally-aided cross-linguistic analysis

1. Introduction

Idioms are one of the most studied linguistic concepts that broadly can be defined as multi-word expressions that are often fixed in terms of their syntactic and lexical aspects, while they usually carry meanings that cannot be directly deduced from the meaning of individual words they contain [1, 2, 3, 4]. Given their syntactic and structural fixedness and non-compositional aspects, they were perceived as peripheral, supplementary, or appendixes to language grammars in earlier approaches to idioms [5, p.504]. However, with the increasing interest in corpus studies of language, it has been observed that much of human linguistic production is routinized and prefabricated [6, 7, 8]. Multi-word expressions with a high degree of conventionality do not seem to be marginal or limited linguistic constructions, as they play an important role in our everyday life [9, 10, 11]. In addition, they seem to be used in communication across various contexts, from novels to political debates and therapeutic dialogues [12]. Given their characteristics and their conventionalized meanings, they pose many challenges to language speakers, especially non-native language speakers [13].

However, their characteristics also make them a good case study in different experimental linguistics settings. Recent advancements in Large Language Models (LLMs) and their widespread application have prompted linguists to investigate the performance of these models across various linguistic concepts, including idioms [14, 15, 16]. In addition, in the case of multi-lingual models, an interesting research area is how these models encode the different languages on which they are trained [17, 18].

In this study, a categorization of English and German idioms based on three cross-linguistic degrees of similarity is proposed. One category includes idioms that have similar formal and semantic aspects in these languages; the second includes idioms with formal similarities but different semantic aspects; and the third category includes idioms with similar semantic aspects but different formal aspects. The goal of our work is to consider how cross-linguistic similarities among idioms affect the representation of idioms in mBERT. More specifically, the questions underlying the following experiment were:

1. Does cross-linguistic similarity have a significant impact on the representation of idioms in mBERT?
2. Does the degree of cross-linguistic similarity and the representation of the model change across the 12 layers of mBERT?

We hypothesized that mBERT's performance would depend on how it utilizes its multilingual training data. Namely, if mBERT draws from a collective pool of all

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

aria.rastegar@fau.de (A. Rastegar); pegah.ramezani@fau.de (P. Ramezani)

0009-0006-1056-1663 (A. Rastegar)

languages, it should perform consistently across all cross-linguistic categories, similar to how it represents idioms from the language it has been given, that is English in this case. However, if it primarily retrieves data from specific languages, we expect to observe significant performance differences among the categories, potentially mirroring some of the patterns seen in cross-linguistic studies with second language speakers. That is, identical cross-linguistic idioms should be represented almost similarly to the control idioms (in this case, English idioms), and idioms with formal and lexical correspondence could both be represented similarly and, in some cases, more differently from the control idioms. Finally, idioms with only corresponding semantics and different formal aspects should be the most differently represented idioms compared to the control group. Furthermore, given the proposed categorizations based on formal and semantic similarities, we anticipated varying performance across mBERT's 12 layers. Particularly, in lower layers, we expect less differentiation among categories, as these layers typically capture more surface-level features. While in higher layers, which represent more of the semantic aspects, we anticipate more varying trends and larger differences among the categories. Mostly because we are primarily focused on the figurative meaning of idioms across different categories.

2. Related Works

Studies on idiomatic expressions generally focus on two main comparisons: the understanding of idioms by participants (literal or figurative understanding of the phrases), and the difference between understanding idioms and non-idiomatic or novel phrases [19, 13, 2]. The figurative meaning of an idiom is usually conventionalized and relatively fixed; therefore, native speakers seem to simply access it. However, its literal interpretation can be logical, nonsensical, or somewhere in between. For instance, as [13] explains, while it is possible that someone is bathing in the example of being 'in hot water' (with an idiomatic or figurative interpretation denoting "in trouble"), in the idiom 'to be on cloud 9' (with a figurative interpretation of "being very happy") there is no likely, logical interpretation in the real world in which a person can be found on a cloud called "9". Furthermore, when considering the literal interpretation of an idiom, research can remain at the phrasal level or can consider access to the literal meaning of the constituent parts. When again considering the idiom 'in hot water', focus on access to the figurative interpretation is possible, "in trouble," access to the whole interpretation of the literal phrase, "to be in heated water such as a bath or hot springs," or access to the meanings of the individual constituent words such as "hot" or "water" is expected. In cross-linguistic studies on idioms, one of the aspects that have been studied is the concept of cross-linguistic similarity or translatability. Among language speakers, the degree of translatability of an idiom in their L1 and L2 seems to play a significant role in how they interpret and understand the idioms [20, 21, 22, 23, 13, 24, 25]. In one of the earliest investigations of translatability's effect on L2 idiom comprehension, [20] examined how advanced Venezuelan learners of English understood and produced English idioms with varying degrees of translatability from Spanish. Using multiple tasks (multiple-choice recognition, open-ended definition, discourse completion, and translation), Irujo found that idioms with identical expressions in both languages (e.g., "point of view" / "punto de vista") were easiest to comprehend and produce. In contrast, idioms representing equivalent concepts without direct translations (e.g., "to pull his leg" vs. *tomarle el pelo* "to take to him the hair") posed the greatest challenge. The study also found a negative interference in the form of transfer errors, when participants producing partially matching idioms (e.g., "to catch him red-handed" vs. *cogerle con las manos en la masa* "to catch him with the hands in the dough"). Irujo [20] concluded that L1 knowledge can be both beneficial and detrimental to L2 idiom processing. For idioms with direct translations, L1 knowledge facilitates both comprehension and production in L2. However, for idioms with partial similarities between languages, L1 knowledge can lead to transfer errors. Additionally, a study by [21], which focused on 3rd-year learners of Spanish, French, and German, found that the translatability of idioms was a key factor in predicting the speed and accuracy of their production, both with and without context. Furthermore, [21] observed that translation is one of the most common strategies employed by L2 users to comprehend idioms, as indicated by learners' written reflections. Also, [23] discovered that idioms that could be translated literally from Latvian and Mandarin Chinese into English were better comprehended by participants. Furthermore, they observed that regardless of the overall similarity of the studied languages to English, if the idioms were similar or if they were decomposable, they would be understood by the participants. Although these studies are focused on language learners and speakers, and they may include more variables, we can argue that, such cross-linguistic similarities, can affect how idioms are represented, in multi-lingual contexts.

In the case of large language models, the way they embed and encode idioms and multi-word expressions has been an ongoing debate [26, 27, 28, 16, 14]. Most studies focusing on how language models encode idioms examine the task of identifying idiomatic expressions in a text. In early works on this task, researchers developed expression-specific models that can capture the idiomatic expressions in a text [29], while more recent approaches have demonstrated that more generic models

such as BERT and mBERT [30] are also able to capture idioms [26, 27, 28]. Studies on the internal mechanisms of how transformer-based language models process idioms demonstrated that BERT, Multilingual BERT, and DistilBERT represent idioms distinctively compared to literal language [16]. These studies also observed that the semantic meaning of idioms is captured more effectively in deeper layers of the models. They found that words within idioms receive less attention from other words in the sentence compared to words in literal contexts. However, [14] argue that LLMs capture MWE semantics inconsistently, as shown by reliance on surface patterns and memorized information. MWE meaning is also strongly localized, predominantly in early layers of the architecture. They also discuss that representations benefit from specific linguistic properties, such as lower semantic idiosyncrasy and ambiguity of target expressions.

Moving from LLMs and idioms, there are different arguments on how models such as BERT work [31], and in the case of multi-lingual approaches, how multilingual they are [17, 18, 32]. Works on the mechanisms of BERT demonstrate that it captures significant linguistic information, with lower layers focusing on local syntactic relationships and higher layers encoding more complex linguistic features. The self-attention heads in BERT show specialization for certain linguistic functions, though many exhibit redundant patterns, suggesting overparameterization. While BERT demonstrates some ability to capture world knowledge, its reasoning capabilities appear limited. Despite impressive performance on many NLP tasks, BERT shows limitations in handling negation, numerical reasoning, and complex inference, often relying on shallow heuristics [31]. Investigations on mBERT across 39 languages found that it performs well on high-resource languages but struggles with low-resource languages. For languages with limited Wikipedia data (which was used to train mBERT), performance drops significantly, especially for tasks like named entity recognition. This suggests that the quality of representations learned by mBERT is not uniform across all 104 languages it supports [32]. Additionally, [18] conducted a series of probing experiments to understand mBERT's cross-lingual abilities. They found that mBERT performs surprisingly well on zero-shot cross-lingual model transfer, even between languages with different scripts. Their analysis suggests that mBERT learns multilingual representations that go beyond simple vocabulary memorization. However, they also note that transfer works best between typologically similar languages, indicating some limitations in mBERT's ability to generalize across very different language structures.

3. Dataset

To investigate our research questions concerning the impact of cross-linguistic similarity on the representation of idioms in mBERT and how this representation changes across the model's 12 layers, a list of idiomatic expressions was compiled. the dataset consists of 72 idioms: 54 from German and 18 from English, the latter serving as a control group. The German idioms are classified based on their similarity with English idioms, using three categories of cross-linguistic correspondence. The first category includes idioms with the highest degree of formal and semantic similarity. These idioms, such as *die Ruhe vor dem Sturm*, have a corresponding form in English when translated word-for-word, e.g., *the calm before the storm*. In addition to the formal similarity, the meaning of the idiom in the target language is also similar to that of the originating language, in this case referring to a period of calmness before argument or trouble. The second category focuses on formal similarities without semantic correspondence. For instance, *jemanden ausnehmen wie eine Weihnachtsgans* ('to gut someone like a Christmas goose') refers to financially exploiting someone. In English, there is an idiom that contains the word "goose" - *to cook one's goose* - but it refers to sabotaging someone's plans, demonstrating some degrees of formal and lexical similarity without semantic alignment. The third category encompasses idioms with semantic similarities but no formal correspondence. For example, the German idiom *Den Löffel abgeben* ('to pass the spoon') and the English idiom *to kick the bucket* both convey the meaning of dying, while sharing no formal similarities. After categorizing the idioms, the German idioms were literally translated into English. We literally translated the idioms to ensure all expressions can be fed to the model in a single language. This approach allows us to control for the language space in which idioms are presented, given that in more complex tasks different subsets of mBERT can affect how idioms are represented [33]. Additionally, for each idiom, a brief entity or description is selected reflecting its figurative meanings. For example, for "the calm before the storm", "episodic tranquility" is chosen, which refers to the figurative interpretations of the idiom. Table 1, summarizes the proposed categorizations, the original and translated idioms, along with their figuratively related entities.

4. Model, and Experiment

For analyzing the embeddings of the studied idioms and their figurative meanings, the dataset was processed using the "bert-base-multilingual-uncased" model [34] without any fine-tuning. This model consists of 12 hidden layers, each containing 768 neurons, and the activity of

Table 1
Examples of idioms in each category. SI: Similar Idiom (formal and semantic similarity), SL: Similar Lexicon (formal similarity only), SM: Similar Meaning (semantic similarity only).

German Idiom	English Translation	Figurative Meaning	Category
die Ruhe vor dem Sturm	the calm before the storm	episodic tranquility	SI
der ball liegt bei dir	the ball lies with you	responsibility	
jemanden ausnehmen wie eine Weihnachtsgans	to gut someone like a Christmas goose	financially exploit	SL
auffallen wie ein bunter Hund	stand out like a colorful dog	noticeable	
Den Löffel abgeben	give away the spoon	death	SM
Einen Vogel haben	have a bird	acting strange	
–	It rains cats and dogs	heavy rain	Control
	it costs an arm and a leg	expensive	

each layer was extracted for the CLS token. Embeddings for the CLS token from each of the 12 layers for every idiom and its associated meanings were extracted. The model is pretrained on the 102 languages with the largest Wikipedias, which includes both German, the language from which our idioms are derived, and English, which is the target language for the translation of the idioms and used for deriving the embeddings. For each sample, the embeddings of the [CLS] token from all 12 layers of mBERT are extracted. The [CLS] token was chosen because it is designed to capture sentence-level semantics in BERT models [35]. Using the [CLS] token's embedding from models can be used as a powerful method for semantic comparison of texts, which can then be compared using similarity measures.

4.1. Similarity Calculation

In the next step to measure how similar BERT's understanding is of each idiom, the similarity of embeddings for each idiom with its figurative meanings was calculated. Cosine similarity is used, a widely used method because of its effectiveness and it is mainly used to determine how similar or related two words are based on their vector representations [36, 37].

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\mathbf{v}\text{Idiom} \cdot \mathbf{v}\text{Meaning}}{|\mathbf{v}\text{Idiom}||\mathbf{v}\text{Meaning}|} \quad (1)$$

In Equation 1, $\mathbf{v}$ stands for word embedding, which is a vector with a length of 768. To interpret the result of the cosine similarity in the context of word embedding, a score of 1 means the vectors are identical, 0 means the vectors are orthogonal (no similarity), and -1 means the vectors are opposed.

5. Results

After deriving the CLS embeddings from all layers of mBERT for the translated idioms and their corresponding figurative meanings, the cosine similarities among the derived embeddings were calculated. Figure 1 illustrates the cosine similarities across different layers of mBERT for each idiom category. As it can be seen, the first layer of mBERT showed identical cosine similarities (equal to 1) for all idioms, representing the entry point of the model. Therefore, this layer is excluded from subsequent analyses to avoid skewing our results.

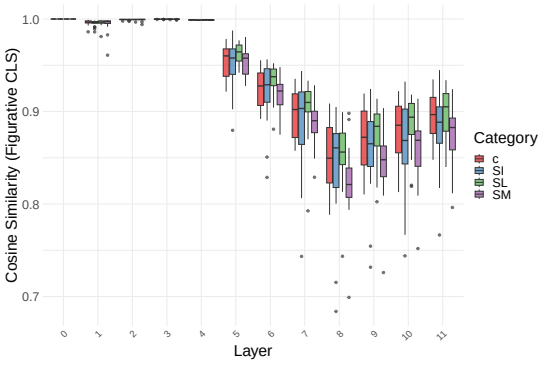

Figure 1: Cosine similarities between idiom embeddings and their figurative meanings across mBERT layers for different cross-linguistic categories. C: Control, SI: Similar Idiom (formal and semantic similarity), SL: Similar Literal (formal similarity only), SM: Similar Meaning (semantic similarity only).

Additionally, as the graph in Figure 1 indicates, the cosine similarities exhibited notable variations across layers. Layer 3 demonstrated the highest cosine similarities across all categories; while layer 8 showed the lowest cosine similarities for all four categories. In addition, as it can be seen from Layer 5 onwards, we observed a general decrease in cosine similarities, suggesting increasing differentiation between CLS representation of idioms and their corresponding figurative meanings in higher layers.

To test our hypothesis on how the embeddings of mBERT would change given the proposed cross-linguistic

similarity categorizations, a linear mixed effects model analysis using the `lme4` [38] package in R [39] was conducted. The model considered layers and categories as fixed effects, with individual idioms as random effects. To analyze the effects a treatment contrast was employed, [40], using the control (C) category as the reference level for categories and the second layer of mBERT (Layer1) as the reference for layers. It is important to note that the model showed high multicollinearity, particularly for the Layer variable and interaction terms (VIF > 10), primarily due to the minimal changes in cosine similarities in the initial layers. While this does not invalidate our results, it does warrant cautious interpretation, especially for the layer effects.

As the figure 2 indicates, and can be seen in table 2 the main effects of Category (SI, SL, SM) were not statistically significant (all $p > .05$), suggesting no overall difference in Cosine Similarity across categories when compared to the baseline category (C). In addition, considering the main effect of the layers it can be observed that there is a significant effect from Layers 5 through 11 (all $p < .001$). The coefficients were increasingly negative for higher layers, indicating a decrease in Cosine Similarity as moving to higher layers this can be seen also in.

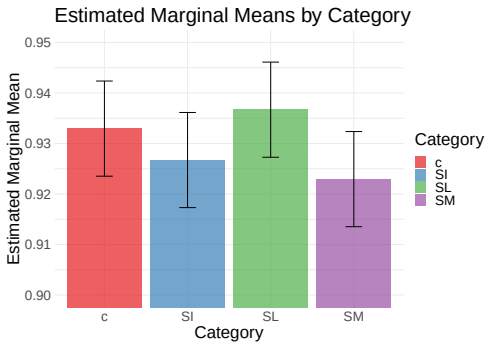

Figure 2: The estimated marginal means for the effect of each of the cross-linguistic categories.

To examine the predicted Cosine Similarity of Figurative CLS representations for each combination of Category and Layer, the estimated marginal means using the `emmeans` package [41] in R computed. In this analysis, the changes in the cosine similarities were compared among the categories, in different layers. The results of the pair-wise comparisons indicate that, For Layers 1-7, there are no significant differences between categories (all p-values > 0.05), this can be also seen in figure 3, in which almost until the 7th layer all of the lines align with each other. However, from layer 8 a significant difference can be seen between the control category and category SM that represents the idioms with cross-linguistically similar semantics (estimate = 0.0272, p = 0.0179). In addition, in layer 8 there is a significant difference between the category SL and SM (estimate = 0.0310, p = 0.0049), and this significant difference continues until layer 10 with estimate = 0.0294, p = 0.0085, and estimate = 0.0248, p = 0.0373.

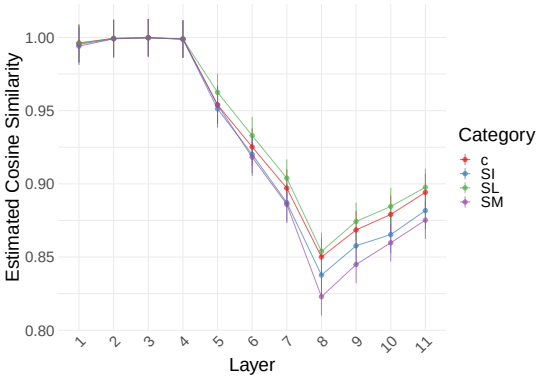

Figure 3: The estimated marginal means for the changes of the cosine similarities for each of the cross-linguistic categories among the layers of mBERT.

6. Discussion and Conclusion

Our study investigated how cross-linguistic similarities among idioms affect their representation in mBERT, with a focus on English and German idioms categorized based on three degrees of cross-linguistic similarity. This study aims to answer two main research questions concerning whether cross-linguistic similarity has a significant impact on the representation of idioms in mBERT, and how the degree of cross-linguistic similarity and the representation of the model change across the 12 layers of mBERT. Our findings provide insights into these questions and our initial hypotheses. Contrary to our initial hypothesis, we found that cross-linguistic similarity does not have a uniformly significant impact on the representations of idioms across all layers of mBERT. The main effects of our translated idioms categorized into cross-linguistic categories (SI: formal and semantic similarity, SL: similar lexicon, SM: Similar Meaning), were not statistically significant when compared to the control category (English idioms) in the early and middle layers of the model. This result may suggest that mBERT might be utilizing knowledge from all languages in its training data as a collective pool, at least in the case of the studied idioms. This aligns with the idea that mBERT learns multilingual representations that go beyond simple vocabulary memorization, as suggested by Pires et al. [18]. However, the emergence of significant differences in higher layers (particularly from Layer 8 onwards) might indicate that mBERT's processing of idioms becomes more nuanced as information

propagates through the network. This finding partially supports our hypothesis that mBERT might show different performances for each cross-linguistic categorization, but suggests that these differences are more significant in the model's deeper layers. Although there are no significant differences among all categories, in Figure 2 there is a continuous trend in different layers showing more similarity first for the SL category, then Control, followed by SI, and finally the SM category. This trend indicates that BERT represents almost all categories similarly, and when there are more literal hints, BERT tends to perform better which aligns with the findings of multi-lingual transfer of Pires et al. [18]. Moreover, for idioms with semantic similarities, the model demonstrates the lowest cosine similarity between the representations of idioms and their figurative meanings, which might suggest that idioms with only semantic correspondence across the studied languages pose a greater challenge for mBERT in capturing the figurative meanings of idioms.

Our second research question focused on how the representation of idioms changes across mBERT's 12 layers. In this analysis, distinct patterns were observed. In early layers (1-4) the cosine similarity for CLS embedding derived from mBERT for the idioms and their corresponding figurative meaning was high and relatively uniform across all categories, suggesting a more general representation, we believe high similarity in early layers can be related to similarity in the syntax of samples and the provided figurative entities since these layers capture more formal and syntactic information. Layer 3 demonstrated the highest cosine similarities, while from Layer 5 onwards, a general decrease in cosine similarities was observed, suggesting increased differences between literal and figurative meanings in higher layers. Layer 8 showed the lowest cosine similarities and marked the beginning of significant differences between categories, particularly for semantically similar idioms (SM category). These findings contribute to our hypothesis that we would observe different performances among the layers of mBERT given the formal and semantic similarities of idioms.

6.1. Limitations and future research

This research also has limitations, that can be tackled in the further and future studies. One of the primary limitations of our study is the size of the dataset. However, the dataset has a good variety of samples but a bigger dataset may improve the generalizability and robustness of our findings. Future research should aim to include a more extensive dataset to confirm and extend these findings. Moreover, literally translating the idioms and the figuratively related entities, can affect on the representations of the model, and the derived cosine similarities; therefore, in further studies, it can be insightful to compare also, how the representations of the model change if the idioms are fed to the model in their original language. In addition, German and English are both Germanic languages and can be considered typologically similar. In future studies, it would be intuitive to compare the categorizations from two more distinct languages to observe how the effect of cross-linguistic similarities changes without the possible influence of typological similarities.

References

[1] J. Pustejovsky, O. Batiukova, The lexicon, Cambridge University Press, 2019.

[2] M. R. Libben, D. A. Titone, The multidetermined nature of idiom processing, Memory & cognition 36 (2008) 1103–1121.

[3] B. Abel, English idioms in the first language and second language lexicon: A dual representation approach, Second language research 19 (2003) 329–358.

[4] R. W. Gibbs Jr, N. P. Nayak, Psycholinguistic studies on the syntactic behavior of idioms, Cognitive psychology 21 (1989) 100–138.

[5] C. J. Fillmore, P. Kay, M. C. O'connor, Regularity and idiomaticity in grammatical constructions: The case of let alone, Language (1988) 501–538.

[6] M. H. Christiansen, I. Arnon, More than words: The role of multiword sequences in language learning and use, 2017.

[7] R. Jackendoff, Précis of foundations of language: Brain, meaning, grammar, evolution„ Behavioral and Brain Sciences 26 (2003) 651–665. doi:10.1017/S0140525X03000153.

[8] J. Sinclair, Corpus, Concordance, Collocation, Describing English language, Oxford University Press, 1991. URL: https://books.google.de/books?id=L8l4AAAAIAAJ.

[9] A. Siyanova-Chanturia, K. Conklin, N. Schmitt, Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers, Second Language Research 27 (2011) 251–272.

[10] S. Wulff, Rethinking idiomaticity, Rethinking Idiomaticity (2008) 1–256.

[11] A. Siyanova, N. Schmitt, Native and nonnative use of multi-word vs. one-word verbs, International Review of Applied Linguistics in Language Teaching 45 (2007) 119–139. URL: https://doi.org/10.1515/IRAL.2007.005. doi:doi:10.1515/IRAL.2007.005.

[12] T. C. Cooper, Processing of idioms by l2 learners of english, TESOL quarterly 33 (1999) 233–262.

[13] S. D. Beck, A. Weber, Bilingual and monolingual idiom processing is cut from the same cloth: The role of the l1 in literal and figurative meaning activation, Frontiers in psychology 7 (2016) 1350.

[14] F. Miletić, S. S. i. Walde, Semantics of multiword expressions in transformer-based models: A survey, Transactions of the Association for Computational Linguistics 12 (2024) 593–612.

[15] M. TAN, J. JIANG, Does bert understand idioms? a probing-based empirical study of bert encodings of idioms.(2021), in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Virtual Conference, September, ????, pp. 1–3.

[16] Y. Tian, I. James, H. Son, How are idioms processed inside transformer language models?, in: Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023), 2023, pp. 174–179.

[17] H. Gonen, S. Ravfogel, Y. Elazar, Y. Goldberg, It's not greek to mbert: inducing word-level translations from multilingual bert, arXiv preprint arXiv:2010.08275 (2020).

[18] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, arXiv preprint arXiv:1906.01502 (2019).

[19] C. Cacciari, P. Tabossi, The comprehension of idioms, Journal of memory and language 27 (1988) 668–683.

[20] S. Irujo, Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language, tesol Quarterly 20 (1986) 287–304.

[21] J. I. Liontas, Killing two birds with one stone: Understanding spanish vp idioms in and out of context, Hispania (2003) 289–301.

[22] D. Titone, G. Columbus, V. Whitford, J. Mercier, M. Libben, Contrasting bilingual and monolingual idiom processing. (2015).

[23] H. Bortfeld, Comprehending idioms cross-linguistically., Experimental psychology 50 (2003) 217.

[24] M. S. Senaldi, D. A. Titone, Less direct, more analytical: Eye-movement measures of l2 idiom reading, Languages 7 (2022) 91.

[25] M. S. Senaldi, J. Wei, J. W. Gullifer, D. Titone, Scratching your tête over language-switched idioms: Evidence from eye-movement measures of reading, Memory & cognition 50 (2022) 1230–1256.

[26] V. Nedumpozhimana, F. Klubička, J. D. Kelleher, Shapley idioms: Analysing bert sentence embeddings for general idiom token identification, Frontiers in Artificial Intelligence 5 (2022) 813967.

[27] G. Salton, R. Ross, J. Kelleher, Idiom token classification using sentential distributed semantics, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 194–204. URL: https://aclanthology.org/P16-1019. doi:10.18653/v1/P16-1019.

[28] V. Nedumpozhimana, J. Kelleher, Finding bert's idiomatic key (2021).

[29] A. Fazly, P. Cook, S. Stevenson, Unsupervised Type and Token Identification of Idiomatic Expressions, Computational Linguistics 35 (2009) 61–103. URL: https://doi.org/10.1162/coli.08-010-R1-07-048. doi:10.1162/coli.08-010-R1-07-048.

[30] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[31] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, 2020. URL: https://arxiv.org/abs/2002.12327. arXiv:2002.12327.

[32] S. Wu, M. Dredze, Are all languages created equal in multilingual bert?, arXiv preprint arXiv:2005.09093 (2020).

[33] J. Libovickỳ, R. Rosa, A. Fraser, How language-neutral is multilingual bert?, arXiv preprint arXiv:1911.03310 (2019).

[34] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[36] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[37] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[38] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67 (2015) 1–48. doi:10.18637/jss.v067.i01.

[39] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2023. URL: https://www.R-project.org/.

[40] D. J. Schad, S. Vasishth, S. Hohenstein, R. Kliegl, How to capitalize on a priori contrasts in linear (mixed) models: A tutorial, Journal of memory and language 110 (2020) 104038.

[41] F. M. S. S. R. Searle, G. A. Milliken, Population marginal means in the linear model: An alternative to least squares means, The American Statistician 34 (1980) 216–221. URL: https://www.tandfonline.com/doi/abs/10.1080/00031305.1980.10483031.

doi:10.1080/00031305.1980.10483031.

A. Appendix A. LMER Model full summary

Fixed Effects					
	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.00	0.01	226.52	153.20	**0.00**
CategorySI	-0.00	0.01	226.52	-0.11	0.91
CategorySL	-0.00	0.01	226.52	-0.03	0.97
CategorySM	-0.00	0.01	226.52	-0.24	0.81
Layer2	0.00	0.01	680.00	0.48	0.63
Layer3	0.00	0.01	680.00	0.55	0.59
Layer4	0.00	0.01	680.00	0.42	0.68
Layer5	-0.04	0.01	680.00	-6.37	**0.00**
Layer6	-0.07	0.01	680.00	-10.70	**0.00**
Layer7	-0.10	0.01	680.00	-14.95	**0.00**
Layer8	-0.15	0.01	680.00	-22.04	**0.00**
Layer9	-0.13	0.01	680.00	-19.26	**0.00**
Layer10	-0.12	0.01	680.00	-17.66	**0.00**
Layer11	-0.10	0.01	680.00	-15.39	**0.00**
CategorySI:Layer2	0.00	0.01	680.00	0.09	0.93
CategorySL:Layer2	0.00	0.01	680.00	0.02	0.98
CategorySM:Layer2	0.00	0.01	680.00	0.20	0.84
CategorySI:Layer3	0.00	0.01	680.00	0.10	0.92
CategorySL:Layer3	0.00	0.01	680.00	0.03	0.98
CategorySM:Layer3	0.00	0.01	680.00	0.22	0.82
CategorySI:Layer4	0.00	0.01	680.00	0.09	0.93
CategorySL:Layer4	0.00	0.01	680.00	0.03	0.98
CategorySM:Layer4	0.00	0.01	680.00	0.22	0.82
CategorySI:Layer5	-0.00	0.01	680.00	-0.19	0.85
CategorySL:Layer5	0.01	0.01	680.00	0.94	0.35
CategorySM:Layer5	0.00	0.01	680.00	0.22	0.82
CategorySI:Layer6	-0.00	0.01	680.00	-0.44	0.66
CategorySL:Layer6	0.01	0.01	680.00	0.86	0.39
CategorySM:Layer6	-0.00	0.01	680.00	-0.51	0.61
CategorySI:Layer7	-0.01	0.01	680.00	-0.95	0.34
CategorySL:Layer7	0.01	0.01	680.00	0.76	0.45
CategorySM:Layer7	-0.01	0.01	680.00	-0.95	0.34
CategorySI:Layer8	-0.01	0.01	680.00	-1.22	0.22
CategorySL:Layer8	0.00	0.01	680.00	0.43	0.66
CategorySM:Layer8	-0.03	0.01	680.00	-2.67	**0.01**
CategorySI:Layer9	-0.01	0.01	680.00	-1.05	0.30
CategorySL:Layer9	0.01	0.01	680.00	0.65	0.52
CategorySM:Layer9	-0.02	0.01	680.00	-2.28	**0.02**
CategorySI:Layer10	-0.01	0.01	680.00	-1.36	0.17
CategorySL:Layer10	0.01	0.01	680.00	0.61	0.54
CategorySM:Layer10	-0.02	0.01	680.00	-1.83	0.07
CategorySI:Layer11	-0.01	0.01	680.00	-1.22	0.22
CategorySL:Layer11	0.00	0.01	680.00	0.40	0.69
CategorySM:Layer11	-0.02	0.01	680.00	-1.79	0.07
Random Effects					
Groups	Variance	Std. Dev.			
idiom	0.00	0.02			
Conditional R^2: 0.908 Marginal R^2: 0.824					

Table 2
summary of linear mixed effects model: The categorizations are Control which is considered as the reference and is not present in the model's summary; SI: Similar Idiom (formal and semantic similarity), SL: Similar Lexicon (formal similarity only), SM: Similar Meaning (semantic similarity only).