Assessing BERT's sensitivity to idiomaticity

Li Liu, François Lareau
OLST, Université de Montréal
Montréal, Canada
{li.liu.2, francois.lareau}@umontreal.ca

Abstract

BERT-like language models have been demonstrated to capture the idiomatic meaning of multiword expressions. Linguists have also shown that idioms have varying degrees of idiomaticity. In this paper, we assess CamemBERT's sensitivity to the degree of idiomaticity within idioms, as well as the dependency of this sensitivity on part of speech and idiom length. We used a demasking task on tokens from 3,127 idioms and 22,551 tokens corresponding to simple lexemes taken from the French Lexical Network (LN-fr), and observed that CamemBERT performs distinctly on tokens embedded within idioms compared to simple ones. When demasking tokens within idioms, the model is not proficient in discerning their level of idiomaticity. Moreover, regardless of idiomaticity, CamemBERT excels at handling function words. The length of idioms also impacts CamemBERT's performance to a certain extent. The last two observations partly explain the difference between the model's performance on idioms versus simple lexemes. We conclude that the model treats idioms differently from simple lexemes, but that it does not capture the difference in compositionality between subclasses of idioms.

Keywords: phraseology, idioms, idiomaticity, multiword expressions (MWEs), language models

1. Introduction

Multiword expressions (MWEs) are characterized by the constrained selection of their components and their partial or complete lack of compositionality (Mel'čuk, 2023). In this paper, we focus on idioms, a prominent category of MWEs known for their non-compositional nature which have long presented a significant challenge for natural language processing (NLP) (Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017).

Idioms cannot be understood simply by the regular combination of the meanings of their components, e.g., *spill the beans* means 'disclose a secret', which cannot be obtained from 'spill'+'beans'. However, while all idioms violate compositionality, some idioms do include the meaning of some or even all of their components, making them more or less semantically transparent. Hence, compositionality in idioms falls on a continuum. According to the degree of inclusion of the meaning of their components, Mel'čuk (2023) classifies idioms into **weak idioms**, which include the meaning of all of their components along with some arbitrary meaning, as in (1), **semi-idioms**, which include the meaning of some but not all of their components along with some arbitrary meaning, as in (2), and **strong idioms**, which are completely non-compositional, as in (3). This is illustrated below with French idioms.

(1) étoile de mer
star of sea
'starfish' = 'star-shaped marine animal'

(2) fruit de mer
fruit of sea
'seafood' = 'food that comes from the sea'

(3) noyer le poisson
drown the fish
'obfuscate things'

The contextualized language model BERT (Devlin et al., 2019), pre-trained on extensive linguistic data, has been widely used and has shown exceptional performance across diverse NLP tasks. Given the high degree of conventionality of idioms (Calzolari et al., 2002), there is a natural expectation for BERT to be good at handling them. Indeed, Tan and Jiang (2021) have validated the model's ability to distinguish between the literal and idiomatic usage of potential idiomatic expressions. Nedumpozhimana and Kelleher (2021) have shown that BERT incorporates information from idioms and their surrounding context to process them. Tian et al. (2023) have demonstrated that BERT-like language models represent idioms differently from their literal counterparts at both sentence and word levels, with words in idioms receiving less attention than words in non-idiomatic contexts. Clearly, BERT has a strong ability at handling idioms. However, one question remains: **is BERT sensitive to the degree of idiomaticity of idioms?**

Our hypotheses are that:

1. CamemBERT should be better at predicting tokens within idioms as opposed to simple lexemes, because tokens within idioms are more strongly constrained.
2. Tokens within idioms with higher idiomaticity should be more likely to be accurately predicted compared to tokens within idioms with lower idiomaticity.

As far as we know, there has been limited research into this question. The closest research was

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 14–23. May 25, 2024.

by Garcia et al. (2021b), who conducted a series of probing tasks to examine whether and to what extent vector space models, including BERT, can appropriately represent idiomaticity in noun compounds (NCs) in English and Portuguese. However, their results do not address the following questions: Does BERT distinguish different degrees of idiomaticity in NCs and other types of idioms? What kinds of tokens within an idiom are more predictable? Does the length of an idiom influence BERT's ability to predict tokens within it?

In this paper, we try to answer these questions by focusing on semantic idiomaticity in French idioms. We took our data from the French Lexical Network (LN-fr), a handcrafted lexical resource containing 3,127 idioms, 22,551 simple lexemes, and 47,395 contextual sentences for these entries. Our experiment used CamemBERT-base (Martin et al., 2020), a pre-trained BERT-derived model for French, in a demasking task on both simple lexemes and tokens embedded within idioms from our dataset.

We compared the prediction results of simple lexemes and tokens within idioms to observe performance differences under different conditions, thereby inferring the model's representation of different level of idiomaticity. Moreover, we analyzed the effect of token part of speech (POS) and idiom length on performance.

2. Related work

In recent years, attention has been focused on detecting and representing idiomaticity. Handling a MWE within a context requires first recognizing its non-compositional nature and then accurately conveying its idiomatic meaning in this context. Currently, the primary approach involves generating embeddings for components of the MWE and then merging them using diverse composition functions to construct a comprehensive representation of the MWE. Ultimately, the idiomaticity can be evaluated by computing the cosine similarity between the merged vector and the vector representing the expression (Cordeiro et al., 2019).

To represent idiomatic meaning in MWEs, recent approaches typically utilize contextualized language models. Among these models, Shwartz and Dagan (2019) found that BERT outperforms other contextualized models implemented in classifiers for creating embeddings in tasks related to lexical composition. However, Nandakumar et al. (2019) and Garcia et al. (2021a,b) indicated that pre-trained contextual models cannot effectively encode idiomaticity in MWEs. In comparison, static models like word2vec perform better (King and Cook, 2018; Nandakumar et al., 2018, 2019; Cordeiro et al., 2019; Sarlak et al., 2023). Nevertheless, supervised approaches leveraging contextualized models tend to outshine in tasks specific to certain languages and types of MWEs with ample resources, as these models offer representations that encode linguistic features and contextual cues (Fakharian and Cook, 2021).

Idiomaticity has also become a topic of recent NLP conference tasks. For instance, SemEval-2022 task 2 (Tayyar Madabushi et al., 2022) focuses on idiomaticity detection and sentence embedding containing multilingual MWEs. Results of these tasks show that the models got better performance with available training data. Although the best-performing methods are based on deep neural models independent of the linguistic features of MWEs, mixed approaches are generally believed to be worth exploring. Additionally, the PARSEME shared task on automatic identification of verbal MWEs (Ramisch et al., 2020), particularly with the *Seen*2020 system (Pasquer et al., 2020), underscores the significance of incorporating linguistic features in MWE-related tasks as well.

In our study, we focused on evaluating language models' sensitivity to idiomaticity. For this, we observed the contextualised model CamemBERT's performance in a classic fill-mask task with simple and idiomatic tokens in French.

3. Experiment

3.1. Data

We extracted our data form LN-fr v3 (Polguère, 2009; Lux-Pogodalla and Polguère, 2011; Polguère, 2014; ATILF, 2023), released in October 2023. It is an extensive, openly accessible lexical resource constructed manually following the methodological principles of explanatory combinatorial lexicology (ECL), the lexicological branch of Meaning-Text Theory (MTT) (Mel'čuk and Polguère, 1987; Mel'čuk et al., 1995; Apresjan, 2000). Every entry in LN-fr is a disambiguated lexical unit, i.e., either a simple lexeme or an idiom with a specific meaning, and each idiom is classified as a weak idiom, a semi-idiom or a strong idiom (see §1). Since our study follows MTT's definition and classification of idioms, and because LN-fr contains explicit information about the idiomaticity level of idioms, it suited our purpose very well.

Each lexical unit has a POS tag, and that of an idiom is determined by its internal syntactic head rather than its function within a sentence (Mel'čuk, 2006). For instance, *bien sûr* ('of course', lit. 'well sure'), because its head *sûr* is an adjective, is described as an adjectival idiom despite functioning as an adverb in sentences. There are a total of 11 POS tags for idioms in our dataset (see Table 2).[1]

[1] Interjective idioms are expressions that function as

Lexical unit	Idiomaticity	POS	Examples
pomme	simple lexeme	N	À la fin du repas, on a parfois droit à un petit morceau de brie et, en guise de dessert, selon la saison, des pommes, des noix, quelques fraises écrasées avec du sucre qu'on étale sur une tartine.
pomme de terre	weak idiom	N Prep N	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une pomme de terre, du fromage blanc. Pierre avait peine à soulever des sacs de pommes de terre de 40 kg, quant à moi je fis un véritable travail de garçon de ferme.

Table 1: Sample data from LN-fr

Idiom type	Example	Count
Nominal	*coup de soleil* lit. 'blow of sun' 'sunburn'	1579
Prepositional	*à propos* lit. 'at purpose' 'by the way'	730
Verbal	*faire la tête* lit. 'make the head' 'sulk'	619
Conjunctive	*quand même* lit. 'when even' 'anyway'	93
Adjectival	*bien sûr* lit. 'well sure' 'of course'	42
Phrasal	*Un ange passe.* lit. 'An angel passes.' 'awkward silence'	27
Adverbial	*pas mal* lit. 'not bad' 'quite good'	23
Propositional	*qui se respecte* lit. 'who respects oneself' 'self-respecting'	5
Numeral	*un à un* lit. 'one to one' 'one by one'	5
Pronominal	*ici et là* lit. 'here and there' 'here and there'	2
Interjective	*Tonnerre de Dieu!* lit. 'thunder of God' 'Good heavens!'	2
Total		**3127**

Table 2: Idiom types in the dataset

The POS of the tokens that are embedded within an idiom is not annotated directly in LN-fr, but one can retrieve it from the idiom's syntactic pattern, which is a string representing a sequence of POS tags. For example, *pomme de terre* ('potato', lit. 'apple of ground'), has the pattern `N Prep N`, so we know that the first and last tokens are nouns and the second is a preposition. We extracted from these patterns the POS tags for most of the embedded tokens. As some idioms did not have a syntactic pattern, we were not able to automatically retrieve the POS for their embedded tokens, which represent about 3.8% of all the tokens in our dataset; these tokens were not included in our second analysis (§4.2).

Each lexical unit has one or more lexicographic examples taken from corpora. These examples have been meticulously selected by lexicographers to reflect the authentic usage of a lexical unit. They aim to showcase various constructions that are possible for the lexical unit, to illustrate its usage and its syntactic and semantic selection (Lux-Pogodalla, 2014). Moreover, the annotation explicitly gives the position, within each sentence, of the tokens that belong to the lexical unit at hand. Note that a lexical unit may appear more than once in the same example; we counted those separately (which is why we have more tokens than examples even for simple lexemes in Table 3). We had in our dataset a total of 47,395 such sentences, with an average of 1.5 examples per idiom and 2 per simple lexeme, each sentence having around 38 tokens on average.

Finally, we counted the length in tokens of each lexical unit. For simple lexemes the length is 1; for idioms, we segmented by spaces and punctuations.

In total, we extracted from LN-fr 25,678 lexical units: 3,127 idioms and 22,551 simple lexemes. Table 3 breaks down these numbers. Compared to the NCs dataset used by Garcia et al. (2021b) covering 9,220 naturalistic and neutral sentences for 280 NCs in English and 180 NCs in Portuguese, our dataset encompasses a broader spectrum of idioms and a larger quantity of contexts.

Our dataset is available at `https://github.com/liliulng/idiomaticity-dataset`.

3.2. Methodology

Our experiment consists in taking the sentences associated with a lexical unit in LN-fr and masking, one at a time in the case of idioms, the tokens that correspond to that lexical unit. We then submit these sentences to CamemBERT for demasking. The model predicts the masked token and provides

independent sentences, like interjections such as *Wow!*

Type	Lexical units	Examples	Tokens
Simple lexeme	22551	42849	45563
Idiom	3127	4546	13529
Weak idiom	592	916	2425
Semi-idiom	589	899	2408
Strong idiom	1946	2731	8696
Total	**25678**	**47395**	**59092**

Table 3: Quantitative overview of our dataset

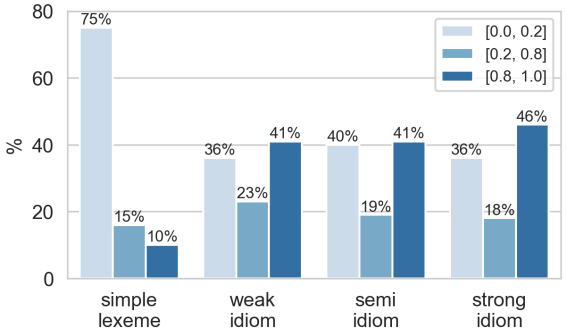

Figure 1: Score distribution

a list of candidates, each with a softmax score reflecting the model's confidence in it being the missing token. We record the confidence score returned by the model for the correct answer (the masked token) and note whether the correct answer was ranked as the first candidate (R1). This is illustrated in Table 4. The R1 candidate is the model's best guess and should be viewed as its "answer". Its score tends to be close to 1 (indeed, the model is optimized for this), but sometimes it can be lower, which reflects the model's confidence in its answer (or lack thereof). We want to take this into account, so if the masked token is guessed at rank 1, we note its score, and we will refer to it as "score@R1" in the rest of this paper.

We did not fine-tune the model because we aimed to evaluate the model's ability to learn idioms without being explicitly trained for it. We used the model as-is with its default parameters.

CamemBERT, as a contextualised model, provides predictions of a masked token based on its context. In our case, the contexts are the sentences retrieved from LN-fr that illustrate the usage of simple lexemes and idioms. Because we mask each token within idioms one by one, the other tokens inside a given idiom are visible and are part of the context. Nedumpozhimana and Kelleher (2021) suggested that BERT's ability to understand an idiom primarily relies on the idiom itself, so context inside idioms is crucial for CamemBERT to predict masked idiomatic tokens.

We utilized the model's tokenizer to segment the tokens, guaranteeing that our tokenization was consistent with the model's vocabulary. In cases where a token was segmented into subtokens, such as the token *tigers* being tokenized into *_tiger* and *s*, we conducted the masking experiment for each subtoken and calculated the product of all subtokens' confidence scores as the confidence score for that token. Furthermore, if the model correctly predicted each subtoken, we marked the whole token as correctly predicted as well.

We analysed the distribution of confidence scores of tokens, scores at rank 1 (scores@R1) and the percentage of correct predictions for masked tokens belonging to simple lexemes and idioms with different idiomaticity degrees, in order to determine how much the model's prediction is related to masked token's contextual idiomaticity degree. We further conducted statistical tests to validate the conclusions drawn from our observations.

4. Results and Discussion

In this section, we explore the impact of idiomaticity, POS, and idiom length on the model's performance. We examine the confidence scores, scores@R1, and the probability of achieving correct predictions token (expressed as a percentage of R1). When analyzing the scores and scores@R1, we take into account the median and mean for tokens across various categories. These are represented, respectively, by a thick line and a triangle in our figures. When there is a notable difference between them, our focus will be on the median.

4.1. Does CamemBERT distinguish different levels of idiomaticity?

Figure 1 shows that 75% of non-idiomatic tokens score below 0.2, with only 10% achieving a high score above 0.8. Conversely, over 40% of idiomatic tokens are predicted with scores exceeding 0.8, highlighting the model's significant challenge in predicting non-idiomatic tokens. Regarding idiomatic tokens, the model's confidence scores for correct answers often fall into polarized categories of high or low scores. However, discerning between varying levels of idiomaticity remains difficult, as indicated by similar score distributions across the three types of idioms.

The Kruskal-Wallis test proved the significant difference between the confidence score distribution for tokens corresponding to simple lexemes and that of tokens belonging to idioms ($p < 0.01$, $\eta^2 = 0.15$). There is no significant difference between scores for tokens in the three types of idioms ($p < 0.01$, but with negligible effect size $\eta^2 < 0.01$).

When comparing the mean and median confidence scores (Figure 2), we further notice a

Lexical unit	Token	POS	Sentence	Score	R1
pomme	pommes	N	À la fin du repas, on a parfois droit à un petit morceau de brie et, en guise de dessert, selon la saison, des **<mask>**, des noix, quelques fraises écrasées avec du sucre qu'on étale sur une tartine.	0.10	F
pomme de terre	pomme	N	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une **<mask>** de terre, du fromage blanc.	0.99	T
	de	Prep	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une pomme **<mask>** terre, du fromage blanc.	0.99	T
	terre	N	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une pomme de **<mask>**, du fromage blanc.	0.99	T

Table 4: Sample fill-mask inputs and results

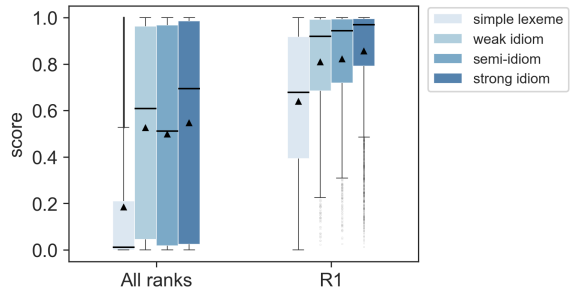

Figure 2: Score given to the masked token at all ranks and at R1

	All	Content	Function
Simple lexemes	25	24	50
Weak idioms	62	55	86
Semi-idioms	58	48	83
Strong idioms	62	49	81

Table 5: Percentage of correctly predicted tokens for content and function tokens

significant difference between idiomatic and non-idiomatic tokens. Idiomatic tokens consistently exhibit higher median and mean scores, typically around 0.5 or above. Still, there is no substantial distinction among the three classes of idioms, as tokens within each category demonstrate fairly similar median and mean scores. However, it is worth noting that score@R1, which represents the model's overall confidence in its predictions, tends to correlate positively with the degree of idiomaticity, which aligns with our previous hypothesis. Additionally, tokens within strong idioms consistently receive the highest median and mean scores, compared to other idiomatic tokens.

R1 predictions: The model correctly guesses the masked token around 60% of the time for tokens within idioms, compared to only 25% for simple lexemes. There is no significant difference between the three types of idioms: 62% for weak idioms, 58% for semi-idioms and 62% for strong idioms. This reveals again the model's higher capacity in predicting tokens within idioms than simple lexemes.

Statistical analysis: We calculated the Spearman's ρ correlation to unveil the dependence of the model's prediction results (confidence scores and scores@R1) on tokens' idiomaticity levels.

Between the free versus idiomatic nature of masked tokens and their prediction results, there is a moderately positive correlation that confirms the model's capability to distinguish tokens on these two general levels, with $p < 0.01$, Spearman's $\rho = 0.36$ for scores and $p < 0.01$, Spearman's $\rho = 0.39$ for score@R1. Specifically for all the four levels of idiomaticity (simple lexeme, weak idiom, semi-idiom, strong idiom), this moderately positive correlation still exists between idiomaticity levels and the prediction results (with $p < 0.01$, Spearman's $\rho = 0.36$ for confidence scores and $p < 0.01$, Spearman's $\rho = 0.38$ for score@R1). As observed in Figure 2, no significant correlation is found between the scores and the three subtypes of idioms ($p = 0.04$, $\rho = 0.02$).

This indicates again that, in general, the model is unable to differentiate between varying levels of idiomaticity within idioms, although it effectively distinguishes between free and idiomatic tokens. A chi-squared test between the idiomaticity levels and correct prediction aligns with this conclusion: $p < 0.01$ and a moderate effect size Cramér's $V = 0.3$ for all idiomaticity levels and the generally free and idiomatic levels, but $p < 0.01$, Cramér's $V = 0.03$ between the three types of idioms.

4.2. What kinds of tokens are more predictable within idioms?

We aimed to pinpoint which kinds of tokens present greater predictive challenge and to understand how this might contribute to the observations above. To accomplish this, we broke down our data by the POS of both free and idiomatic tokens. This data was readily available in LN-fr, which distinguishes

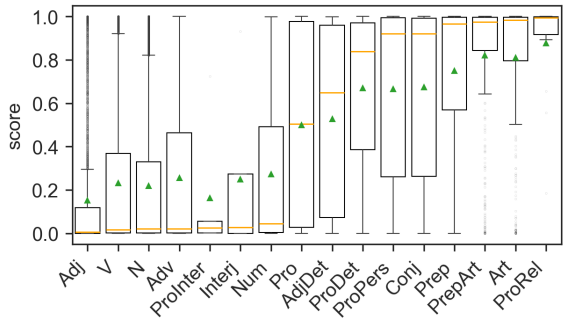

Figure 3: Score by token POS

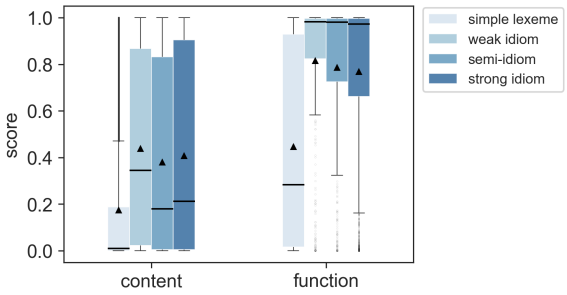

Figure 4: Scores for content and function words

a total of 16 POS tags (distinct from the 11 for idioms listed in Table 2) that can be divided into two categories: content and function tokens. Content tokens represent 94% of the tokens in our dataset and include nouns (N), verbs (V), adjectives (Adj), adverbs (Adv), numerals (Num), interrogative pronouns (ProInter) and interjections (Interj). Function tokens represent the other 6% and include pronouns (Pro), prepositions (Prep), articles (Art), preposition-article amalgams (PrepArt), conjunctions (Conj), personal pronouns (ProPer), pronominal determiners (ProDet), adjectival determiners (AdjDet) and relative pronouns (ProRel). Three of these categories had very low counts, namely Interj (4 occurrences), ProInter (14) and ProRel (5), so the scores reported here for those categories are to be taken with a grain of salt (this explains why the mean is outside of the box for ProInter).

As Figure 3 shows, the median and mean scores for all function tokens are notably higher than those for content tokens, exceeding 0.5. Conversely, the median and mean confidence scores for content tokens are low, with mean scores below 0.3 and median scores below 0.1. This suggests that overall, disregarding idiomaticity, the model excels in predicting function tokens. The score@R1 exhibits the same trend, hence we omit the graph here.

R1 predictions: 82% of function tokens were correctly predicted, against only 28% of content tokens.

Statistical analysis: Spearman's ρ test demonstrated a moderately positive correlation between predictions and type of POS (content or function token): with $p < 0.01$, $\rho = 0.31$ for confidence scores and $p < 0.01$, $\rho = 0.39$ for score@R1. The chi-squared test also detected a certain level of dependence between the correct prediction of tokens and their POS status ($p < 0.01$, Cramér's $V = 0.3$)

These results are not surprising, because function words belong to closed classes, thus there are far fewer options for the model to choose from. However, given the model's adeptness at managing function tokens, we wondered if this could explain its better performance on idioms. Indeed, there is a stark contrast between the distribution of content and function tokens in simple lexemes versus idioms: function tokens comprise only 0.5% of the simple lexemes, while they account for 28.6% of the tokens within idioms. This is because idioms are phrases, so they often contain function words, especially in French, where compounds are much less common than in some other languages such as English or Chinese. Hence, could this imbalance account for the elevated median and mean scores observed for tokens within idioms reported in Figure 2?

We analyzed separately the confidence scores of content and function tokens with varying degrees of idiomaticity. As shown in Figure 4, regarding content tokens, the median and mean scores of idiomatic tokens generally fall below 0.5 but still remain significantly higher than those for simple lexemes. Similarly, there is no substantial disparity in scores among tokens in different types of idioms for content tokens. As for function tokens, those within idioms receive higher confidence scores overall, with mean scores surpassing 0.7 and median scores nearing 1. The variance among different types of idioms is minimal. Conversely, scores for simple function tokens are notably lower than those for idiomatic function tokens, below 0.5. Thus, regardless of the degree of idiomaticity, the model's prediction of function tokens consistently outperforms that of content tokens. As for content tokens, the model's prediction of tokens within idioms surpasses that of simple lexemes, and its prediction ability for tokens within idioms with varying degrees of idiomaticity remains stable. This corresponds to our previous conclusion in the first analysis (see §4.1).

R1 predictions: The percentages of correct predictions for content tokens across various levels of idiomaticity further support our findings (see Table 5). Specifically, more than 50% of the content tokens within idioms were correctly predicted, compared to only 24% for simple content tokens. In addition, while roughly half of simple function tokens were correctly predicted, this figure exceeded 80% for idiomatic function tokens.

Statistical analysis: We conducted the same statistical analysis for prediction results across idiomaticity levels for content and function tokens separately. Spearman's ρ correlation between idiomaticity levels and confidence scores or score@R1 always yielded $p < 0.01$ but with no significant ρ values. The chi-squared test showed only modest dependence between correct prediction and idiomaticity level (either considering all four levels or only free versus idiomatic), for both content and function tokens: $p < 0.01$, Cramér's $V = 0.2$. There is no clear dependence between prediction results and the three idiomaticity levels across idiom subtypes ($p < 0.01$, Cramér's $V = 0.05$). Thus, the moderate correlation between idiomaticity levels and correct prediction observed in the first analysis no longer exists when we separate content and function tokens. This suggests that the variation in prediction performance of the model between free and idiomatic tokens may actually be at least partly due to the differing proportions of content and function words in these tokens.

No specific POS within content or function tokens appears to significantly influence the model's performance. The primary types of content words include nouns, verbs, and adjectives. In both simple lexemes and idioms, nouns comprise most of the words, accounting for approximately 61% in simple lexemes, 75% in weak idioms, 78% in semi-idioms, and 66% in strong idioms. Verbs represent a similar portion in simple lexemes (21%) and strong idioms (16%), while they only make up 4% and 6% in weak idioms and semi-idioms. There is no significant difference in the proportion of adjectives across simple lexemes and idioms, ranging from approximately 12% to 18%. Confidence scores for nouns, verbs, and adjectives do not show significant differences. As for function tokens, pronouns (59%), conjunctions (24%), and personal pronouns (10%) are the primary function token types in simple lexemes, while prepositions constitute the main portion of the function tokens in idioms, comprising 74% in weak idioms, 78% in semi-idioms, and 60% in strong idioms. Additionally, preposition-articles are the second major type, accounting for 17%, 14%, and 13% respectively in the aforementioned subtypes of idioms. Notably, the proportion of articles in strong idioms is higher at 15% compared to weak and semi-idioms (2% and 4%).

To sum up, function words tend to be accurately predicted by the model in all types of expressions regardless of the level of idiomaticity, because they belong to closed classes with a small number of members. In free context, their predictability arises from governing syntactic relations and sentence coherence. Meanwhile, within idioms, they contribute to idiomaticity by maintaining the structural integrity and idiomatic meaning of the expression.

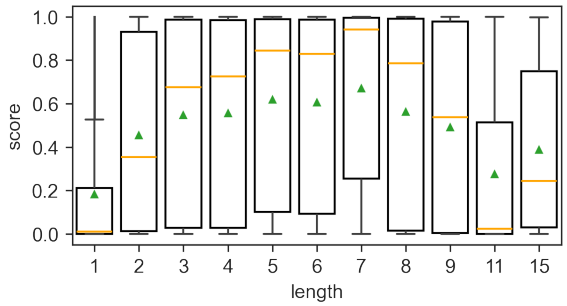

Figure 5: Scores by lexical unit length

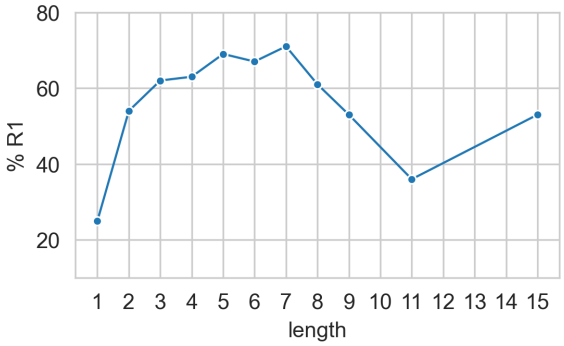

Figure 6: Percentage of R1 by lexical unit length

4.3. Is CamemBERT sensitive to the length of idioms?

When a token in an idiom is masked, CamemBERT utilizes contextual information to predict the masked one, and that context includes the remaining tokens in the idiom. Therefore, the more tokens an idiom contains, the more context it provides. Consequently, does CamemBERT achieve better prediction results for tokens within longer idioms? In our dataset, 99% of idioms comprise 7 tokens or fewer, whereas longer idioms amount to only 171 occurrences, representing only 1% of the idioms. Most idioms, specifically 91% of weak idioms, 87% of semi-idioms, and 59% of strong idioms, consist of 2 or 3 tokens. Additionally, a small proportion (8% of semi-idioms and 20% of strong idioms) extend to 4 tokens, while another 10% of strong idioms span 5 tokens. No statistically significant relation is found between the level of idiomaticity and the length of idioms.

We compared the score and score@R1 for tokens in lexical units of varying lengths. Here again, the results for score@R1 are not different, so we only present the results for confidence scores in Figure 5. They suggest that as the length of lexical units increases, both the mean and median confidence scores tend to rise (we disregard the drop for lengths over 7 tokens, which we attribute to the scarcity of data in that range).

R1 predictions: Similarly, as shown in Figure 6, when the length of idioms is 7 tokens or fewer, there is a generally increasing trend between idiom length and the percentage of correct predictions.

Statistical analysis: With $p < 0.01$, the Spearman's ρ coefficient between lexical unit length and scores is 0.36, while it is 0.4 for score@R1, suggesting a moderate positive correlation. Similarly, correct prediction displays a moderate positive association with idiom length in the chi-squared test ($p < 0.01$, Cramér's $V = 0.32$). These findings suggest that the length of idioms significantly impacts CamemBERT's prediction of idiomatic tokens. The model evidently demonstrates sensitivity to the length of idioms when interpreting tokens within them.

Due to the small proportion (1%) of idioms with lengths exceeding 7 tokens, and despite their proportion of correct predictions not aligning with the general trend, their impact has been disregarded in our analysis.

5. Conclusion

We aimed to assess CamemBERT's ability to capture varying degrees of idiomaticity within idioms. We measured this by comparing the model's off-the-shelf performance on fill-mask tasks with tokens pertaining either to simple lexemes or idioms, further distinguishing three levels of idiomaticity among idioms: weak idioms, semi-idioms and strong idioms. We collected 59,092 tokens with illustrative examples from LN-fr, including 45,563 simple lexemes and 13,529 idiomatic tokens from more than 3,000 idioms.

In §1, we posited two hypotheses:

1. CamemBERT should be better at predicting tokens within idioms as opposed to simple lexemes.
2. Tokens within idioms with higher idiomaticity should be more likely to be accurately predicted.

Our main observations are:

1. The model is significantly better at predicting tokens that belong to an idiom as opposed to simple lexemes.
2. It is not sensitive to varying levels of idiomaticity among subtypes of idioms.
3. It exhibits a heightened performance in predicting function words, regardless of idiomaticity.
4. There is a positive correlation between idiom length and performance.

These observations validate our first hypothesis (see §1), but invalidate the second.

Our findings corroborate those of Garcia et al. (2021b), who showed that vector space models, including BERT, cannot capture the semantic overlap between idiomatic NCs and one or none of their components. Furthering their research, we additionally considered weak idioms, which have a semantic overlap with all of their components, as well as a broader range of idioms, not only NCs.

Our analysis of the effects of POS and the length of idioms suggest that these factors may at least partially explain the model's heightened proficiency at predicting tokens within idioms compared to tokens corresponding to simple lexemes. Nonetheless, this does not explain why CamemBERT is not sensitive to varying levels of idiomaticity among idioms. The very notion of idiomaticity is ambiguous, and the distinction between various types of idiomaticity is often overlooked and tends to be conflated into semantic aspects, i.e., non-compositionality. In our study, we explored both lexical and semantic idiomaticity. Lexical idiomaticity implies that idiomatic tokens exhibit stronger constraints on lexical selection compared to free tokens, i.e., they cannot be replaced by their synonyms while preserving their idiomatic meaning and grammatical correctness. On the other hand, the varying degrees of idiomaticity are indicative of their semantic idiomaticity, which denotes the contribution of internal components to their overall semantic meaning. So CamemBERT's performance in our experiment suggests that in fact the model is more sensitive to lexical idiomaticity than semantic idiomaticity.

This raises questions about other aspects of idiomaticity. Indeed, idioms exhibit idiomaticity on multiple levels simultaneously: lexical, semantic, syntactic, morphological, etc. For instance, *faire la tête* ('sulk', lit. 'make the head') is a strong idiom in French that exhibits not only lexical and semantic idiomaticity, but also prohibits syntactic operations like passivisation, dislocation, etc., as well as morphological inflection to tokens other than the head *faire*. While there is no theoretical consensus on the classification of idiomaticity, our experience may offer valuable insights to address the matter.

In future research, we would like to refine our experiment, extend it to other types of MWEs and explore other forms of idiomaticity. Moreover, we intend to carry out further analyses on language model representations of idiomaticity, exploring additional potential influencing factors such as idiom frequency, or extending our investigation to more complex tasks. We also aim to replicate our experiments with different language models and available datasets in other languages.

6. Acknowledgements

We express our gratitude to the anonymous reviewers for their valuable and constructive feed-

back. We would like to thank ATILF for the original dataset and our colleagues at OLST for engaging in helpful discussions. Li Liu acknowledges the financial support of the China Scholarship Council (#202008310177).

7. Bibliographical References

Juri Apresjan. 2000. *Systematic Lexicography*. Oxford University Press, Oxford.

ATILF. 2023. Réseau lexical du français (rl-fr). ORTOLANG (Open Resources and TOols for LANGuage)—www.ortolang.fr.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing*, 2nd edition, pages 267–292. CRC Press, Boca Raton, FL, USA.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, volume 2, pages 1934–1940, Las Palmas, Spain. ELRA.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL'19: Human Language Technologies*, volume 1, pages 4171–4186, Minneapolis, MN, USA. ACL.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. ACL.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, volume 1, pages 2730–2741, Online. ACL.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. ACL.

Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 345–350, Melbourne, Australia. ACL.

Veronika Lux-Pogodalla. 2014. Integrating lexicographic examples in a lexical network (intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014*, volume 2, pages 586–591, Marseille, France. ATALA.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In *First International Workshop on Lexical Resources (WoLeR 2011)*, pages 54–61, Ljubljana, Slovenia.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219, Online. ACL.

Igor A. Mel'čuk, André P Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot.

Igor A. Mel'čuk and Alain Polguère. 1987. A formal lexicon in meaning-text theory (or how to do lexica with words). *Computational linguistics*, 13:261–275.

Igor A. Mel'čuk. 2006. Parties du discours et locutions. *Bulletin de la Société de Linguistique de Paris*, 101:29–65.

Igor A. Mel'čuk. 2023. *General phraseology: Theory and practice*. John Benjamins.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models

capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA 2018)*, pages 71–76, Dunedin, New Zealand. ALTA.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. ACL.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3333–3345, Barcelona, Spain (Online). ICCL.

Alain Polguère. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43:41–55.

Alain Polguère. 2014. From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4):396–418.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX)*, pages 107–118, online. ACL.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference (CICLing)*, pages 1–15, Mexico. Springer.

Mahtab Sarlak, Yalda Yarandi, and Mehrnoush Shamsfard. 2023. Predicting compositionality of verbal multiword expressions in Persian. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 14–23, Dubrovnik, Croatia. ACL.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Minghuan Tan and Jing Jiang. 2021. 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)*, pages 1397–1407, Online. INCOMA.

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

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 174–179, Toronto, Canada. ACL.