Probing, Generalization and Application of Metaphorical Knowledge in Pre-trained Language Models

Abstract

Human languages are full of metaphorical expressions. Metaphors help people understand the world by connecting new concepts and domains to more familiar ones. Large pre-trained language models (PLMs) are therefore assumed to encode metaphorical knowledge useful for NLP systems when processing language. In this paper, we investigate this hypothesis for PLMs by probing the metaphorical knowledge in their encodings, by measuring the cross-lingual and cross-dataset generalization of this knowledge, and by analyzing the application of this knowledge when generating metaphorical expressions. We present studies in multiple metaphoricity detection datasets and four languages (i.e., English, Spanish, Russian, and Farsi). Our extensive experiments suggest that contextual representations in PLMs do encode metaphorical information, and mostly in their middle layers, and the knowledge is transferrable between languages and datasets in most cases. Finally, we show that PLMs face more challenges in generating metaphors, especially as their novelty increases. Our findings give helpful insights for both cognitive and NLP scientists.

1 Introduction

Pre-trained language models (PLMs) (Peters et al., 2018; Devlin et al., 2019), are now used in almost all NLP applications, e.g., machine translation (Li et al., 2021), question answering (Zhang et al., 2020), dialogue systems (Ni et al., 2021) and sentiment analysis (Minaee et al., 2020). They are the foundation models of NLP systems (Bommasani et al., 2021), causing huge impact in research and industry.

Metaphors are important aspects of human languages. In conceptual metaphor (CM) theory (Lakoff and Johnson, 2008), metaphor is mapping a concept in one domain (target) to a concept in another domain (source). Modeling metaphors is essential in building human-like computational systems that can relate new concepts to the old and familiar ones. The creativity and problem solving (i.e., generalization to new problems) also depend on the analogies and metaphors a cognitive system relies on.

We intuitively guess that PLMs must encode some information about metaphors due to their great performance in language processing tasks. Confirming that experimentally is a question that we try to focus on in this paper. So far, there has been no comprehensive analysis of how PLMs represent metaphorical information. The recent works in metaphor detection using PLMs, e.g., Choi et al. (2021), are related, but those are focused on achieving the best final performance.

We follow three paths in our study of metaphors in PLMs. First, we apply probing methods to understand the distribution of encoded metaphorical information, and mostly in their middle layers, and the knowledge is transferrable between languages and datasets in most cases. Finally, we show that PLMs face more challenges in generating metaphors, especially as their novelty increases. Our findings give helpful insights for both cognitive and NLP scientists.

Second, to evaluate the metaphorical knowledge in their transferability and generalization, we design zero-shot cross-lingual and cross-dataset experiments. Four languages and four datasets are considered in this evaluation.

Finally, we explore how PLMs apply their metaphorical knowledge to generate or score a metaphorical expression. We take the novelty of metaphors and observe the differences in the fill-in-the-blank generation with that respect.

The LCC dataset (Mohler et al., 2016) is our main resource, containing annotations in four languages: English, Russian, Spanish and Farsi. We also experiment with three other English metaphor datasets, TroFi (Birke and Sarkar, 2006), VUA pos, and VUA verbs (Steen, 2010).
To summarize, we aim to answer questions such as (i) How do different layers of Transformers contribute to distinguishing literal from metaphorical usages of words? (ii) Do different PLMs encode metaphorical knowledge differently? (iii) Is the encoded information about metaphoricity transferrable across languages and datasets? (iv) How do PLMs generation of a word in a context depend on its metaphoricity?

Our findings and contributions could be summarized as: (i) we show that metaphors are encoded differently across the layers of PLMs, but the trend is similar in three popular PLMs: BERT, RoBERTa, and ELECTRA. This is the first probing study of metaphors in PLMs. (ii) We run generalization analysis for PLMs in their out-of-distribution transferability of metaphorical knowledge. Our experiments confirm this generalization in most cases of cross-lingual and cross-dataset setups. (iii) We show that generating metaphors, especially more novel ones, is challenging for PLMs, confirming that PLMs behave similarly to humans in this respect.

2 Related Work

Metaphor detection using PLMs. The metaphoricity detection task (Mason, 2004; Birke and Sarkar, 2007; Shutova et al., 2013) is a good fit for most of our studies, even though we focus on analyzing the metaphorical knowledge and not on achieving the best detection results. Using PLMs for metaphoricity detection is normal in recent years, resulting in state-of-the-art results, indicating implicitly that PLMs do represent metaphoricity information. Choi et al. (2021) introduces a new architecture that integrates metaphor theories with BERT. Similarly, Song et al. (2021) presents a new perspective on metaphor detection task by framing it as relation classification, focusing on the verbs. These approaches beat other earlier works of using PLMs (Su et al., 2020; Chen et al., 2020; Gong et al., 2020), RNN-based (Wu et al., 2018; Mao et al., 2019) and feature-based approaches (Turney et al., 2011; Shutova et al., 2016).

Tsvetkov et al. (2014) present cross-lingual metaphor detection models using linguistic features and word embeddings. They make use of bilingual dictionaries to map between languages. The datasets they employ are quite small (1000 training and 200 testing examples), making them unsuitable for a statistically robust evaluation.

Probing methods in NLP. Probing is an analytical tool used for assessing word representations linguistic knowledge. In probing, the information richness of the representations is determined by the quality of a supervised model in predicting linguistic properties solely based on the representations (Kohn, 2015; Gupta et al., 2015; Yaghoobzadeh and Schütze, 2016; Conneau et al., 2018; Tenney et al., 2019b,a; Hewitt and Manning, 2019).

A popular probing method introduced by Tenney et al. (2019b) is edge probing (Figure 2). They propose a suite of span-level tasks, including Part-of-speech tagging and coreference resolution. They demonstrate that BERT understands core NLP linguistic knowledge better than its uncontextualized counterparts and indicate the approximate position where each linguistic knowledge is encoded in BERT layers (Tenney et al., 2019a).

Despite the widespread use of edge probing and
other conventional probes, the question of whether the probing classifier is learning the task itself or it is actually identifying the linguistic knowledge in the representations raises concerns, see Belinkov (2021) for more discussions. An Information theoretic view can solve this issue (Voita and Titov, 2020) by reformulating probing as a data transmission problem. They consider the effort needed to extract linguistic knowledge as well as the final quality of the probe, showing that this approach is more informative and robust than normal probing methods. We employ both edge probing and MDL probing in this work.

**Probing multi-lingual PLMs.** The application of probing methods in NLP is extended to multi-lingual PLMs as well (Pires et al., 2019; Eichler et al., 2019; Ravishankar et al., 2019a; Choenni and Shutova, 2020). Choenni and Shutova (2020) introduce probing tasks for typological features of multiple languages in multi-lingual PLMs. Ravishankar et al. (2019a,b) extend the probing tasks of Conneau et al. (2018), to few other languages. Pires et al. (2019) study the generalization of multilingual-BERT across languages when performing cross-lingual downstream tasks. Here, as part of our study, we evaluate the generalization of multi-lingual PLMs in the representation of metaphoricity. We employ this setting as one of the main ways to understand the quality of encoded metaphoricity information in PLMs.

**Out-of-distribution generalization.** There has been no earlier work on studying or evaluating the out-of-distribution generalization in metaphor detection systems. Out-of-distribution generalization refers to scenarios where testing data comes from a different distribution from training data (Duchi and Namkoong, 2018; Hendrycks et al., 2020a,b). Here, we have scenarios where testing data is in different languages or in different domain / dataset. These are challenging evaluation scenarios for the generalization of encoded information (metaphoricity in our case).

### 3 Metaphorical knowledge in PLMs

#### 3.1 Probing

Based on conceptual metaphor theory, one domain (e.g., ARGUMENT) is explained using another domain (e.g., WAR). For example, in “We won the argument”, ARGUMENT is linked to WAR, and the word “won” is used as its “non-literal” or “metaphorical” meaning. The same word “won” in a sentence like “China won the war” refers to its “literal” meaning. Here, our goal is to distinguish the metaphorical and literal words in given contexts. In other words, we want to detect if, given a context

<table>
<thead>
<tr>
<th>VUA Verbs</th>
<th>He [finds]$_1$ it hard to communicate with people, not least his separated parents. → 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>VUA POS</td>
<td>They picked up power from a [spider]$_1$’s web of unsightly overhead wires. → 1</td>
</tr>
<tr>
<td>TroFi</td>
<td>“Locals [absorbed]$_1$ a lot of losses,” said Mr. Sandor of Drexel → nonliteral</td>
</tr>
<tr>
<td>LCC</td>
<td>Lawful [gun ownership]$_2$ is not a [disease]$_1$. → 3.0</td>
</tr>
</tbody>
</table>

**Table 1:** Examples of sentences, spans, and target labels for each probing dataset.
and a token span, the span refers to a metaphor or not.

As these examples show, metaphoricity is a highly contextualized phenomenon. The clues in the context might give us some hints. Here, our datasets take a sentence as the context. This might not be enough for many cases where larger contextual information is needed, but it is a reasonable simplification of the task to obtain large datasets.

Contextualized representations obtained by PLMs are shown to encode the contextual meanings of words relatively well (Devlin et al., 2019; Zhao et al., 2020). One word might have one or more metaphorical meanings and, likewise, one or more literal meanings. Our task is to distinguish between only two classes: metaphorical and literal.

We aim to answer a general question about metaphors in PLMs: do PLMs understand metaphors? We do not attempt to achieve the best metaphoricity detection results but to analyze and study how PLMs represent the necessary information to perform this task. In trying to answer this question, we apply probing methods, discussed as follows, to focus on the representation itself and not the fine-tuning task learning.

Methods We employ edge probing (Tenney et al., 2019b) and MDL (Voita and Titov, 2020). Edge probing consists of a classifier in which word representations obtained from PLMs are fed to it as inputs\(^1\). The quality of the classifier illustrates how well the representations encode a specific linguistic knowledge. This method is designed for span-level tasks, i.e., the classifier can only access the representations of a limited part of the input sentence specified in the dataset. The Edge Probing has two pooler sections for making fixed-sized vectors; one pools representations across the words in the span and the other pools representations across the layers.

The MDL probe is based on information theory and combines the quality of the classifier and the amount of effort needed to achieve this quality. Voita and Titov (2020) propose two methods for computing MDL: "variational coding" and "online coding." The former computes the complexity of the classifier with a Bayesian model. In the latter, the classifier is trained gradually on different portions of the dataset and the code length will be the sum of the cross-entropies, each for a data portion.

Voita and Titov (2020) show that the two methods’ results are consistent with each other. Accordingly, we opted for the "online coding" method since it is more straightforward in implementation. Since the code length is related to the size of the dataset, we report the compression, which is equal to 1 for a random classifier and more than 1 for better models. See extra details in (Voita and Titov, 2020)

3.2 Generalization

To see if PLM representations encode metaphoricity well, we evaluate them in settings where testing data comes from a different distribution from training. We explore transferability analysis across both languages and datasets. From the perspective of generality, these explorations also show how well PLMs generalize metaphorical information and how well a classifier can detect them across distributions. These important and helpful experiments are understudied in metaphor detection research. We adopt the term “domain” to refer to both “languages” and “datasets”.

3.2.1 Cross-lingual

Multi-lingual encoders project the representations in multiple languages into a shared space so that semantically similar words and sentences across languages end up close to each other. To answer questions about the transferability of metaphor information across languages, if we use a multi-lingual PLM model for representations, and our classifier shows that representations in language S are informative about metaphoricity, what happens if we apply this classifier to the representations in language T? If the representations are rich in both languages, information is accessible similarly across them, and metaphoricity can be transferred, then the classifier would be able to predict metaphoricity in language T from what it leans in S.

3.2.2 Cross-dataset

The great performance of PLMs is repeatedly related to their success in learning the existing heuristics in the training datasets, rather than the actual tasks (McCoy et al., 2019). By testing on a test set drawn from a different dataset compared to the training set, we can better measure the generalization of the knowledge encoded in PLMs. Therefore, another generalization dimension we consider is cross-dataset transfer, i.e., when training on dataset S and testing on dataset T. S and T are annotated in different groups with possibly different goals.
Table 2: Edge probing f1 results for various metaphoricity datasets in BERT, RoBERTa, and ELECTRA. The edge probing results are the average of three runs. The compression result is the best across layers, and the subscript denotes the best layer.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline F1</th>
<th>Baseline Comp.</th>
<th>BERT F1</th>
<th>BERT Comp.</th>
<th>RoBERTa F1</th>
<th>RoBERTa Comp.</th>
<th>ELECTRA F1</th>
<th>ELECTRA Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC (en)</td>
<td>75.78</td>
<td>1.08_8</td>
<td>88.24</td>
<td>1.93_6</td>
<td>88.03</td>
<td>2.02_6</td>
<td>89.49</td>
<td>2.11_6</td>
</tr>
<tr>
<td>TroFi</td>
<td>65.79</td>
<td>1.05_9</td>
<td>67.21</td>
<td>1.12_3</td>
<td>67.87</td>
<td>1.16_3</td>
<td>68.79</td>
<td>2.46_5</td>
</tr>
<tr>
<td>VUA POS</td>
<td>48.11</td>
<td>1.79_6</td>
<td>67.58</td>
<td>2.40_5</td>
<td>68.14</td>
<td>2.39_4</td>
<td>69.29</td>
<td>1.57_6</td>
</tr>
<tr>
<td>VUA Verbs</td>
<td>52.33</td>
<td>1.24_5</td>
<td>70.54</td>
<td>1.50_4</td>
<td>69.29</td>
<td>1.57_6</td>
<td>72.14</td>
<td>1.55_5</td>
</tr>
</tbody>
</table>

Table 3: Statistics of the datasets. Percentage of metaphors (%M) and number of instances for train / test sets are given for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>%M</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC (en)</td>
<td>45.2 / 46.8</td>
<td>39,769 / 4,465</td>
</tr>
<tr>
<td>LCC (fa)</td>
<td>43.7 / 44.8</td>
<td>18,496 / 2,070</td>
</tr>
<tr>
<td>LCC (es)</td>
<td>34.6 / 34.4</td>
<td>29,002 / 2,209</td>
</tr>
<tr>
<td>LCC (ru)</td>
<td>45.1 / 34.4</td>
<td>17,492 / 1,932</td>
</tr>
<tr>
<td>TroFi</td>
<td>42.6 / 42.3</td>
<td>5,770 / 666</td>
</tr>
<tr>
<td>VUA Verbs</td>
<td>27.8 / 30.0</td>
<td>17,240 / 5,873</td>
</tr>
<tr>
<td>VUA POS</td>
<td>15.2 / 17.9</td>
<td>72,611 / 2,2196</td>
</tr>
</tbody>
</table>

4 Experimental setup and results

4.1 Datasets and setup

Datasets We use three metaphor detection datasets in our study. One of these datasets, i.e., LCC, contains annotations in four languages: English, Russian, Spanish, and Farsi. The other two, TroFi and VUA, are in English only. The statistics of the datasets are shown in Table 3. Some example sentences with the metaphor annotation can be seen in Table 1.

The annotations of LCC (Mohler et al., 2016) are done according to Conceptual Metaphor theory, with source and target domains, in four languages. TroFi (Birke and Sarkar, 2006) TroFi dataset consists of metaphoric and literal usages of 51 English verbs from the Wall Street Journal. VU Amsterdam (Steen, 2010) corpus consists of words in the academic, fiction, and news subdomains of the British National Corpus (BNC), annotated with five annotators as figurative (specifying metaphor/personification/other) or literal.

Setup In implementing the edge probe, following Tenney et al. (2019b), we use 32 as the batch size, learning rate of 5e-5, and 256 projection di-
For the MDL probe, the same structure of edge probing is employed. We apply a logarithm to the base in cross-entropy loss to have all the obtained code lengths in bits. See extra details in Voita and Titov (2020)

4.2 Probing
We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) to represent our PLMs. Due to resource limitations, we conduct our experiments formed on the base version of the models (12 layers, 768 hidden size, 110M parameters) implemented in HuggingFace’s Transformers library (Wolf et al., 2020). We employ edge probing for evaluating overall metaphorical knowledge in our selected PLMs. However, as discussed by Belinkov (2021) and Fayyaz et al. (2021), edge probing is not reliable for layer-wise cross-model experiments, and so we leverage the MDL probe to this end.

Table 2 shows the edge probing F1 and MDL probing compression results for our three PLMs. Accordingly, RoBERTa and ELECTRA are shown to encode metaphorical knowledge better than BERT on both metrics. This can be attributed to the two models’ better performance on various tasks, acquired by having better pre-training objectives and enjoying more extensive pre-training data.

MDL probing compression across layers is demonstrated in Figure 3. Except for RoBERTa results in TroFi and VUA Verbs, we see the numbers increase at the first 3 to 6 layers, depending on the dataset, but it decreases afterwards. In other words, metaphorical information is more concentrated in the middle layers. The representations in the middle layers are relatively contextualized but not as much as higher layers. This indicates that detecting metaphoricity of a span is a contextualized task but it usually needs only a few layers of contextualization.

4.3 Generalization
As our PLMs, we use XLM-RoBERTa (Conneau et al., 2020) for cross-lingual and BERT for cross-dataset experiments. We apply the edge probing architecture as in the probing experiments. As we mentioned in Section 3.2, we sometimes refer to both language and dataset as domain for simplicity.

For each case of a source domain S and a target domain T, we run two experiments: one with the PLM and one with a randomized version of the PLM where weights are set to random values. Randomly initialized PLM that did not experience the pre-training process is a commonly used baseline in the community. So, the difference between the two gives evidence about the helpfulness of the encoded knowledge in PLMs for metaphoricity detection. When S = T, this effect is measured for in-domain generalization and when S ≠ T, for out-of-domain generalization. Comparing results of in-domain (e.g., training and testing on English data) and out-of-domain (e.g., training on Spanish and testing on English) setups demonstrates how generalizable the metaphoricity knowledge is across domains.

4.3.1 Cross-lingual
The LCC datasets in four languages are used here. We train on one source (e.g., English) and test on a target (e.g., Spanish) language. We subsample from the datasets to have the same number of examples in each dataset (17,492 which is the size of Russian dataset).

The results of our cross-lingual experiments are shown in Table 4. The random baseline is acquired using a randomly initialized XLM-RoBERTa. We observe that in all cases, the pre-trained PLM
Table 4: Cross-lingual metaphoricity detection F1 results. The multi-lingual PLM, i.e., XLM-RoBERTa, is better than random by large margins in most cases. For each test language, we bold its in-domain (e.g., en → en), and underline the best out-of-domain (e.g., ru → en) numbers.

Table 5: Cross dataset edge probing F1 results on BERT shown in pairs: pre-trained model and, in the parenthesis, the randomly initialized model. We set the training size to the minimum among datasets, i.e., TroFi. For each test dataset, we bold its in-domain (e.g., TroFi → TroFi), and underline the best out-of-domain (e.g., LCC(en) → TroFi) numbers.

outperforms the random. This shows that some metaphorical knowledge, learned during the pre-training phase, is transferable across languages.

Further, some language pairs (English & Russian and Russian & Farsi) seem to have higher transferrability than others. In some cases, interestingly, one way is much better than the other (e.g., Farsi → English is 78.1 but English → Farsi is 71.4).

Finally, we can also compare the in-domain random PLM with the out-of-domain pre-trained PLM results. For English as test, the best out-of-domain result, i.e., 80.9 for Russian → English, is better than the in-domain random result of 72.3.

4.3.2 Cross-dataset

Similar to the cross-lingual evaluations, here we have four datasets as sources and targets. We set the size of all datasets to the minimum, i.e., 5770. For each pair, we run two experiments: one with random and one with XLM-RoBERTa as our PLM. In total, here we run $4 \times 4 = 16$ experiments. Results are shown in Table 5.

As expected, VUA Verbs and VUA POS achieve the best results when mutually tested since they are from fairly the same distribution. Similar distributions seem to be impactful when testing on the VUA Verbs dataset as well, where we can see that training on the TroFi dataset outperforms LCC. This can be attributed to the fact that TroFi only has verbs as metaphorical spans, just like VUA Verbs, whereas LCC has both verbs and nouns. This observation can be validated where we test on the VUA POS dataset. In this setting, the opposite happens, and the LCC dataset outperforms TroFi, which can be related to having both nouns and verbs like VUA POS rather than just verbs.

4.4 Application

LCC dataset provides metaphoricity scores including 0 as no metaphoricity, 2 likely/conventional metaphor, and 3 clear metaphor. We leverage these scores to study PLM’s understanding of the commonality of the metaphors and their resemblance to human annotators as the novelty changes.

Table 6 shows the average predicted probability for the masked spans in different metaphoricity classes across four languages. Our results demonstrate that, as the metaphoricity score increases, the model’s prediction probability for the desired word decreases consistently over languages.

---

2 It is possible/weak metaphor and as Mohler et al. (2016) describes metaphors with $0.5 \leq \text{score} < 1.5$ as unclear we ignore this score.
Figure 4: XLM-RoBERTa MLM prediction recall in LCC dataset over four languages. Recall is the proportion of the entries predicted within the ranking limit to all the entries. Scores are from the LCC dataset including no metaphoricity, likely/conventional metaphor, and clear metaphor.

Table 6: XLM-RoBERTa MLM average probability of the source concept in the LCC datasets over four languages. Metaphoricity scores are from the LCC dataset, and 0: no metaphoricity, 1: possible/weak metaphor, 2: likely/conventional metaphor, and 3: clear metaphor. The probabilities are acquired after applying softmax.

<table>
<thead>
<tr>
<th>Language</th>
<th>Metaphoricity Score</th>
<th>0</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td></td>
<td>.20</td>
<td>.17</td>
<td>.08</td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td>.27</td>
<td>.19</td>
<td>.12</td>
</tr>
<tr>
<td>Farsi</td>
<td></td>
<td>.28</td>
<td>.15</td>
<td>.13</td>
</tr>
<tr>
<td>Russian</td>
<td></td>
<td>.37</td>
<td>.27</td>
<td>.15</td>
</tr>
</tbody>
</table>

We show that the trend and conclusion are consistent in Farsi, Spanish, and Russian as well.

5 Conclusion

In this paper, we shed light on how metaphorical knowledge is encoded in PLMs, through probing, cross-dataset and cross-language analysis and generation. We ran novel scenarios on metaphor detection and generation and presented findings helpful for both NLP and cognitive sciences.

We showed that metaphorical knowledge in PLMs is somewhat generalizable across languages. This could be an interesting direction to follow for further investigations within both linguistics and NLP. Our evaluation also demonstrated that PLMs generate metaphors and even novel ones but with more hardship than literal expressions. This is an indication that they might face issues understanding them as well, and more work is needed to equip them with better metaphorical information. Metaphors are important in human cognition, and if we seek to build cognitively inspired or plausible language understand systems, we need to work more in their integration in the future.
References


A Appendices
Table 7: Examples of sentences, spans, and annotations for LCC dataset in Farsi, Spanish, and Russian.

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
<th>Score</th>
<th>Src Concept</th>
<th>Target Concept</th>
<th>Polarity</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa</td>
<td>اما امریکا در افغانستان از همان اگز. با سلاح [، دموکراسی [، آمده است.</td>
<td>3.0</td>
<td>WAR(3.0)</td>
<td>DEMOCRACY</td>
<td>NEUTRAL</td>
<td>1.0</td>
</tr>
<tr>
<td>es</td>
<td>[atorado]₁ en la [deuda]₂ pública y sin avances en Estado de Derecho</td>
<td>3.0</td>
<td>BARRIER(3.0)</td>
<td>DEBT</td>
<td>NEGATIVE</td>
<td>2.0</td>
</tr>
<tr>
<td>ru</td>
<td>Мировые [деньги]₁ [мечутся]₁ , не зная , куда вложить .</td>
<td>3.0</td>
<td>MOVEMENT(3.0)</td>
<td>MONEY</td>
<td>NEGATIVE</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Table 8: XLM-RoBERTa MLM ranking average of the source concept in LCC dataset over four languages. Scores are from the LCC dataset, 0 indicating no metaphoricity, 1 possible/weak metaphor, 2 likely/conventional metaphor, and 3 clear metaphor.