Do LLM's Know the Prerequisites of Figurative Language?

Sivan Almogy **Avi Kabra** **William Palmer**

Abstract

Figurative language is very common in English, allowing speakers to characterize referents using points of comparison, for example with metaphors and similes. Some prior research, including Liu et al. (2022), has shown that large language models do not desirably predict sentences with felicitous uses of figurative speech to be more probable than infelicitous uses. Several factors may be responsible for this, including failure to compositionally interpret figurative language constructions or having a lack of world knowledge to apply the qualities of one referent to another. In this paper, we test the latter of these two possibilities as a potential bottleneck and find that lack of requisite knowledge may be a significant hindering factor in the ability of large language models to correctly interpret metaphorical language.

1 Introduction

The ability of humans to understand and interpret figurative language is a hallmark of human cognition, yet it remains a persistent challenge for large language models (LLMs) (Tong et al., 2024; Li et al., 2024). Figurative language consists of constructions which characterize referents using a point of comparison, known as the *vehicle* of the construction. Canonical examples of figurative language include metaphors and similes. These constructions, as cognitive-linguistic phenomena, rely on both the ability to linguistically parse the construction as well as the crucial world knowledge of how certain referents are exemplars for certain attributes. Although humans intuitively grasp these relationships, language models must overcome significant barriers in representation and reasoning to achieve a similar level of understanding. This study focuses on a foundational question in this domain: Do LLMs possess the prerequisite content knowledge of attribute-vehicle relationships independent of figurative language structures?

Understanding metaphors involves two primary components. Firstly, it is necessary to possess the content knowledge regarding the relevant attribute-vehicle relationships, for example understanding that *steel* is associated with characteristics like strength and durability. Secondly, it is also necessary to be able to apply this knowledge to a given figurative construction, for example interpreting the metaphor *her resolve is steel*, by applying the contextually key attributes of *steel* to the referent *her resolve*.

Previous research by Liu et al. (2022) investigates this ability in LLMs by creating Fig-QA, a Winograd-style task to test the precision of metaphor interpretation. Given a starting sentence that uses a metaphorical phrase, models must predict which of two given possible endings is likelier to follow, with one stating a correct interpretation of the metaphor and the other contradicting it. Liu et al. (2022) found that without fine-tuning on the dataset, the models perform above the chance level baseline but well below human accuracy. While they note the the need for a variety of knowledge to interpret the metaphors and provide annotations for the category of needed knowledge for each metaphor, they do not directly test whether missing knowledge is a primary cause of the poor performance.

Building on these findings, this paper seeks to address this dimension of figurative language processing, focusing on the knowledge necessary to do well on the dataset of Liu et al. (2022). If LLM's are shown to know the necessary world knowledge, but do poorly on the dataset of Liu et al. (2022), then that would imply LLM's are unable to compositionally derive the intended meanings of figurative language. On the other hand if LLM's are shown to not have the requisite knowledge for the metaphor interpretation, this would call into question whether the results of Liu et al. (2022) actually do imply that LLMs struggle with the composi-

tional and non-literal aspects of the figurative language interpretation task and would suggest that additional experimentation is required to measure these capabilities.[1]

2 Linguistic Background

Metaphors and other forms of figurative language are extremely prevalent in spoken and written English, functioning as mechanisms for conceptualizing one domain in terms of another. These constructions enable humans to structure abstract ideas through concrete associations (Lakoff and Johnson, 1980). A metaphor operates by mapping attributes from a *source* onto a *target*, a process that relies on understanding the way in which vehicles exemplify certain attributes.

According to Papafragou (1996), instances of figurative language are not "mere linguistic devices", but rather correspond to mental figures grounded in cognition. Papafragou (1996) emphasizes the importance of conventionalism in determining the meaning of metaphor, with this meaning encoded in the relational structure between the vehicle and the referent being characterized by the metaphor, also known as the *tenor*. It is this relational nature that introduces a layer of ambiguity: the specific aspect of the relationship being highlighted can remain open to interpretation. This is to say that the precise way a vehicle exemplifies an attribute, whether positively or negatively, is not derivable in a compositional way from the lexical items themselves.

The vehicle serves as the figurative anchor of a metaphor, embodying specific qualities that bridge the conceptual gap between the source and target domains. For example, describing someone as having a *heart of gold* involves mapping the vehicle *gold* onto a person's character. Here, *gold* introduces attributes like purity, value, and rarity, all of which influence the intended meaning. Such relationships depend heavily on cultural knowledge–interpreting *coal* as *dark* or *winter* as *cold* assumes shared conceptual associations grounded in human experience.

Cultural and conceptual knowledge is vital for understanding metaphors because these comparisons often emerge from shared experiences and collective perception. Deeply embedded in linguistic communities, these shared associations make metaphor comprehension second-nature for humans. For machines, however, the lack of real-world experience or embodied cognition poses a formidable challenge. Without direct access to cultural grounding, LLMs must infer these relationships solely from textual data, which may lack the context or depth required to build such associations reliably.

Attribute-vehicle relationships further complicate metaphor processing because their relevance is context-dependent and often polysemous. A single vehicle can convey multiple attributes depending on usage. *iron* can symbolize strength as in *iron will* or rigidity as in *iron rule*, depending on the desired contextual mapping. Recognizing these variations requires information about both lexical semantics and pragmatic context, a task at which LLMs traditionally falter. Unlike humans, who resolve ambiguity through experience and intuition, machines rely on statistical patterns, which may inadequately capture the nuances of these associations.

The challenges faced in this arena stem from a reliance on data-driven learning without embodied knowledge. Though models are excellent at predicting the statistics of language, they often struggle to infer the deeper semantic relationships that underpin metaphorical meaning. This limitation highlights a fundamental gap between human and machine cognition: where humans draw on cultural context, conceptual grounding, and lived experience, LLMs must approximate these processes through patterns in text. Consequently, evaluating whether LLMs possess the content knowledge necessary for recognizing attribute-vehicle relationships independent of metaphorical structure is a critical step in understanding their capabilities and limitations.

2.1 Literature Review

Liu et al. (2022) explores whether models could identify and map the relevant attributes of vehicles to target concepts in figurative expressions. With a controlled experimental paradigm, Liu et al. (2022) presents LLMs with sentences containing figurative constructions and compares the probability of continuation implicated by the construction with the probability of a continuation which is not implicated by the construction, conditioning on the occurrence of the figurative construction. The authors found that while LLMs performed ad-

[1] All the code and data for this project can be found at https://github.com/avikabra/ling_final_project.

equately on frequent, well-documented metaphors, they struggled with novel or less conventional mappings. These findings suggested that while LLMs may learn some figurative relationships found in their training datasets, they often fail to generalize them effectively across new contexts. Importantly, Liu et al. (2022). left open the question of whether LLMs possess the foundational knowledge of these relationships in isolation, which this study seeks to address.

Other studies have similarly highlighted challenges in metaphor comprehension for machines. Veale and Hao (2008) examined metaphor generation and recognition in computational models, emphasizing the importance of structured knowledge bases for capturing conceptual associations. Their *Metaphor Magnet* system utilized lexical databases like WordNet to identify plausible mappings between vehicles and targets. However, their approach relied heavily on pre-structured knowledge, limiting its generalizability to novel metaphors. Tsvetkov et al. (2014) similarly explored metaphor detection through cross-lingual model transfer, leveraging multilingual data to train computational models capable of identifying metaphorical expressions across languages. The study did reveal limitations; while cross-linguistic methods improved metaphor identification, performance declined when encountering metaphorical expressions that required cultural grounding. Tsvetkov's findings underscore the ongoing challenge of equipping computational systems with the flexible, knowledge-driven reasoning.

On the semantic side, Shutova et al. (2013) introduced a data-driven angle to metaphor processing using corpus-based methods. By leveraging statistical models trained on large datasets, the authors aimed to identify metaphorical mappings based on word co-occurrences. While their method demonstrated improved performance over rule-based systems, it struggled with interpreting metaphors requiring deeper conceptual knowledge or cultural context. Bizzoni and Lappin (2017) took a parallel approach, focusing on neural models' ability to process metaphorical meaning particularly within the framework of distributional semantics. Their findings showed that while LLMs excel at identifying surface-level similarities between words, they often fail to encode the richer, non-literal associations necessary for metaphor interpretation. This limitation highlights a gap in how LLMs represent abstract conceptual relationships. Huguet Cabot et al. (2020) extended this line of research by analyzing LLMs' capacity to integrate pragmatic context when interpreting metaphors. Their study emphasized the role of real-world grounding in metaphor comprehension, arguing that LLMs lack the embodied knowledge humans use to disambiguate figurative expressions.

Collectively, these studies highlight significant progress and persistent limitations in computational metaphor processing. While previous research has explored metaphor application, far less attention has been given to evaluating whether LLMs possess the underlying content knowledge of attribute-vehicle relationships. This gap is critical: without verifying that LLMs can recognize these relationships in isolation, it remains unclear whether failures in metaphor comprehension stem from knowledge deficits or from difficulties in applying existing knowledge within figurative structures.

3 Methods

3.1 Overview

We introduce an adaptation of the dataset and methods employed by Liu et al. (2022), who employed sentence pairs containing metaphors to test whether models could correctly identify the conceptual mappings required for metaphor interpretation.

Our approach utilizes some of the same models as Liu et al. (2022), but introduces a critical modification to isolate content knowledge from its figurative application. Specifically, we adapt the sentences from their development split[2] by rearranging the sentences for our new task format. In the original paradigm, models were tasked with interpreting a metaphorical sentence (e.g. *The burrito has the heat of an iceberg* $\rightarrow$ *The burrito is cold* rather than *The burrito is hot*). They did this my comparing the autoregressive probability of the second sentence occurring, conditioning on the metaphor. In our adaptation, the sentences are restructured to test whether the model recognizes the attribute-vehicle relationship independently. For instance, the task above becomes determining whether *an iceberg is cold* or *an iceberg is hot* is more likely/correct. This adaptation

[2]We chose to adapt the development set since this was the only set labeled with both the correct ending index and the category of knowledge required to interpret each simile. Liu et al. (2022) also recommends testing on their development set.

removes the figurative structure from the evaluation, allowing us to determine whether the LLMs' difficulty in interpreting the similes of Liu et al. (2022) can be attributed in part to the models lacking the requisite knowledge.

3.2 Dataset Adaptation

To generate our new sentences, we first find the most frequent structures for the starting phrases (containing the similes) in the dataset of Liu et al. (2022). We automated the detection of the vehicle noun phrases within sentences matching these structures by using spaCy's DependencyMatcher (Honnibal et al., 2020). For each starting sentence and corresponding pair of ending sentences that had the same subject, we plugged in the vehicle noun phrase in place of the original subject for each of the endings. Examples for three of these templates can be found in Table 1.

As a result of the simple subject replacement method, some of the modified endings did not have noun-subject agreement. Since this was a small portion of the sentences and incorrect agreement might lower the assessed likelihood of a semantically correct but ungrammatical ending, we decided to remove all entries with a noun-verb mismatch. Overall, after removing sentences that did not match any of the templates or did not have noun-verb agreement, we were able to produce 728 sentence pairs from the 1094 entries in the original development dataset. We call this task the object identification task since a model must assign a higher probability to the continuation with the correct attribute, i.e. the one which the subject exemplifies.

3.3 Models and Experimental Procedure

To evaluate LLMs' knowledge of attribute-vehicle relationships, we model our procedure after that of Liu et al. (2022) so that our results are as comparable as possible to theirs. To that effect we also include GPT-2 in our experiment for the sake of comparison. Liu et al. (2022) also tested GPT-3, however since it is no longer possible to access the probability distributions GPT-3 produces, we instead test Llama 3.2 with 1 billion parameters, another modern language model.

For our primary knowledge evaluation task, we test our modified pairs with the same framework Liu et al. (2022) used to evaluate auto-regressive LLMs (both GPT-2 and Llama are auto-regressive). In particular, the model is given the initial shared portion of the phrase—both items of the pair have the same subject—-and the auto-regressive probability of each continuation was compared. In practice, the probability of each entire sentence was measured, but this will maintain he same ordering as the relative likelihood since the subjects are identical. If the probability was higher for the subject with its exemplifying predicate compared to with the contradictory predicate, as determined by the labels in Liu et al. (2022), then the model was labeled as correct, i.e. having the necessary knowledge for that pair.

The sum of log probabilities used to compare the likelihoods of the paired sentences is sensitive to the number of tokens in the predicate, since this is equivalent to the number of log probabilities being summed. As a result, we decided to retest Llama with the average log probability (same as taking geometric mean of the product of probabilities) in order to normalize any bias towards some sequences based on length, but there was no substantial difference in performance, and in fact a slightly lower accuracy (61.8%).

Another concern we had about the likelihood measurement was that it could be greatly affected by factors unrelated to determining the more semantically accurate sentence, such as the overall grammaticality or the respective probabilities of the unmodified endings (most of which are kept within their modified counterparts). To try to get a less biased measure of the models' knowledge, we used a subject identification evaluation, in which a predicate was paired with each associated metaphor vehicle, rather than pairing each vehicle with its two predicates. Then, the same likelihood evaluation was used, but in this case, all of the evaluated tokens were the same, the only difference being the two different objects in the preambles. A model "predicts" that a given predicate is associated with the object that produces the higher conditional probability for the predicate.

4 Results

Having obtained the binary results on the datasets for both models, we now turn to analysis. Figure 1 visualizes the mean accuracies for the experiment described in the previous section, which we will now break down. For convenience, we also provide Table 2, which contains the values represented by Figure 1 for the primary knowledge evaluation task for GPT-2 and Llama 3.2. We give the mean accuracy for the whole dataset and the accuracy

Original Preamble	Original Endings	Adapted Sentences
The music was loud like a siren. (SUBJECT) . . . LIKE (VEHICLE)	The music was very loud. The music was very quiet.	A siren was very loud. A siren was very quiet.
He was as dangerous as a viper. SUBJECT [BE] (QUALITY) as VEHICLE	He was very dangerous. He was not dangerous.	A viper was very dangerous. A viper was not dangerous.
The razor had the sharpness of a sword. (SUBJECT) [HAVE] (QUALITY) of (VEHICLE)	The razor was sharp The music was very quiet.	The razor was dull. A siren was very quiet.

Table 1: Representative examples of sentences in the original dataset fitting three of our templates with their generated adapted sentences. The first column shows the original unadapted starting sentences, beneath which we represent the matching template. The second column provides the original pair of possible continuations. The third column gives the pair of adapted sentences, with the vehicle replacing the original subject. Each test example consists of one such pair.

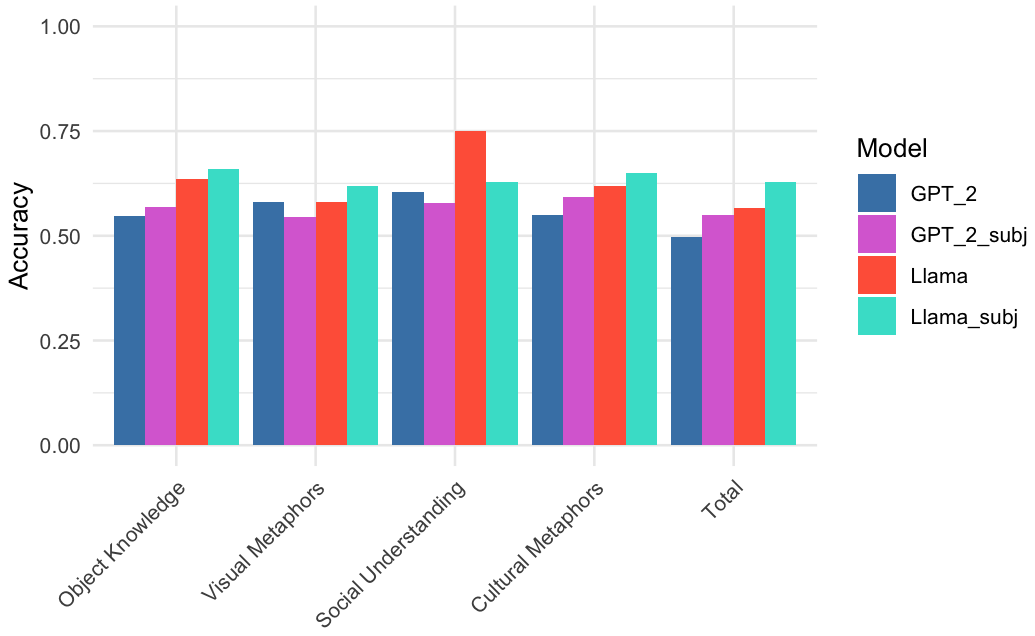

Figure 1: Bar graph of results for visual comparison. Each bar represents the mean accuracy for a particular model on a particular test set for a given category of metaphor. *GPT_2* and *Llama* represent the performance of the models on the object identification dataset. *GPT_2_subj* and *Llama_subj* represent the performance of the models on the subject identification dataset. Chance baseline is drawn at 0.50.

Categories	GPT-2	Llama
Object Knowledge	0.548	0.635
Visual Metaphors	0.604	0.750
Social Understanding	0.496	0.566
Cultural Metaphors	0.581	0.581
Total	0.549	0.619

Table 2: Knowledge evaluation accuracies for GPT-2 and Llama for each type of metaphor; baseline is at 0.50.

Categories	GPT-2	Llama
Object Knowledge	0.546	0.619
Visual Metaphors	0.593	0.651
Social Understanding	0.568	0.659
Cultural Metaphors	0.579	0.629
Total	0.55	0.629

Table 3: Subject identification accuracies for GPT-2 and Llama 3.2 for each type of metaphor and overall; Chance baseline is at 0.50.

within four subcategories which Liu et al. (2022) assigned based on the knowledge required to interpret each metaphor. These categories are: *object knowledge*–which includes physical properties of common objects and animals, *visual metaphors*, *social understanding*–such as expected human behaviors and emotions, and *cultural metaphors*.

With this paired predicate comparison for each vehicle, GPT-2 performs above chance level, but not by much at 54.9% accuracy on the whole data set. Llama does better than GPT-2, with a total mean accuracy of 61.9%. For both models we see that visual metaphors have the highest accuracy reaching as high as 75% for Llama.

Table 3 contains the results for the object identification task for GPT-2 and Llama. We find very similar results to the object identification task with mean accuracies of 55% and 62.9% for GPT-2 and Llama respectively.

5 Discussion

The accuracies achieved by the LLMs on our adapted dataset are similar to those of comparable models tested on Fig-QA. In particular, Liu et al. (2022) reported that GPT-2 without fine-tuning had a 53.93% accuracy–slightly below its 54.9% accuracy on the knowledge evaluation task. Additionally, Liu et al. (2022) also reported a 56.89% accuracy for GPT-neo–which was designed to replicate GPT-3's architecture–with 1.3B parameters on Fig-QA, which is only a bit lower than the 61.9% accuracy attained by Llama 3.2 with 1B parameters on the knowledge task.

Although these similar outcomes are notable, we cannot directly compare the results due to the loss of some of the development set data during the data adaption process (removed ∼350 out of 1100 pairs) as well as the uncertainty of whether similar overall accuracies indicate similar performance for the models on each entry in the original dataset and its counterpart in the adapted dataset. While it does seem likely that poor performance on the knowledge task would be highly associated with poor performance on the respective metaphor interpretation test, further testing is required to validate this assumption. However, the relatively low accuracies achieved by GPT-2 and Llama 3.2 on our knowledge tasks do indicate that a lack of necessary world knowledge is an underlying problem that the models face in interpreting Liu et al. (2022)'s metaphors. These results highlight that the results of Liu et al. (2022) do not necessarily point to a misunderstanding of figurative language in the steps of identifying metaphorical structures or applying referenced knowledge, since the limited accuracies can be attributed at least in part to a lack of the requisite knowledge.

Several challenges remain for future research. First, this study evaluates attribute-vehicle relationships in controlled, context-free tasks. However, metaphor processing in natural language is inherently influenced by cultural and linguistic nuances. Future research must explore whether LLMs can generalize these relationships across diverse languages, domains, or creative uses of language, such as poetry or idiomatic expressions. Second, the influence of model architecture on metaphor understanding requires further investigation. Exploring how architectures like multimodal transformers integrate semantic knowledge from non-textual sources (e.g., images) presents a promising direction.

Future research should also test the impact of augmenting models with explicit knowledge in their prompts. If this study reveals that LLMs lack attribute-vehicle knowledge, providing this information directly in the prompt may help assess whether failures arise from missing knowledge or an inability to retrieve and apply it. Additionally, fine-tuning models on knowledge-rich datasets or leveraging few-shot and zero-shot learn-

ing paradigms could enhance performance on metaphor-related tasks, particularly for novel and culturally specific associations.

Finally, opportunities exist for developing new benchmarks and tasks that evaluate the application of attribute-vehicle knowledge in complex, real-world contexts. These tasks could combine syntactic and pragmatic challenges, pushing models to approximate human-like reasoning more effectively. Addressing these broader limitations in natural language processing—including difficulties with non-literal language and embodied reasoning—will be essential for advancing LLMs' ability to handle metaphor comprehension and related cognitive-linguistic phenomena.

Limitations

There are also several limitations in the methods used to generate data and interpret results for this paper. In generating the adapted dataset, there was a trade-off between creating more general match conditions, that could catch more starting phrases and therefore generate more sentences, and making templates specific enough to prevent generalization errors. For example, entries were only adapted if the subject of the starting phrase matched the subject of the ending phrase, so that the ending phrase subject could be swapped with the metaphor vehicle object. However, an exception was introduced to catch additional cases where the ending subject was a pronoun. In these cases, there was no straightforward analytical method to determine if the pronoun in the ending sentence referred to the subject of the first, but it was assumed that given the structure of the metaphor interpretation task, this would likely be true in most cases (e.g. The bear was as hungry as a lion → It was starving). Of the 66 subject-ending pairs included by this exception, 21 (∼31.8%) appeared to have a mismatch between the starting subject and the object referenced by the ending pronoun. Although this is a relatively small portion of our generated data, there are likely to exist other such artifacts as a result of the automated adaption. Future studies would benefit from either a more extensive set of conditions for the data generation, or ideally, a manual conversion of the data.

Additionally, there appear to be many cases of errors or lack of clarity in the Fig-QA data, such as the following:

- *Typos*: "He is a dice missing a side." → "His is unlucky", "he has talant"
- *Confusion over subject / point of view*: "We're as good of friends as two dogs fighting over the last bone." → "They're not really good friends at all and are desperate."

These errors likely arose due to the data for Fig-QA being crowd-sourced, and their inclusion suggests that further cleaning of the original dataset should be performed, as it could harm both LLM performance on Fig-QA as well as on the adapted task by introducing noise that could affect the assessed likelihood of either ending in a pair.

Another limitation of our methods is the indirect comparison of the accuracy on Fig-QA with the accuracy on the adapted knowledge task. While the results of the models tested by Liu et al. (2022) on Fig-QA are not publicly accessible, it should be possible to reproduce these results for at least GPT-2 using the code they released. Comparing the result of GPT-2 on each Fig-QA entry with the result on each adapted knowledge evaluation entry could offer much more concrete evidence that the availability of knowledge plays a direct role in hindering the model's ability to interpret the Fig-QA metaphors.

References

Yuri Bizzoni and Shalom Lappin. 2017. Deep learning of binary and gradient judgements for semantic paraphrase. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The Journal of Philosophy*, 77(8):453–486.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Finding challenging metaphors that confuse pretrained language models.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language.

Anna Papafragou. 1996. Figurative language and the semantics-pragmatics distinction. *Language and Literature*, 5(3):179–193.

Ekaterina Shutova, Barry Devereux, and Anna Korhonen. 2013. Conceptual metaphor theory meets the data: A corpus-based human annotation study. *Language Resources and Evaluation*, 47.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Tony Veale and Yanfen Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors.