# Figurative Language and Mobilization to Action: a Multimethod Approach

MIT PRIMES Computer Science
Student: Raj Saha  Mentors: Dr. Ann Kronrod and Dr. Ivan Gordeliy

January 15, 2025

**Abstract**

This paper seeks to analyze the effect of figurative on mobilization to action and empowerment to act. The approach toward this hypothesis is multimodal; on the one hand, we will computationally analyze peer advice datasets, and on the other hand, we will experimentally discern the influence of non-literal language. The first computational approach necessitates an accurate figurative language detection tool. Following previous literature, we plan to leverage Large Language Models for this task; BERT has already been effective for preliminary tests on metaphor detection. After developing the computational tool to detect figurative language in text, we will analyze queries and replies in online peer advice communities such as health forums, and test the influence of a reply's figurative language on the adherence with the advice given. We will then augment this analysis with two controlled experiments. The first will test whether figurative language in advice and directives increases compliance due to trust, and the second will investigate what mental processes underlie the ratio of figurative language in giving advice and responding to it.

**Introduction**

Figurative language is undeniably a central part of spoken and written discourse, both concerning its ubiquity and its influence on achieving effective communication. An array of such communicative goals has been documented – ranging from efforts to be conventional to attempts at showing positive emotion or even being humorous (Roberts and Kreuz 1994). Thus, due to the frequency of figurative language, the task of its detection remains critical for an accurate understanding of discourse (Lai, Toral, and Nissim 2023). However, the very nature of figurative language makes this an arduous task. By definition, there is a difference between the form and meaning of text containing figurative language. In addition to this inherent disparity, figurative language can be further categorized into many types – including metaphor, sarcasm, irony, etc. – that each behaves uniquely, making a general-use tool more challenging to implement (Reyes, Rosso, and Buscaldi 2012). The majority of previous attempts to develop a computational method to detect figurative language focused on one of these aspects (e.g., Potamias, Siolas, and Stafylopatis 2019). The first goal of the current project is to test existing methods that detect figurative language as a whole, and to develop an improved detection tool.

The second goal of this paper is to examine the influence of figurative language on mobilization to action. Our hypothesis is grounded in past literature. First, figurative language is known to impact mental imagery (Carston 2018), and imagery increases thought concreteness. Further, according to Construal Level Theory, more concrete thought can decrease the perceived

psychological distance in both time and space, eliciting a sense of "here and now" (Carston 2018; Hansen and Wänke 2010). Finally, this closeness sensation encourages increased urgency and, thus, mobilization for action. Although each of these pairwise causal relationships has been tested before, we seek to connect these separate findings into one process, thereby linking figurative language with mobilization to action.

This connection could significantly affect the text present in a wide array of domains. For example, in marketing, figurative language in advertisements would compel consumers to buy the product more than literal language (Kronrod and Danziger 2013). Additionally, writers could use this theory to make their stories more engaging for the reader – resulting from the decreased psychological distance between the novels' content and the receiver. Even politicians and activists would benefit from more non-literal language, which may better mobilize supporters. The second purpose of the current work is therefore to experimentally test a conceptual model hypothesizing that figurative language is linked to action via imagery and thought concreteness.

**Past Approaches and Literature Review**

Most approaches to figurative language detection involve separately detecting the subtypes. For example, Rolandos-Alexandros et al. designed their classifier by employing individual deep learning techniques – BiLSTMs, AttentionLSTMs, and Dense layers – for sarcasm, irony, and metaphor (Potamias, Siolas, and Stafylopatis 2019). Furthermore, their data was collected from Twitter, isolating tweets with specific hashtags, including "#sarcasm" and "#humour." The authors reported an F1 score of 0.73 for an irony dataset and 0.87 for a sarcastic dataset.

To simplify the detection of figurative language, recent studies have undertaken the task of solely identifying metaphors. There are several reasons for this choice. First, metaphoric language accounts for a large quantity of figurative language and even language as a whole. Some research reveals that nearly one-third of language in a typical corpus will be metaphoric (Martin 2006). Like figurative language, the frequency of metaphors is matched by its influence on various NLP tasks. Sentiment analysis, machine translation, and language generation have all benefited from Computational Metaphor Processing (CMP), the direction of research towards improving machine understanding of metaphor (Li et al. 2023; Rai and Chakraverty 2021). Second, discrete procedures – notably, the Metaphor Identification Procedure (MIP) – have been established, allowing for a more formal computational approach (Pragglejaz Group 2007).

Just as with figurative language, however, the task of metaphor detection remains difficult. The root complexity stems from the MIP, which recommends that both the basic and the contextual meaning of words be identified. Obtaining and analyzing these meanings proved to be a strenuous manual task before the arrival of neural networks and word embeddings (Pragglejaz Group 2007). More recently, approaches leveraging models with the transformer architecture have been most successful and remain the gold standard for metaphor detection (Ptiček and Dobša 2023).

Central to the emergence of Neural Networks' use for metaphor detection is the concept of word embeddings. These multi-dimensional vector representations of words allow words of similar semantic or syntactic meaning to share similar vectors (Ptiček and Dobša 2023; Mikolov et al. 2013). The use of word embeddings - as opposed to arbitrary tagging of tokens - drastically improves the performance of NLP tasks in deep learning models. However, until the advent of Large Language Models, these embeddings were "static." In this way, the vectors were not dependent on context.

Metaphor detection attempts using static word embeddings produced mediocre results. Notably, Do Dinh and Gurevych published "Token-Level Metaphor Detection using Neural Networks," where they used a feedforward artificial neural network and word2vec word embeddings for identification. Having trained and evaluated the model on VUA Metaphor Corpus, they received F1 scores of about 0.6 (n.d.-a). In 2018, Swarnkar et al. similarly used static embeddings – this time GloVe, rather than word2vec – in combination with a neural network; their F1 scores on the VUA dataset were 0.57 (Swarnkar and Singh 2018).

On the other hand, previous approaches leveraging contextual word embeddings yielded more promising results. Although many studies have employed Large Language Models in recent years, perhaps the first such study was that of Gao et al. in 2018. Rather than using static embeddings like GloVe and word2vec, the authors opted to use ELMo word embeddings, which adapted to surrounding tokens. As they articulated in the results, the presence of these embeddings, in combination with their bidirectional LSTM network, greatly influenced the accuracy. Without the use of ELMo, the F1 score was 0.617 on the VUA dataset, and with the contextual word embeddings, the F1 score rose to 0.704 on the same dataset. This accuracy, at the time, surpassed the previously most successful approach of Wu et al., which used word2vec embeddings (Gao et al. 2018; Ptiček and Dobša 2023).

Research exhibiting a direct causative relationship with figurative language and mobilization to action is rare. Some studies, however, have presented a similar connection in specific domains. For instance, McMullen et al. indicated that the presence of figurative language correlated positively with the success of psychotherapy cases. These sessions require the client to identify and cope with their personal thoughts – a form of initiating an action (McMullen 1989). Moreover, Yang et al. used Brain Topography to reveal the direct effects of figurative language on the organ. Ultimately, strong activations were found in the right parahippocampal gyrus, which is related to spatial processing, and the left inferior frontal gyrus, known for playing a role in "executive function and social cognition" (Yang and Shu 2016).

## Methods

### *Training Dataset*

Following the success of past literature, we began the process of metaphor detection by seeking to leverage transformer-based models. Our initial choice of dataset largely depended upon availability, size, and effectiveness in earlier studies. The three regularly occurring such datasets were the TroFi Example Bank (TroFi), MOH-X, and the VUA Metaphor Corpus (VUA) (Ptiček and Dobša 2023).

The data from TroFi was collected first, consisting of sentences from the Wall Street Journal. All verbs categorized as metaphorical were annotated; the dataset documents their literal and metaphorical meanings. Moreover, TroFi includes the broader context for each verb, usually containing the full sentence. The MOH-X dataset, created ten years later, is an extension of TroFi, providing a more refined annotation process than its predecessor with higher reliability and accuracy (n.d.-b). Finally, the VUA Metaphor Corpus is a large-scale dataset containing metaphorical and non-metaphorical uses of words. Text is taken from four genres of British English texts, allowing for analysis of metaphor in different types of discourse. Furthermore, the annotation procedure follows the Metaphor Identification Procedure VU University Amsterdam (MIPVU), commonly accepted as an effective means of detection. Perhaps most notable about

the VUA, however, is its scope. The corpus contains over 200,000 words, with about 50,000 words for each genre (Steen et al. 2010). On the other hand, MOH-X, which is larger than TroFi, contains less than 700 sentences (n.d.-b). For these reasons, many metaphor detection models have used the VUA dataset.

Similarly, we first proceeded by leveraging the VUA-18 dataset, a subset of VUA, commonly used for computational tasks. The corpus is publicly available through VU University Amsterdam as an XML file, though we used a version provided by the authors of MelBERT in CSV format (Choi et al. 2021). Specifically, the first column, titled "index," offered a reference for the row's specific sentence. Next, the "label" column indicates whether or not the word indicated is categorized as a metaphor; a "1" signifies a metaphor, and a "0" signifies no metaphoric meaning. The "POS" column further informs the word's part of speech. Finally, the "w_index" gives the location – with zero-based indices – of the word in question (see Figure 1).

| index | label | sentence | POS | w_index |
|---|---|---|---|---|
| b1g-fragment02 841 | 0 | If it now seems self-evident that monitoring of the global environment is necessary, indeed is even vital, the prec | ADP | 42 |
| fef-fragment03 667 | 0 | Which equation should we use in a practical case, the equation for the vector potential, Ampre's law, Biot-Savar | NOUN | 10 |
| as6-fragment01 76 | 0 | It was initiated partly in response to the furore caused by Enoch Powell's 1969 rivers of blood speech, much as | ADP | 10 |
| ew1-fragment01 108 | 1 | You fully know as an old pressman the difficulty of dealing with a big speech late at night | VERB | 10 |
| fpb-fragment01 1152 | 0 | It was a condition of her gift to you the ten thousand pounds capital to start KITS. | ADJ | 5 |
| bpa-fragment14 2162 | 0 | In the stroboscopic view, the giant pistons were the only things moving until a figure detached itself from the wa | DET | 8 |
| ew1-fragment01 8 | 0 | Two strong groups emerged behind Long and Chamberlain, about equal in numbers, but there were few backers | VERB | 14 |
| ea7-fragment03 809 | 0 | The usual source of supply was town militias Suger mentions those of Rheims, Chlons, Laon, Soissons, Orlans, | VERB | 9 |
| kcc-fragment02 40 | 0 | So I thought well in case I can't get them anywhere else the market and he said well we'll have one from here, tv | NUM | 20 |

Figure 1: The first nine entries of the VUA-18 dataset used for training. This version is publicly available by Mao et al.

However, to detect a broad range of types of figurative language, we sought to train on a dataset that did not exclusively contain metaphors. Thus, we used the FLUTE dataset generated by researchers at Columbia (Chakrabarty et al. 2022). Although the data was created to solve the textual entailment task, we leveraged FLUTE for figurative language detection. Most importantly, the dataset includes over 1500 sentences; about 7500 of them are literal and the remaining sentences, which amount to over 7500, are figurative. These figurative sentences are further labeled as either sarcasm, simile, metaphor, or idiom. Since the data is organized in entailment (or contradiction) pairs, most figurative sentences have a corresponding literal sentence that is either similar (entailment) or contradictory (contradiction) in meaning. We encoded every type of figurative language as figurative language rather than maintaining the specificity of the provided labels, as our tool served solely to classify a sentence as figurative or literal.

Although the FLUTE dataset offered a large sum of data reflecting a broad range of figurative language, the sentences remained highly structured and, therefore, unlike potential sentences written in a public health forum. Thus, we augmented the FLUTE training data with an annotated dataset of 729 sentences from the patient.info health forum in the "Cancer" group. We used three annotators who were students with experience in linguistics. To ensure accurate annotations, we asked for them first to confirm that they understood the meaning of the sentence. Then, to aid their detection of figurative language, we asked, "do the words comprising the sentence literally denote the intended meaning of the sentence?" Once we collected all annotated data, we generated a new dataset containing the classification shared by the majority. All three

annotators were in agreement over 50% of the time. Our total training dataset, therefore, included 15,797 total sentences, containing both sentences from the FLUTE database and annotated sentences from the patient.info health forum.

*Using the Generative Pre-trained Transformer (GPT)*

As indicated by the success of most recent approaches towards Metaphor detection, we opted to employ Large Language Models. We began by downloading OpenAI's GPT-3.5 turbo via an API key. For each of the 200 test sentences, which were randomly selected from the VUA-18 dataset, we ran the following prompt:

"Detect if there is a metaphor in this sentence: {sentence}. If there is, display 'Metaphor'. Otherwise, display 'no_Metaphor'"

Notably, the results of this approach were poor. Across several runs, for less than 100 of the 200 test sentences, GPT accurately detected the presence, or lack thereof, of a metaphor. Understandably, the model was not fine-tuned for this purpose. The initial results, however, were indicative of a larger pitfall in using GPT to detect metaphors. GPT – literally "Generative Pre-trained Transformer" – focuses primarily on text generation (Brown et al. 2020). Thus, tokens are processed unidirectionally, which makes analyzing the nuances of text with full context a more difficult task.

*Using Bidirectional Encoder Representations from Transformers (BERT)*

Informed by our preliminary results with GPT, we sought to use BERT, another Large Language Model. This mode, unlike GPT, is widely used in literature for metaphor detection (Su et al. 2020; Gong et al. 2020; Chen et al. 2020). BERT processes text in both directions, allowing for more success at NLP classification tasks that necessitate deep semantic understanding. Furthermore, although GPT can be fine-tuned, its architecture is primarily geared towards text generation. On the other hand, BERT was designed to be fine-tuned. The model even contains [CLS] tokens, specifically beneficial for classification tasks, as well as segment embeddings that help delineate different parts of a sentence (Brown et al. 2020; Devlin et al. 2018).

Our preliminary results with BERT have provided empirical evidence for the predictions based on its architecture. BERT is available in two sizes – BERTbase and BERTlarge – which have 12 and 24 transformer layers, respectively. Since we planned to train on less than 1000 sentences, we opted to use the BERTbase model (uncased), requiring fewer computational resources than its counterpart (Devlin et al. 2018).

After 800 rows of the VUA-18 dataset were randomly selected, we first tokenized the input text data using the BERT tokenizer. Next, having completed batching and shuffling of the dataset, we trained the model for five epochs. Following the training, the BERT model was evaluated using 200 other randomly selected rows of the dataset. As desired, the accuracy drastically increased in comparison to GPT. On average, about 165 of the 200 test sentences were accurately identified regarding metaphor status. With little training (800 sentences), the BERT approach significantly outperformed the GPT approach.

However, when we fine-tuned BERT with our second dataset including both FLUTE data and annotated data, our accuracy dramatically increased. We proceeded with the same procedure;

first, we split 80% of our data into a train set, then we used the BERT tokenizer for tokenization, and finally, we trained for five epochs. In this case, we tested the model on 3012 sentences and achieved an accuracy of 94.2% for detecting sentences as either literal or figurative. We specifically detected literal sentences with 96.5% accuracy and figurative sentences with 91.9% accuracy.

*Patient.info Dataset*

We scraped public health forum data from the patient.info dataset (patient.info/forums). The site contains hundreds of groups that each focus on a specific disease. Furthermore, within each each group are discussion threads. These threads begin with an original post written by the original poster. Typically, this post describes an issue that the author is facing related to the disease covered in the broader group. Discussion threads also include titles, which are created by the original poster. In response to this post are replies. Authors from the public community of patient.info users are all able to respond. In some cases, moreover, users may respond to replies—an event that we denote as a response. Particularly intriguing to our study are cases when the original poster responds to a reply, giving a response. Only one layer of nested comments is permitted; that is, a user is unable to respond to a response. Finally, the content of comments, which include original posts, replies, and responses, can consist of some combination of images and text,

We employed the BeautifulSoup python package to parse the HTML from the patient.info/forums URL. For each group, we collected the group name, the community name, the author name, the recipient name, the title, the contents of the comment, the number of likes, the number of replies, the time stamp, the URL, the discussion ID, the comment ID, the author ID, the category ID, the group ID, and the number of followers. After completing this original scraping process, we included columns describing when an original poster wrote a response as well as the ID of the reply to which they responded. We also ran our figurative language detection model on each comment and recorded figurative language scores. We normalized these scores by the number of sentences, dividing the number of sentences with figurative language detected by the total number of sentences.

**Results**

*Comparison between Skin Cancer and Acne Datasets*

Our first computation analysis considered figurative language's influence on the number of replies for a given original post. Specifically, we considered the differences in these correlations between comments concerning a more severe medical condition—skin cancer, in our case—and a more mild condition—acne. We extracted about 3500 original posts from each condition, along with their corresponding figurative language scores. Then, we calculated the Pearson correlation coefficient between both sets of data.

Among Skin Cancer comments, there was a slight positive correlation between the figurative language score and the number of replies, with a correlation coefficient of 0.005. On the other hand, we identified a more negative correlation coefficient for acne comments: -0.016. Although both scores are numerically close to 0 and differ only slightly, the large sample size (~3500 comments for each condition) makes the results more significant.

We also performed our experimental study comparing the same two datasets. The study began by asking respondents to read a health advice 'post.' The specific post indicated that an increased coffee intake resulted in headaches. Notably, it is deliberately an example of a 'mild' condition, much like the acne comments served to represent. For a randomly selected half of the respondents they read a post written entirely in literal language. The other half received a post with identical meaning but with sentences written as different forms of figurative language. Each sentence of the literal post directly corresponds to its figurative counterpart in the figurative post. Then, we asked respondents to answer the following question on a scale of 1 to 7, with '1' being 'not at all likely' and '7' being 'definitely likely': "How likely are you to follow the advice?" The mean response score for those given the literal post was 5.19, and the average for those given the figurative post was 4.88. Again, we noticed that, with the context of a mild medical condition, an increase in the amount of figurative language decreased the mobilization to action, measured here as the likelihood of following the advice.

*Comparison between Bowel Cancer and Irritable Bowel Syndrome (IBS) Datasets*

Our second stage of computation analysis compared comments related to bowel cancer and others related to irritable bowel syndrome. In this study, we hoped to better understand the differences between comments about mild conditions like IBS and those about more severe conditions like cancer. Thus, we began by scraping 4371 comments from the IBS group of patients.info and 2931 comments from the bowel cancer group. Each subset of the data included about 700 cases when the original poster offered a response to a reply. In addition to comparing mild and severe conditions, we also drew distinctions between original posts, replies, and responses. We labeled these post types '1', '2', and '3' respectively.

We noted a statistically significant difference between the mean figurative language scores of bowel cancer comments and IBS comments; the mean was 0.1545 in the former and 0.1788 in the latter ($p<0.05$). The prevalence of figurative language in cancer comments indirectly supports the result that figurative language is more effective in empowering responders to act in severe settings than in mild ones. The average emotionality, on the other hand, was higher for IBS comments than for cancer comments ($p=0.05$). Furthermore, the positive emotionality was significantly higher for IBS comments ($p<0.05$). For other metrics, such as extremity and certainty, we did not observe statistically significant differences between the two groups.

Concerning different post types, however, we noted significant statistical differences across a range of variables. For both IBS and cancer comments, the figurative language scores increased from type 1 to type 2 and from type 2 to type 3. The differences between both groups across the post types were not significant. Word count, as we expected, also differed between different post types. The number of words dropped by an average of 70 from original posts (type 1) to replies (type 2). Responses maintained the lower word count of the replies. Notably, the average extremity was statistically different across post types ($p<0.05$), unlike between the two groups. In cancer comments, the extremity dramatically increased following the original post and subsequently remained the same for the response (type 2 and 3). However, in IBS comments, the extremity remained similar to the original post and the reply but dropped in the responses. The difference in average extremity between the two groups across all three post types was statistically significant ($p<0.05$). Finally, the average emotionality also differed significantly across post types. In both IBS and cancer comments, the emotionality score decreased in the

replies and then increased in responses, though not to the level of the original post. Due to the similarities in their respective progressions, there was not a statistically significant difference between the average emotionality of the two groups for all post types.

**Future Work**

Our future work centers around data collection, fine-tuning our figurative language detection model, and an additional experiment. Although we have created the scraping scripts, they run at a pace that is too slow to scrape all the data from patient.info in a few months. Thus, we are working on improving the code to allow for this efficiency. We will specifically seek to utilize several cores at once to perform more parallel processing and cut down on unnecessary scraping steps. Once the new code is complete, we will hope to complete the groups that have not yet been collected. Additionally, we plan to find two subforums that accurately represent a mild and severe medical condition, which may narrow the focus of our scraping task.

To improve our detection model, we will fine-tune it on more unstructured data. Although the accuracy is above 90% with the current version, these results are skewed by the highly positive results from testing the highly structured FLUTE data. When tested solely on unstructured data, the model performs far more poorly. Therefore, we will create a new subset of sentences for manual labeling of figurative language or, in other words, extend our previously annotated dataset. Following the acquisition of this data, we will re-finetune the model for figurative language detection and assess the accuracy. Once the model is completed, our computational analyses can be quickly re-run.

Regarding the experimental approach, we may run a similar study to our first but instead set the context of a more severe condition. This will allow for a clearer comparison between the experimental and computational approaches to identifying the relationship between figurative language and the mobilization to act in the two domains. Our second study builds upon the first. In the context of a social network or forum thread like Reddit, we will write a hypothetical post from a peer containing varying levels of figurative language and analyze the effect this post has on the language of the participant's response; extending the first study's results, this language may be affected by the changing mental states, including compliance. Thus, for the second study, attitudes and compliance are new mediators, in addition to trust, imagery, and pleasure. Since this study also analyzes the participant's textual response, we will use similar textual features.

## References


Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." https://doi.org/10.48550/ARXIV.2005.14165.

Carston, Robyn. 2018. "Figurative Language, Mental Imagery, and Pragmatics." *Metaphor and Symbol* 33 (3): 198–217.

Chakrabarty, Tuhin, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. "FLUTE: Figurative Language Understanding through Textual Explanations." *arXiv [cs.CL]*. https://doi.org/10.48550/ARXIV.2205.12404.

Chen, Xianyang, Chee Wee (ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. "Go Figure! Multi-Task Transformer-Based Architecture for Metaphor Detection Using Idioms: ETS Team in 2020 Metaphor Shared Task." In *Proceedings of the Second Workshop on Figurative Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.32.

Choi, Minjin, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. "MelBERT: Metaphor Detection via Contextualized Late Interaction Using Metaphorical Identification Theories." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.141.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." https://doi.org/10.48550/ARXIV.1810.04805.

Gao, Ge, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. "Neural Metaphor Detection in Context." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/d18-1060.

Gong, Hongyu, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. "IlliniMet: Illinois System for Metaphor Detection with Contextual and Linguistic Information." In *Proceedings of the Second Workshop on Figurative Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.21.

Hansen, Jochim, and Michaela Wänke. 2010. "Truth from Language and Truth from Fit: The Impact of Linguistic Concreteness and Level of Construal on Subjective Truth." *Personality & Social Psychology Bulletin* 36 (11): 1576–88.

Kronrod, Ann, and Shai Danziger. 2013. "'Wii Will Rock You!' the Use and Effect of Figurative Language in Consumer Reviews of Hedonic and Utilitarian Consumption." *The Journal of Consumer Research* 40 (4): 726–39.

Lai, Huiyuan, Antonio Toral, and Malvina Nissim. 2023. "Multilingual Multi-Figurative Language Detection." In *Findings of the Association for Computational Linguistics: ACL 2023*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.589.

Li, Yucheng, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. "Metaphor Detection via Explicit Basic Meanings Modelling." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.9.

Martin, James H. 2006. “A Corpus-Based Analysis of Context Effects on Metaphor Comprehension.” In *Corpus-Based Approaches to Metaphor and Metonymy*, 214–36. Berlin, New York: Mouton de Gruyter.
McMullen, Linda M. 1989. “Use of Figurative Language in Successful and Unsuccessful Cases of Psychotherapy: Three Comparisons.” *Metaphor and Symbolic Activity* 4 (4): 203–25.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” https://doi.org/10.48550/ARXIV.1301.3781.
Potamias, Rolandos Alexandros, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2019. “A Transformer-Based Approach to Irony and Sarcasm Detection.” https://doi.org/10.48550/ARXIV.1911.10401.
Pragglejaz Group. 2007. “MIP: A Method for Identifying Metaphorically Used Words in Discourse.” *Metaphor and Symbol* 22 (1): 1–39.
Ptiček, Martina, and Jasminka Dobša. 2023. “Methods of Annotating and Identifying Metaphors in the Field of Natural Language Processing.” *Future Internet* 15 (6): 201.
Rai, Sunny, and Shampa Chakraverty. 2021. “A Survey on Computational Metaphor Processing.” *ACM Computing Surveys* 53 (2): 1–37.
Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2012. “From Humor Recognition to Irony Detection: The Figurative Language of Social Media.” *Data & Knowledge Engineering* 74 (April):1–12.
Roberts, Richard M., and Roger J. Kreuz. 1994. “Why Do People Use Figurative Language?” *Psychological Science* 5 (3): 159–63.
Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging Evidence in Language and Communication Research 14. Amsterdam, Netherlands: John Benjamins Publishing.
Su, Chuandong, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. “DeepMet: A Reading Comprehension Paradigm for Token-Level Metaphor Detection.” In *Proceedings of the Second Workshop on Figurative Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.4.
Swarnkar, Krishnkant, and Anil Kumar Singh. 2018. “Di-LSTM Contrast : A Deep Neural Network for Metaphor Detection.” In *Proceedings of the Workshop on Figurative Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-0914.
Yang, Jie, and Hua Shu. 2016. “Involvement of the Motor System in Comprehension of Non-Literal Action Language: A Meta-Analysis Study.” *Brain Topography* 29 (1): 94–107.
n.d.-a. In . https://doi.org/10.18653/v1/W16-1104.
n.d.-b. In . https://doi.org/10.18653/v1/S16-2003.