Effects of Source Language on Idiom Interpretation

Anjali Ragupathi

March 21, 2024

1 Introduction

1.1 Background and Motivation

We begin this work by defining some key terms. An idiom is a non-compositional, conventionally used phrase whose meaning cannot be easily deduced from a combination of the meanings of its component words. A metaphor, by contrast, is neither conventionalized nor standardized, which means that the same phrase can have multiple meanings. These are further distinguished from proverbs which are sayings (usually sentences) that speak some truth and cannot be modified in constructions; idioms, by contrast, have some leeway in modifying their construction (e.g. "the beans were spilled", "she spilled the beans"). (Citron et al., 2016)

Many figurative expressions that are metaphorical are said to possess an underlying conceptual metaphor ((Lakoff & Johnson, 1980), (Lakoff, 1993)) which is an association between a source domain and a target domain (that is, we try to understand one idea in the form of another idea). For instance, arguments can be understood in the form of war, leading to the conceptual metaphor ARGUMENT is WAR. This leads to many figurative expressions like "winning an argument", "attacking weak points", and "presenting a defense" which relate to specific associations within the domain of argumentation. Idioms are special cases of metaphorical expressions and so can be said to possess some underlying conceptual metaphor.

Newer theories like the Conventional Figurative Language Theory by Dobrovol'skij and Piirainen (2018) illustrate how conceptual metaphor theory alone cannot explain the mapping between the literal and figurative forms of idioms. This theory in particular states that in addition to the mental image provided through the conceptual metaphor itself, other motivating factors allow for some idioms to be interpreted more easily than others. Some of these motivations can be cognitive (more strongly focused on conceptual metaphor) while others rely on shared cultural symbols such as bread as a means of livelihood to create additional links between the source and target domains (Dobrovol'skij & Piirainen, 2021).

Motivated by these theories, we wish to examine two things in this study: the first is the likelihood of predicting the expected figurative meaning of an unfamiliar idiom without any surrounding context or relevant cultural knowledge; the second is the effect of the source language of the idiom on the prediction outcomes.

1.2 Hypotheses and Assumptions

To address the first question, our null hypothesis is that participants cannot predict the expected meaning of the idiom 50% of the time or above, while the alternative hypothesis argues in favor of participants' abilities. To address the second question, our null hypothesis is that there is no effect of the source language of an idiom on the likelihood of predicting the expected meaning, whereas the alternative hypothesis states that such an effect exists.

We assume that in the absence of any context or cultural knowledge, when presented with an unfamiliar idiom, participants will try to decompose the literal translation of the idiom into its constituents and

map it to the target domains expressed in the options for the potential figurative meaning of the idiom. The first part of the process of idiom decomposition for unfamiliar or L2 (second-language) idioms was examined by Senaldi and Titone (2022).

1.3 Links to GitHub and OSF Preregistration

The GitHub repository for this project can be found at this link.
The experiments conducted for this project have been preregistered with the Open Science Foundation at this link.

2 Experiment

2.1 Prior work

This project is part of a larger thesis, the objective of which is to determine if the idiom interpretation capabilities of large language models are comparable to those of humans and to what extent an underlying conceptual metaphor drives the predictions. Large language models (LLMs) are deep neural networks with billions of parameters that are trained to predict the next token or a masked token by learning statistical co-occurrences between words. Since this is done over many hours with large datasets and high compute, LLMs can "learn" to approximate (or model) the language they are trained on.

Post-training, LLMs are often fine-tuned on more specific downstream tasks like question answering or machine translation. A more recent class of LLMs including ChatGPT and Gemini includes "instruction-tuned" models where the inputs during the fine-tuning process are phrased as instructions or prompts. This allows for more natural interactions between humans and the models at inference time and can encompass most of the previously defined downstream tasks with additional strategies like chain-of-thought prompting (allowing LLMs to reason from the input to the predicted output step-by-step)((Wei et al., 2023)) or few-shot prompting (which allows the user to provide examples of input-output pairs at inference which the model can use for improved answer quality).

For the task at hand - idiom interpretation, we compiled a dataset of 80 non-English idioms translated literally into English, the process of which is described in Section 2.2.1. The choice of translating the idioms into English was made because English could act as an anchor language; this meant that the LLM would be able to "see" these idioms from the point of view of a primarily English-speaking listener. Additionally, English was the more suitable option as most commercial LLMs are trained on English data and it would be easier to integrate English versions of these idioms into the prompting and analysis pipeline. Finally, an important reason for this choice was that we wanted to test the extent to which the surface features of the idiom itself (i.e., constituent words and their types, associated sentiment, etc.) influenced idiom interpretation - not necessarily the ability to recall idioms (an effect of familiarity) or the influence of contextual cues from surrounding dialogue (which would mean that the listener would rely heavily on context and not on the idiomatic phrase itself). These factors motivated the choice of language and inclusion of only the idiom phrase without the context of usage in conversation or larger bodies of text.

We used three different large language models for this test: HuggingChat (the default model under the hood is Llama-70B), Llama-13B, and PaLM-2 Text-Bison-001). HuggingChat is a free LLM provided by HuggingFace and we used an unofficial open-source Python API to interface with it (API Link). PaLM-2 is a predecessor of Gemini provided by Google and Llama is provided by Meta; both were trained on diverse, multilingual datasets and evaluated for potential harms.

Eight variants of the same prompt were constructed, including "Interpret the figurative meaning of this idiom", "Retrieve the figurative meaning of this idiom", "This is the literal translation of an idiom: $\langle IDIOM \rangle$. What is its figurative meaning?", as well as versions which asked the model to provide step-by-step reasoning for its prediction, gave the model some room for experimentation by asking what the model "thought" the figurative meaning was (models were less likely to hedge their answers) and asked the model to make the prediction ignoring the language of origin (to potentially

eliminate any shortcuts the model could take by just retrieving the meaning and not coming up with plausible answers). The step-by-step prompt results were subsequently excluded from further analysis to standardize the result format and make subsequent analyses easier, but they were retained for potential future analyses.

Each model ran inference on the entire dataset of 80 idioms thrice to account for variation in response and to make the resulting evaluation strategies more robust. Some of these free-form responses were then used as confound options for the current project.

2.2 Methods

2.2.1 Materials

A list of 80 idioms was prepared by collecting the literal translations (into English) and corresponding figurative meanings of different idioms from Russian, Spanish, Hindi, and Mandarin Chinese from language-learning websites and publicly available digital school textbooks. 20 idioms were included from each language for balance.

For each idiom, in addition to the expected figurative meaning, three confounds were prepared. In reference to the outputs of the previous experiment (see Section 2.1), confounds were chosen from among the plausible responses produced by the large language models. A confound here is defined as a phrase that conveys a plausible potential meaning as either produced by a large language model or one that relates to a relevant underlying conceptual metaphor. For cases where large language models produced largely consistent results, the latter method of defining confounds was used and the experimenter manually created entries for those confounds.

In addition to the 80 cross-lingual idioms, 4 English idioms - namely, "once in a blue moon", "to face the music", "to break the ice", and "to feel under the weather" - were selected as attention check idioms. They served as a means of standardizing participants' understanding of what an idiom was and if they were paying attention during the experiment instead of randomly selecting options.

2.2.2 Participants

A pilot study with 10 participants was run to determine the estimated completion time and any difficulties with the conceptual and technical components of the experiment. Participant recruitment was done on Prolific and restricted to the US, UK, EU, and India. Based on the results of this experiment, the initial instructions were modified to include an example of an idiom, i.e. "kick the bucket", so that participants would be primed toward the format of the experiment. The estimated completion time was also reduced to 12 minutes from the originally expected 15 minutes after the pilot study.

The main study recruited 100 participants on Prolific and included participants across the US, UK, EU, India, Singapore, Switzerland, and Mexico, to encourage a diversity of participants to take part in the study. The main study was released in two batches (50 participants each) at two different time slots to gather data from participants in different global time zones. Nevertheless, the primary demographic of the final participant pool was residents of the United Kingdom and the United States.

2.2.3 Procedure

Each participant was shown 20 idioms - 4 from each source language and 4 attention check idioms - randomly shuffled and presented in a sequence. Each idiom was presented along with the 4 options (the expected figurative meaning and the three confounds) which were also shuffled. Participants saw one idiom at a time and were required to select an option before moving on, with no option to go back and modify their response. Towards the end of the experiment, participants were asked to fill out a survey with demographic information like age, gender, level of education, and languages they were natively proficient in (that is, language(s) they grew up speaking or possessed the ability to fluently write, read, and speak in at a native level).

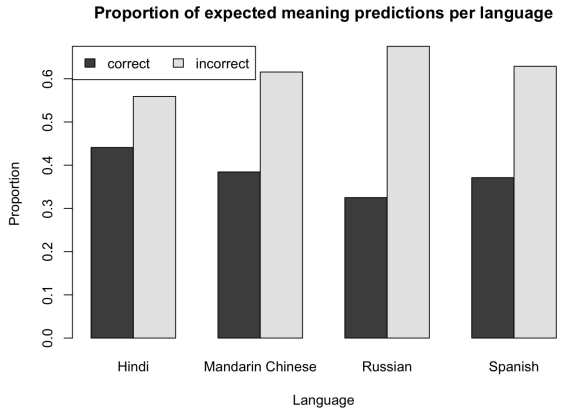

Figure 1: Proportion of Expected Meaning Predictions by Language

In the case of the participants as well as the LLMs, subjects were not informed that the idioms were translated from another language. This was done in an attempt to invoke idiom processing based on conceptual metaphor without priming the subjects to also think about potential languages of origin. Additionally, this task differed from the prior LLM task where the latter involved free responses and the former constrained choice to four options. This was done to reduce the degrees of freedom involved in the study, leading to the advantage of lower variability and simplified analysis. The tradeoff, however, was the loss of expressiveness in the experiment so we could not observe the direct functioning of idiom interpretation through a more creative, free-form response.

2.3 Results

2.3.1 Outcomes of the Experiment

For this study, we defined "correct" to be a situation where the participant predicted the expected meaning (corresponding figurative meaning of the translated idiom in its original language) and "incorrect" to be a situation where they predicted a different meaning. Naturally, these were simplifications to the actual nature of such an experiment and we did not aim to test if people could predict a correct meaning as these idioms simply do not exist in English; it would be an invalid assumption to make that people can predict the correct meaning of a non-existent idiom.

Figure 2.3.1 shows the proportion of judgments that predicted the expected meaning of the idiom presented, grouped by each of the source languages in the study. We observed that Russian had the highest proportion of incorrectly predicted idioms while Hindi had the highest proportion of correctly predicted ones. This demonstrated that, at least at a baseline level, idioms from certain languages were more difficult to predict. Surprisingly, though Mandarin Chinese and Spanish are considered the two most spoken languages in the world, it appeared that idioms from these languages were still unfamiliar to a primarily English-speaking population suggesting that idioms did not often travel across languages. These results also provided evidence where we can fail to reject the null hypothesis for Question 1 (predicting expected idiom meaning with a 50% chance or more).

Figure 2.3.1 shows the proportion of judgments evaluated correctly and incorrectly per-item (for each idiom). We observed that some idioms like "the first pancake is always a blob" were predicted with perfect accuracy while idioms like "to spit at the ceiling" and "to shoe a flea" were predicted with 0% accuracy despite all three being from the same source language of Russian. From a cursory comparison of these results with the free-form responses generated by LLMs in 2.1, there seemed to be a small correlation between idioms that were predicted correctly by humans and idioms that were consistently correctly predicted across prompts in LLMs, but a slightly stronger correlation with the incorrect predictions (the LLMs also tended to predict those idioms incorrectly). However, this comparison is not substantiated with concrete evidence and is a topic for future phases of this project.

2.3.2 Mixed-Effects Logistic Regression Model Results and Discussion

In Figure 2.3.1, the proportion of expected meaning predictions is shown. In our experiment, we aimed to study if there was an effect of the source language of the idiom on the likelihood of predicting the expected meaning. A complete mixed-effects logistic regression model (Model 1) with language as the fixed effect and a random intercept by item, as well as a random slope explaining variation in language effects on outcome between participants, was fit to the data and was able to model the distribution with 78% accuracy. The statistics of this model are shown in Figure 2.3.2, indicating that the model was able to explain around 45% of the total variance in the data with the random effects. However, after accounting for by-item and by-subject variability (including the variability of language effects on expected prediction across participants), it was observed that source language did not have a main effect on the likelihood of prediction of the expected idiom meaning.

However, it was observed that variance was fairly high between idioms (items) compared to the variance between participants (subjects), indicating that the idioms themselves were relatively ambiguous in their meaning and naturally, different idioms would be predicted with varying levels of perceived success. For the fixed effect of language by-subject, we saw that participants tended to vary less within their own predictions compared to other participants, i.e., they were more likely to consistently predict either the expected or the unexpected figurative meaning, but this effect was not particularly strong either.

If we did not account for per-item variability (Model 2), we could observe a small effect of language on the outcome. Model 2, however, only explained 6% of the variance in the data with the random slope effect and the language fixed effect and so would not have been an accurate model of the data. Here, we noted the stronger correlation between language and outcome, which showed that Russian idioms were more likely to be predicted incorrectly compared to the reference Hindi idioms (and overall)(beta = 0.49, sd = 0.17, p = 0.004). This trend is observed in Model 1 too, but with a very weak and insignificant effect. We also saw that there was a higher likelihood of these models overfitting because of the introduction of varying levels of fixed and random effects, indicated by the fact that the models found it difficult to converge.

Overall, we could conclude that the source language of the idiom did not have a strong effect on the likelihood of predicting the expected figurative meaning, failing to reject the null hypothesis for Question 2.

3 Discussion and Conclusion

The results of the study showed that there was no strong effect of idiom source language on the likelihood of predicting the expected figurative meaning. Perhaps the language itself is too variable and broad of a category by which to explain the meaning of an idiom. This does not indicate that idioms in general cannot be interpreted without context or familiarity; there may be other factors that can influence the likelihood of predicting the expected meaning. For instance, Citron et al. (2016) discussed affect or emotion playing a role in idiom interpretation (in terms of arousal and valence) while Williams, Bannister, Arribas-Ayllon, Preece, and Spasić (2015) demonstrated that idioms were key to building better sentiment analysis systems as the presence of idioms in a sentence usually indicated some non-neutral emotional component (generally negative). Under this assumption, we could postulate that people might be more likely to select a potential figurative meaning that has a more negative affect. This could be a direction for future studies.

Further, we have other hypotheses that rely on the presence of natural or artificial kinds in the idiom, as well as the thematic roles of different components of the idiom. We could also think about the participants themselves: is there an effect for participants who speak languages within the same immediate language family as the idiom's source language or even among dialects within the same language (Cantonese vs Mandarin) compared to participants who culturally and geographically share similar spaces but speak languages from a different language family (for example, Hindi and Tamil are both spoken in India - they do not share many words in common, being from different languages, but have common cultural and geographical boundaries)?

More immediate lines of work would involve exploring how LLMs perform on the same constrained choice task and if their results are comparable to those of human participants. We could also recruit participants to do the free-form response task to examine trains of thought that occur in idiom interpretation.

References

Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M. (2016). When emotions are expressed figuratively: Psycholinguistic and affective norms of 619 idioms for german (panig). *Behavior research methods*, *48*, 91–111.

Dobrovol'skij, D., & Piirainen, E. (2021). *Figurative language: Cross-cultural and cross-linguistic perspectives* (Vol. 350). Walter de Gruyter GmbH & Co KG.

Dobrovol'skij, D., & Piirainen, E. (2018). Conventional figurative language theory and idiom motivation. *Yearbook of Phraseology*, *9*(1), 5–30. Retrieved 2023-05-09, from https://doi.org/10.1515/phras-2018-0003 doi: doi:10.1515/phras-2018-0003

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202–251). Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). Conceptual metaphor in everyday language. *The Journal of Philosophy*, *77*(8), 453–486. Retrieved 2023-05-09, from http://www.jstor.org/stable/2025464

Senaldi, M. S. G., & Titone, D. A. (2022). Less direct, more analytical: Eye-movement measures of l2 idiom reading. *Languages*, *7*(2). Retrieved from https://www.mdpi.com/2226-471X/7/2/91 doi: 10.3390/languages7020091

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models.*

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., & Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, *42*(21), 7375-7385. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417415003759 doi: https://doi.org/10.1016/j.eswa.2015.05.039

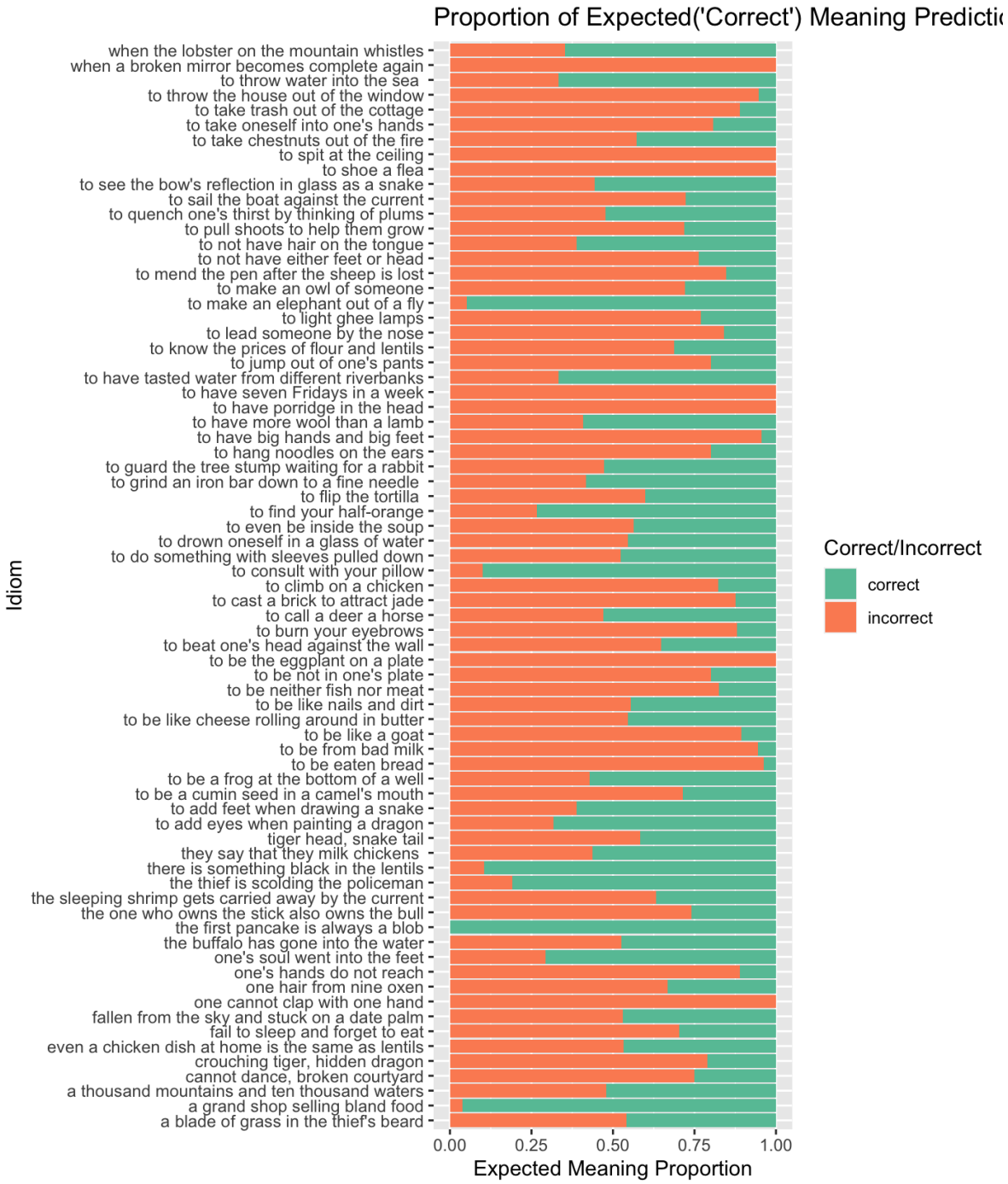

Figure 2: Proportion of Expected Meaning Predictions by Idiom

	result		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.47	0.66 – 3.26	0.348
language [Mandarin Chinese]	1.34	0.46 – 3.93	0.591
language [Russian]	1.99	0.68 – 5.83	0.212
language [Spanish]	1.40	0.47 – 4.17	0.548
Random Effects			
σ^2	3.29		
$\tau_{00\ workerid}$	0.43		
$\tau_{00\ idiom}$	2.23		
$\tau_{11\ workerid.languageMandarin\ Chinese}$	0.35		
$\tau_{11\ workerid.languageRussian}$	0.36		
$\tau_{11\ workerid.languageSpanish}$	0.66		
$\varrho_{01\ workerid.languageMandarin\ Chinese}$	-0.44		
$\varrho_{01\ workerid.languageRussian}$	-0.82		
$\varrho_{01\ workerid.languageSpanish}$	-0.78		
ICC	0.44		
$N_{\ workerid}$	100		
$N_{\ idiom}$	73		
Observations	1468		
Marginal R^2 / Conditional R^2	0.010 / 0.442		

Figure 3: Model statistics of mixed effects model