Seeing the Unsaid: Visualizing English Idioms with Text-to-Image Generation

Irene Russo (ILC CNR), Paola Vernillo (Universiyt of Bologna)

Introduction

The accessibility and widespread availability of generative Large Language Models (LLMs) have enabled the creation of learner-centered, adaptive materials for second language learners [1]. Experimental studies have explored the use of these models to tailor texts to learners' proficiency levels by constraining generated content to specific grammatical structures [2] or producing personalized reading comprehension exercises for middle school English learners [3]. Thus far, these efforts have focused solely on textual outputs. However, the impressive capabilities of generative multimodal models, in particular text-to-image (T2I) models, suggest new opportunities to enhance language learning with visual content. Visual aids such as images, animations, and videos can help illustrate the meaning of expressions and lexical items effectively [4]. In particular, images have been shown to support the acquisition of foreign language vocabulary [5], while illustrated dictionaries improve comprehension by visually reinforcing verbal definitions and supporting memory retrieval [6].

Traditionally, the creation of illustrated dictionaries and lexicons has relied on illustrators providing pictorial representations of individual words or broader semantic units. Such illustrations are especially useful for idioms, where visual representations can significantly aid both short-term and long-term retention of their form and meaning [7].

Research Questions

This study focuses on two research questions:

- **RQ1** *Can text-to-image models effectively visualize the meanings of English idioms?*
- **RQ2** *How do the semantic properties of the idiom affect the quality of the generated images?*

To answer these questions, we evaluated the performance of four text-to-image models —Stable Diffusion 2.1, PixArt-LCM, ChatGPT's Dall-E implementation, and Midjourney 6.1 —on a dataset of 51 idioms. These idioms, structured as "verb-X-noun" phrases and accompanied by example sentences (see example (1)), are part of the experimental data used by [8] in their study about individual differences in the processing of idioms.

(1) a. *play with fire*
They have both been playing with fire and it's no surprise that things have gone so badly.

b. *eat your words*
He is going to have to eat his words and everyone will see now that he was mistaken.

Methodology

0.1 Dataset

The dataset was based on the work of [8] and is made up of 51 triplets structured in the form of "verb-X-noun" phrases, each accompanied by a sentence illustrating the meaning of the idiom. We enrich the dataset with scores related to the semantic and conceptual features of the idiom. For each idiom, we include the following:

- **semantic transparency**: following previous works in computational linguistics [9], the overall semantic transparency of an idiom was computed using `bert-base-uncased` semantic vectors to represent the figurative meanings by considering the example sentence in which the idiom is used and obtaining as literal meaning the mean of the word semantic vectors constituting the idiom. The similarity between the figurative and literal meanings is measured using cosine similarity;

- **concreteness**: concreteness ratings from [10], retrieved for 46 idioms;
- **imageability**: sum of imageability ratings for content words (i.e., verbs and nouns) retrieved from [11].

Concerning semantic transparency, *smell a rat* is the most transparent, while *under the weather* is the least transparent (mean = 0.54, std = 0.06). The idioms in the dataset differ with respect to concreteness, with *rock the boat* and *save your skin* being the most and least concrete, respectively (mean = 2.16, std = 0.46). Concerning imageability, broadly defined as the degree to which a word or phrase evokes a mental image, arousing mental imagery [12], the maximum value characterizes *rock the boat*, while the lower value is found for *drown your sorrows* (mean = 890, std = 216).

We expect semantic transparency and concreteness to be positively correlated with the overall abilities of the models to generate images that codify the literal meaning, even if each model can show different performances. Imageability is the sum of ratings for the words composing the idiom, and, as a consequence, we cannot exclude that the overall rating for the whole phrase - holistically evaluated - would be different. As such, we do not formulate a hypothesis about this property.

0.2 Images Generation

In this work, our goal is to understand which text-to-image model is better at visualizing the meanings of English idioms. Acknowledging that the landscape of available and testable models is constantly evolving — and that the results presented here may soon be improved — we select two classes of text-to-image models, open-source generative models and proprietary generative models. Among open-source models, we choose PixArt-alpha and Stable Diffusion. Among the proprietary models, we choose ChatGPT's implementation of Dall-E, and Midjourney 6.1 [1]. For each model, we collect 3 images as output because we want to understand whether each model generates consistent results for the same idiom, or whether the inherent variability of the automatic generation process leads, in some cases, to inconsistent performance. None of the models were configured with fixed random seeds, to explore natural variability. Importantly, no images were filtered or discarded post hoc, preserving the integrity of the data for model comparison. All generations were performed in a controlled time frame to reduce the influence of model updates or server-side drift.

The final dataset is composed of 612 generated images (153 for each model). We complemented this dataset by generating the same number of images for the sentences provided as examples in Carrol2023, for a total of 1024 images. The dataset will be released to foster future comparisons.

0.3 Annotating Generated Images

The automatically generated images were subjected to a manual annotation process carried out by the two authors. The annotation guidelines were inspired by the work of [7]. Specifically, each image is labeled with the following tags:

- LF: the image depicts elements of both the literal and figurative layers of meaning;
- F: only the idiomatic meaning is illustrated;
- L: only the literal reading is represented;
- F2: the image depicts a figurative meaning that differs from the idiomatic meaning, but it is inferrable from it.

Preliminary Findings

RQ1 *Can text-to-image models effectively visualize the meanings of English idioms?*

In our annotation experiment, inter-annotator agreement was assessed using Cohen's Kappa (), which accounts for agreement occurring by chance. The two annotators labeled a total of N items, and the resulting confusion matrix provides a detailed breakdown of their labeling decisions. The observed Cohen's Kappa was = 0.49 for idioms and 0.57 for example sentences, indicating moderate agreement according to the Landis and Koch (1977) scale. The confusion matrix reveals that most disagreements occurred about category F2, the most subjective, with one annotator more consistently identifying symbolic meanings associated with the idiom's figurative interpretation compared to the other. These results support the reliability of the annotation process, though future rounds may benefit from improved guidelines or additional training to reduce ambiguity between specific labels.

[1] Version 6.1, released on July 30, 2024, generates more coherent images with enhanced detail and improved textures.

In total, the two annotators agreed on 185 images for idioms and 275 for example sentences, with the majority of them identified as literal realizations for idioms (%65), while for example sentences is the figurative aspect the dominant one (%54).

Concerning the comparison between models, in accordance with [7] findings, we identify as the most promising one the model 1) that more frequently output images that are recognised by both annotators as expressing literal and figurative meanings at the same time 2) that output at least one literal and figurative images for each idioms, for the biggest set of idioms. Looking at the results, ChatGPT's implementation of Dall-E emerges as the most promising model. In general terms, all the models perform better when the generated image is based on sentences while the idiomatic expression, when considered in isolation as a prompt, forces some models (PixArt-LCM and Midjourney 6.1) into a symbolic interpretation that is unrelated to the idiom's actual meaning.

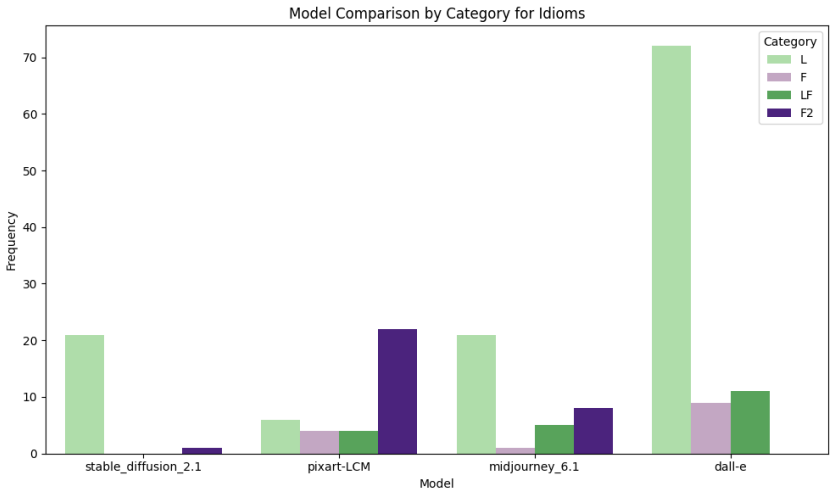

Figure 1: Distribution of annotated categories across all models for images generated from idioms.

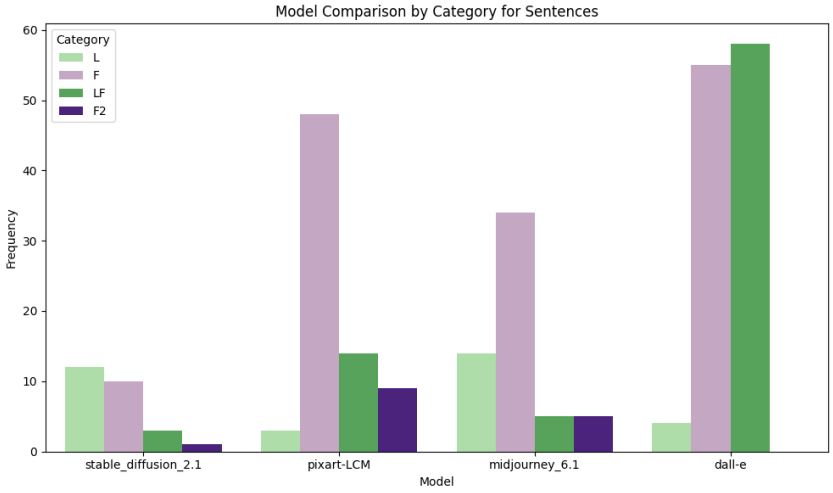

Figure 2: Distribution of annotated categories across all models for images generated from sentences.

RQ2 *How do the semantic properties of the idiom affect the quality of the generated images?*

While only images with annotation agreement were considered for the overall evaluation of model performance, all annotations were taken into account to better understand how the semantic properties of idioms affect the ease with which their literal or figurative meanings are visually represented. Each idiom is assigned four scores, corresponding to the percentage of images annotated as L, F, LF, or F2.

We found a significant negative correlation (r = -0.34, p ¡ 0.05) between the semantic transparency of the idiom and the frequency with which only its figurative meaning is depicted. Concerning example sentences, there is a positive correlation between the concreteness of the idiom and the frequency of LF (r = 0.36) and a negative correlation between the frequency of F and LF imag (r = - 0.47).

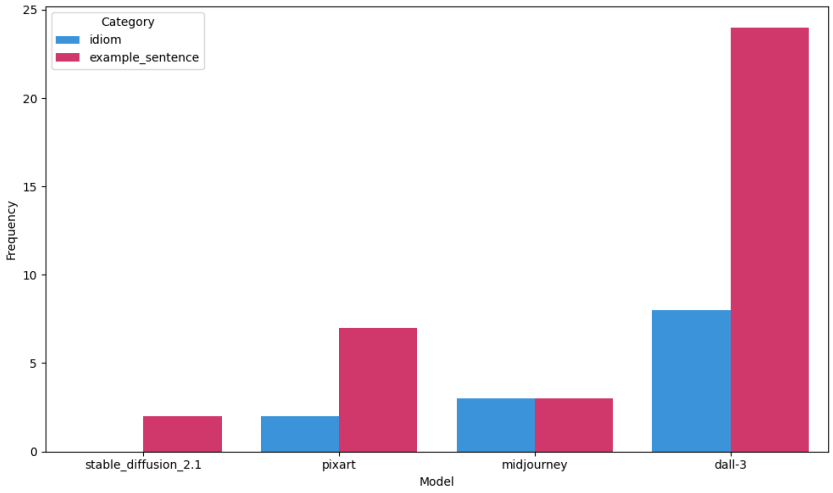

Figure 3: Distribution of images annotated as LF across all models.

In terms of group comparison, we checked if idioms with or without a specific image type have significantly different rating means. The only significant result concerns idioms that have F images and semantic transparency (T-test: t=-2.453, p=0.019).

References

[1] Euan Bonner, Ryan Lege, and Erin Frazier. "Large Language Model-Based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching". In: *Teaching English With Technology* (2023).

[2] Dominik Glandorf and Walt Detmar Meurers. "Towards Fine-Grained Pedagogical Control over English Grammar Complexity in Educational Text Generation". In: *Workshop on Innovative Use of NLP for Building Educational Applications*. 2024.

[3] Changrong Xiao et al. "Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications". In: *Workshop on Innovative Use of NLP for Building Educational Applications*. 2023.

[4] Abbas Pourhosein Gilakjani. "Visual, Auditory, Kinaesthetic Learning Styles and Their Impacts on English Language Teaching". In: *Journal of Studies in Education* 2 (2011), pp. 104–113.

[5] Shana K. Carpenter and Kellie M Olson. "Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not." In: *Journal of experimental psychology. Learning, memory, and cognition* 38 1 (2012), pp. 92–101.

[6] Anna Dziemianko. "The usefulness of graphic illustrations in online dictionaries". In: *ReCALL* 34 (2021), pp. 218–234.

[7] Renata Szczepaniak and Robert Lew. "The role of imagery in dictionaries of idioms". In: *Applied Linguistics* 32 (2011), pp. 323–347.

[8] Gareth Carrol and Katrien Segaert. "As easy as cake or a piece of pie? Processing idiom variation and the contribution of individual cognitive differences". In: *Memory & Cognition* 52 (2023), pp. 334–351.

[9] Hui Gao et al. "Consistency Rating of Semantic Transparency: an Evaluation Method for Metaphor Competence in Idiom Understanding Tasks". In: *International Conference on Computational Linguistics*. 2025.

[10] Emiko J. Muraki et al. "Concreteness ratings for 62,000 English multiword expressions". In: *Behavior Research Methods* 55 (2022), pp. 2522–2531.

[11] Michael Wilson. "MRC Psycholinguistic Database". In: 2001.

[12] Allan Paivio, John C. Yuille, and Stephen A. Madigan. "Concreteness, imagery, and meaningfulness values for 925 nouns." In: *Journal of experimental psychology* 76 1 (1968), Suppl:1–25.