Approximating Cognitive Representations Using Space

John R. Starr (jrs673@cornell.edu), Ashlyn Winship,
& Marten van Schijndel[a]

[a] *Department of Linguistics, Cornell University*

Abstract

In everyday life, people intuitively use space to make meaningful distinctions between objects. In this paper, we present a novel, free-to-use on-line experimental paradigm that capitalizes on these intuitions: GRIS (Generating Representations in Space). In GRIS experiments, participants manipulate a set of objects (text, audio, images) and place them on canvases. Following an introduction to the paradigm, we present three studies which demonstrate how experiments in the GRIS paradigm can both a) replicate prior psycholinguistic results and b) reveal nuanced insights about human and computational representations.

Keywords: representations, psycholinguistics, acceptability, typicality, paradigm

1. Introduction

In our daily lives, we use space to make and represent meaningful relationships between objects: we separate different kinds of clothes into different compartments, read menus that spatially group items on the page according to their broader classifications, and press elevator buttons that are vertically ordered to reflect the structure of their buildings. In these ways and many more, humans intuitively use space to simplify choice, perception, and computation (Kirsh, 1995), allowing us to navigate and represent complex structures and relations with ease.

One of the primary approaches to studying space in the cognitive sciences is through its relationship with language. Previous research has shown that people construct mental representations that encode spatial relationships (Taylor and Tversky, 1992; Bryant, 1997; Kemmerer, 1999), and that these relationships are marked on a schematic level of varying detail (Talmy, 1983; Landau and Jackendoff, 1993; Hayward and Tarr, 1995; Tversky and Lee, 1998). Studies in this domain often focus on how language organizes our cognitive representations of objects and their locations, both in discourse and in the real world. Other work on the relationship between space and language suggests that we transfer linguistic information onto mental spaces of the world (consisting of information of referents, their beliefs, actions, etc.), which we then blend together to understand the relevant discourse (Fauconnier, 1994; Sweetser, 1999; Fauconnier et al., 2007).

In this paper, we demonstrate that an alternative perspective on the relationship between space and language is also fruitful: space as a tool to contextualize our understanding of language, and, more broadly, human cognition. To show the utility of this alternative perspective, we present an experimental paradigm – GRIS (**G**enerating **R**epresentations **I**n **S**pace) – which a) capitalizes on the way humans intuitively use space, b) approximates representations of language and other cognitive phenomena, and c) does so in a way that is easily comparable to embedding representations from computational models, allowing us to further probe the matches and mismatches between humans and models. At a high level, participants in GRIS experiments can move objects (text, image, audio) onto a canvas and use space in a meaningful way, where information is incrementally collected about which objects were moved, when they were moved, and where they moved to. To briefly highlight our results, we demonstrate how GRIS experiments can 1) both replicate results from other psycholinguistic paradigms and provide further contextual nuance to such results, 2) develop multi-dimensional graphs that can be used for computational modeling, and 3) facilitate and simplify experimental designs that require multiple complex (pairwise) comparisons. More broadly, we argue that GRIS allows participants to use their natural intuitions about space to inform our understanding of how people represent various kinds of linguistic information.

1.1. Article Organization

In the following section, we provide further motivation for developing a paradigm that uses space meaningfully. In section 3, we outline the GRIS paradigm, introducing its key functionalities and structure. In sections 4-6, we present three GRIS experiments,[1] demonstrating how the paradigm can a)

[1] All items, data, and analysis code can be found at the following anonymized link: https://osf.io/94gck/?view_only=1ed03a3757fe44ba9a036510be60b7c6.

replicate prior results across a number of cognitive domains, and b) capture more nuanced relationships between representations than other experimental paradigms and computational models of linguistic structure.[2] In section 7, we discuss the general implications of GRIS and present possible directions for future work. In section 8, we conclude.

2. On Space

2.1. Space & Psycholinguistic Paradigms

From a design perspective, many standard psycholinguistic paradigms[3] minimally engage with space: standard rating and judgment tasks often present an item in isolation (or near isolation) alongside a scale or a drop-down box, and forced-choice tasks only capture the pairwise differences between one or two items. Some experimental paradigms do inherently use space as a metric for psycholinguistic effort, such as measuring how participants' eyes move to different locations on a screen when using eye-tracking in the visual world paradigm (Cooper, 1974; Tanenhaus et al., 1995), or following the trajectory of a participant's mouse/cursor across the screen using a mouse-tracking paradigm (Freeman and Ambady, 2010; Wilcox et al., 2024) However, these paradigms do not fully capitalize on the possible utilities of space: the locations of objects in the visual world paradigm are often optimized to be distinct and do not carry inherent meaning themselves, and mouse-tracking uses space as a proxy for processing difficulty instead of as a representational, organizational mechanism.

We propose that the use of space can simplify psycholinguistic designs for both researchers and participants. As an example, consider a rating task where a participant is asked to rate the difference between item pairs on a scale (according to some metric), for a total of four unique items. We visualize two possible iterations of this experiment in Figure 1.

In Figure 1A, participants are asked to directly measure the difference between all possible combinations of the four items (either in one trial or across

[2]In this paper, we focus on *linguistic* representations, though GRIS can be easily extended to approximate other kinds of cognitive structures and relationships such as in vision or acoustics.

[3]As will be described later in this article, GRIS is not designed to captured on-line processing. Accordingly, we do not elaborate on the use of space in psycholinguistic tasks such as self-paced reading (Just et al., 1982) or (Forster et al., 2009).

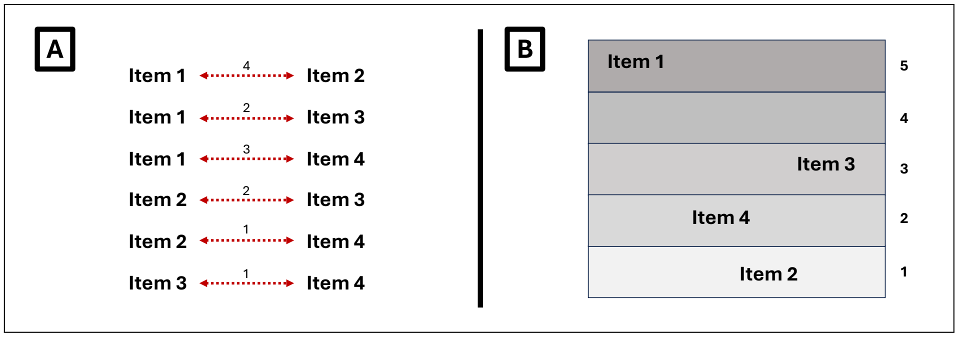

Figure 1: Sample rating tasks for four items. **(A)** Ratings in a pairwise comparison layout, where the number above the arrows reflects the scalar differences between two items. **(B)** Ratings in a space-motivated layout. The overall scalar ratings are identical between (A) and (B).

several trials). For these items, participants maintain six pairwise comparisons that are all in relation to one another. Alternatively, the experiment in Figure 1B present a version of the experiment that better capitalizes on human spatial intuitions: participants are asked to use space to distinguish between all possible combinations, where larger separation between items reflects larger differences. Note that the absolute value of the ratings are identical between both iterations of the experiment.

While both versions require the participant to make the same number of (underlying) pairwise comparisons, we offer that experiment (B) is more informative than experiment (A), for a number of reasons. First, participants are able to concretize the relative relations. Rather than needing to maintain an implicit scale of differences in experiment (A) – which may lead to possible inconsistencies as the number of comparisons increases – the relations between items are explicitly visualized in a manner that is easy to manipulate. Second, the directionality of differences is transparently coded: while Item 4 is one away from both Item 2 and Item 3, experiment (B) easily captures the direction of the effects, whereas experiment (A) does not. Third, experiment (B) contextualizes the different items and their relative relations, allowing participants to quickly set the bounds of the underlying scale(s) that they are using to distinguish between items.

2.2. Space & Computational Models

Vector spaces that are generated by modern computational models of language are often used as proxies for human linguistic structure: for example, high-dimensional vectors for the words “cat” and “dog” are typically near one another in computational vector spaces, whether such vectors are computed using word co-occurrence statistics (e.g., Pennington et al., 2014) or more complex, contextual operations (e.g., Radford et al., 2019). Accordingly, close proximity in computational vector spaces[4] is often interpreted as human-like similarity,[5] at all levels of linguistic structure, including phonetic information (Parrish, 2017; Zouhar et al., 2023), phonological segments (Silfverberg et al., 2018), phrases (Passos et al., 2014), and others.

However, while this interpretation about the relationship between human and model representations holds true generally, prior work has noted some mismatches: for example, model representations have been shown to occupy a narrow region of the embedding space (a phenomenon known as *anisotropy*; Mimno and Thompson, 2017; Ethayarajh, 2019), have “rogue” dimensions that dominate similarity metrics (Timkey and van Schijndel, 2021), and fail to be robust to minor orthographic noise (Matthews et al., 2024). [6] Moreover, no current psycholinguistic methodology – to our knowledge – approximates human representational spaces of linguistic structure in a manner that is comparable to those generated by off-the-shelf computational models, making it difficult to align human and model representations.

2.3. Desiderata for GRIS

Given this overview of psycholinguistic paradigms and computational representations, we present the fundamental motivations behind GRIS:

1. A flexible experimental paradigm that uses space to construct meaningful, interpretable relations between objects.
2. A tool that allows researchers to quickly build experiments.
3. An experimental interface that participants can intuitively use.
4. Output data that approximates human cognitive representations that are easily aigned to representations from computational models.

[4]We use “proximity” as a catch-all term for similarity in the vector space, given that there are a variety of metrics – Euclidean distance, cosine similarity, Spearman’s ρ, etc. – to determine representational similarity.

[5]See Apidianaki (2023) for an overview.

[6]Some research in activation & representation engineering (e.g., Turner et al., 2023; Wu et al., 2024) demonstrates how these representations can be fine-tuned to perform better on down-stream tasks; we do not discuss these approaches in detail, though we do address them in the discussion.

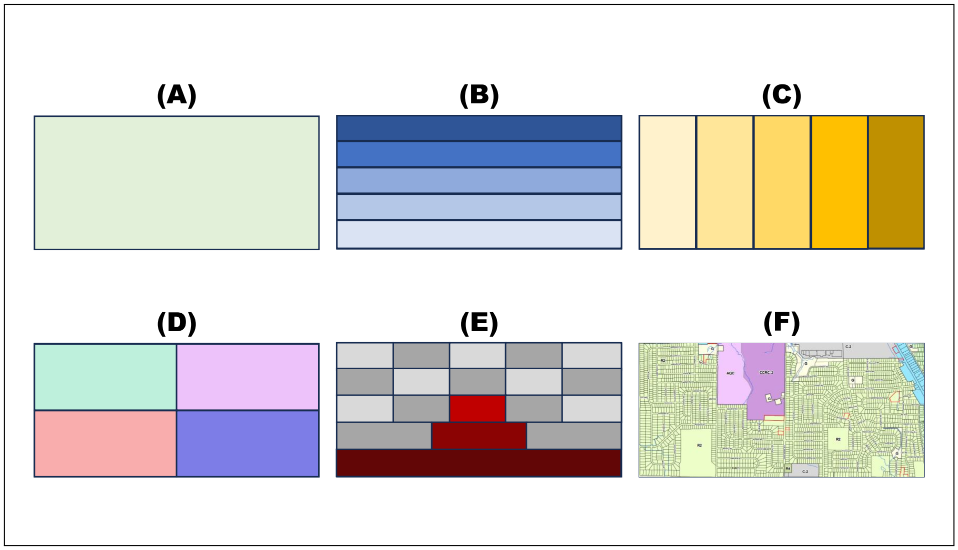

Figure 2: Sample GRIS canvases. Canvases can be blank (A), split into categories in both cartesian dimensions (B, C) simultaneously (D) and irregularly (E), or placed under an image (F).

3. GRIS: A Walkthrough

The core idea of the GRIS paradigm is to provide participants with objects that they can drag and drop onto a labeled canvas. In the following two subsections, we overview the structure of a GRIS experiment and demonstrate how participants navigate through a trial.

3.1. Structuring a GRIS Experiment

GRIS experiments have two fundamental components: 1) objects that can be placed, and 2) a canvas to place objects on.

Objects can be text, images, or audio; these objects are distinct targets that can be individually moved. By default, objects are located in a reservoir at the bottom of the screen, though their initial positions can be changed to accommodate relevant research questions.

Canvases, from a participant's perspective, can be either blank or split into different categories; some sample canvases are provided in Figure 2. From a researcher's perspective, canvases are built of individual, labeled canvas blocks that are either square or rectangular. By default, canvas blocks are labeled using a four-point coordinate system (*x-cat, y-cat, x-abs, y-abs*), where the first two dimensions are used to mark the category that the block belongs to, and the last two dimensions are used to mark the absolute position of the block on the overall canvas; note that canvas labels can be modified to accommodate other systems. Beyond labeling, each canvas block can be independently specified for height, width, and color. Finally, images can be overlaid on the canvas, allowing for additional designs beyond those possible by combinations of squares and rectangles.

For ease of use for other researchers, we have developed the GRIS toolkit,[7] which provides instructions on how to build, run, and analyze GRIS experiments.

3.2. Participating in a GRIS Experiment

GRIS experiments are designed to be simple and intuitive for participants. To explain how a participant navigates through a GRIS experiment, we provide a sample, partially-completed trial in Figure 3. In this sample trial, the participant has access to five objects – different shapes – which begin in the reservoir (B) below the blank canvas (A). The instructions at the top of the screen indicate that the participant should order the objects in a line, where the leftmost shapes are the "roundest" according to their intuitions. The participant first placed the star on the right boundary of the screen, then placed the oval on the left boundary; the abnormal shape was placed between these two shapes. Once they have placed all five objects on the canvas, the participant will be prompted to continue to the next trial, though they can continue to re-arrange the objects at any point in time throughout the trial.

For each drag-and-drop, GRIS collects 1) which object was moved, 2) the object's original location, 3) its new location, 4) and the timestamps for both the initial drag and the final drop. Data are also collected about when each trial begins and ends, as well as the final positions for all objects at the conclusion of each trial.

3.3. Interim Summary

In summary, GRIS is a simple – yet flexible – experimental paradigm that can accommodate a wide variety of research questions and designs. To

[7]The GRIS toolkit is publicly-available on GitHub at the following link: https://anonymous.4open.science/r/gris-toolkit-demo-923F/. Currently, GRIS experiments are run on PC Ibex (Zehr and Schwarz, 2018), a free, on-line research platform intended for experiments in psycholinguistics and cognitive science.

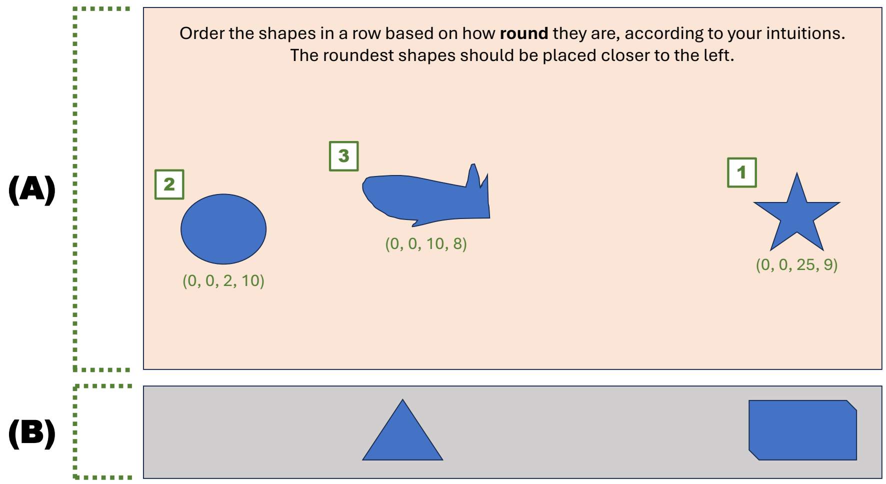

Figure 3: Sample of a partially-completed GRIS trial. Participants see a canvas (A) and a reservoir of objects (B); participants do not see anything marked in green. In this example, the objects are shapes. Participants are presented instructions – either during the trial (as in this figure) or prior to the trial – which guide how the participant should manipulate the objects. For each recorded drag-and-drop action, data is collected about the time/order of that action (boxed red numbers in the figure) and both the original and new coordinates for that object; only the dropped coordinates are presented in the figure.

validate GRIS' effectiveness and demonstrate the kinds of analyses that it permits, we present a series of three experiments which capitalize on many of the features offered by the paradigm.

4. Experiment 1: Sentence Acceptability

Acceptability judgments probe what structures are (un)acceptable in a language: these structures can range from low-level judgments of phonological structure to high-level judgments of multi-sentence, multi-speaker discourses. In this section, we focus on sentence acceptability judgments in English.

For explanatory purposes, consider the sentences in (1)-(3). We adopt standard conventions for marking degrees of acceptability and grammaticality from linguistic research, where * indicates a sentence is ungrammaticality, and # indicates a sentence is odd or slightly marked.

(1) *An girls is hungry.

(2) Randy wanted to write a novel.

(3) #?Want to write, Randy did a novel.

While ungrammatical sentences like (1) are rated toward the boundaries of the acceptability spectrum, others display more gradient judgments: for example, sentence (2) is often preferred over sentence (3), even though both are grammatical sentences of English. Previous research primarily collects acceptability judgments using Likert scales (Gibson et al., 2011), forced-choice tasks (Mahowald et al., 2016), or response times (Konieczny, 2000). In isolation, such sentence acceptability judgments appear to be robust across experimental paradigms, suggesting that people have consistent preferences about the internal structure of their language (Sprouse, 2011; Sprouse et al., 2013). However, these measures do not always capture the relative relationship of sentence acceptability across structures. For example, people express consistent preferences: generally speaking, (2) > (3) > (1). But, each of these pairwise preferences reflects a different underlying scale: while (1) is less acceptable than (3) and (3) is less acceptable than (2), the former distinction is motivated by differences in grammaticality, while the latter distinction is motivated by differences in frequency and syntactic complexity.

Moreover, isolated syntactic judgments may also conflate degrees of acceptability: a rating of 3 for one construction may not be comparable to a rating of 3 for another construction, even though the ratings are identical. Capturing the contextual organization of syntactic acceptability across phenomena would help us understand the broader organization of human language understanding and cognition.

In this study, we use GRIS to replicate large-scale sentence acceptability judgments from prior work, while also showing how the acceptability difference between sentence pairs can strongly vary depending on the context that they appear in.

4.1. Design & Procedure

4.1.1. Stimuli

All stimuli were drawn from Sprouse et al. (2013), which randomly sampled informal (i.e. not experimentally-tested) acceptability judgments of English sentence pairs from *Linguistic Inquiry*, a well-established journal in theoretical linguistics. After sampling these sentence pairs, Sprouse et al. (2013) collected acceptability ratings for each sentence within each pair to test whether the informal judgments were valid for larger populations; we will use these ratings to confirm that our findings correlate with prior work.

We sampled 72 pairs from the Sprouse et al. (2013) dataset. All 72 sentence pairs were classified according to the general linguistic phenomenon that their original paper tested; these classifications were drawn from the abstracts of the papers themselves. By labeling the linguistic phenomenon that each pair tests, we can then combine pairs of different classifications to understand how different syntactic phenomena influence sentence acceptability across structures, allowing us to obtain a broader understanding of the organizational preferences of acceptability judgments. Some sample classifications of phenomena are listed below in (4):

(4)
a. WORD ORDER:
Fred mowed the green lawn. > Fred mowed the lawn green.[8]
b. DEFINITES:
This is a table. > This is table.

From this set of 72 sentence pairs, we randomly selected 24 sentence pairs to serve as our target pairs: all participants saw each of these 24 sentence pairs. To test the impact of context on making these acceptability judgments, the remaining 48 items were broken into two sets of 24 sentences, each of which was paired with the 24 example items so that each target pair could appear in context with different phenomena. In sum, this process led to two sets of 24 items with four sentences (two pairs) each.

4.1.2. Procedure

See Figure 4 for a sample trial for Experiment 1. Participants saw four sentences below a gradiently-colored canvas, where the color gradient reflected a 5-point Likert scale. Participants were instructed to move the sentences from the bottom of the screen onto the canvas according to how "acceptable" the sentences were, according to their intuitions. Participants were told that the "most acceptable" sentences should be placed at the top of the canvas (5, on a standard Likert scale), while the "least acceptable" sentences should be placed at the bottom (1, on a standard Likert scale). They were

[8] While the example provided here does introduce a resultative construction, the primary arguments of the original paper discuss the construction's implications on word order.

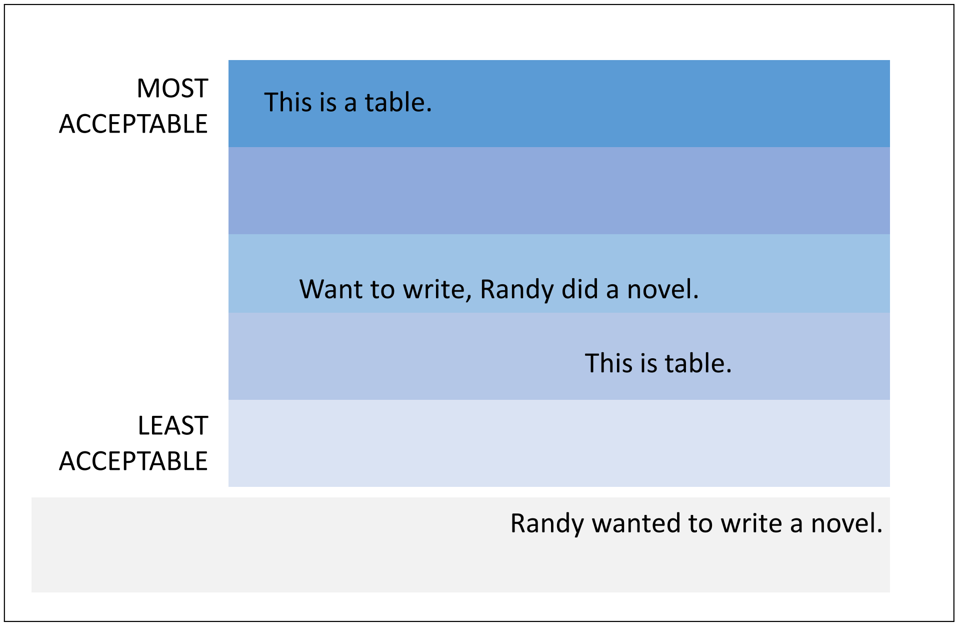

Figure 4: Sample trial for Experiment 1. Font has been enlarged for readability.

also told that multiple sentences could occupy the same level on the scale. Sentence positions below the canvas were randomized for each item.

4.1.3. Participants

Twenty-five participants were recruited using the online research platform Prolific. Participants were all native speakers of English between the ages of 18 and 55.

4.2. Results

4.2.1. Base Acceptability

To measure sentence acceptability judgments within each trial, we collected the final position of all sentences once the trial was complete. We z-scored acceptability ratings by participant to ensure that responses were compared on similar scales.

Results for Experiment 1 are visualized in Figure 5. To test whether unacceptable sentences were rated significantly lower than acceptable ones, we fit a linear mixed-effects model to the z-scored acceptability rating, with a fixed effect of sentence TYPE (acceptable/unacceptable), and random in-

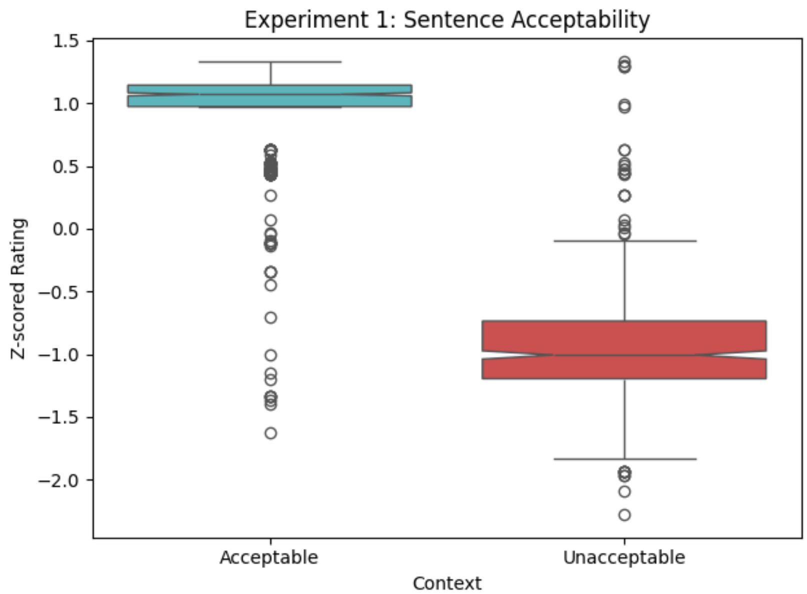

Figure 5: Base acceptability results for Experiment 1. Notches indicate 95% bootstrapped CIs.

tercepts for participants and items.[9] Participants rated the UNACCEPTABLE sentences as significantly less acceptable than the ACCEPTABLE ones ($\hat{\beta}$ = -0.184, SE = 0.031, t=-58.80, p <0.001); these sentence ratings also strongly correlate (r=0.88) with those found by Sprouse et al. (2013).

4.2.2. Contextual Acceptability

In addition to the basic acceptability analyses in the previous section, we measured how acceptability differences varied within each target pair according to the classification of the context pair that was present in the trial. To do so, we calculated the difference between each sentence in the target pair, then averaged the ratings within each context classification.

Results for contextual acceptability differences are shown in Figure 6. We find that some phenomena display similar levels of acceptability (< 0.4 Likert difference) regardless of context (e.g., *Agreement, Definites*), while others show significant variation (e.g., *Movement, Word Order, Clause*). For exam-

[9]The complete model formula was: Z-SCORED RATING ∼ TYPE + (1 | item) + (1 | participant). The baseline was the "Acceptable" condition.

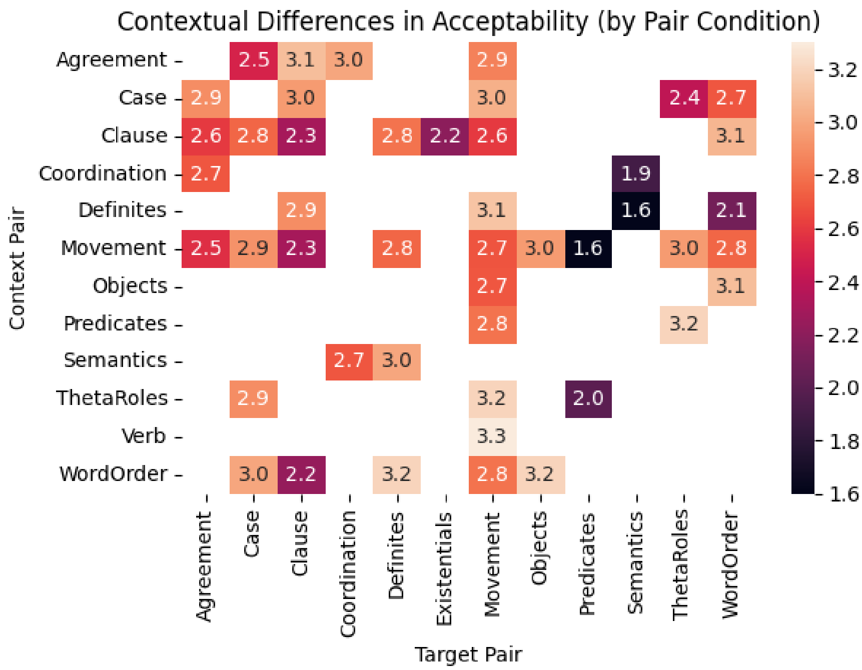

Figure 6: Contextual acceptability results for Experiment 1. X-axis represents the classification for the target pair. Y-axis represents the classification of the context pair. Cells indicate difference between acceptable and unacceptable sentences within each target pair; darker colors indicate smaller differences on a 5-point Likert scale.

ple, consider the *Word Order* classification for the target pair from (4-a): *Fred mowed the green lawn* > *Fred mowed the lawn green*. When placed in the context of a sentence pair that modulates *Definites*, the difference between the *green lawn* and *lawn green* sentences was approximately 2.1 on a 5-point Likert scale; but, when placed in the context of a sentence pair that modulates *Object*hood, the difference between the *green lawn* and *lawn green* sentences was approximately 3.1. These varying differences have significant consequences on how researchers interpret acceptability judgments: a difference of ∼3 points on a 5-point Likert scale easily distinguishes an acceptable sentence (5) from an unacceptable one (2), whereas a difference of ∼2 points could be the distinction between a totally acceptable sentence (5) and a moderately acceptable one (3).

4.3. Discussion

The results of this task show that GRIS can be used to reliably replicate prior experimental results involving pairwise comparisons, while also systematically capturing the variability of sentence acceptability in different contexts. More specifically, GRIS reveals how previous sentence acceptability judgments in isolation may not serve as reliable representations of overall sentence acceptability in context.

5. Experiment 2: Category Typicality

Category typicality assesses how “typical” an object is within a broader category (Rosch, 1975; Farmer et al., 2006). For example, “robins” and “sparrows” are found to be more typical representations of birds than “toucans” and “penguins” across cognitive domains, including language (Rosch, 1975; Meints et al., 1999) and vision (Maxfield et al., 2014). Traditionally, category typicality has been measured using rating or decision tasks (Rosch, 1975), production tasks (Rosch et al., 1976), or inductive-reasoning tasks (Osherson et al., 1990), all of which ask the participant to consider a specific word in relation to the broader category label. Recent computational work also suggests that computational models of language may learn some aspects of category typicality from the statistical usage distributions of everyday language (Misra et al., 2021), though these analyses focus on probability estimates from pre-trained language models rather than representational analyses.

In this experiment, we build a typicality-rating experiment using GRIS, finding that manipulating words in space both 1) replicates previous category typicality effects and 2) allows us to directly compare representational spaces between humans and models.

5.1. Design & Procedure

5.1.1. Stimuli

We used eight of the original ten categories from Rosch (1975): *fruits, vehicles, weapons, vegetables, tools, birds, sports, clothing.* All items were in English. Each category has a list of approximately 50-60 words, where each word has a typicality rating that was averaged across 209 subjects; we use these ratings as our ground truth. To test whether the presence of different words modified typicality ratings, we constructed eight items that used ten words from each category; we did not use all of the words from Rosch (1975), as there would be too many words for participants to move on the screen.

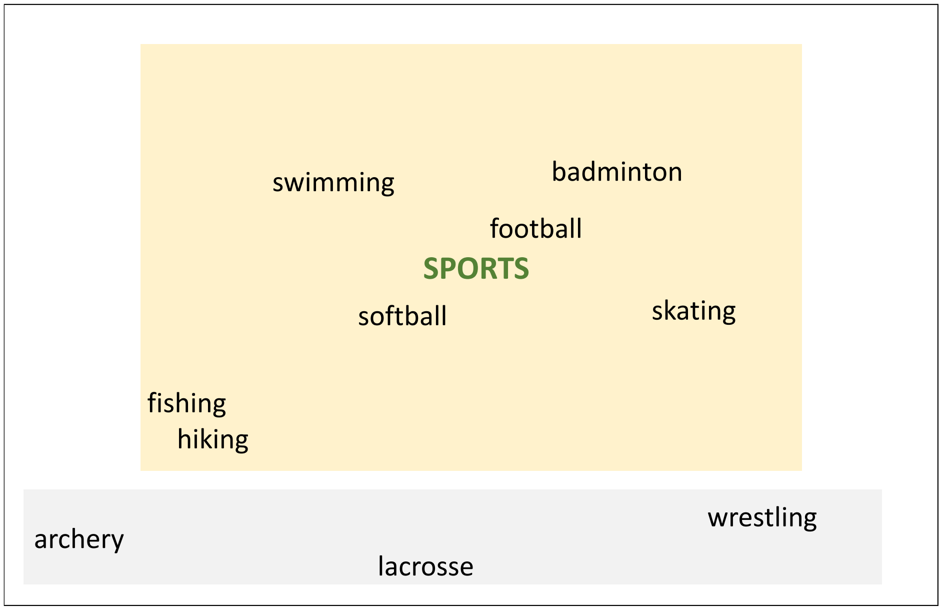

Figure 7: Sample trial for Experiment 2 (Typicality); font size enlarged to improve figure readability. Category label is marked in the center in green.

5.1.2. *Procedure*

A sample item for Experiment 2 is visualized in Figure 7. Participants saw a canvas with a word bank below. In the middle of the canvas was a bolded category label (i.e. **SPORTS**). Participants were told to move words from the bank onto the canvas according to how "typical" an example the word was of the category: words that were more typical examples of the category should be placed closer to the category label.

5.1.3. *Participants*

As in Experiment 1, twenty-five participants were recruited using the online research platform Prolific. Participants were all native English speakers between the ages of 18 and 55.

5.2. *Results*

As in Experiment 1, we collected the final positions for all words once the trial was complete. For each trial, we calculated every word's distance from the center; we z-scored these distances by participant to ensure that all participants were comparable in how they used the space. Finally, following

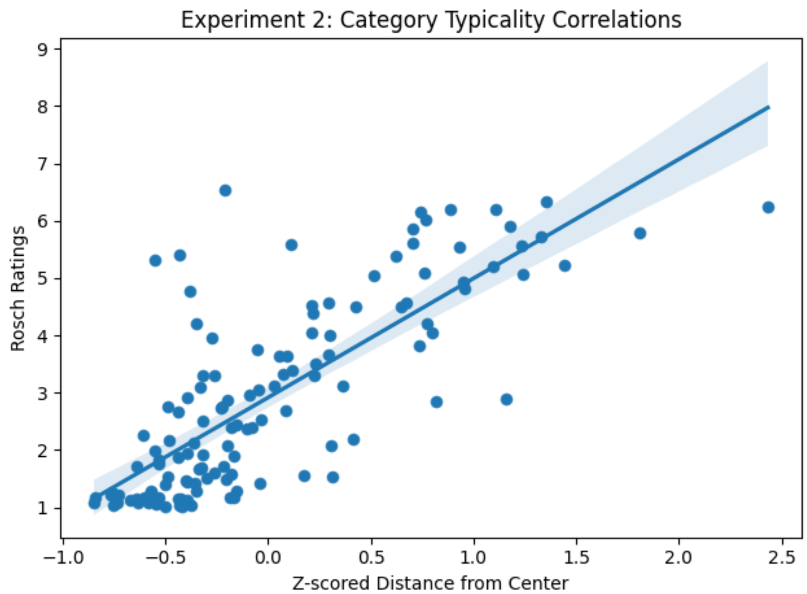

Figure 8: Correlation results for Experiment 2. X-axis indicates the Z-scored distance from center for a word. Y-axis indicates the original ratings from Rosch (1975).

the rating averaging from Rosch (1975), we meaned the distances for each word across participants.

Experimental results are visualized in Figure 8. We find a strong correlation (r= 0.78) between the original rankings from Rosch (1975) and the distance of each word from its category label in our study, indicating that GRIS can be used to replicate prior category typicality results.

5.2.1. Computational Analyses

For our computational analyses, we extracted vector representations of words from three models: GLoVe 6B.300D (Pennington et al., 2014), BERT (Devlin, 2018), and GPT2 (Radford et al., 2019). For the non-contextual model (GLoVe), we gathered the raw vectors for both the word and the category label. Following Misra et al. (2021), for both of the contextual models (BERT & GPT2), we framed each word X with its category label Y in the following way: *A(n) X is a typical Y.*; instead of gathering the probability of each word X in the sentence, we extracted the vector representations of both the word and the label using the `minicons` Python package (Misra, 2022). Approaching our computational analyses in this way allows us to

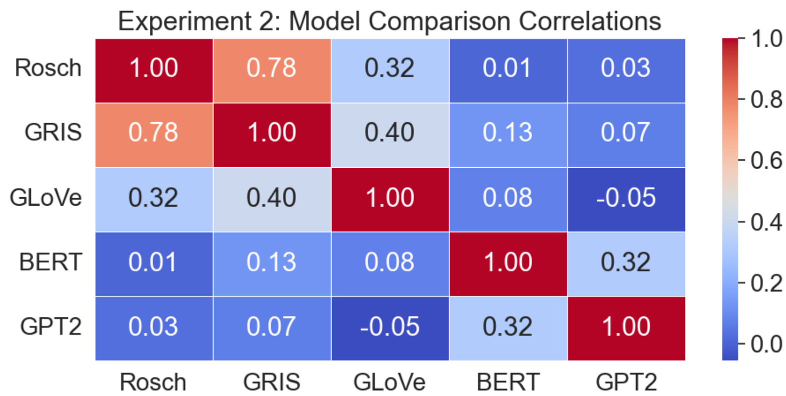

Figure 9: Correlation metrics between model representations and experimental results. Each cell corresponds to the Pearson's correlation coefficient between the models and experimental measures on the x- and y-axes.

most directly compare the representational spaces constructed in the human experiment with those generated by computational models of language; our approach differs from that of Misra et al. (2021), in directly comparing model similarities to human similarity judgments rather than mapping model log-probabilities to human behavioral responses.

For each of the three models, we computed the Euclidean distance between the vectors for every word and its corresponding category label.[10] We then calculated the Spearman's correlation for all possible model comparisons.

Results for these multiple-correlation analyses are visualized in Figure 9. We find that GRIS is the only set of representations that connect a word to its category label in a manner that strongly correlates with the original rankings from Rosch (1975); the distances between words and their labels for GLoVe representations only weakly correlate with the original Rosch rankings, though there is a slightly stronger correlation between GloVe distances and our experimental data. We note that representational distances in BERT and GPT2 weakly correlate with one another, but fail to display any strong correlations with GLoVe or either set of experimental data. We also note that

[10]Analyses using standardized cosine similarity and Spearman's rank correlation coefficient were also conducted; Euclidean distance performed best in the correlation analyses.

GRIS also has the highest average correlation coeffecient across comparisons.

5.3. Discussion

In this experiment, we replicated prior typicality representations for eight categories. Experiments 1 and 2 show how GRIS can reliably replicate prior results; this experiment also demonstrates how GRIS builds constructs representational spaces more accurately than a number of well-established computational models. These findings differ from Misra et al. (2021), likely due to the fact that we are conducting *representational* analyses and not *behavioral* ones: while previous computational work has shown that behavioral measures moderately align with human behavior, our work demonstrates that studies of human representations cannot simply rely on vectors generated by these models.

6. Experiment 3: Multi-dimensional Similarity

In the previous two experiments, we demonstrated how GRIS can be used to both replicate and provide further detail about prior studies. In this experiment, we showcase how GRIS can be used to advance new questions within an established literature in cognitive science: pattern recognition.

For decades, cognitive scientists have studied how people recognize patterns across a variety of cognitive domains (Chater and Vitányi, 2003; Reed, 1972; Edelman, 1999; Edelman and Duvdevani-Bar, 1997). We contribute to this literature by examining how one form of pattern recognition – similarity assessments – arises during language processing.

Prior work suggests that the cognitive sources of similarity are a concept's familiarity (strength in memory), association (relationships with other concepts), and inherent perceptual likeness (surface appearance); see Hiatt and Trafton (2017) for an overview. Linguistic similarity, broadly defined, has also been shown to influence pattern recognition. For example, semantic similarity is well-known to produce priming effects (McNamara, 2005; Neely et al., 1989; Shelton and Martin, 1992), and, while less studied, syntactic similarity has shown similar effects (Lester et al., 2017). Orthographic similarity improves recall accuracy in a probed serial-recall task (Lin et al., 2015), and phonological similarity has been shown to facilitate the learning of novel words (Papagno and Vallar, 1992).

While each of these features contributes to overall perception of similarity between linguistic units, how do people balance the multiple avenues of sim-

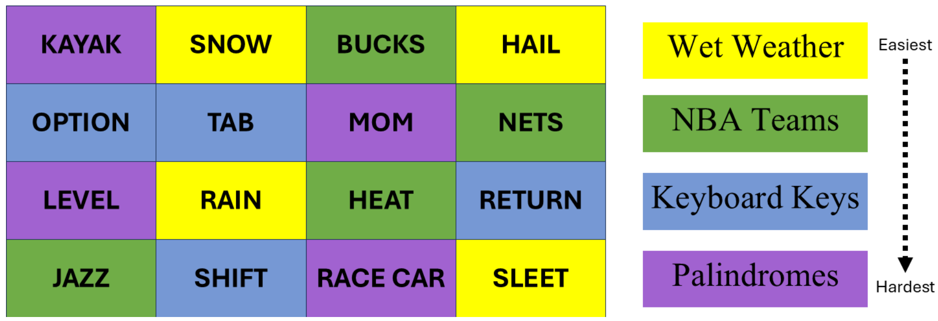

Figure 10: Sample Connections puzzle (left) with categories (right); puzzle in original format does not have colors. Colors reflect difficulty, as determined by the editors of the publication: yellow is the easiest, green is the second-easiest, blue is the second-hardest, and purple is the hardest.

ilarity to determine a single sense of similarity? Importantly, this research question would be difficult to test with standard paradigms, as it involves significant numbers of pair-wise comparisons that would be both costly to run and difficult to interpret. In this experiment, we demonstrate how the drag-and-drop functionality of GRIS-based experiments easily allows us to determine how different types of similarity are represented and prioritized among each other.

6.1. Stimuli

Materials for this experiment come from *Connections*, a free, publicly-available game hosted by *The New York Times*. In this game, players see a grid of 16 words and are told to separate the words into four distinct groups that are labeled; each item belongs to only one group. Importantly, each group of four words forms a labeled category, and these categories have varying difficulty: yellow groups are the easiest, green groups the second-easiest, blue groups the second-hardest, and purple groups the most difficult.[11] A sample item and its corresponding solution are shown in Figure 10.

For 300 puzzles, two annotators categorized each group of words into one of three broader similarity categories: *Semantic Association* (e.g., "wet weather": hail, rain, sleet, snow), *World Knowledge* (e.g., "NBA teams":

[11] These difficulties are suggested by *The New York Times*; we do not focus on whether these difficulties are accurate, instead studying the cognitive question surrounding similarity comparisons.

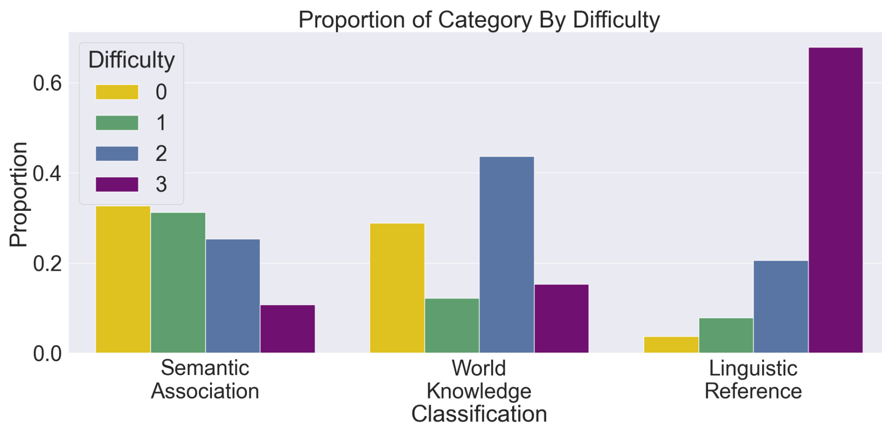

Figure 11: Distribution of similarity categories by difficulty. Difficulty levels closer to 0 are considered easier.

bucks, heat, jazz, nets), and *Linguistic Reference* (e.g., "palindromes": kayak, level, mom, race car). As visualized in Figure 11, we see that indeed some similarities are considered more difficult than others: semantic association groups tend to occupy the easier categories, world knowledge groups tend to occupy the middle difficulties, and abstract linguistic reference groups tend to occupy the most challenging difficulties.

6.2. Design & Procedure

6.2.1. Stimuli

From our annotated data, we selected 10 puzzles that had at least two of the similarity categories. Given that we are using puzzles generated by the publication, we were unable to perfectly balance the different similarity categories across all puzzles.[12]

6.2.2. Procedure

Similar to Experiment 2, participants saw a blank canvas with a word bank of words below. Participants were instructed to move these words onto the canvas according to how similar they were; similar words should be placed

[12]Instead, categories were balanced to be approximately 40% semantic association, 30% world knowledge, and 30% linguistic reference.

closer together. Participants were instructed to use as much of the canvas as they felt was appropriate.

To train them on the task but to avoid biasing their decisions, participants completed two practice trials prior to the experiment where they grouped both shapes and numbers.

6.2.3. Participants

Nineteen native speakers of English between the ages of 18 and 55 were recruited on Prolific.

6.3. Results

For each trial, we collected the final position for all words. For every group within each trial, we computed two distance comparisons. WITHIN GROUP distances were computed by calculating the average distance between every word within each group with other members of that same group. OUTSIDE GROUP distances were computed by calculating the average distance between every word within a group with every other word not in that group.

Results are visualized in Figure 12. To determine how people used distance to group similar words together, we fit a linear mixed-effects regression model that predicted DISTANCE, with fixed effects of COMPARISON (within group/outside group), CATEGORY (semantic association/world experience/linguistic reference), and their full interactions, along with random intercepts for participants, items, and puzzle difficulty.[13] We find a main effect of COMPARISON, such that WITHIN GROUP comparisons are significantly closer together than OUTSIDE GROUP comparisons ($\hat{\beta}$ = -2.323, SE = 0.772, t=-3.263, p <0.01). Additionally, we report a significant interaction between COMPARISON and CATEGORY, such that SEMANTIC ASSOCIATION groups clustered significantly closer together than LINGUISTIC REFERENCE groups in the WITHIN GROUP comparison ($\hat{\beta}$ = -3.085, SE = 0.884, t=-3.491, p <0.001).

6.4. Discussion

In this experiment, we showed that certain similarity patterns are easier to find than others. More specifically, this experiment showed that groups

[13]The complete model formula was: DISTANCE ∼ COMPARISON*CATEGORY + (1 | item) + (1 | participant) + (1 | difficulty). The baseline conditions were the OUTSIDE GROUP and LINGUISTIC REFERENCE groups, respectively.

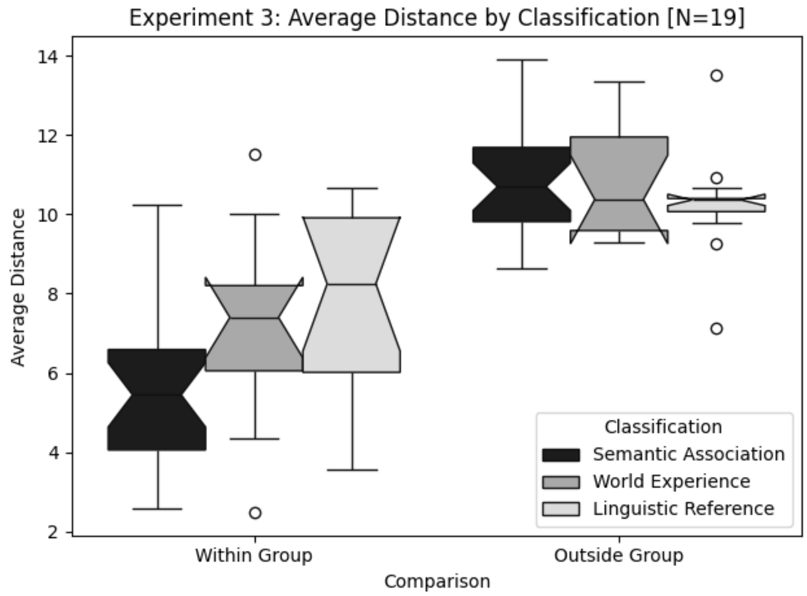

Figure 12: Average distance by category for Experiment 3. Notches indicate bootstrapped 95% CIs.

of words that pattern according to semantic association are easiest to find. These findings may derive from the fact that semantic association requires less reasoning to identify possible clusters of words, compared to other, more abstract groupings.

Beyond these results, we argue that the drag-and-drop paradigm of GRIS-based experiments works well to investigate the complex relationships between representations and reasoning: other paradigms – including rating tasks, forced-choice tasks, and priming tasks – would require significantly less transparent pairwise comparisons to accomplish the results of this study.

7. General Discussion

In this paper, we have shown how GRIS allows researchers across the cognitive sciences to use space as a way of approximating human representational spaces, allowing experimenters to model representational spaces both within class (Experiment 2) and across classes (Experiment 3), while also providing information about the relative relationships between objects on the grid across participants. These findings align with prior work which demonstrate how similarity and difference is highly individualized (Simmons and Estes, 2008).

Additionally, we have shown how GRIS can the use of space can easily contextualize psycholinguistic findings: we found that acceptability differences between sentence pairs can vary greatly according to the context that they appear in (Experiment 1). We hope that future work using GRIS can expand the relative comparisons between different stimuli modalities (e.g., text, image, audio).

7.1. What kinds of analyses does GRIS support?

In this subsection, we introduce four broad categories for analyzing future GRIS data, each of which are tied to specific kinds of research questions. These broad categories are:

1. Location-based
2. Graph-based
3. Timing-based
4. Trial-based

Location-based analyses suit questions about ordering or categorical distinctions between objects. For example, the sample trial in Figure 3 studies the linear order of shapes, where each object's position on the x-axis reflects the object's relative roundness, according to the participant: as a result, an analysis for this sample trial would likely focus on the y-axis information for each object, unless otherwise specified in the question. Canvases with categorical splits – like those in Figure 2(B)-(E) – also likely use location-based analyses. We demonstrated location-based analyses in Experiments 1 and 2.

Graph-based analyses fit questions that investigate the relative relationship between objects. Given that the tool collects information about the individual position of each object over the course of the trial, each GRIS trial builds a fully-connected weighted graph, where each object is a node, and the distance between two objects serves as the weighted edge between these objects. For example, a graph-based analysis would align with an experiment involving unsupervised clustering of objects. We demonstrated a graph-based analysis in Experiment 3.

Timing-based analyses address questions that involve the order of individual movements and how long each movement took. For example, a timing-based analysis could indicate which objects were most salient to participants (i.e. which objects were moved first), or whether certain objects were more difficult to place (i.e. took longer to drop) in relation to the relevant research question.

Finally, trial-based analyses address questions about participant- and item-level behaviors. For example, a trial-based analysis might study whether people how similar representational spaces are between people; an analysis of this kind might construct a large-scale network of object relations for each participant, and then apply transformations to such networks to determine if certain clusters emerge across participants.

8. Why Use GRIS?

We conclude the paper by collecting our broader arguments for how GRIS can help further our understanding of the human mind.

First, GRIS relies on natural human intuitions around space to build contextual and interpretable approximations of cognitive representations. In this paper, we demonstrate three possible ways that space can be meaningfully used to advance questions in the cognitive sciences; we hope that future work further develops this approach to understanding the mind.

Second, as has been mentioned previously, GRIS is very flexible and can be used to answer a range of questions in the cognitive sciences; the paradigm provides a sandbox for both researchers and participants alike to play in. GRIS is supported for desktop, laptops, and tablets.

Third and finally, GRIS creates multi-dimensional representations that are easily comparable to popular computational models of language, such as Large Language Models (LLMs). These representations can be used to further explore mismatches between humans and models to help understand what aspects of human cognition are not determinable from data alone.

In summary, we note the centrality of spatial reasoning and language to cognition, and how unifying them can 1) make an experiment more intuitive, 2) yield more holistic and contextually-relevant results, and 3) construct representations that facilitate comparisons between humans and computational models.

References

Apidianaki, M., 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. Computational Linguistics 49, 465–523. doi:10.1162/coli_a_00474.

Bryant, D.J., 1997. Representing space in language and perception. Mind & Language 12, 239–264. doi:10.1111/j.1468-0017.1997.tb00073.x.

Chater, N., Vitányi, P., 2003. Simplicity: a unifying principle in cognitive science? Trends in Cognitive Sciences 7, 19–22. doi:10.1016/s1364-6613(02)00005-0.

Cooper, R.M., 1974. The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology 6, 84–107. doi:10.1016/0010-0285(74)90005-X.

Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Edelman, S., 1999. Representation and recognition in vision. MIT Press.

Edelman, S., Duvdevani-Bar, S., 1997. A model of visual recognition and categorization. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 352, 1191–1202. doi:10.1098/rstb.1997.0102.

Ethayarajh, K., 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings, in: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 55–65. URL: https://aclanthology.org/D19-1006/, doi:10.18653/v1/D19-1006.

Farmer, T.A., Christiansen, M.H., Monaghan, P., 2006. Phonological typicality influences on-line sentence comprehension. Proceedings of the National Academy of Sciences 103, 12203–12208. doi:10.1073/pnas.0602173103.

Fauconnier, G., 1994. Mental spaces: Aspects of meaning construction in natural language. Cambridge University Press.

Fauconnier, G., et al., 2007. Mental spaces. The Oxford Handbook of Cognitive Linguistics , 351–376.

Forster, K.I., Guerrera, C., Elliot, L., 2009. The maze task: Measuring forced incremental sentence processing time. Behavior Research Methods 41, 163–171. doi:10.3758/BRM.41.1.163.

Freeman, J.B., Ambady, N., 2010. Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods 42, 226–241. doi:10.3758/BRM.42.1.226.

Gibson, E., Piantadosi, S., Fedorenko, K., 2011. Using mechanical turk to obtain and analyze english acceptability judgments. Language and Linguistics Compass 5, 509–524. doi:10.1111/j.1749-818X.2011.00295.x.

Hayward, W.G., Tarr, M.J., 1995. Spatial language and spatial representation. Cognition 55, 39–84. doi:10.1016/0010-0277(94)00643-y.

Hiatt, L.M., Trafton, J.G., 2017. Familiarity, priming, and perception in similarity judgments. Cognitive Science 41, 1450–1484. doi:10.1111/cogs.12418.

Just, M.A., Carpenter, P.A., Woolley, J.D., 1982. Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General 111, 228. doi:10.1037//0096-3445.111.2.228.

Kemmerer, D., 1999. "near" and "far" in language and perception. Cognition 73, 35–63. doi:10.1016/s0010-0277(99)00040-2.

Kirsh, D., 1995. The intelligent use of space. Artificial intelligence 73, 31–68. doi:10.1016/0004-3702(94)00017-U.

Konieczny, L., 2000. Locality and parsing complexity. Journal of Psycholinguistic Research 29, 627–645. doi:10.1023/a:1026528912821.

Landau, B., Jackendoff, R., 1993. Whence and whither in spatial language and spatial cognition? Behavioral and Brain Sciences 16, 255–265. doi:10.1017/S0140525X00029927.

Lester, N., Feldman, L., del Prado Martín, F.M., 2017. You can take a noun out of syntax...: Syntactic similarity effects in lexical priming., in: CogSci.

Lin, Y.C., Chen, H.Y., Lai, Y.C., Wu, D.H., 2015. Phonological similarity and orthographic similarity affect probed serial recall of chinese characters. Memory & Cognition 43, 538–554. doi:10.3758/s13421-014-0495-x.

Mahowald, K., Hartman, J., Graff, P., Gibson, E., 2016. Snap judgments: A small n acceptability paradigm (snap) for linguistic acceptability judgments. Language , 619–635doi:10.1353/lan.2016.0052.

Matthews, J., Starr, J., Schijndel, M., 2024. Semantics or spelling? probing contextual word embeddings with orthographic noise, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 4495–4504. URL: https://aclanthology.org/2024.findings-acl.266/, doi:10.18653/v1/2024.findings-acl.266.

Maxfield, J.T., Stalder, W.D., Zelinsky, G.J., 2014. Effects of target typicality on categorical search. Journal of Vision 14, 1–1. doi:10.1167/14.12.1.

McNamara, T.P., 2005. Semantic priming: Perspectives from memory and word recognition. Psychology Press.

Meints, K., Plunkett, K., Harris, P.L., 1999. When does and ostrich become a bird? the role of typicality in early word comprehension. Developmental Psychology 35, 1072. doi:10.1037//0012-1649.35.4.1072.

Mimno, D., Thompson, L., 2017. The strange geometry of skip-gram with negative sampling, in: Palmer, M., Hwa, R., Riedel, S. (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark. pp. 2873–2878. URL: https://aclanthology.org/D17-1308/, doi:10.18653/v1/D17-1308.

Misra, K., 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. arXiv preprint arXiv:2203.13112 .

Misra, K., Ettinger, A., Rayz, J.T., 2021. Do language models learn typicality judgments from text? Proceedings of the Annual Meeting of the Cognitive Science Society 43.

Neely, J.H., Keefe, D.E., Ross, K.L., 1989. Semantic priming in the lexical decision task: roles of prospective prime-generated expectancies and retrospective semantic matching. Journal of Experimental Psychology: Learning, Memory, and Cognition 15, 1003. doi:10.1037//0278-7393.15.6.1003.

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., Shafir, E., 1990. Category-based induction. Psychological Review 97, 185. doi:10.1037/0033-295X.97.2.185.

Papagno, C., Vallar, G., 1992. Phonological short-term memory and the learning of novel words: The effect of phonological similarity and item length. The Quarterly Journal of Experimental Psychology 44, 47–67. doi:10.1080/14640749208401283.

Parrish, A., 2017. Poetic sound similarity vectors using phonetic features, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, pp. 99–106.

Passos, A., Kumar, V., McCallum, A., 2014. Lexicon infused phrase embeddings for named entity resolution. arXiv preprint arXiv:1404.5367 .

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Reed, S.K., 1972. Pattern recognition and categorization. Cognitive Psychology 3, 382–407. doi:10.1016/0010-0285(72)90014-X.

Rosch, E., 1975. Cognitive representations of semantic categories. Journal of Experimental Psychology: General 104, 192. doi:10.1037/0096-3445.104.3.192.

Rosch, E., Simpson, C., Miller, R.S., 1976. Structural bases of typicality effects. Journal of Experimental Psychology: Human Perception and Performance 2, 491. doi:10.1037/0096-1523.2.4.491.

Shelton, J.R., Martin, R.C., 1992. How semantic is automatic semantic priming? Journal of Experimental Psychology: Learning, Memory, and Cognition 18, 1191. doi:10.1037//0278-7393.18.6.1191.

Silfverberg, M.P., Mao, L., Hulden, M., 2018. Sound analogies with phoneme embeddings. Society for Computation in Linguistics 1.

Simmons, S., Estes, Z., 2008. Individual differences in the perception of similarity and difference. Cognition 108, 781–795. doi:10.1016/j.cognition.2008.07.003.

Sprouse, J., 2011. A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. Behavior Research Methods 43, 155–167. doi:10.3758/s13428-010-0039-7.

Sprouse, J., Schütze, C.T., Almeida, D., 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. Lingua 134, 219–248. doi:10.1016/j.lingua.2013.07.002.

Sweetser, E., 1999. Compositionality and blending: semantic composition in a cognitively realistic framework. Cognitive Linguistics: Foundations, Scope, and Methodology 129162.

Talmy, L., 1983. How language structures space, in: Spatial Orientation: Theory, Research, and Application. Springer, pp. 225–282.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634. doi:10.1126/science.7777863.

Taylor, H.A., Tversky, B., 1992. Spatial mental models derived from survey and route descriptions. Journal of Memory and language 31, 261–292. doi:10.1016/0749-596X(92)90014-O.

Timkey, W., van Schijndel, M., 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality, in:

Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 4527–4546. URL: https://aclanthology.org/2021.emnlp-main.372/, doi:10.18653/v1/2021.emnlp-main.372.

Turner, A.M., Thiergart, L., Leech, G., Udell, D., Vazquez, J.J., Mini, U., MacDiarmid, M., 2023. Activation addition: Steering language models without optimization. arXiv e-prints , arXiv–2308.

Tversky, B., Lee, P.U., 1998. How space structures language, in: Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge. Springer, pp. 157–175.

Wilcox, E.G., Ding, C., Sachan, M., Jäger, L.A., 2024. Mouse tracking for reading (motr): A new naturalistic incremental processing measurement tool. Journal of Memory and Language 138, 104534. doi:10.3929/ethz-b-000676855.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C.D., Potts, C., 2024. Reft: Representation finetuning for language models. Advances in Neural Information Processing Systems 37, 63908–63962.

Zehr, J., Schwarz, F., 2018. PennController for Internet Based Experiments (IBEX).

Zouhar, V., Chang, K., Cui, C., Carlson, N., Robinson, N., Sachan, M., Mortensen, D., 2023. Pwesuite: Phonetic word embeddings and tasks they facilitate. arXiv preprint arXiv:2304.02541 .