Generating Representations In Space with GRIS

John R. Starr (jrs673@cornell.edu)
Department of Linguistics, Cornell University

Ashlyn Winship (alw329@cornell.edu)
Department of Linguistics, Cornell University

Marten van Schijndel (mv443@cornell.edu)
Department of Linguistics, Cornell University

Abstract

When conducting experimental research, the research questions are often inherently linked (and limited) to the paradigm that is used. In this paper, we present a new experimental tool – GRIS (Generating Representations in Space) – that builds experiments where participants can manipulate objects on a screen. Through a series of three experiments on sentence acceptability, category typicality, and multi-dimensional similarity, we demonstrate how GRIS-based experiments allow cognitive scientists to approximate representational spaces for a variety of cognitive phenomena, expanding the set of possible research questions that cognitive scientists may ask.

Keywords: acceptability, typicality, similarity, paradigm, representations

Introduction

For decades, cognitive scientists have conducted experiments to further our understanding of the mind, using a variety of behavioral paradigms (e.g., Stroop task, memory tasks, reading tasks, etc.), brain-imaging techniques (e.g., EEG, fMRI), and computational models (e.g., recurrent networks, Bayesian models, etc.), among others. However, such researchers are constrained by the tools, methodologies, and resources that are available to them. For example, while vector representations of computational models of language are often assumed to approximate linguistic meaning in humans,[1] we do not have access to the comparable representations that humans may use. **In this paper, we introduce a new experimental tool that approximates representational spaces for cognitive phenomena.**

GRIS[2] (Generating Representations In Space) builds web-based human experiments which allow participants to explicitly construct representational spaces for many different kinds of stimuli. As we will show in a series of three experiments, GRIS-based experiments can be used to both replicate prior studies and allow for novel investigations that would have been challenging to construct using established behavioral paradigms.

An Overview of GRIS

In simplest terms, GRIS is a tool that builds experiments where participants drag-and-drop objects on a canvas. Objects can be text, audio, and images; multi-modal designs (e.g., experiments that move both text and images in the same trial) are also supported. Both objects and the canvas can be labeled according to the researcher's relevant question: we demonstrate how canvases can be both region-defined (Experiment 1) and position-defined (Experiments 2 & 3). Importantly, GRIS can accommodate a wide variety of experimental paradigms and designs beyond those proposed in this paper: we discuss some possible avenues of investigation in the General Discussion section.

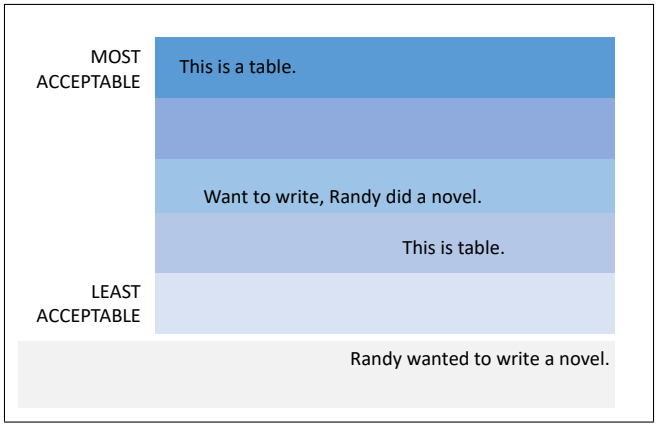

Figure 1: Sample trial for Experiment 1 (Acceptability); font size enlarged to improve readability. More acceptable sentences should be placed towards the top; less acceptable sentences should be placed near the bottom.

In the following three sections, we describe a series of three experiments which display how GRIS can be used to better approximate cognitive representational spaces: Experiment 1 studies sentence acceptability, Experiment 2 investigates category typicality, and Experiment 3 probes multi-dimensional similarity relationships.

Experiment 1: Sentence Acceptability

Sentence acceptability judgments probe what sentences are (un)acceptable in a language. While some sentences are rated toward the boundaries ("*An girls is hungry"), others display gradience: "Randy wanted to write a novel" is judged as more acceptable than "Want to write, Randy did a novel", even though both are acceptable. Prior work mostly

[1] See Wang, Wang, Chen, Wang, and Kuo (2019) for an overview.

[2] Use the GRIS-toolkit to build GRIS experiments: https://github.com/johnstarr-ling/gris-toolkit. Currently, GRIS builds experiments for PC Ibex (Zehr & Schwarz, 2018).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

collects acceptability judgments using Likert scales (Gibson, Piantadosi, & Fedorenko, 2011), forced-choice tasks (Mahowald, Hartman, Graff, & Gibson, 2016), or response times Konieczny (2000): in isolation, sentence acceptability judgments appear to be robust across experimental paradigms (Sprouse, 2011; Sprouse, Schütze, & Almeida, 2013). However, these measures do not always capture the relative relationship between sentences: in a forced-choice task, people express consistent preferences ("Randy wanted…" > "Want to…" > "*An girls…"), but each pairwise preference may reflect a different underlying scale.

In this study, we use GRIS to replicate sentence acceptability judgments from prior work, while also showing how placing sentence pairs in different contexts with other sentence pairs can significantly alter their acceptability ratings.

Design & Procedure

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 linguistics journal. 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 determining the linguistic phenomenon that each pair tests, we can then combine pairs of different classifications to understand how different syntactic phenomena influence sentence acceptability. Sample classifications are: *Word Order, Definites, Movement*.

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. 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.

Procedure See Figure 1 for a sample trial for Experiment 1. Participants saw four sentences below a gradiently-colored canvas, where the color gradient reflects 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 also told that multiple sentences could occupy the same level on the scale. Sentence positions below the canvas were randomized for each item.

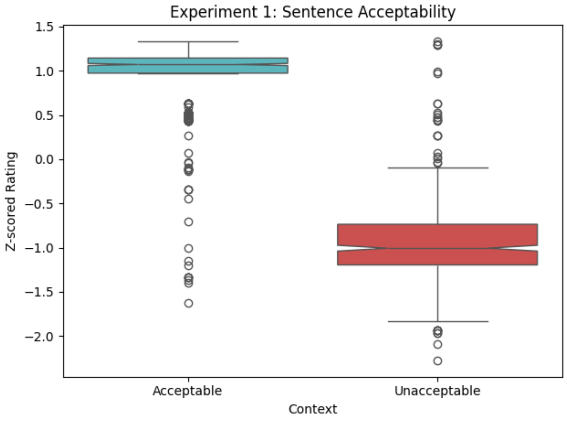

Figure 2: Base acceptability results for Experiment 1. Notches indicate 95% bootstrapped CIs.

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.

Results

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 participants were being compared on similar scales.

Results for Experiment 1 are visualized in Figure 2. To test whether acceptable sentences were rated significantly higher than unacceptable 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 intercepts for participants and items.[3] We find that participants rated the ACCEPTABLE sentences as significantly more acceptable than the UNACCEPTABLE ONES (β = -0.184, SE = 0.031, t=-58.80, p <0.001); these sentence ratings also strongly correlate with those found by Sprouse et al. (2013, r=0.88).

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 3. We find that some phenomena display similar levels of acceptability (<0.5 difference) regardless of context (e.g. *Agreement, Definites*), while others show significant variation (e.g. *Movement, Word Order, Clause*). For example,

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

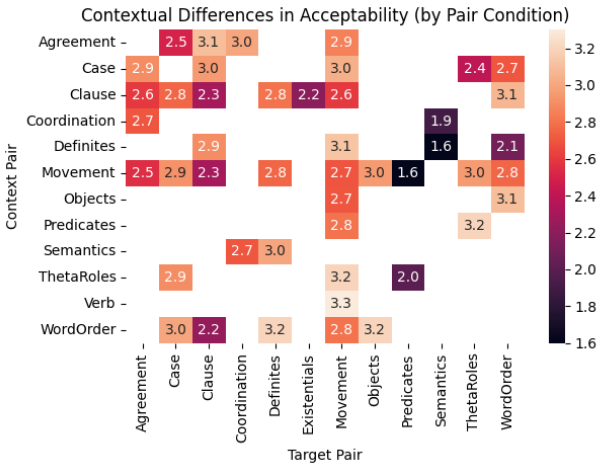

Figure 3: Contextual acceptability results for Experiment 1. X-axis represents the classification for the target pair. Y-axis represents the classification of the context pair. Cells indicate difference between acceptable and unacceptable sentences within each target pair; darker colors indicate smaller differences on a 5-point Likert scale.

consider the *Word Order* classification for the target pair (e.g., *Fred mowed the green lawn* > *Fred mowed the lawn green*).[4] When placed in the context of a sentence pair that modulates *Definites* (e.g. *This is a table* > *This is table*), 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 *Movement* (e.g., *Randy wanted to write a novel* > *Wanted to write, Randy did a novel*), the difference between the *green lawn* and *lawn green* sentences was approximately 2.8. 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 an average sentence (3).

Discussion

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

Experiment 2: Category Typicality

Category typicality assesses how "typical" an object is within a broader category (Farmer, Christiansen, & Monaghan, 2006; Rosch, 1975). For example, "robins" and "sparrows" are found to be more typical representations of birds than "toucans" and "penguins" across cognitive domains, including language (Meints, Plunkett, & Harris, 1999; Rosch, 1975) and vision (Maxfield, Stalder, & Zelinsky, 2014). Traditionally, category typicality has been measured using rating or decision tasks (Rosch, 1975), production tasks (Rosch, Simpson, & Miller, 1976), or inductive-reasoning tasks (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990), all of which ask the participant to consider a specific word in relation to the broader category label. Recent work also suggests that computational models of language learn category typicality from the statistical usage distributions of everyday language (Lake & Murphy, 2023; Misra, Ettinger, & Rayz, 2021).

In this experiment, we build a typicality-rating experiment using GRIS, finding that manipulating words in space both 1) replicates previous work and 2) provides more accurate representations of category typicality than a number of computational models.

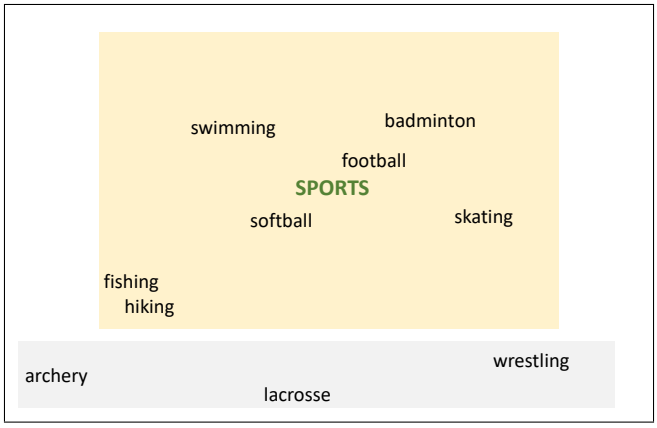

Figure 4: Sample trial for Experiment 2 (Typicality); font size enlarged to improve figure readability. Category label is marked in the center in green.

Design & Procedure

Stimuli We used eight of the original ten categories from Rosch (1975); all items were in English. Each category has a list of approximately 50-60 words with corresponding typicality ratings; 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.

Procedure A sample item for Experiment 2 is visualized in Figure 4. 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

[4]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.

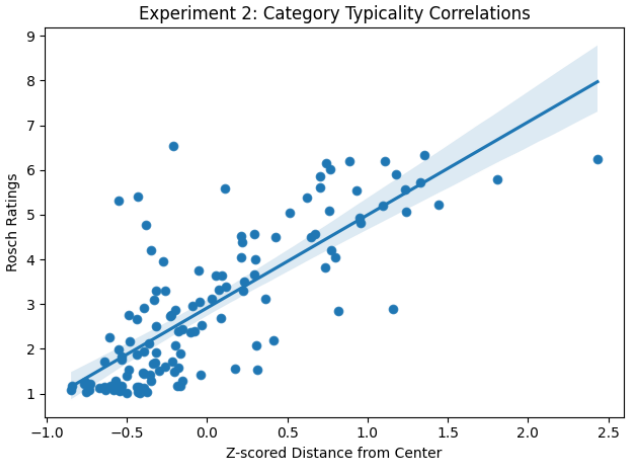

Figure 5: Correlation results for Experiment 2. The x-axis indicates the *Z*-scored distance from center for a word, and the y-axis indicates the original ratings from Rosch (1975).

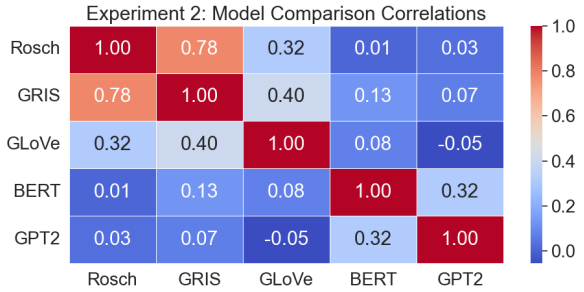

Figure 6: Correlation metrics between model representations and experimental results. Each cell corresponds to the Pearson's correlation coefficient between the models and experimental measures on the x- and y-axes.

the category label.

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.

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.

Experimental results are visualized in Figure 5. 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.

Computational Analyses For our computational analyses, we extracted vector representations of words from three models: GLoVe 6B.300D (Pennington, Socher, & Manning, 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 sentence label Y in the following way: *A(n) X is a typical Y.*; however, 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 these ways allows us to 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), which conducts behavioral analyses using model log-probabilities.

For each of the three models, we computed the Euclidean distance between the vectors for every word and its corresponding category label.[5] We then calculated the Spearman's correlation for all possible model comparisons.

Results for these multiple-correlation analyses are visualized in Figure 6. We find that GRIS is the only set of representations that connect a word to its category label in a manner that is comparable to Rosch (1975); the distances between words and their labels for GLoVe representations weakly correlate with the original Rosch rankings and our experimental data.

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 align well with human behavior, our work demonstrates that studies of human representations cannot rely on vectors generated by these models.

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 & Vitányi, 2003; Edelman, 1999; Edelman &

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

Figure 7: Sample Connections puzzle (left) with solution (right). 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.

Duvdevani-Bar, 1997; Reed, 1972). 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, Keefe, & Ross, 1989; Shelton & Martin, 1992), and, while less studied, syntactic similarity has shown similar effects (Lester, Feldman, & del Prado Martín, 2017). Orthographic similarity improves recall accuracy in a probed serial-recall task (Lin, Chen, Lai, & Wu, 2015), and phonological similarity has been shown to facilitate the learning of novel words (Papagno & Vallar, 1992).

While each of these features contributes to overall perception of similarity between linguistic units, what kinds of similarity do people optimize for? 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.

Data

Data 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.[6] A sample item and its corresponding solution are shown in Figure 7.

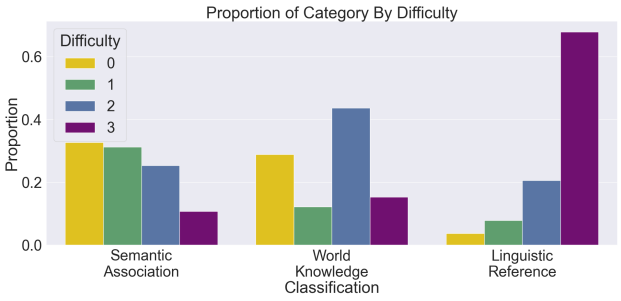

Figure 8: Distribution of classifications by difficulty. Difficulty levels closer to 0 are considered easier.

For 300 puzzles, two annotators categorized each group of words into one of three possible classifications: *Semantic Association* (e.g., "wet weather": hail, rain, sleet, snow), *World Knowledge* (e.g., "NBA teams": bucks, heat, jazz, nets), and *Linguistic Reference* (e.g., "palindromes": kayak, level, mom, race car). As visualized in Figure 8, 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.

Design & Procedure

Stimuli From our annotated data, we selected 10 puzzles that had at least two of the classifications. Given that we are using puzzles generated by the publication, we were unable to perfectly balance the different classifications across all puzzles.[7]

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 are; similar words should be placed closer together. Participants were instructed to use as much of the canvas as they felt was appropriate.

[6]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.

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

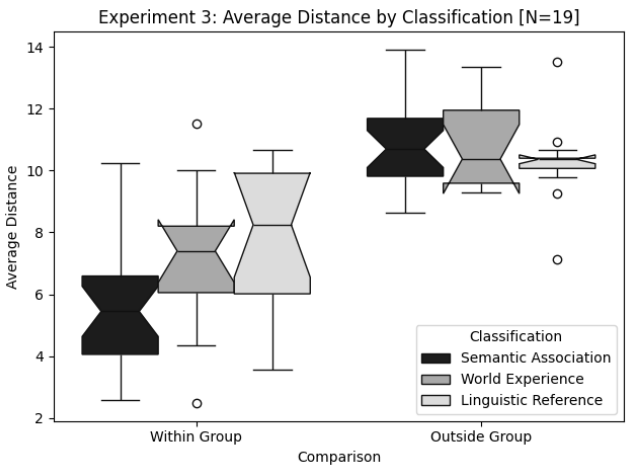

Figure 9: Average distance by classification for Experiment 3. Notches indicate bootstrapped 95% CIs.

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.

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

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 9. 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), CLASSIFICATION (semantic association/world experience/linguistic reference), and their full interactions, along with random intercepts for participants, items, and puzzle difficulty.[8] We find a main effect of COMPARISON, such that WITHIN GROUP comparisons are significantly closer together than OUTSIDE GROUP comparisons (β = -2.323, SE = 0.772, t=-3.263, p <0.01). Additionally, we report a significant interaction between COMPARISON and CLASSIFICATION, such that SEMANTIC ASSOCIATION groups clustered significantly closer together than LINGUISTIC REFERENCE groups in the WITHIN GROUP comparison (β = -3.085, SE = 0.884, t=-3.491, p <0.001).

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

Discussion

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

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

Why Use GRIS?

In this paper, we have shown how GRIS allows researchers across the cognitive sciences to test a large amount of data simultaneously (Experiment 1), while also evaluating the relative comparisons between different kinds of stimuli (e.g. text, image, audio). For example, GRIS allows experimenters to capture individual differences in representational spaces. Perception of similarity and difference is highly individualized (Simmons & Estes, 2008); GRIS provides a representation of the idiosyncrasies of an individual's 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.

Additionally, GRIS studies mimic the representational spaces constructed by computational models of language. Much work in cognitive science has shifted toward computational approaches, especially when evaluating representations (Achananuparp, Hu, & Shen, 2008; Deudon, 2018; Hiatt & Trafton, 2017; Mandera, Keuleers, & Brysbaert, 2017). Models for evaluating relationships between words use distributional spaces (e.g. Deudon, 2018) or representational analyses of embeddings to obtain similarity measurements of words or sentences. These measurements are often considered proxies to human representation spaces, which researchers cannot actually access. In this work, we show that these computational representations are not reliable proxies of human similarity, but that GRIS does provide an efficient way to collect such judgments.

References

Achananuparp, P., Hu, X., & Shen, X. (2008). The evaluation of sentence similarity measures. In *Data warehousing and knowledge discovery: 10th international conference, dawak 2008 turin, italy, september 2-5, 2008 proceedings 10* (pp. 305–316).

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Deudon, M. (2018). Learning semantic similarity in a continuous space. *Advances in neural information processing systems*, *31*.

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*(1358), 1191–1202.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203–12208.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical turk to obtain and analyze english acceptability judgments. *Language and Linguistics Compass*, *5*(8), 509–524.

Hiatt, L. M., & Trafton, J. G. (2017). Familiarity, priming, and perception in similarity judgments. *Cognitive Science*, *41*(6), 1450–1484.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of psycholinguistic research*, *29*, 627–645.

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological review*, *130*(2), 401.

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.

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

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Maxfield, J. T., Stalder, W. D., & Zelinsky, G. J. (2014). Effects of target typicality on categorical search. *Journal of vision*, *14*(12), 1–1.

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*(4), 1072.

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? *arXiv preprint arXiv:2105.02987*.

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*(6), 1003.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *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*(1), 47–67.

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*(8), 9.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, *3*(3), 382–407.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, *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*(4), 491.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, memory, and cognition*, *18*(6), 1191.

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition*, *108*(3), 781–795.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, *43*, 155–167.

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.

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, *8*, e19.

Zehr, J., & Schwarz, F. (2018). *Penncontroller for internet based experiments (ibex)*.