Lexical Search Dynamics in Taxonomic, Thematic, and Ad hoc Categories

Channing E. Hambric (c.hambric@bowdoin.edu)
Department of Psychology, Bowdoin College
5600 College Station, Brunswick, ME 04011 USA

Michael N. Jones (jonesmn@iu.edu)
Program in Cognitive Science and Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Abhilasha A. Kumar (a.kumar@bowdoin.edu)
Department of Psychology, Bowdoin College
5600 College Station, Brunswick, ME 04011 USA

Abstract

Searching through semantic memory involves navigating across clusters of related items, but the current body of work has primarily focused on search within hierarchically structured taxonomic categories (e.g., *Animals*). The present work explored search behavior via the verbal fluency task in thematic categories, where relatedness is construed via complementary roles in a shared environment, and ad hoc categories, where relatedness is construed via shared service to an external goal. We found strong differences across domains within each type of category, but compared to *Animals*, responses in thematic and ad hoc categories had higher average phonological similarity and word frequency. Phonology-inclusive process models provided the best account of search in taxonomic categories, but overall model performance was poor for ad hoc categories. There were also important differences in the contribution of lexical sources to within and between-cluster transitions across domains. These results underscore the necessity of exploring lexical search and validating computational models in categorical domains with different compositions and types of relatedness.

Keywords: memory search; verbal fluency; computational modeling; word embeddings; language models, categorization

Introduction

Generating words from memory involves navigating across a complex semantic landscape. A large body of work has focused on using the semantic fluency task (Bousfield & Sedgewick, 1944) to identify the structures and processes that guide search through semantic memory. In this task, participants are presented with a category prompt (e.g., *Animals*) and asked to generate as many exemplars as possible. Fluency tasks have been widely used in basic science research but are also an integral component of neuropsychological testing batteries. A common finding in fluency tasks is that exemplars tend to be generated in related *clusters*, which have in turn been used to approximate the underlying structure of semantic memory. Early work by Troyer (2000) used membership in a shared subcategory to indicate if an item was member of a given cluster or if the production indicated a *switch* to a new search area. Alternatively, Hills et al. (2012) used distributional semantic models (DSMs) to establish inter-item similarity. Mimicking statistical learning in human language acquisition, these models learn semantic relatedness from large natural language text corpora to generate high-dimensional vector representations of concepts (see Kumar et al., 2021 for review). Hills et al. (2012) also implemented a set of computational process models to simulate semantic search within and across clusters. Specifically, they proposed that search over an internally structured semantic space follows the marginal value theorem, in which new areas of search are sought when the marginal value of resource intake is equal to the overall average intake over all clusters. Their *optimal foraging* models used a combination of word frequency and semantic similarity to predict the probability of producing a given item and successfully accounted for search patterns in the fluency task. More recent work by Kumar et al. (2022) found that phonological similarity was associated with the production of a higher number of items in the semantic fluency task and that optimal foraging models that incorporated phonological similarity during search within clusters provided the best account of the behavioral data. Kumar et al. (2024) thus expanded both the switch methods and the foraging models developed by Hills et al. (2012) to incorporate phonological similarity.

Despite the substantial progress made using these methods, structural and process accounts of semantic fluency (as well as the majority of clinical assessments) have been largely limited to the outcomes from the ubiquitous *Animals* domain. It is easy to see why - *Animals* as a domain is large, well known, and acquired early in development (Eimas & Quinn, 1994). The implicit and scientifically-constructed taxonomy of this category also allows for ready delineation into subcategories (e.g., Rosch, 1975; Nelson, 1988; Barsalou, 1991). Indeed, such *taxonomic* categories, which can be defined as categories in which relatedness is measured via overlapping features, tend to dominate behavioral and computational research in memory retrieval (e.g., Estes et al., 1989; Anderson et al., 1994; Ursino et al., 2011; Hills et al., 2015). However, taxonomic categories differ in relational structure and are neurologically dissociable, such that *Animals* may not be a sufficient approximation of memory search in all hierarchically structured categories (Blundo et al., 2006; Mahon &

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

Caramazza, 2009; Anderson et al., 2014; Le et al., 2024).

Moreover, other *types* of categories also play a significant role in typical cognition. For example, say you are tasked to pack for a beach vacation. What types of semantic knowledge might you need to access? A single taxonomy cannot be used to guide your packing - rather, one must identify the things that might be needed for a tropical locale, such as a *swimsuit* and *beach towel*, as well as serve relevant external goals, such as *Suitable for a Carry-on Suitcase*. Here, we focused on two other types of categories that have been extensively studied in the categorization literature. *Thematic* categories can be defined as groups of relations that perform complimentary roles in the same scenario (Estes et al., 2011; Mirman et al., 2017; Elman & McRae, 2019; Lin & Murphy, 2001). For example, *sunscreen* and *sunglasses* have complimentary but distinct roles of *sun protection* in the shared environment of *Beach*. Lastly, *ad hoc* (literally translating to "for this" in Latin) categories can be defined as categories that are created spontaneously in service to some external goal (Barsalou, 1983; Barsalou, 1991, 2003; Estes et al., 2011; Abdel Rahman & Melinger, 2009). The items in these categories typically share very few features and do not correspond to preexisting categories in memory (Barsalou, 1983). Rather, relatedness is construed purely through the shared property that serves the external goal. For example, *jewelry* and *cat* share little to no taxonomic features and serve no complimentary roles, but they would both be considered *valuables* in the ad hoc category of *Things to Rescue from a Burning Home*.

Surprisingly, to date there have been little to no direct investigations of free lexical search within thematic and ad hoc categories. Thematic and ad hoc categories are occasionally included in category production norm datasets (Battig & Montague, 1969; Van Overschelde et al., 2004; Castro et al., 2021; Banks & Connell, 2023), where researchers record how frequently a given exemplar is produced in response to a category prompt (e.g., how often is *money* produced for the prompt *Things to Rescue from a Burning Home*). However, researchers have yet to investigate the lexical properties of the generated items or explore the underlying structures that guide lexical search and retrieval in these types of categories. Some relevant findings can be found in the language production and processing literature. Broadly, the current body of work points to substantial differences in lexical retrieval across different types of categories. For example, both Schwartz et al. (2011) and Xu et al. (2018) found that taxonomic and thematic processing are associated with different brain regions (the anterior temporal lobe and temporoparietal junction, respectively), which the authors suggest point to key differences in the underlying representations. Taxonomic and thematic categories have also been compared using various picture naming methodologies, but the current consensus is complicated by methodological differences across naming paradigms. For example, while some work has found that priming pictures with a thematically related word facilitates lexical retrieval (Xavier Alario et al., 2000; de Zubicaray et al., 2013; Hambric & O'Séaghdha, 2023), others have shown that, in other production contexts, thematic relations induce interference in much the same way as taxonomic categories (Abdel Rahman & Melinger, 2011; Rose & Abdel Rahman, 2016).

Work on ad hoc categories is somewhat more limited. Although these categories display the same typicality gradient (in which more typical items are generated first) as taxonomic categories, their coactivation is fully mediated by activation of the relevant external goal (Barsalou, 1983; Barsalou, 2003). Together, these results highlight important structural differences across types of categories. In the context of the semantic fluency task, there are thus reasons to believe that the differing topology of these categories may lead to distinct search behavior.

The present pre-registered study investigated the commonalities and differences in lexical search dynamics in taxonomic, thematic, and ad hoc categories. Specifically, we examined lexical metrics, clustering behavior, and process models of search for two taxonomic categories (*Animals* and *Occupations*), as well as two thematic categories (*Things at the Beach* and *Things in a Classroom*) and two ad hoc categories (*Excuses for Being Late* and *Things to Rescue from a Burning Home*). The two thematic categories represent two common, well-scripted thematic scenarios, whereas the two ad hoc categories correspond to goal-derived scenarios unlikely to already have a stable existing representation in memory. We hypothesized that in categories with less explicit hierarchical structure, individuals may produce fewer items overall, produce smaller clusters of related items, and switch to new search areas more often. In addition, since functionality is often linked lexically (e.g., *pencil* and *pencil sharpener*), we also predicted a greater degree of phonological similarity among responses, especially in thematic categories. We explored these questions through a gamified web version of the semantic fluency task (https://semantigories.research.bowdoin.edu/). Each participant completed a fluency task for each of the above categories. The fluency lists were then evaluated using the python package *forager*, which extracts lexical metrics and automatically calculates a variety of switch methods and optimal foraging models (Kumar et al., 2024). Examples of items generated in each domain and an overview of the computational procedure can be seen in Figure 1.

Methods

The study was preregistered, and full descriptions of the design, hypotheses, and analytic approaches can be found on Open Science Framework (`https://osf.io/qn49c/?view_only=56e698d2030044008a93442aef6a8c86`).

Participants Participants were recruited from Prolific and an introductory psychology course at Bowdoin College. Data were only analyzed for participants who successfully completed all experimental procedures (see below) and for those who learned English prior to age four. In addition, only

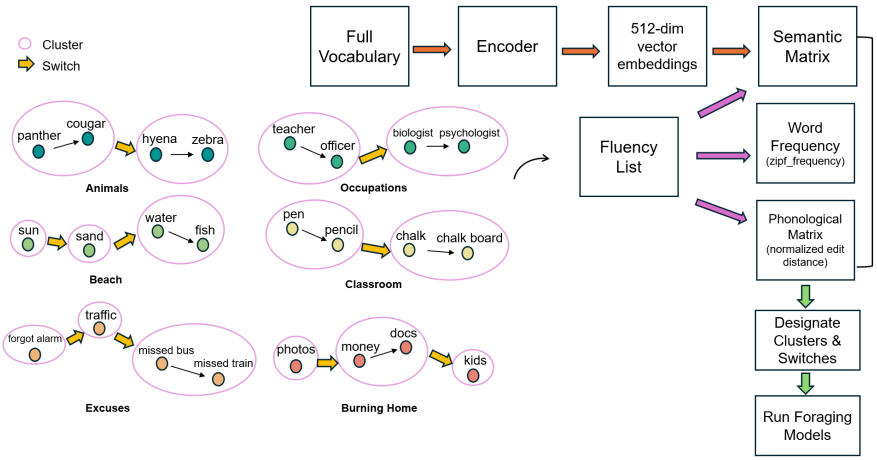

Figure 1: Sample clusters and procedure overview. A priori, lexical metrics were extracted for the full vocabulary for each domain. Using *forager* (Kumar et al., 2024), semantic similarity (derived from Universal Sentence Encoder), phonological similarity (normalized edit distance) and word frequency (zipf_frequency) were generated for each fluency list and used to compare lexical metrics, automatically designate clusters, and run optimal foraging models.

lists that contained at least four valid entries were analyzed ($n_{Animals}$ = 610, $n_{Occupations}$ = 596, n_{Beach} = 599, $n_{Classroom}$ = 616, $n_{Excuses}$ = 501, $n_{BurningHome}$ = 570).

Verbal fluency The semantic fluency task was administered through a gamified online platform. Participants first completed a series of semantic fluency tasks. In each round, participants were presented with one category prompt (*Animals, Occupations*, *Things at the Beach*, *Things in a Classroom, Excuses for Being Late,* or *Things to Rescue from a Burning Home*) and had one minute to type in as many exemplars as possible. Following this task, participants rated the similarity of pairs of items from self and peer-generated lists. Participants also completed a Remote Association Test and a Media Consumption Questionnaire. The present work only reports on the results of the semantic fluency task.

Structure-level models Vocabularies were generated from a combination of available fluency data (Hills et al., 2015), online datasets, category norm data bases ((Van Overschelde et al., 2004; Castro et al., 2021; Banks & Connell, 2023), and the current data prior to any analysis ($n_{Animals}$ = 2082, $n_{Occupations}$ = 6017, n_{Beach} = 900, $n_{Classroom}$ = 596, $n_{Excuses}$ = 1000, $n_{BurningHome}$ = 458). Semantic representations were obtained from the Universal Sentence Encoder (Cer et al., 2018). This model is trained on a large corpus of natural language and transforms variably sized text input into a 512-dimension vector. A semantic similarity matrix was then constructed for each domain to obtain pairwise semantic similarity for each possible pair of adjacent items in a given fluency list. Pairwise phonological similarities were computed via normalized edit distances between phonetic transcriptions obtained from CMUDict (Lenzo, 2014). Zipf word frequencies were obtained from the *wordfreq* package (Speer, 2022).

Process Models We examined different clustering methods alongside process models that utilize lexical sources at different stages of the search process based on prior work (see Kumar et al., 2024). The first method for designating clusters and switches used hand-coded norms developed by Troyer (2000). This method, which is only available for *Animals* in the current list of domains, assigns a specific subcategory to each animal. A cluster is designated for items that share at least one subcategory label (e.g., *hyena* and *zebra* both belong to the subcategory *African*, see Figure 1), and a switch marks a shift to a new subcategory. We also used methods that draw on distributional semantic representations. These include: *similarity drop* (Hills et al., 2012), in which an entry is designated as a switch if there is a drop in semantic similarity followed by an immediate rise; *delta* similarity (Lundin et al., 2023), which is similar to *similarity drop* with the exception that the fall and rise in similarity must exceed a specific threshold, *multimodal* similarity (Kumar et al., 2024), in which an item is designated as a switch if there a drop in the weighted sum of semantic and phonological similarity (with the respective weights of either adjusted by the tuning parameter α), and *multimodal-delta* similarity (Kumar et al., 2025), which is similar to the *multimodal* method except that it adopts the threshold parameters of the *delta* similarity method.

A series of process models were tested in conjunction with each of these switch methods that emphasized different lexical sources for within and between-cluster transitions. These variations included: the *static* foraging model, which uses semantic similarity and word frequency to calculate the probability of retrieving any item, the *dynamic* foraging model (Hills et al., 2012), which uses semantic similarity and word frequency to predict retrieval for within cluster designations

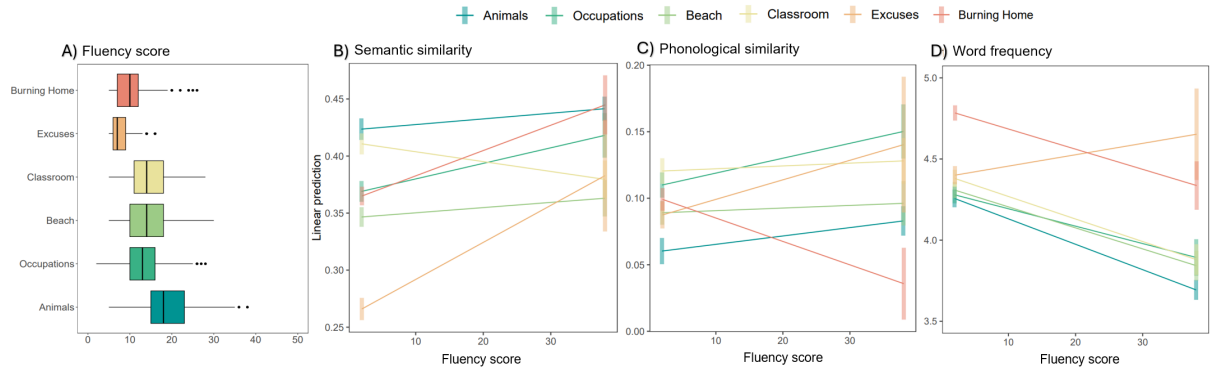

Figure 2: Average number of generated items, semantic similarity, phonological similarity, and word frequency across domains. Model predicted values are plotted for the lexical metrics to account for the number of generated items.

.

and only frequency if a given item is designated as a switch, and *phonology inclusive* models (Kumar et al., 2022, 2024), which explore the impact of phonology alongside semantic similarity and word frequency. These models also include *static* and *dynamic* variations, the latter of which incorporate phonological similarity for both cluster and switch designations (global), only for within cluster designations (local), or only for switch designations (switch). More extensive details can be found in Kumar et al. (2024).

Obtaining Best-Fit Models Data evaluation, automated switch designations, and modeling were conducted using the *forager* python package (Kumar et al., 2024). Each process model was fully crossed with each switch method and likelihoods for each item within a fluency list were obtained. To determine the best model for each domain, we used the output negative log likelihood of the participant data to compute the delta Bayesian Inference Criteria (BIC), where a higher value reflects improved model performance in comparison to the random model. To account for skewness, we used the median delta BIC for each domain to identify the best-fit model.

Results

Lexical Metrics First, we examined fluency performance for each domain. Consistent with our predictions, significantly more exemplars were generated for the taxonomic domains compared to the average of thematic (b = -1.58, t = -6.95, $p < .001$) and ad hoc categories (b = -6.99, t = -30.40, $p < .001$), see Figure 2A. We then examined key lexical metrics, including average semantic similarity, phonological similarity, and word frequency. Note that semantic and phonological similarity measures reflect the average degree of similarity between adjacent items in each fluency list. Each lexical metric varied with fluency performance. Higher fluency scores were generally associated with higher average semantic and phonological similarity, but a higher fluency score was associated with *lower* average word frequency, see Figure 2B-D. Although there were some domain-specific deviations, these findings replicate previous work (Kumar et al., 2022).

Collapsing across domains, consecutive responses in taxonomic categories were more semantically similar than in ad hoc (b = -.10, t = -16.68, $p < .001$) categories, but did not significantly differ from thematic categories (b = -.009, t = 1.45, $p = .148$), see Figure 2B. However, there were differences across domains within each category type. Specifically, items in the *Animals* category were more semantically similar to one another than in *Occupations* (b = .04 , t = 16.33, $p < .001$), items in *Classroom* were more similar to one another compared to *Beach* (b = .05, t = 20.96, $p < .001$), and items in *Burning Home* were more similar to one another compared to *Excuses*, b = -.09, t = -16.14, $p < .001$. Across domains, consecutive responses in taxonomic categories were also more phonologically similar than in thematic (b = -.02, t = -3.23, $p = .001$) and ad hoc (b = -.02, t = -3.21, $p = .001$) categories. However, this pattern was qualified by significant domain differences, where responses in *Occupations* had a much higher average phonological similarity compared to *Animals* (b = -.06, t = -19.27, $p < .001$), responses in *Classroom* were more similar to one another compared to *Beach* (b = -.03, t = -12.89, $p < .001$), and responses in *Excuses* were more similar to one another than in *Burning Home* (b = -.02, t = -4.37, $p < .001$), see Figure 2C. Lastly, items produced in ad hoc categories were associated with a greater average word frequency than those in taxonomic (b = -.48 , t = -33.18 , $p < .001$) and thematic (b = -.41, t = -28.94 , $p < .001$) categories, and this pattern was relatively stable across domains, see Figure 2D.

Foraging Models Foraging models that included phonology best predicted the data in taxonomic categories, although the specific nature of this contribution to the search process varied by domain (see Table 1). The best-fit model for the *Animals* domain was the model where phonology is only included during local, within-cluster transitions, while the global variation (in which phonology is used in both within and between-cluster transitions) best accounted for *Occupations* fluency. All thematic and ad hoc domains were best accounted for by standard dynamic models, which did not include phonology as a search cue. However, as can be seen

Category Type	Domain	Best-fit Switch Method	Best-fit Foraging Model	Median Δ BIC
Taxonomic	Animals	Similarity Drop	Phonological (Local)	69.72
Taxonomic	Occupations	Similarity Drop	Phonological (Global)	32.21
Thematic	Beach	Similarity Drop	Dynamic	16.12
Thematic	Classroom	Multimodal ($\alpha = 0.9$)	Dynamic	18.52
Ad hoc	Excuses	Similarity Drop	Dynamic	3.03
Ad hoc	Burning Home	Multimodal ($\alpha = 0.8$)	Dynamic	9.23

Table 1: Best-fit switch method, foraging model, and median delta BIC for each domain.

by the low median delta BIC values, particularly for the ad hoc categories, these models were not a vast improvement in comparison to the random models, suggesting that the current approach is limited in its ability to capture search in these categories. We return to this issue in the discussion.

Clustering Analyses We then examined the size of clusters and the number of switches using the best-fitting switch method for each domain. Aside from taxonomic categories, where the *similarity drop* method provided the best fit for both domains, there were domain differences in the best-fit switch method within each category type. The *similarity drop* method provided the best fit for *Beach* and *Excuses*, while the *multimodal* method provided the best fit for *Classroom* and *Burning Home*. When including fluency score as a covariate, the average size of clusters in taxonomic categories did not significantly differ from thematic (b = -.09, t = -1.05 , p = .295) or ad hoc categories (b = .03, t = 0.33, p = .739), but individuals switched to new clusters less often in ad hoc categories compared to taxonomic categories, b = .19, t = 2.03, p = .042. We also compared lexical metrics within clusters versus at cross-cluster switch points, see Figure 3. As expected, semantic similarity was universally higher within clusters than at switch points, b = .12, t = 57.25, $p < .001$. Though much less extreme, word frequency was higher within clusters than at switch points, b = .14, t = 8.99, $p < .001$. Phonological similarity was also higher within clusters compared to at switch points (b = .007, t = 2.53, p = .011), but the degree differed across specific domains, with items in the thematic domain *Classroom* (b = .06, t = 18.75, $p < .001$) displaying significantly greater phonological similarity within clusters. As can be seen in the examples in Figure 1, this is likely due to the overlapping word stems in generated exemplars (e.g., *chalk* and *chalk board*).

To better understand how common it was to use shared word stems to facilitate search across categories, we examined the average number of consecutive shared phonemes in each domain. To examine the prevalence of exploiting shared word stems, we identified all pairs of adjacent items that shared at least three consecutive phonemes. The average number of consecutive shared phonemes and the percentage of adjacent items with shared word stems can be seen in Table 2. Consistent with the earlier analysis of phonological similarity, items in *Occupations*, *Classroom*, and *Excuses* shared more consecutive phonemes on average and a greater percentage of adjacent items shared word stems. Interestingly, the *type* of shared stem differed across domains. In *Occupations*, participants produced items with both shared whole words (e.g., *physical therapist* and *massage therapist*) as well as shared suffixes (e.g., *biologist* and *psychologist*). In *Classroom*, items with shared stems typically comprised functional pairs (e.g., *chalk* and *chalk board*), and in *Excuses*, participants often generated successive items using the same verb phrase (e.g., *missed bus* and *missed train*).

Category Type	Domain	Average Consecutive Shared Phonemes	Pecentage Shared Word Stems
Taxonomic	Animals	0.56 (0.65)	1.36%
Taxonomic	Occupations	1.08 (0.89)	4.53%
Thematic	Beach	0.77 (0.78)	3.41%
Thematic	Classroom	0.93 (0.88)	7.32%
Ad hoc	Excuses	1.15 (0.97)	4.69%
Ad hoc	Burning Home	0.74 (0.75)	2.29%

Table 2: Average number of consecutive shared phonemes and percentage of adjacent items with a shared word stem per domain. Standard deviations in parentheses.

Discussion

In this work, we explored the characteristics and mechanisms of lexical search in taxonomic, thematic, and ad hoc categories. Although recent methods have leveraged distributional semantic representations and optimal foraging process models to better understand the nature of semantic retrieval, this work has been limited to taxonomic categories. However, other types of categories, such as thematic and ad hoc relations, also play critical roles in cognition, and it is unknown how well the conclusions drawn from the study of taxonomic search map onto other types of semantic relatedness. Overall, we found strong domain differences and few consistent patterns across broad category types. Of particular interest, items generated in the taxonomic domain *Occupations*, the thematic domain *Classroom*, and the ad hoc domain of *Excuses* were much more phonologically similar to one another than in their same-category-type counterparts. Clustering analyses revealed that semantic and phonological similarity were generally greater within clusters than at switch points. Similar to Kumar et al. (2022), we also observed a positive linear relationship between fluency score and the average semantic and phonological similarity of generated items, as well as a negative linear relationship between average word frequency and fluency score, though there were some notable outliers. Computational modeling of search responses indicated that phonology-inclusive models provided the best account of search in taxonomic categories while the standard dynamic model provided the best account for thematic and ad hoc categories. Crucially, the latter findings are

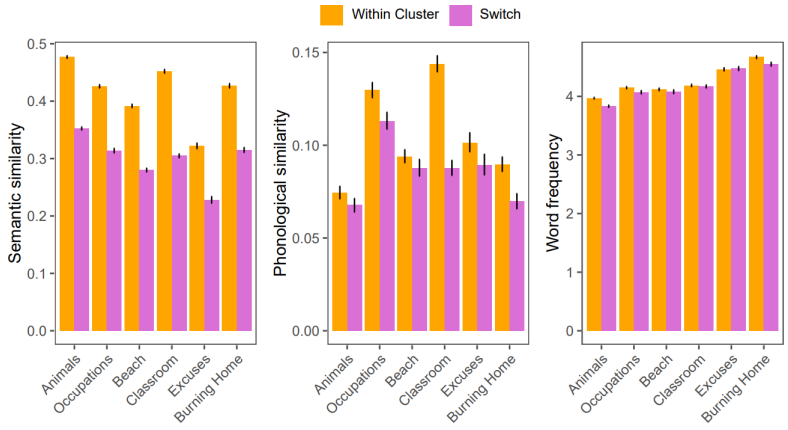

Figure 3: Average semantic similarity, phonological similarity, and word frequency for switch trials versus within cluster transitions. Error bars represent bootstrapped 95% confidence intervals.

constrained by the fact that model performance was relatively poor in these categories. Together, these findings support our hypothesis that non-taxonomic categories are associated with distinct structures and search processes, but one size does not fit all for broad category types; rather, search is shaped by the intrinsic lexical structure of each individual domain.

As noted above, there were two outliers to the linear relationship between lexical metrics and fluency. First, word frequency was *positively* associated with fluency performance in the *Excuses* domain. We suggest that, in comparison to the other domains, *Excuses* is quite abstract, such that frequency may be a particularly helpful cue. In addition, fluency in *Burning Home* exhibited a *negative* linear relationship with phonological similarity. Closely related items often share phonological units, but items in ad hoc categories typically share few surface level features (Barsalou, 1983), making it unlikely that fluency would be enhanced by relying on phonological similarity as a search cue. On the other hand, the ad hoc domain *Excuses* does display a positive linear relationship between phonological similarity and fluency. As discussed above, this is due to participants using shared verb phrases and syntactic structures to guide retrieval in this domain. We also note that Excuses actually had the greatest average number of consecutive shared phonemes out of all domains, which paints a slightly different picture of phonological similarity than that provided through normalized edit distance calculations. It is possible that the multi word nature of entries in this domain introduces unforeseen complexities, such as over-inflating the role of surface level differences in responses. For example, although *forgot alarm* and *forgot to set my alarm* share two key words, edit distance similarity would reflect the added presence of less-central linguistic elements.

In addition, recall that the best-fit switch method and foraging model for the *Excuses* domain did not include phonology, despite the high phonological similarity among items in this domain. We posit that this discordance reflects that distributional semantic models do not appropriately tap into the highly conditional nature of ad hoc similarity. For example, in the *Excuses* domain, *snow storm* and *ice storm* had a sensible semantic similarity score of 0.64, while *slept in* and *forgot to set alarm* only had a similarity score of 0.24, despite being intuitively semantically related. Recent investigations into the ability of natural language models to capture non-taxonomic similarity appear to support this suspicion. Edinger & Goldstone (2022) examined cued generation across a variety of cutting-edge natural language processing models. Of particular relevance, they asked each model to generate exemplars in response to taxonomic categories and in response to ad hoc *Family Feud* prompts (e.g., *Things children often lose*) and compared responses to those generated by actual participants. Overall model performance was shockingly low; only 1 correct response was generated out of 20 for ad hoc prompts. Regardless, they found that BERT (Devlin, 2018) provided the best approximation of ad hoc lexical search, which the authors attributed to the attention mechanism's ability to use contextual information to identify relevant properties of concepts belonging to multiple categories. Future work should investigate search using versions of BERT that explicitly compare sentence-level information, such as Sentence-BERT (Reimers & Gurevych, 2019), alongside more sophisticated measures of phonological similarity.

In sum, the present work sheds light on lexical search in non-taxonomic categories. We found several significant differences between how individuals search through taxonomic, thematic, and ad hoc categories, suggesting an opportunity to revisit how search is conceptualized within a broader and more dynamic lexicon. Future work could also examine individual differences in search, as these may be particularly exacerbated in non-hierarchical categories. Additionally, for ad hoc categories, comparing cluster designations derived from language models to those generated by goal-conscious human raters (Kumar et al., 2025) could be a promising future direction. Ultimately, integrating evidence of similarities and differences across different semantic categories will be important for developing a complete theory of lexical search.

References

Abdel Rahman, R., & Melinger, A. (2009). Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes*, *24*(5), 713–734.

Abdel Rahman, R., & Melinger, A. (2011). The dynamic microstructure of speech production: semantic interference built on the fly. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 149.

Anderson, A. J., Murphy, B., & Poesio, M. (2014). Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*, *26*(3), 658–681.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063.

Banks, B., & Connell, L. (2023). Category production norms for 117 concrete and abstract categories. *Behavior Research Methods*, *55*, 1292-1313.

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*(5-6), 513–562.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In *Psychology of Learning and Motivation* (Vol. 27, pp. 1–64). Elsevier.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, *80*(3p2), 1.

Blundo, C., Ricci, M., & Miller, L. (2006). Category-specific knowledge deficit for animals in a patient with herpes simplex encephalitis. *Cognitive Neuropsychology*, *23*(8), 1248–1268.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology*, *31*, 149.

Castro, N., Curley, T., & Hertzog, C. (2021). Category norms with a cross-sectional sample of adults in the united states: Consideration of cohort, age, and historical effects on semantic categories. *Behavior Research Methods*, *53*, 898-917.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... others (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

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

de Zubicaray, G. I., Hansen, S., & McMahon, K. L. (2013). Differential processing of thematic and categorical conceptual relations in spoken word production. *Journal of Experimental Psychology: General*, *142*(1), 131.

Edinger, A., & Goldstone, R. (2022). Getting situated: Comparative analysis of language models with experimental categorization tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, *65*(3), 903–917.

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*(2), 252.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: a comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 556.

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic thinking: The apprehension and consequences of thematic relations. In *Psychology of Learning and Motivation* (Vol. 54, pp. 249–294). Elsevier.

Hambric, C. E., & O'Séaghdha, P. G. (2023). The unseen, the seen, and the spoken: Latent and overt priming in cyclic picture naming. *Quarterly Journal of Experimental Psychology*, *76*(10), 2410–2430.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, *7*(3), 513–534.

Kumar, A. A., Apsel, M., Zhang, L., Xing, N., & Jones, M. N. (2024). forager: A python package and web interface for modeling mental search. *Behavior Research Methods*, 1–17.

Kumar, A. A., Lundin, N. B., & Jones, M. N. (2022). Mouse-mole-vole: The inconspicuous benefit of phonology during retrieval from semantic memory. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44).

Kumar, A. A., Lundin, N. B., & Jones, M. N. (2025). What's in my cluster? evaluating automated clustering methods to understand idiosyncratic search behavior in verbal fluency. *Journal of Memory and Language*, *141*, 104606.

Kumar, A. A., Steyers, M., & Balota, A., David. (2021). A critical review of network-based and distributional approaches to semantic memory structure and processes topics in cognitive science. *Topics in Cognitive Science*, *45*. doi: https://doiorg/101111/tops12548

Le, T. T., Luong, D. A. Q., Joo, H., Kim, D., & Woo, J. (2024). Differences in spatiotemporal dynamics for processing specific semantic categories: An eeg study. *Scientific Reports*, *14*(1), 31918.

Lenzo, K. A. (2014). *Carnegie mellon pronouncing dictionary (cmudict)-version 0.7 b.* Nov.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*(1), 3.

Lundin, N. B., Brown, J. W., Johns, B. T., Jones, M. N., Purcell, J. R., Hetrick, W. P., ... Todd, P. M. (2023). Neural evidence of switch processes during semantic and phonetic foraging in human memory. *Proceedings of the National Academy of Sciences*, *120*(42), e2312462120.

Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: a cognitive neuropsychological perspective. *Annual Review of Psychology*, *60*(1), 27–51.

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological Bulletin*, *143*(5), 499.

Nelson, K. (1988). Where do taxonomic categories come from? *Human Development*, *31*(1), 3–10.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192.

Rose, S. B., & Abdel Rahman, R. (2016). Cumulative semantic interference for associative relations in language production. *Cognition*, *152*, 20–31.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., ... Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences*, *108*(20), 8520–8524.

Speer, R. (2022, September). *rspeer/wordfreq: v3.0.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7199437` doi: 10.5281/zenodo.7199437

Troyer, A. K. (2000). Normative data for clustering and switching on verbal fluency tasks. *Journal of Clinical and Experimental Neuropsychology*, *22*(3), 370–378.

Ursino, M., Cuppini, C., & Magosso, E. (2011). An integrated neural model of semantic memory, lexical retrieval and category formation, based on a distributed feature representation. *Cognitive Neurodynamics*, *5*, 183–207.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(3), 289-335.

Xavier Alario, F., Segui, J., & Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(3), 741–764.

Xu, Y., Wang, X., Wang, X., Men, W., Gao, J.-H., & Bi, Y. (2018). Doctor, teacher, and stethoscope: neural representation of different types of semantic relations. *Journal of Neuroscience*, *38*(13), 3303–3317.