Fast and frugal memory search for communication

Collin J. Kovacs
Indiana University
cokovacs@iu.edu

Jasper M. Wilson
Washington University in St. Louis
jaspermwilson@wustl.edu

Abhilasha A. Kumar
Indiana University
kumaraa@iu.edu

Abstract

Communication involves searching for optimal utterances within memory and then evaluating those utterances against a target goal. This task is substantially harder when information about multiple concepts has to be communicated, such as describing how *music* and *tides* are similar. Whether the search process for this challenging communicative task converges onto the optimal response relatively quickly, or involves more strategic decision-making to evaluate different candidates remains understudied. In this work, speakers generated single word "clues" that would enable a listener to correctly identify a *pair* of words among several distractor words. Speakers and listeners generated candidates before producing final responses. Each player was biased towards the first candidate(s) they generated, even when this candidate was suboptimal compared to other candidates, as was the case for less related concepts. Furthermore, straying away from the initial semantic "patch" of responses decreased accuracy in the game. Overall, these findings suggest that individuals tend to identify the relevant semantic cluster early on during semantic search, and are likely to employ the "take-the-first" strategy for selecting utterances in ambiguous, ill-defined semantic contexts.

Keywords: semantic retrieval; communication; memory search; take-the-first heuristic; reference games

Introduction

An extensive body of work suggests that generating optimal utterances for the purpose of communication involves complex search and decision-based processes (Goodman & Frank, 2016; Olson, 1970). However, how do search and retrieval occur within communicative contexts? Does this process involve relatively automatic spreading activation-type processes (Collins & Loftus, 1975), or is it mediated by more conscious, attentional (Neely, 1977) mechanisms? One possibility is that when individuals are searching for the right words to convey an intended message, this process occurs quickly and individuals are able to rapidly arrive at the most optimal utterance within a given context. Another possibility is that individuals deliberate between several different possible utterances and ultimately choose the most optimal one. Although some work has examined how speakers choose between different potential utterances (Jara-Ettinger & Rubio-Fernandez, 2021), the process by which such utterances are generated in the first place remains understudied.

Understanding the intricacies of how individuals search through memory to generate potential candidate responses within the context of communication requires a rich experimental paradigm that would elicit a broad range of utterances and be able to capture the variability in response selection. Reference games, where speakers are asked to produce utterances that would enable a listener to identify a target amongst several distractors have been used to study various aspects of communication (Olson, 1970; Dale & Reiter, 1995), such as why speakers provide redundant information to listeners (Degen et al., 2020) and how partners engage in perspective-taking and divide effort during communication (Goodman & Frank, 2016; Hawkins et al., 2020). These studies typically involve identifying a *single* target item within the context of several distractors. However, humans routinely refer to groups of items with varying degrees of relatedness, such as when communicating about items belonging to a well-defined natural category (e.g., *lion* and *tiger* are *predators*), selecting items based on ad-hoc categories (e.g., Alice would grab her *laptop* and *passport* in the event of a *fire*, Barsalou, 1983), finding similarities and/or differences between seemingly unrelated concepts via analogies and metaphors (e.g., *life* is like a *box* of *chocolates*), or identifying abstract relationships between concepts on an intelligence test (e.g., How are a *poem* and *statue* alike? How are *music* and *tides* similar?). Conveying the meaning of multiple concepts is therefore a relatively common but understudied communicative challenge.

Recently, Kumar, Steyvers, & Balota (2021) explored how speakers generate optimal utterances when *multiple* concepts need to be identified, through a two-player word game called Connector. They found that speakers were able to converge onto similar utterances even when the target set contained unrelated words (such as *cave* and *knight*), and showed that random walk-based associative models provided the best account of the behavioral patterns. This work was extended by Kumar, Garg, & Hawkins (2021) to show that speakers may be employing some level of pragmatic inference to produce these utterances, although the nature of search processes employed and different possibilities considered by speakers was not thoroughly investigated.

A primary goal of the current work was to better understand how individuals search for, generate, and select responses to distinguish a set of target items from an array of distractors within a communicative task. Previous work on search within semantic memory has employed a variant of the think-aloud procedure, where participants are asked to provide any responses that come to mind as they think of the answer. For example, in a variant of the remote associates test,

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

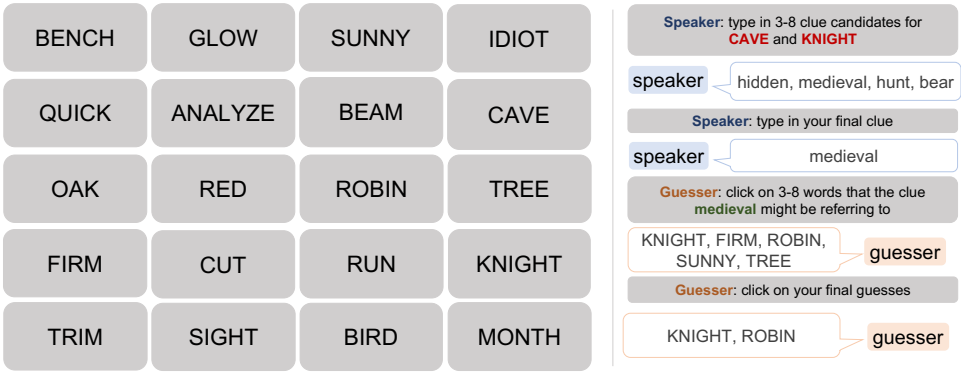

Figure 1: An experiment trial between the speaker and guesser in the game. Speakers first typed in 3-8 potential clues ("candidates") before submitting a final clue, and guessers clicked on 3-8 potential answers before submitting their final guesses.

Davelaar (2015) asked participants to type in as many possible answers as they could come up with before they arrived at the final solution. We employed a similar methodology to better understand the search processes that occur when speakers are attempting to find words that would best communicate a set of targets to a listener, and when listeners were attempting to identify the target words. Specifically, we modified the Connector paradigm to let speakers and guessers generate a set of **candidate** words before selecting a final response to send to their partner. In this way, we were able to examine the specific candidates generated for different target item sets and evaluate whether individuals engage in rapid or strategic decision-making during communication.

The strategic selection hypothesis would predict that individuals would consider a variety of candidates before arriving at an optimal response. If so, we may observe that speakers and guessers choose their final response from the later candidate responses. On the other hand, the rapid selection hypothesis would predict that highly relevant words are activated relatively automatically, and the optimal response would then be chosen from among these initially activated candidate words. In this work, we evaluated whether speaker and listener patterns showed evidence for strategic or rapid selection processes, and whether this behavior was correlated with successful task performance [1].

Method

Participants

We recruited 57 dyads ($N = 114$) from the undergraduate psychology subject pool at Washington University in St. Louis, who were compensated via course credit. Twenty two dyads were unable to finish the task due to technical difficulties, out of which 4 dyads had less than 15 trials and were therefore excluded. The final sample thus consisted of 53 dyads.

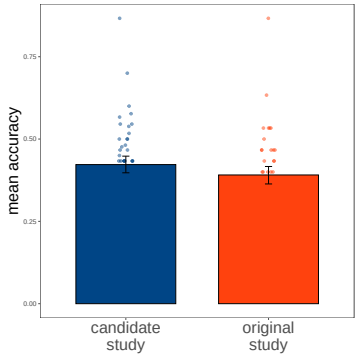

Figure 2: Average guesser accuracy across experiments. Error bards indicate bootstrapped 95% confidence intervals.

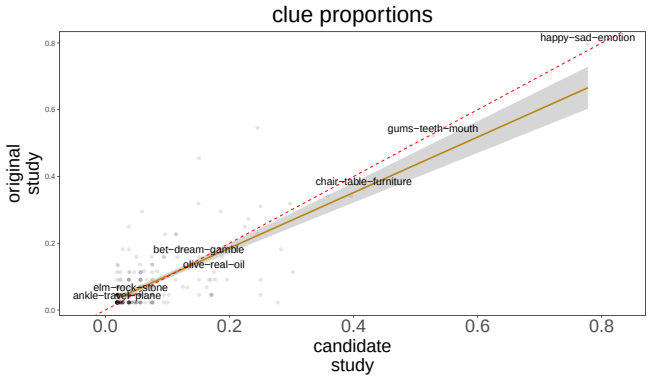

Figure 3: Proportion of times a clue was generated across the original and candidate study. Labels refer to target words (happy-sad) and the chosen clue (emotion). Dotted red line indicates a perfect fit and gold line indicates the empirical fit.

[1] All data and analysis scripts are available at https://github.com/cjk5642/CogSci2022-Connector

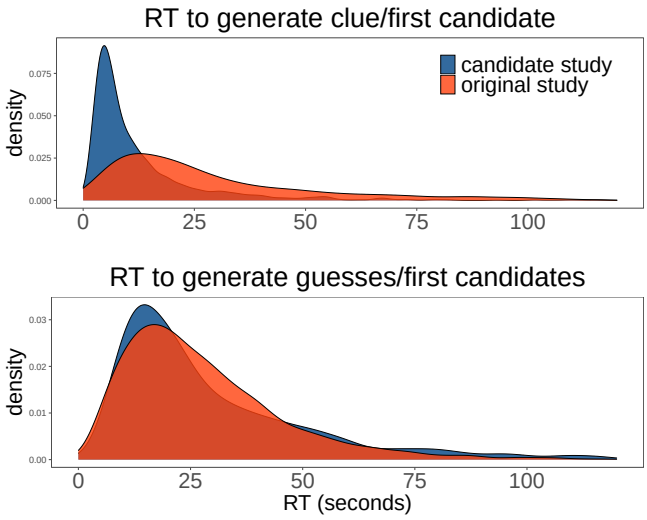

Figure 4: Response times to generate candidate clues and guesses across the original and candidate generation study.

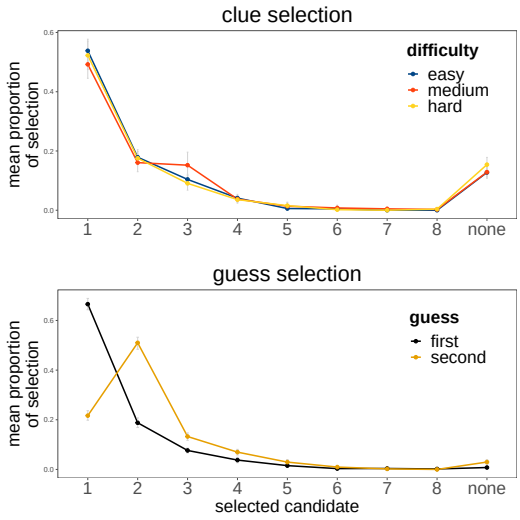

Figure 5: Proportion of times a candidate was chosen as the final clue or guess. "none" corresponds to cases when the final response was not among the initial candidates. Error bars indicate bootstrapped 95% confidence intervals.

Design and procedure

Figure 1 displays the general paradigm used in the study. The game was programmed in nodeGame (Balietti, 2017) and played online. Each participant was randomly assigned the role of a speaker or guesser for the duration of the game. Speakers were provided a word pair (e.g., *cave-knight*) from a 20-word board and had to find a one-word clue that would enable the guesser to identify that word pair from the same board. The game design and target words were identical to Kumar, Steyvers, & Balota (2021; Experiment 2) with one significant change: before generating the clues and guesses, speakers and guessers were asked to generate 3-8 candidates. During this candidate selection phase, speakers were encouraged to type in any word that popped into mind before selecting a final clue, whereas guessers were asked to click on as many words on the board as they deemed to be possible correct answers. Participant dyads played 30 rounds across 10 different boards, with 3 word pairs on each board of varying difficulty.

Results

Task comparison

First, we compared the patterns observed in the current "candidate-generation" study with those observed in original study by Kumar, Steyvers, & Balota to compare the types of responses generated in each experiment and evaluate whether the process of generating candidates altered the search process in any way.

Figure 2 shows the mean accuracy of the guesser across both datasets. A linear mixed effects model with random intercepts for participants and word pairs, revealed no significant effect of dataset (p=.435). Therefore, game accuracy was in the same range across both experiments.

Second, we examined whether the proportion of times a particular clue was selected for a given word pair differed across the two experiments. Figure 3 displays the clue proportions for some word pair and final clue combinations across both datasets. As shown, clue proportions were highly similar across the studies, with no significant effect of the dataset ($p > .05$).

Finally, we examined whether the manipulation of asking participants to type in candidates had any influence on the total time taken on each trial. Figure 4 (top panel) displays the response time (RT) to generate the first candidate in the candidate study, and RT to generate the clue in the original study[2]. As shown, speaker RTs were considerably faster in the candidate study ($b = 15.96$, $z = 6.96$, $p < .001$), which suggests that speakers were indeed typing any words that came to mind. On the other hand, as shown in Figure 4 (bottom panel), guesser RTs did not differ across the experiments ($p = .29$), which may be a function of pooling the RTs for the two guesses given that we had only one estimate of guesser RT from the original study, or it could mean that the task of visually scanning the board for potential answers was perceived to be similar across both studies by participants.

Taken together, these results suggest that the slightly different experimental procedure did not significantly alter the basic behavioral patterns in the experiment, although it did encourage speakers to respond faster and type in candidate words as they came to mind.

Response selection

Next, we examined the extent to which participants chose the different candidates as their final response and whether guesser accuracy varied as a function of which candidate was

[2]RTs above 120 seconds were excluded

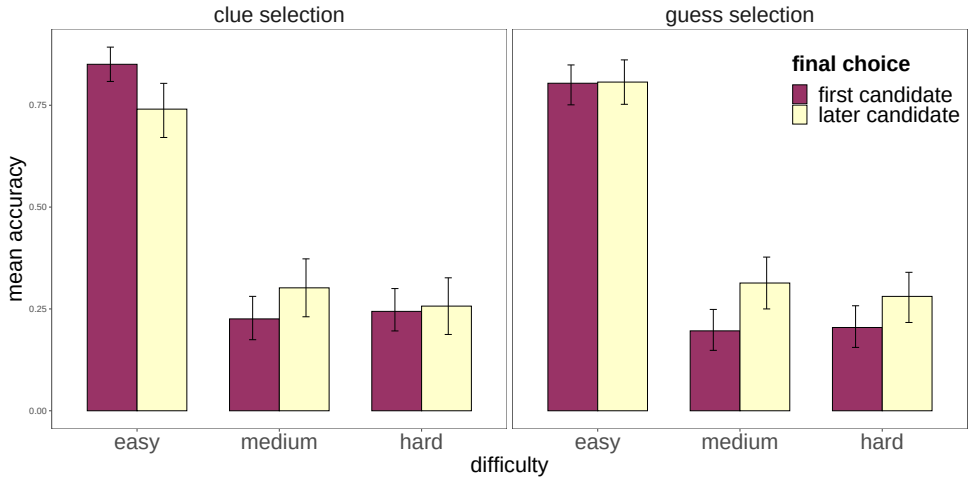

Figure 6: Mean guesser accuracy as a function of which clue and guess candidate was ultimately chosen by the speaker (left) and guesser (right). Error bars indicate bootstrapped 95% confidence intervals.

chosen in the candidate generation study.

Selection behavior Figure 5 (top-panel) displays the proportion of times different candidates were selected as the final clue by the speaker in the current study. As shown, participants selected the first word they thought of as their final clue over 50% of the times, and the likelihood of selecting later candidates was very low. There were no significant differences in selection across different levels of difficulty ($p > .05$). Figure 5 (bottom-panel) displays the proportion of times different candidates were selected as the final answers by the guesser. As shown, guessers were most likely to select their first candidate as their first final answer, and their second candidate as their second final answer. These effects again did not interact with difficulty (p's $> .05$).

Accuracy across selections Although participants generally selected their final responses among the first few candidates (most often the first one), it is not clear whether this was the best strategy in the current game. Therefore, we investigated whether selecting the first or one of the later candidates had any significant impact on overall task performance for both the speaker and guesser. We focused on the first four candidates in these analyses given that likelihood of selection sharply declined after this point. Specifically, for the speaker we scored whether the final clue was the first candidate typed in by the speaker or among the "later" three candidates. Similarly, for the guesser, we scored whether the final first answer chosen was indeed the first candidate clicked by the guesser, and whether the final second answer was the second candidate clicked by the guesser, or among the other three candidates.

Figure 6 (left panel) displays the accuracy of the guesser in selecting the correct word pair as a function of which clue candidate was selected and word pair difficulty. Accuracy was highest for *easy* word pairs when the first candidate was selected as the final clue. However, this pattern was reversed for *medium* word pairs, resulting in a significant interaction between difficulty and choice of candidate ($b = 1.13$, $z = 3.26$, $p = .001$), such that later candidates produced slightly higher accuracy. Therefore, speakers were biased towards selecting the first candidate as their final clue, even though they may have benefited from choosing one of the subsequent candidates for the "medium" word pairs in the task. "Hard" word pairs did not show any reliable effect of candidate choice. Along similar lines, as shown in Figure 6 (left panel), selecting later candidates for the medium word pairs was also beneficial in terms of guesser selections, which was again confirmed by a significant interaction between accuracy for easy and medium pairs' selection choices ($b = 1.20$, $z = 2.48$, $p = .013$). Selection choices for the harder word pairs showed a similar trend ($b = 0.51$, $z = 1.20$, $p = .23$)

Semantic search analysis

To better understand how search may be occurring within the context of the game, we next analyzed the semantic neighborhood of the candidates generated by speakers.

Patch construction We divided the candidates produced by speakers into items that fell within the word pair's *patch* and items that fell outside this patch. Patches were constructed for each word pair based on the methodology adopted by Davelaar (2015), such that all unique first candidates produced for a given word pair across all participants formed the patch for that word pair. Therefore, some word pairs may have smaller patches than other pairs if most participants produced similar first candidates. Figure 7 shows words within the patch constructed for the word pair *cave-knight*.

Next, we classified each subsequent candidate generated by participants as being inside or outside the constructed patch. Therefore, transitions across candidates could be from within the patch to within the patch (in-in), to outside the

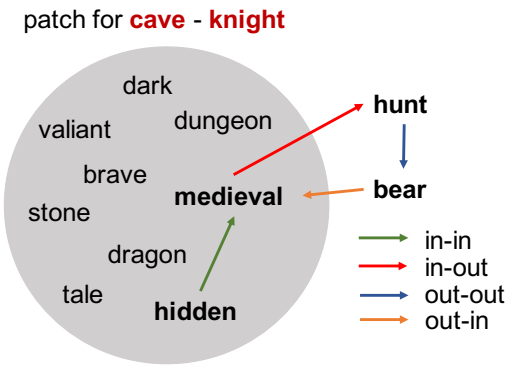

Figure 7: An example of transitions inside and outside the patch. The speaker generated candidates hidden→ medieval→hunt→bear and chose *medieval* as the final clue.

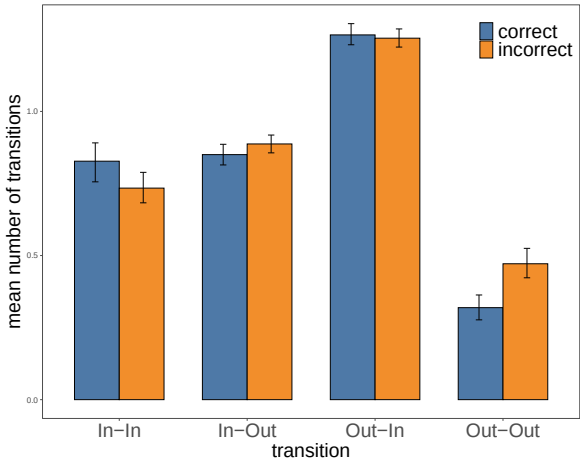

Figure 8: Correct trials had greater in-in and fewer out-out transitions compared to incorrect trials. Error bars indicate bootstrapped 95% confidence intervals.

patch (in-out), from outside the patch to inside the patch (out-in) and to outside the patch (out-out). Note that the transition from the first candidate to the second candidate could only be in-in or in-out, given that the patch was defined based on the first candidates. Figure 7 also shows examples of these transitions for the word pair *cave-knight*.

Solved and unsolved trials After constructing the patch and classifying the different responses into transitions in or out of the patch, we examined whether the frequency of different types of transitions varied as a function of whether a trial was successful or unsuccessful. As shown in Figure 8, incorrect trials had greater out-out and fewer in-in transitions, compared to correct trials. A linear mixed effects model with random intercepts for participants and word pair, and random slopes for transition type and accuracy revealed a significant interaction between transition type and accuracy ($p < .001$). Therefore, straying away from patch of initial candidates that came to mind led to unsolved trials.

Clustering Finally, we examined whether the different candidates produced by the speaker showed any evidence of clustering, as one might expect in semantic retrieval tasks. Clustering was measured via the Adjusted Ratio of Clustering (ARC; Roenker et al., 1971), which estimates the expected distribution of items across a given set categories. ARC values close to 1 are considered evidence for perfect clustering, values close to 0 are considered to be at chance, and negative values are considered evidence for anti-clustering. We calculated ARC values using the MemoryOrg package in R (Greeley, 2021) for each unique sequence of candidates generated by individuals as a function of whether the candidate was semantically closer to one of the target words. Semantic similarity was operationalized via an random walk associative model based on the Small World of Words (SWOW) dataset (De Deyne et al., 2019). For example, for a given target word pair *cave-knight* and a candidate sequence hidden→ medieval→hunt→bear, we classified each candidate as being closer to either *cave* (1) or *knight* (2) based on the SWOW model. The candidate sequence above would therefore have the assigned labels 1, 2, 1, and 1, which was then used to calculate an ARC value of -1. This process was repeated for each individual candidate sequence generated by participants. These ARCs were then averaged across all trials, and separately for solved and unsolved trials, and difficulty.

First, the mean ARC across all responses was -0.009 (*SD*=1.33), which was not reliably different from 0, $t(1114) = .24$; p=.81, therefore suggesting that candidate responses were not systematically clustered towards any one of the target words. ARC also did not reliably differ across solved and unsolved trials, or across different difficulty levels (p's>.05). Therefore, candidates generated did not show any reliable clustering with respect to the two target words.

Discussion

In this work, we explored how individuals search through memory for related concepts when they have a communicative goal in mind. We employed a "candidate-generation" procedure, by which we asked speakers and guessers to generate a set of potential candidates before they made their final choice. Our main findings were: (1) speakers and guessers were both biased towards selecting their first responses in the task, even when later responses may have been more effective, and (2) the initial candidates generated across all speakers represented an optimal patch, such that when speakers produced candidates that were outside this patch, this led to lower performance overall. We now discuss the broader implications of each of these findings.

When examining the selection choices of the speaker and guesser, we found that both players were most likely to select their final choices from among the *first* candidates they generated. This bias towards the first response may be indicative of two possibilities: speakers may be employing a simple "take-the-first" heuristic (TTF; Johnson & Raab, 2003) and

sticking with their initial "hunch", a strategy that has been shown to be effective in other contexts, such as sports, where decisions have to be made with limited knowledge about familiar, but ill-defined situations (Hepler & Feltz, 2012). It is possible that in the face of uncertainty about how their partner may perceive different choices, speakers tend to rely on such a heuristic and offload some of the inferential work on their partner. Indeed, given that previous work has shown that associative models generally best capture the types of responses generated in this task (Kumar, Garg, & Hawkins, 2021; Kumar, Steyvers, & Balota, 2021), participants may indeed be relying on these initial associations that come to mind.

On the other hand, it is possible that the current manipulation of asking participants to generate candidates as they came to mind may not have been fully effective, and participants simply responded *after* their search for optimal candidates had ended. Although this is possible, given that speakers were overall faster at responding in this experiment compared to the Kumar, Steyvers, & Balota (2021) experiment (see Figure 4), and that over 40% of the final responses were *not* the initial response produced by the speaker (see Figure 5), we believe that participants did in fact follow task instructions and were responding with potential candidate words as they came to mind. Future work could explore other methods of probing participants in such complex tasks, such as eye-tracking, which may provide deeper insights into the types of search processes involved in complex reference games. Eye movements may be particularly useful in understanding how speakers and guessers emphasize similar or different aspects of the context (i.e., the board) and whether this correlates with their initial candidate responses. Furthermore, comparing eye movements between partners could provide further insights into whether speaker-listener preferences align with each other.

Another interesting finding from this work was that participants chose their initial responses even when the later candidates may have been more effective (see Figure 6). This was especially true for "medium" word pairs, and to some extent also for harder word-pairs. This suggests that the TTF strategy may not be optimal in *all* contexts. Specifically, "medium" word pairs were those that were slightly related to each other, such as *cage-glass*, *sun-bowl*, *army-drum*, etc. In such cases, it is possible that the first word that comes to mind may be a strong associate for one of the target words but not the other, but the subsequent associates may indeed be better suited to both targets. Table 1 shows examples of such cases, where the first associate generated by the speaker was often sub-optimal, but the later candidates were better clues and led to successful trials. Importantly, these "later" candidates were still among the initial four candidates generated by the speaker. Therefore, these patterns indicate that while the "fast and frugal" TTF strategy may be optimal for easy associations, there could also be benefits to evaluating some of the later candidates in cases when the initial associate is not optimally related to both targets. However, we did not find a reliable effect of "hard" word pairs, which may indicate that when concepts are truly unrelated, coming up with related concepts is a difficult and unfamiliar task, given that most conversational situations require us to make connections between somewhat related concepts. Future work should examine the specific conditions under which simple heuristics guide semantic search and retrieval, and the types of relationships that individuals draw upon when engaging in communication about multiple concepts.

word pair	**clue candidates**	**final choice**	**accuracy**
hand-birth	Sistine, creation	Sistine	0
	baby, small		
	doctor, baby,	give	1
	give		
army-drum	instrument,	instrument	0
	kill, soldier		
	rhythm,	marching	1
	marching, band		
cage-glass	delicate,	delicate	0
	dangerous, box		
	zoo, window	exhibit	1
	exhibit		

Table 1: Examples of "medium" word pairs where later clue candidates were closer associates to *both* target words and resulted in higher accuracy.

Finally, our analyses of the semantic "patch" and different transitions for each word pair revealed that most of the words either stayed within the patch or, if they traveled outside the patch, they stayed outside the patch (Figure 8). Furthermore, generating words that were outside the patch increased the likelihood of an unsuccessful trial. Indeed, the final clue chosen by the speakers was within the patch 78% of the time. This begs the question of what kinds of words were in this patch, and whether we can determine how this patch of highly clustered initial responses is generated. For example, a fast spreading activation-type method may be responsible for producing these highly clustered initial items, and future work could explore different implementations of this mechanism to generate the optimal semantic patch. Alternate corpora and underlying semantic models, such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020) could also be used to compare multiple patch methods.

The present work showed that speakers and listeners are biased towards initial responses when tasked with generating optimal utterances in a reference game with multiple targets. This "take-the-first" strategy is generally successful, although it can lead to sub-optimal responses for less related concepts. Overall, this work indicates that individuals are able to converge onto the most relevant responses relatively quickly in semantic retrieval tasks with loosely-defined constraints. Future work should investigate the specific mechanisms underlying the fast activation processes that mediate this search behavior.

References

Balietti, S. (2017). nodegame: Real-time, synchronous, online experiments in the browser. *Behavior Research Methods*, *49*(5), 1696–1715.

Barsalou, L. (1983). Ad hoc categories. *Memory Cognition*, *11*(2), 211–227. doi: https://doi.org/10.3758/BF03196968

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Davelaar, E. J. (2015). Semantic search in the remote associates test. *Topics in Cognitive Science*, *7*(3), 494–512.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, *51*(3), 987–1006.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science*, *20*(11), 818–829.

Greeley, G. D. (2021). Memoryorg: (a few) retrieval organization metrics for human memory research. Retrieved from `https://github.com/ggreeley` (R package version 0.0.0.9000)

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, *44*(e12845).

Hepler, T. J., & Feltz, D. L. (2012). Take the first heuristic, self-efficacy, and decision-making in sport. *Journal of Experimental Psychology: Applied*, *18*(2), 154.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science advances*, *7*(47).

Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes*, *91*(2), 215—229.

Kumar, A. A., Garg, K., & Hawkins, R. D. (2021). Contextual flexibility guides efficient communication in a cooperative language game. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, *43*.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). Semantic memory search and retrieval in a novel cooperative word game: A comparison of associative and distributional semantic models. *Cognitive Science*. doi: https://doiorg/101111/cogs13053

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: General*, *106*(3), 226.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, *77*(4), 257.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, *76*(1), 45—48. doi: https://doi.org/10.1037/h0031355