COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13053
© 2021 Cognitive Science Society LLC
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13053

Semantic Memory Search and Retrieval in a Novel Cooperative Word Game: A Comparison of Associative and Distributional Semantic Models

Abhilasha A. Kumar,[a] Mark Steyvers,[b] David A. Balota[c]

[a]*Department of Psychological and Brain Sciences, Indiana University*
[b]*Department of Cognitive Sciences, University of California, Irvine*
[c]*Department of Psychological and Brain Sciences, Washington University in St. Louis*

Received 4 July 2020; received in revised form 30 July 2021; accepted 11 September 2021

Abstract

Considerable work during the past two decades has focused on modeling the structure of semantic memory, although the performance of these models in complex and unconstrained semantic tasks remains relatively understudied. We introduce a two-player cooperative word game, Connector (based on the boardgame Codenames), and investigate whether similarity metrics derived from two large databases of human free association norms, the University of South Florida norms and the Small World of Words norms, and two distributional semantic models based on large language corpora (word2vec and GloVe) predict performance in this game. Participant dyads were presented with 20-item word boards with word pairs of varying relatedness. The speaker received a word pair from the board (e.g., *exam-algebra*) and generated a one-word semantic clue (e.g., *math*), which was used by the guesser to identify the word pair on the board across three attempts. Response times to generate the clue, as well as accuracy and latencies for the guessed word pair, were strongly predicted by the cosine similarity between word pairs and clues in random walk-based associative models, and to a lesser degree by the distributional models, suggesting that conceptual representations activated during free association were better able to capture search and retrieval processes in the game. Further, the speaker adjusted subsequent clues based on the first attempt by the guesser, who in turn benefited from the adjustment in clues, suggesting a cooperative influence in the game that was effectively captured by both associative and distributional models. These results indicate that both associative and distributional models can capture relatively unconstrained search processes in a cooperative game setting, and Connector is particularly suited to examine communication and semantic search processes.

Keywords: Semantic retrieval; Distributional semantic models; Semantic memory; Memory search; Language games

Correspondence should be sent to Abhilasha A. Kumar, 1101 E 10th St, Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, 47405, USA. E-mail: abhilasha.kumar@wustl.edu

1. Introduction

Retrieval from semantic memory is ubiquitous in cognitive tasks. For example, humans retrieve concepts when they describe other concepts (e.g., kittens are young cats), assess similarity or relatedness (e.g., cats and dogs are both pets), and recall items from a particular category (e.g., groceries, animals, etc.). Indeed, one might argue that every instance of communication demands retrieval from semantic memory. Of course, understanding *how* concepts are retrieved from semantic memory requires establishing a structural account of how these concepts are organized in memory. Consequently, significant work in semantic memory research has attempted to model how concepts are learned, stored, and organized, by conceptualizing semantic memory through large-scale semantic networks (for a recent review, see Siew, Wulff, Beckage, & Kenett, 2018) or high-dimensional vector spaces (for a recent review, see Günther, Rinaldi, & Marelli, 2019). Traditionally, computational accounts of semantic memory have assumed *context-free* semantic representations, implying that there is a single representation of each concept in memory that is not influenced by task demands. However, there is now considerable evidence to suggest that retrieval from semantic memory is inherently contextual (for a review, see Yee, Lahiri, & Kotzor, 2017) and depends upon linguistic and task-based contexts within which words are retrieved. Therefore, it is important to investigate how recent computational models of semantic memory accommodate different types of task-based contexts that may influence retrieval and search processes to achieve the task goal.

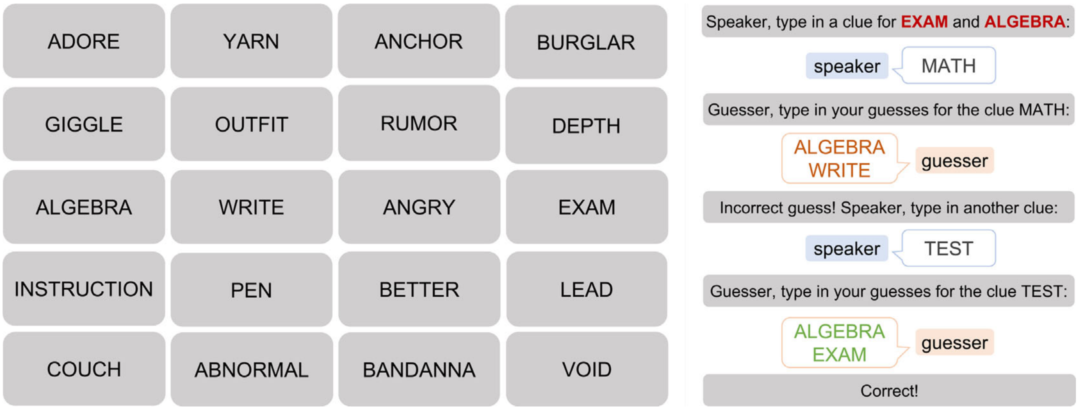
In order to study retrieval from semantic memory, it is critical to formalize the *context* within which retrieval occurs. Some work in this domain has attempted to define the "context" of semantic search in terms of the *semantic category* of the retrieval cue. For example, Hills, Todd, and Jones (2015) showed how search processes in the semantic fluency task (where individuals are asked to produce as many exemplars from a category, e.g., *animals*, as possible within a given time period) mimic patterns of optimal foraging for food found among animals in their natural environment. Other work has examined how "context" in the form of multiple retrieval cues influences search processes, such as finding creative associations in the remote associates test (RAT, e.g., Davelaar, 2015; Smith, Huber, & Vul, 2013) or navigating from one word to another in semantic word games (e.g., Beckage, Steyvers, & Butts, 2012). Although different tasks bring online different processes, formalizing the context of semantic search via word games offers the opportunity to study different types of semantic memory processing and how individuals modify their search and retrieval strategies in response to task demands.

Word games have been considered fundamental tools to study the nature and structure of language (Wittgenstein, 1953) and also have a rich history in natural language processing (for a recent brief review, see Moskvichev & Steyvers, 2019), where the performance and success of artificial systems are often tested through word games, such as Jeopardy (IBM Watson; Ferrucci et al., 2010), chess (Deep Blue; Campbell, Hoane Jr., & Hsu, 2002), and Go (Alpha Go; Silver et al., 2017). As discussed, word games have also proven to be useful tools to explore semantic search processes. For example, Beckage et al. (2012) had participants navigate from a randomly chosen word (e.g., *anything*) to a target word (e.g., *pen*) using a given set of candidate words at each step. They found that individuals were relatively successful at finding such paths (73% success rate) and used the global structure of semantic memory (modeled using word association norms, Nelson, McEvoy, & Schreiber, 2004) to play the game. More recently, Fathan, Renfro, Austerweil, and Beckage (2018) used a similar game to model the search process as a weighted random walk (RW) over nodes in a large memory network.

Game-based studies represent an important case of complex choice or decision-based processes operating over structural representations of semantic memory that likely differs from the mechanisms underlying other semantic retrieval-based tasks, such as the semantic fluency (Hills et al., 2015) or RAT task (Smith et al., 2013). Specifically, game-based tasks allow for more flexibility and creative choices in semantic retrieval, compared to standard semantic retrieval tasks, therefore offering a unique opportunity to investigate explicit semantic retrieval processes in a relatively unconstrained manner. For example, although the search is relatively unconstrained in the RAT (where individuals are asked to produce one word related to three words, e.g., *cottage*, *swiss*, *cake*), there is typically *one* correct answer to RAT problems (e.g., *cheese*). Similarly, although individuals can produce semantically related words in different ways within the semantic fluency task, they are restricted by the semantic *category* (e.g., animals, vegetables, etc.). On the other hand, word games offer the opportunity to relax these constraints and allow individuals to connect concepts in relatively complex and more flexible ways. Indeed, Marrs, Straka, and Beckage (2017) have argued that human performance in word games is not easily explained through random walk models and may require additional considerations driven by semantic similarities between words.

Despite the potential to track complex search and retrieval processes, there are some limitations to the game-based studies discussed above. Specifically, within the semantic word games, attention is often directed to the *navigational* aspects of the game, that is, individuals are specifically instructed to start from a particular word and navigate to the goal/target word, therefore limiting the ways in which individuals may consider two words to be related. For example, an individual may think *virus* and *weather* are connected through *sickness*, but in the MindPaths game (Marrs et al., 2017), their performance depends on necessarily finding a path from *virus* to *weather* through forced choice among a limited set of words (derived from free association norms) that may or may not include the relationship they had originally inferred. The current study proposes an alternative game-based methodology to study retrieval from semantic memory by constraining the overall context through the game structure but also ensuring that the search process remained relatively unconstrained. Importantly, because there are multiple attempts between two agents in the game, we can also explore how individuals interact to achieve the task goal.

The current study introduces the Connector game, a two-player semantic word game based on the popular multiplayer word game, Codenames designed by Vlaada Chvátil in 2015 (see https://czechgames.com/en/codenames/). Codenames won the Game of the Year award in 2016 (Veyra, 2016) and has since been published in 38 languages. Within research in natural language processing, the format of Codenames has been recently adapted to investigate associative creativity (Zunjani & Olteteanu, 2019), compare artificial bot performance on the game when trained on modern distributional models such as word2vec and Global Vectors



ADORE	YARN	ANCHOR	BURGLAR
GIGGLE	OUTFIT	RUMOR	DEPTH
ALGEBRA	WRITE	ANGRY	EXAM
INSTRUCTION	PEN	BETTER	LEAD
COUCH	ABNORMAL	BANDANNA	VOID

Speaker, type in a clue for **EXAM** and **ALGEBRA**:

speaker MATH

Guesser, type in your guesses for the clue MATH:

ALGEBRA
WRITE
guesser

Incorrect guess! Speaker, type in another clue:

speaker TEST

Guesser, type in your guesses for the clue TEST:

ALGEBRA
EXAM
guesser

Correct!

Fig 1. An experiment trial in Connector. The speaker is given two words from a 20-word board (*exam* and *algebra*) and comes up with a one-word clue (*math*) that is delivered to the guesser. If the guesser fails to guess the word pair in the first attempt, the speaker provides another clue (*test*) and the guesser can attempt to identify the word pair again (see "Procedure" section for details).

(GloVe; Kim, Ruzmaykin, Truong, & Summerville, 2019), and also evaluate various models of associative meaning based on distributional principles (Shen, Hofer, Felbo, & Levy, 2018; also see Xu & Kemp's, 2010, work on the closely related Password game). Additionally, Codenames has been used to develop educational tools to teach complex concepts and vocabulary to students (Octaviana, Rahmah, & Puspitasari, 2019; Souza, Morais, & Girardi, 2018).

The Connector game differs from Codenames[1] and previous work in this area,[2] as all interactions between players occur in real time, and both players work cooperatively to arrive at the correct answer. In Connector, two players view a grid of 20 words and are randomly assigned the roles of a speaker and guesser (see Fig. 1). For each trial of the game, the speaker is provided a word pair on the board (e.g., *algebra-exam*) and generates a one-word clue related to both words (e.g., *math*) that would serve to help the guesser identify the two words. This clue is then displayed to the guesser, who uses it to identify which two words on the board the clue was most likely referring to. If the first attempt is unsuccessful, the speaker can provide a second clue to the guesser for a particular word pair, with this process being repeated on a third trial if the guesser is still unsuccessful. In this way, one is able to constrain the *retrieval context* to the word pairs on the board, but the game also allows participants to freely search their semantic space during the task in a social/cooperative context. Specifically, although the first clue is entirely dependent on the search process of the speaker, subsequent second and third clues can vary based on the responses of the guesser; hence, the paradigm allows for investigating these cooperative interactions.

In addition to proposing a novel game-based method for examining retrieval from semantic memory, the current study had three research goals. First, we were interested in evaluating the predictive power of different semantic models in explaining performance in this game. Previous work in this area has examined how associative semantic networks (Steyvers &

Tenenbaum, 2005) created from free association norms (e.g., Nelson et al., 2004) can account for word choice in word games (Beckage et al., 2012; Marrs et al., 2017). Association network models represent the lexicon as a large memory network, where words are represented via nodes, and semantic relationships are represented via edges, in line with early work by Collins and Loftus (1975). The relationships are typically extracted using norms derived from free association tasks where participants produce a single word (e.g., Nelson et al., 2004) or multiple words that come to mind in response to a given cue (e.g., De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019; Kenett, Kenett, Ben-Jacob, & Faust, 2011). Although previous research has compared some network configurations (e.g., weighted vs. unweighted networks, Fathan et al., 2018), no work has compared different associative norms within a game-based task. Indeed, in addition to the widely used University of South Florida (USF)-free association norms (further described below), there now exist more recent and larger databases of association norms, such as the Small World of Words database (SWOW; De Deyne et al., 2019). While the USF norms are based on a relatively smaller set of items (5000 cues) and only contain the primary free association response, SWOW is based on a larger set of items (over 12,000 cues) and also records secondary and tertiary free association responses. Importantly, it remains unknown whether there are differences in the extent to which these norms can predict unconstrained semantic retrieval processes within a game-based setting. Therefore, the present work provides a novel comparison across associative models derived from two different free association databases, USF and SWOW norms, in this complex game task.

In addition, although associative models have been successful in accommodating a considerable number of findings on semantic memory retrieval (for a review, see Siew et al., 2018), they have also been criticized because the associative information itself is generated through semantic retrieval, that is, one is predicting semantic retrieval directly from explicit semantic retrieval norms (Jones, Hills, & Todd, 2015). Therefore, given the potential overlap between the retrieval processes in the free association task and the explicit retrieval involved in word games, one may a priori expect associative models to perform relatively well in the present game tasks. Of course, the counterargument here is that associative models contain unique sources of information that are difficult to capture via purely linguistic corpora (see De Deyne, Perfors, & Navarro, 2016), which may be particularly important within unconstrained semantic tasks. In contrast to the associative models based on free association norms, within the past decade, there has been an explosion of "distributional semantic models" (DSMs) that propose explicit mechanisms for how humans learn word meaning from natural language. DSMs use large-scale language corpora (e.g., Wikipedia database, Google News articles, etc.) to infer semantic representations by applying complex learning algorithms to co-occurrence patterns of words. DSMs have shown unprecedented success at explaining behavioral performance across a variety of semantic tasks (for a recent review, see Günther et al., 2019). Importantly, DSMs represent a significant departure from traditional associative accounts, in which they infer word meaning from distributional information present in language (compared to associative information) and are typically derived from large text corpora (compared to free association norms). Therefore, a major research goal of this study was to evaluate whether modern DSMs (word2vec and GloVe; described in the Methods section) can explain word choice in our relatively unconstrained semantic game task and how they compare to associative models. As noted above, one might a priori expect associative representations to better capture the associative patterns in this game task, compared to DSMs because associative models are directly constructed from human associative information. Therefore, if DSMs can achieve comparable performance or actually provide a better account of observed patterns in this word game, it would suggest that distributional information derived solely from natural language corpora is sufficient to account for conceptual processing in this game. Alternatively, if associative models do indeed outperform the DSMs, this may point to unique meaning-related information present within the associative norms that may not be effectively captured via DSMs trained on text corpora. As discussed earlier, although Shen et al. (2018) and Kim et al. (2019) compared the performance of word2vec and GloVe models in their game-based studies, they did not compare DSMs to associative models and also did not model real-time player interactions in their work. Therefore, the present work adds to previous literature by making novel comparisons across different classes of semantic models in a two-player cooperative word game.

Given that all responses in our game were made in real time, a second goal of this study was to begin investigating the extent to which these semantic models can accommodate response latencies, a relatively unexplored issue in past work using games to explore retrieval from semantic memory. In particular, response latencies may provide useful information about the search process, that is, faster responses may indicate that words are closer in semantic space or there is lesser competition among words, whereas slower responses may indicate that words are more distant in semantic space, or there is more competition among words. Given that the format of the Connector game allows for unconstrained responses in the form of speaker-generated clues, response latencies may be particularly useful in understanding the search and retrieval process of the speaker.

The third goal of this project was to begin examining how interactions contribute to search processes in this task. As discussed earlier, cooperative word games such as Codenames are extremely popular in social settings and not only involve individual search processes but also social skills such as perspective-taking and referential communication in order to retrieve the right concepts from a shared semantic memory space. To understand the extent to which players in Connector engaged in social collaboration and perspective-taking, we explored the clues that the speaker provided the guesser after a failed first attempt. In this case, the speaker is given information by the guesser to better understand the failed search attempt. The speaker can use this information to provide a cue that should facilitate the guesser's retrieval on the next trial. For example, for the word pair *exam-algebra*, first clue *math* and first incorrect guess *algebra-pen*, the speaker could use the knowledge of the first incorrect guess and steer the guesser in the direction of the correct word, by providing a clue that is closer to the word that was not guessed correctly (e.g., provide *testing* as a second clue to steer the guesser closer to *exam*).

We report combined results from two experiments conducted with different sets of items.[3] Across both experiments, we explored whether patterns of responses and response latencies in clue generation and word guessing were predicted by estimates of semantic similarity derived from different computational models of semantic memory.

2. Methods

2.1. Participants

A total of 156 students ($M_{age} = 21.1$ years, $SD = 3.1$) were recruited in dyads ($N = 78$) from undergraduate courses at Washington University in St. Louis and compensated through course credit for their participation. The mean years of education in the sample was 14.5 years ($SD = 1.8$), and the mean score on the Shipley Vocabulary Test was 33.23 ($SD = 3.22$). Twelve participants were non-native English speakers, out of which only two indicated that they learned English after the age of 5 and whose performance differed from the sample. Furthermore, one participant did not comply with the game instructions and gave two-word clues to the guesser. Therefore, we excluded the dyads to which these three participants belonged from our final sample, which consisted of 75 dyads. The game was played in person, and we explicitly asked participants within each dyad to confirm that they did not know each other before playing the game.

2.2. Semantic models

2.2.1. Distributional models

As discussed earlier, DSMs typically use large-scale text corpora to approximate the natural language environment and apply statistical techniques to infer semantic word representations. Word2vec (Mikolov, Chen, Corrado, & Dean, 2013) is a three-layer neural network-based DSM that comes in two versions: The *skip-gram* model is trained to predict four context words before and after a particular target word in a sentence, whereas the *continuous bag-of-words* model reverses this objective. Both versions of the word2vec model refine representations using an error-driven learning mechanism. By training on millions of sentences in a large text corpus, word2vec tends to develop very rich semantic representations, which have proven to be useful inputs for several downstream natural language processing and semantic tasks (Baroni, Dinu, & Kruszewski, 2014; Collobert & Weston, 2008; Mandera, Keuleers, & Brysbaert, 2017). Another popular embedding model, GloVe was introduced by Pennington, Socher, and Manning (2014). Although GloVe is a DSM, unlike word2vec, GloVe constructs a word-by-word co-occurrence matrix and attempts to predict the *ratio* of co-occurrence probabilities between words using a regression model. GloVe has been shown to perform remarkably well at analogy tasks, word similarity judgments, and named entity recognition (Pennington et al., 2014), similar to word2vec. Therefore, we assessed word2vec (*skip-gram* version) and GloVe, compared to each other and the associative models (described below). For all analyses, pretrained word2vec (*skip-gram* version) and GloVe models, both trained on the English Wikipedia 2017 corpus (3 billion tokens) were used, which were available from Kutuzov, Fares, Oepen, and Velldal (2017). These pretrained models were processed via the *pymagnitude* package in Python available via Patel, Sands, Callison-Burch, and Apidianaki (2018), and both models produced 300-dimensional word vector representations.

2.2.2. Associative models

We utilized two different databases of free association norms, namely, the USF norms collected from over 6000 participants for 5018 cues by Nelson et al. (2004), and the more recent SWOW norms collected from over 88,000 participants for 12,217 cues by De Deyne et al. (2019). For each database, three different similarity measures were examined, based on (a) associative strength (S), (b) positive pointwise mutual information (PPMI), and (c) random walk (RW) measures.[4] As described in De Deyne et al. (2019), the S measure refers to the probability of responding with a word given a particular cue. The PPMI measure takes the general pattern of responses across *all* cues into account when considering the similarity between words, thus emphasizing responses that are unique to specific cues and de-emphasizing responses that are produced for several cues. The RW measure considers not only the direct responses produced to given cues but also any indirect paths or neighbors of neighbors as would be consistent with a spreading activation mechanism (Collins & Loftus, 1975). The RW measure is based on a decaying random walk process, which estimates a weighted sum of paths for a given pair of words, by assuming a damping parameter (alpha; fixed at .75 as in De Deyne et al., 2019) that controls the extent to which similarity is driven by shorter or longer paths. For details of exact implementations of these models, the reader is referred to De Deyne et al. (2019). Importantly, the S, PPMI, and RW measures were calculated for 4927 cues in the USF database (removing 90 cues with fewer than 100 responses as in De Deyne et al., 2019) and 12,217 cues in the SWOW database. Furthermore, while the USF database only contained primary responses to the cue, the SWOW database contained primary, secondary, and tertiary responses to a given cue. Given the difference in the total number of words between the USF and SWOW norms, we also explicitly evaluated the differences across the models in additional analyses that restricted the SWOW dataset to the same number of words as the USF dataset, as well as models that examined only primary associations (see the Results section).[5]

2.2.3. Constructing word association spaces

Given that DSMs produce semantic vectors projected onto a high-dimensional space, whereas the associative models produce "similarity" estimates between words based on free association norms, there are limitations to the types of comparisons one can make with these different representations. For example, although networks can be created via associative models (as in Kumar, Balota, & Steyvers, 2020; Steyvers & Tenenbaum, 2005, etc.), the criterion used for defining path length in the majority of associative models is somewhat arbitrary (with the exception of RW models, which consider walks of infinite length) and does not directly map onto the notion of cosine similarity between word vectors within a semantic space as in the DSMs. To address these differences, we converted the information derived from associative models above into Word Association Spaces (WAS; Steyvers, Shiffrin, & Nelson, 2005), a technique that places words within a high-dimensional space by applying multidimensional scaling on free association data. WAS utilizes principles similar to latent semantic analysis (Landauer & Dumais, 1997) and applies singular value decomposition (a factor-analytic technique) to infer direct and mediated paths between words and in this way uncovers "latent" semantic representations of words. In the present work, for each model derived from the

USF/SWOW norms (i.e., S, PPMI, and RW), we computed a 300-dimensional word association space,[6] such that every word within the norms was now represented as a 300-dimensional vector in this space, similar to the vectors produced by the DSMs, to effectively compare these vectors in predicting game performance. All subsequent analyses were based on these semantic word vectors derived via associative models or DSMs.

2.3. Materials

Ten boards of 20 words were created, with each board consisting of 10 word pairs. Word pairs were classified as "close" (e.g., *happy-sad*), "medium" (e.g., *army-drum*), or "distant" (e.g., *cave-knight*) based on tertiles defined via average cosine similarities between the words based on all semantic models. Each board was then constructed to ensure that it contained approximately three to four of these "close," "medium," and "distant" word pairs.[7] Among these 10 word pairs on each board, we randomly selected *one* "close", "medium", and "distant" word pair to serve as the stimuli for the game task. Each board was used for three trials, resulting in 30 word pairs across 10 boards for each participant dyad. Twenty different boards with non-overlapping word pairs were used across the two experiments. At any given time during the task, the boards were displayed page-wise to both the speaker and the guesser using booklets.

3. Procedure

Participants were informed that they would be playing a two-player word game with another participant and were introduced to each other before the experiment. Following this introduction, participants were randomly assigned the role of a speaker and guesser for the rest of the experiment. Both participants were handed a booklet of the boards to be used during the study. Participants were instructed to turn the booklet page or respond *only* when the computer program specified that it was their turn (speaker's or guesser's turn, respectively). Both participants then proceeded to complete the practice session in the same room. After the practice session, one of the participants moved to an adjoining room for the duration of the task and viewed the same (shared) computer screen as their partner for the duration of the experiment.

Each experimental trial began with instructions to both players to turn the page and view the board (see Fig. 1). During this time, participants familiarized themselves with the 20 words on the board. Once the speaker was ready to see the word pair for clue generation, they were instructed to turn their booklet page and press the return/enter key. When the speaker turned their page, two words (e.g., *exam-algebra*) on the same board were highlighted in red ink, and their task was to generate a clue that was related to both words and would serve as a good clue for the guesser (who did not see the words in red ink). The speaker typed their clue (e.g., *math*) and pressed the "9" key to end their response. After pressing 9, the screen turned red for 500 ms to direct the guesser's attention to the screen (to ensure that they were indeed looking at the screen and not the board at this time). The screen then displayed the speaker's clue, instructing the guesser to identify two words on the board that matched this clue. After the guesser typed their guesses and pressed 9, the program evaluated whether the guesses corresponded to the word pairs assigned to the speaker.[8] If the guessed words were correct, the program congratulated the players and moved to the next word pair on the board. If one or both of the guessed words were incorrect, the program informed both players that the words were not guessed correctly (without specifying whether one or both words were incorrect) and instructed the speaker to provide a second clue for the same word pair. This process was repeated a third time if the guesser did not successfully guess the word pair in the second attempt, resulting in at most three total attempts for each word pair. If the word pair was not guessed in the final attempt, or when the word pairs were guessed correctly, the program instructed the speaker to turn the page and view the next word pair for the same board and advance the program. After completing three word pairs on one board, both players were instructed to turn the page and view the next board. The experiment then proceeded in a similar manner for the remaining nine boards.

4. Results

4.1. Game descriptives

Overall, participant dyads were very successful at the game, correctly guessing word pairs with an overall success rate of 87% ($SD = 9\%$) across all three attempts. Given that word pairs were classified as "close," "medium," and "distant" (see Materials section), it is important to evaluate the differences in performance across distance categories. First, "close" word pairs were retrieved in significantly fewer attempts ($M = 1.45$, $SD = 0.79$), compared to "medium" word pairs ($M = 2.12$, $SD = 1.02$), $p < .001$, and "medium" word pairs were retrieved in significantly fewer attempts, compared to "distant" word pairs ($M = 2.40$, $SD = 1.05$), $p = .025$. Table 1 reports the performance of the speaker and guesser across the three attempts and distance levels in this word game. As shown, we also found that response times (RTs) to generate clues and guesses were significantly faster for "close" word pairs, compared to "medium" and "hard" word pairs. However, for both the accuracy and RT measures, the differences were no longer significant in the third attempt, likely due to smaller sample sizes in the third attempt, especially for close items.

Table 2 displays the three most frequent first clues and the proportion of speakers who chose those clues across "close," "medium," and "distant" word pairs across both experiments. As shown, although there was more agreement among speakers for the "close" word pairs, there was still considerable overlap and agreement in the clues even for "medium" and "distant" word pairs, suggesting that individuals were able to converge onto the same clues even when words were not a priori closely related to each other.

4.2. Model accuracy in predicting speaker responses

For all regression-based analyses that follow, we used generalized or simple linear mixed effect models (LME) from the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in the RStudio environment (R version 4.0.2 (2020-06-22), R Core Team, 2020). Total explained

Table 1
Descriptive statistics about performance in the Connector game

Attempt	Overall Guesser Accuracy (*SE*)	Distance Between Words	*N*	Mean response time (RT) to Generate Clue in seconds (*SE*)	Guesser's Mean RT in seconds (*SE*)	Guesser Accuracy (*SE*)
First	0.41 (0.01)	Close	75	19.28 (0.73)	21.67 (0.59)	0.69 (0.02)
		Medium	75	30.80 (1.04)*	27.84 (0.62)*	0.32 (0.02)*
		Distant	75	37.70 (1.29)$^{+}$	32.20 (0.74)$^{+}$	0.21 (0.02)$^{+}$
Second	0.55 (0.01)	Close	73	19.66 (1.28)	20.14 (1.21)	0.67 (0.03)
		Medium	75	22.47 (1.00)$^{\wedge}$	24.23(0.94)*	0.54 (0.02)*
		Distant	75	26.26 (1.18)$^{+}$	26.84 (0.98)$^{+}$	0.50 (0.02)$^{\#}$
Third	0.51 (0.02)	Close	47	26.72 (2.76)	18.46 (2.56)	0.57 (0.06)
		Medium	69	26.13 (1.60)	31.07 (2.84)*	0.56 (0.03)
		Distant	71	29.16 (1.80)	26.95 (1.85)	0.45 (0.03)$^{+}$

Note. *N* denotes total number of participants within each attempt and distance level. Standard errors are indicated in parentheses. Means were calculated across all trials and differences were assessed using mixed effect models. * indicates significant difference between close and medium word pairs, $^{\wedge}$ indicates marginally significant difference between close and medium word pairs, $^{+}$ indicates significant difference between medium and distant word pairs, and $^{\#}$ indicates marginally significant difference between medium and distant word pairs.

variance (R^2) computed via the *r.squaredGLMM* function from the MuMIn package in R (Barton, 2020)[9] was used to estimate the predictive power of the different models. In addition, to assess the variability in the obtained R^2 estimates, bootstrapped confidence intervals were obtained for each fixed-effect R^2 estimate by sampling with replacement across 1000 simulations using the *boot* function (Canty & Ripley, 2020) in R.

In our basic semantic modeling approach, we assumed that the strategy chosen by the speaker was to find the words closest to the average of the two target words. Although this is a fairly simple model of the speaker task, we were interested in understanding whether the models could potentially predict speaker responses via this simple search model. Specifically, to investigate the extent to which each of the semantic models predicted the explicit clues generated by the speaker, for each word pair, we first computed an "average" word vector by averaging the 300-dimensional vectors for the individual words derived from the associative and distributional models. Next, we rank-ordered words that were closest to this average vector by computing cosine similarities between the average vector and each word in the semantic space[10]. For the USF models, this space contained 4927 words, whereas for the SWOW and DSM-based models, this space contained 12,217 words. After rank ordering the closest words to the average vector, three measures were computed: *top clue score*, *average clue score*, and *rank correlation*. The *top clue score* estimated the accuracy of each model in predicting the most frequent clue produced by the speakers for a given word pair (see Table 2). The *average clue score* estimated the overall proportion of clues predicted by each of the models. Specifically, we computed the total number of unique clues (n) produced for each given word pair ($M = 18.27$, $SD = 6.89$) and then evaluated whether the topmost n predictions from each of the models corresponded to these clues (for a similar approach, see Thawani, Srivastava,

Table 2
Percentage occurrence of three most frequent first clues for word pairs in Connector

Type	Word Pair	First Most Frequent Clue 1 (%)	Second Most Frequent Clue 1 (%)	Third Most Frequent Clue 1 (%)
Close	happy-sad	emotion (79.55)	mood (9.09)	feeling (4.55)
	lion-tiger	cat (54.55)	animal (11.36)	feline (9.09)
	teeth-gums	mouth (52.27)	dentist (15.91)	floss (9.09)
	exam-algebra	math (70.97)	test (9.68)	school (6.45)
	collar-pants	clothes (70.97)	shirt, cloth, suit (6.45)	apparel, attire, belt (3.33)
	gold-silver	metal (45.45)	element (13.64)	bronze (6.82)
	aircraft-birds	fly (64.52)	flight (25.81)	sky (6.45)
	chair-table	furniture (36.36)	dine (18.18)	dinner (13.64)
	jump-leap	hop (34.09)	frog (15.91)	bounce (9.09)
	egg-kitchen	cook (48.39)	bake, breakfast (12.90)	fry (6.45)
	bounce-bat	baseball (45.16)	ball (32.26)	fly (6.45)
	candle-wick	flame (31.82)	burn (18.18)	light (15.91)
	old-new	age (31.11)	opposite, time (13.33)	generation (6.67)
	sit-stand	chair (27.27)	position (22.73)	action, desk, leg, etc. (4.55)
	adultery-perjury	crime (38.71)	sin (25.81)	court (6.45)
	apartment-homeless	house (38.71)	shelter (16.13)	live, home (9.68)
	bean-tea	coffee (32.26)	drink (19.35)	café, green (6.45)
	bubble-breeze	blow (22.58)	air (16.13)	float, wind (9.68)
	tree-oak	wood (11.36)	arbor, forest, leaf, plant (9.09)	acorn (6.82)
	dance-circle	spin (12.90)	celebration, moshpit, move, party (6.45)	art, circus, etc. (3.23)
Medium	calorie-famine	food (58.06)	hunger (16.13)	starve (9.68)
	sun-bowl	round (34.09)	breakfast, circle, sphere (9.09)	picnic (4.55)
	almond-lunch	food (45.16)	snack (22.58)	health (9.68)
	feather-heavy	weight (45.16)	opposite (12.90)	light (6.45)
	army-drum	march (29.55)	military (15.91)	band (11.36)
	elm-rock	nature (25.00)	forest, hard, landscape (6.82)	stone, tree, wood (4.55)
	quick-glow	firefly (25.00)	flash (15.91)	light (13.64)
	cigarette-onion	smell (32.26)	gross (9.68)	burn, odor, stink (6.45)
	hand-birth	baby (22.73)	body, midwife (11.36)	deliver (6.82)
	dream-bet	gamble 15.91)	goal (11.36)	lottery (9.09)
	holy-kind	jesus (13.64)	christian, priest, saint (9.09)	angel, religious, righteous (4.55)
	copy-graph	paper (19.35)	excel, homework (9.68)	computer, data, math, spreadsheet (6.45)
	actress-bad	movie (19.35)	boo (9.68)	critic, rotten tomatoes (6.45)

(*Continued*)

Table 2
(Continued)

Type	Word Pair	First Most Frequent Clue 1 (%)	Second Most Frequent Clue 1 (%)	Third Most Frequent Clue 1 (%)
	stop-school	bus (19.35)	crosswalk, graduate (9.68)	graduation (6.45)
	glass-cage	container, zoo (11.63)	aquarium, window (6.98)	barrier, trap (4.65)
	quiet-war	peace (16.13)	fight (9.68)	ceasefire, silent, stealth (6.45)
	abnormal-giggle	snort (12.90)	funny, laugh, weird (9.68)	comedy, hysterical (6.45)
	comedy-tourist	laugh (12.90)	entertain, show, travel (6.45)	activity, cruise, etc. (3.23)
	feet-chapel	jesus, kneel, wash, worship (6.82)	baptism, church, etc. (4.55)	aisle, cleanse, etc. (2.27)
	weird-trauma	shock (6.82)	abnormal, feeling, etc. (4.55)	accident, clown, etc. (2.27)
Distant	astronaut-near	space (48.39)	moon (9.68)	mars, distance (6.45)
	east-short	direction (31.82)	vector (6.82)	asian, small, etc. (4.55)
	economy-hundred	money (41.94)	dollar (9.68)	number, rich (6.45)
	crust-boot	mud (29.55)	dirt (11.36)	shoe, dirty (9.09)
	assist-definition	dictionary (38.71)	help (16.13)	explain (9.68)
	garage-bone	dog (22.73)	doghouse, storage (11.36)	car, skeleton, etc. (4.55)
	travel-ankle	walk (22.73)	basketball. feet, etc. (4.55)	achilles, boot, etc. (2.27)
	cooking-communicate	recipe (29.03)	kitchen (9.68)	bake, chef, show (6.45)
	brake-beginning	start (29.03)	slow, stop, time (6.45)	abrup, car, etc. (3.23)
	dusk-pendulum	time (25.81)	clock (19.35)	swing (16.13)
	cave-knight	dark, dragon (18.18)	medieval (15.91)	dungeon (9.09)
	olive-real	food (15.91)	oil (11.36)	martini (9.09)
	fight-corpse	war (22.58)	death (19.35)	battle (9.68)
	rude-regret	mean (13.64)	negative (11.36)	argument, fight (6.82)
	snake-ash	poison (13.64)	black, death, Pokemon (6.82)	dragon, hiss, etc. (4.55)
	stern-wind	sail (13.64)	boat, ship (11.36)	harsh (9.09)
	giant-subtle	opposite (11.36)	gentle (9.09)	conspicuous, hint, etc. (4.55)
	flat-alike	similar (16.13)	pancake, paper, same (6.45)	angle, bland, etc, (3.23)
	couch-void	lazy (16.13)	alone, empty, etc. (16.15)	comfort, pillow, etc. (3.23)
	dracula-toes	blood (16.13)	vampire (12.90)	count (9.68)

Table 3
Model prediction scores and rank correlations for speaker's first clue

Model	Top Clue Score (%) (*SE*)	Average Clue Score (%) (*SE*)	Rank Correlation (*SE*)
University of South Florida-strength (USF-S)	6.67 (3.25)	18.91 (1.19)	.27 (0.03)
USF- positive pointwise mutual information (PPMI)	6.67 (3.25)	21.22 (1.24)	.29 (0.03)
USF- random walk (RW)	11.67 (4.18)	21.86 (1.26)	.30 (0.03)
Small World of Words (SWOW)-S	15.00 (4.65)	15.88 (1.10)	.24 (0.03)
SWOW-PPMI	18.33 (5.04)	19.80 (1.20)	.27 (0.03)
SWOW-RW	21.67 (5.36)	20.62 (1.22)	.27 (0.03)
Global Vectors (GloVe)	8.33 (3.60)	14.14 (1.05)	.26 (0.03)
word2vec	3.33 (2.33)	10.68 (.93)	.20 (0.03)

& Singh, 2019). Therefore, this measure provided additional information about how well the models captured performance beyond the most frequent clue.[11] Finally, the *rank correlation* estimated the extent to which the *ranking* of clues for a given word pair based on frequency corresponded to the ranking predicted by the semantic models based on cosine similarity. For example, for the word pair *exam-algebra*, the ranking of clues based on speaker probabilities was *math* (0.71), *test* (0.10), *school* (0.07), *calculus* (0.03), *equation* (0.03), *knowledge* (0.03), and *study* (0.03). The ranking of these clues based on cosine similarity of the average vector of the word pair from the SWOW-RW model was *math* (0.89), *calculus* (0.87), *equation* (0.80), *test* (0.72), *school* (0.64), *study* (0.64), and *knowledge* (0.40). Therefore, Kendall's tau correlation of these ranks for *exam-algebra* was $r = .39$. Rank correlations were computed for each word pair within each semantic model and then averaged across participants, to assess whether the models captured the full pattern of responses produced by the speakers.

Table 3 displays the top clue score, average clue score, and rank correlation of each model in predicting the first clues generated by the speaker. As shown, although accuracy in the speaker task was overall low, there were significant differences across the models. Specifically, the SWOW-RW model predicted the greatest proportion of top clues, compared to word2vec ($p < .001$) and GloVe ($p = .006$), as well as USF-S ($p = .002$), USF-PPMI ($p = .002$), and USF-RW ($p = .040$). Differences between SWOW-RW, SWOW-PPMI, and SWOW-S were not significant (ps $> .05$). Finally, there were no significant differences across the USF models, but the USF-RW model outperformed the word2vec model ($p = .035$). Differences across word2vec and GloVe were not significant ($p = .146$). These patterns of top clue score were generally consistent with the average clue scores and rank correlations, although the USF-RW measure performed slightly better than the SWOW-RW model in predicting average clue scores and ranks of these clues, but these differences were again not significant (ps $> .05$). Overall, these findings suggest that the RW-based SWOW model (SWOW-RW) outperformed other models, including the DSMs.

To effectively understand the differences across the USF and SWOW-based norms, some additional analyses were conducted. Specifically, it is possible that the *size* of the normed databases (i.e., 4927 words in USF vs. 12,217 words in SWOW) or the presence of *secondary*

Table 4
Speaker model prediction scores for USF versus SWOW-based models

Dataset Size	Model	Top Clue Score (%; *SE*)	Average Clue Score (%; *SE*)	Rank Correlation (*SE*)
4923 words	USF-S	6.67 (3.25)	19.02 (1.19)	.28 (0.03)
	USF-PPMI	6.67 (3.25)	21.11 (1.24)	.29 (0.03)
	USF-RW	11.67 (4.18)	21.61 (1.25)	.30 (0.03)
	SWOW-R1-S	6.67 (3.25)	18.19 (1.17)	.26 (0.03)
	SWOW-R1-PPMI	13.33 (4.43)	21.98 (1.26)	.28 (0.03)
	SWOW-R1-RW	20.00 (5.21)	22.35 (1.27)	.29 (0.03)
	SWOW-S	11.67 (4.18)	21.14 (1.24)	.27 (0.03)
	SWOW-PPMI	13.33 (4.43)	25.67 (1.33)	.29 (0.03)
	SWOW-RW	18.33 (5.04)	24.75 (1.31)	.30 (0.03)
12,217 words	SWOW-R1-S	8.33 (3.60)	14.06 (1.05)	.24 (0.03)
	SWOW-R1-PPMI	18.33 (5.04)	18.81 (1.18)	.26 (0.03)
	SWOW-R1-RW	23.33 (5.51)	21.64 (1.25)	.28 (0.03)
	SWOW-S	15.00 (4.65)	15.88 (1.10)	.24 (0.03)
	SWOW-PPMI	18.33 (5.04)	19.80 (1.20)	.27 (0.03)
	SWOW-RW	21.67 (5.36)	20.62 (1.22)	.27 (0.03)

and tertiary responses in SWOW may be contributing to the higher prediction accuracy of SWOW-based models, compared to the USF-based models. To discriminate between these possibilities, we compared the USF dataset of free associations to the SWOW dataset of *only* primary associations (SWOW-R1) and the full database of primary, secondary, tertiary responses (SWOW), by restricting these analyses to only the 4923 words common to both norms. In this way, one could assess the predictive power of the different models on the *same* dataset. As shown in Table 4 (top half), although accuracy slightly decreased in the SWOW norms when the dataset was restricted, the SWOW database continued to outperform the USF database. Specifically, the SWOW-R1-RW model had higher accuracy than USF-S ($p =$ .003), USF-PPMI ($p = .003$), and USF-RW ($p = .065$). Interestingly, the models based on SWOW-R1 showed slightly higher accuracy than models based on the full SWOW dataset when the dataset size was smaller, although these differences were not significant (ps $> .05$).

Next, we exclusively compared SWOW-R1 and SWOW on the full database of 12,217 words to assess whether there was any additional contribution of secondary and tertiary responses above and beyond the primary associations contained within SWOW-R1. As shown in Table 6 (bottom half), these analyses revealed that the RW-based models generally outperformed other models (i.e., S and PPMI-based) in predicting the top clue, although there were no reliable differences between SWOW-R1-RW and SWOW-RW when the full database was considered ($p = .720$). Average clue scores and rank correlations followed similar patterns. Taken together, these analyses indicate that the SWOW norms were better able to capture the speaker behavior in this game, compared to the USF norms, even when we controlled for the difference in the dataset sizes across the two norms. This may reflect the recency of the SWOW norms, as well as the potential difference across task demands when asking participants to produce the first word or three words that come to mind. This issue is further

Table 5
Examples of modal clues and model predictions

Word Pair	Modal Clue	Model	Predicted Clue
gold-silver	metal	USF (S, PPMI, RW)	bronze, copper, shiny
		SWOW (S, PPMI, RW)	platinum, shiny, metallic
		GloVe	bronze
		word2vec	bronze
war-quiet	peace	USF (S, PPMI, RW)	tranquil, silent, peace
		SWOW (S, PPMI, RW)	silence, silent, silence
		GloVe	conflict
		word2vec	wary
exam-algebra	math	USF (S, PPMI, RW)	analysis, calculus, calculus
		SWOW (S, PPMI, RW)	mathematics, calculus, math
		GloVe	mathematics
		word2vec	theorem
glass-cage	container	USF (S, PPMI, RW)	shatter, crystal, mug
		SWOW (S, PPMI, RW)	shatter, captivity, trapped
		GloVe	steel
		word2vec	glaze

Table 6
Semantic relationships between first clues and word pairs

Clue Relationship	Clue Percentage (%)	Example 1(*Clue*: Word Pair)	Example 2 (*Clue*: Word Pair)
attributive	46.31	*fly*: birds-aircraft	*flame*: candle-wick
hierarchical	20.53	*animal*: lion-tiger	*furniture*: chair-table
coordinate	13.33	*shirt*: pants-collar	*military*: army-drum
locative	10.88	*court*: perjury-adultery	*ocean*: breeze-bubble
argument	4.62	*bake*: kitchen-egg	*suck*: toes-Dracula
temporal	2.40	*future*: dream-bet	*summer*: school-stop
idiosyncratic	1.91	*ale*: travel-ankle	*lesson*: holy-kind

discussed in the General Discussion. Furthermore, given that there were no significant differences across SWOW-R1 (consisting of only primary associations) and SWOW (consisting of primary, secondary, and tertiary responses), for all analyses that follow, to ensure comparability with previous work based on the SWOW norms (e.g., De Deyne et al., 2019), we only compare models based on the *full* dataset of primary, secondary, and tertiary SWOW responses consisting of 12,217 words (SWOW) and the full dataset of USF norms consisting of 4927 words (USF).

Table 5 displays examples of some word pairs and their modal clues alongside the predictions by the different models. As shown, although model prediction accuracy in this task was relatively low, it is important to emphasize that the models generated reasonable predictions in several cases (e.g., *bronze* for *gold-silver*, *conflict* for *war-quiet*, etc.), even though these predictions did not map onto the modal response very well. Future work will examine how human raters assess the quality of model-generated clues compared to human-generated clues, and whether the game can be successfully played with model-generated clues.

4.3. Predicting different types of semantic relationships

Given that we used a wide range of items in the study, the speaker task provided an opportunity to infer different types of semantic relationships between the words. For example, consider the word pair *lion-tiger*. While one speaker may infer a hierarchical relationship that both are *animals*, another speaker may instead identify another animal of the same category, such as a *leopard* or *jaguar*. In order to further understand the extent to which different types of semantic relationships were successfully inferred by each semantic model, we performed some additional analyses within the speaker task. For each first clue produced in the game, we classified the semantic relationship inferred by the speaker into the following categories: attributive, argument, hierarchical, coordinate, locative, temporal, and idiosyncratic based on Jouravlev and McRae's (2016) classification. Clues were classified as *attributive* if they described a property or feature of one or both of the word pairs, *argument* if they modified or performed an action on one or both of the word pairs, *hierarchical* if they grouped the word pairs into a superordinate category, *coordinate* if they belonged to the same category as one or both of the word pairs, *locative* if they highlighted a location-based aspect of one or both of the word pairs, *temporal* if they highlighted a temporal aspect of one or both of the word pairs, and *idiosyncratic* if the clues did not fall within any of the above categories. Table 6 displays some examples of word pairs and clues that were classified based on this criterion, along with the total percentage of these clues within the dataset.

As is evident, the majority of the clues produced by the speaker reflected *attributive*, *hierarchical*, *coordinate*, or *locative* relationships, and we, therefore, investigated whether there were differences in the predictive accuracy of associative versus distributional models across these semantic categories. For these analyses, we only examined the *average clue score* (that calculates the proportion of n unique clues for a given word pair correctly predicted by the models) because this measure resulted in greater accuracy overall across all models (as in Table 5). As shown in Fig. 2, there were substantial differences in the extent to which semantic models predicted clues corresponding to different semantic relationships.

First, all models were better at predicting *coordinate* clues, compared to other types of clues (ps $< .05$), which may indicate that both associative and distributional models are more likely to emphasize these types of relationships in their semantic representations. Second, qualitatively, the difference in predictive accuracy between associative models versus DSMs was greater for *coordinate* (0.48 vs. 0.35) and *hierarchical* clues (0.18 vs. 008), compared to *locative* (0.12 vs. 0.09) and *attributive* (0.19 to 0.13) clues. However, the Relationship x Model interaction was not reliable ($p = .513$), likely due to uneven distribution of items across the different semantic relationships. Qualitative analyses indicated that these differences were prominently driven by the "close" word pairs used in the game (e.g., *chair-table*, *happy-sad*, *gold-silver*, etc.). Speakers successfully inferred superordinate/hierarchical relationships between close word pairs (e.g., 80% of the speakers chose *emotion* as their first clue for

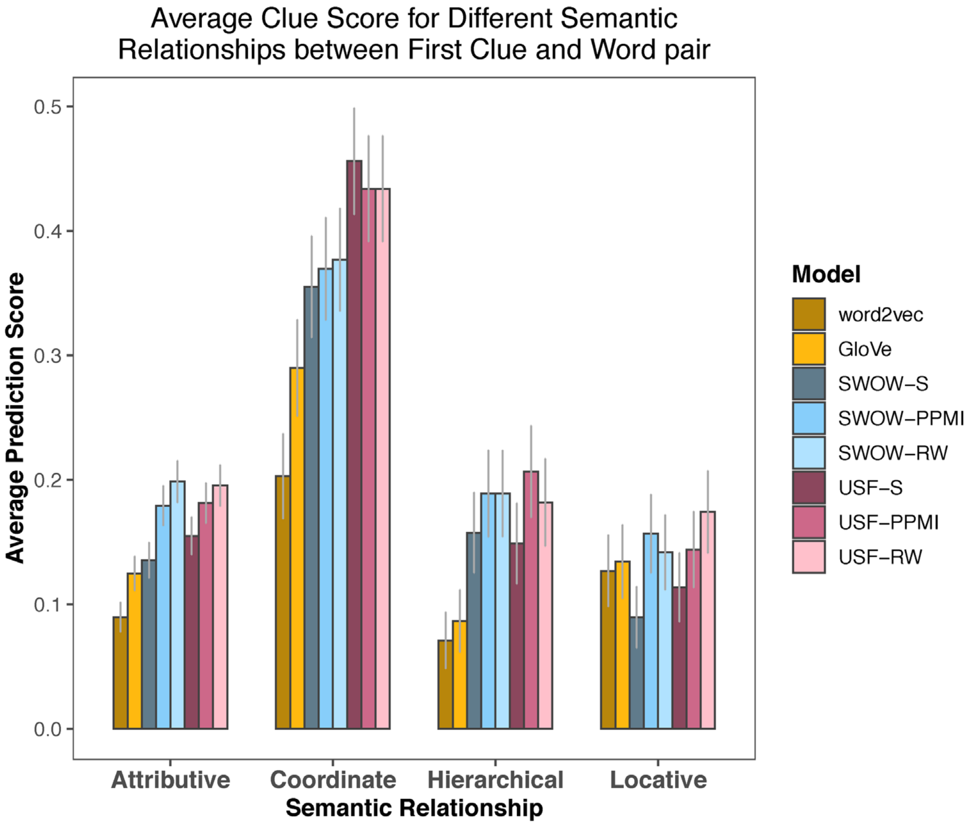


Fig 2. Average clue prediction score within different models as a function of different semantic relationships. Error bars indicate standard errors.

happy-sad) or chose exemplars/coordinate clues that were more related to both words (e.g., selecting *house* for *homeless-apartment*), which were reflected in the WAS created via USF and SWOW. However, the DSMs failed to infer such relationships and their clues leaned more toward words that typically fulfilled similar roles (i.e., coordinates) to only *one* of the words (e.g., *sorry* for *happy-sad*, and *bedroom* for *homeless-apartment*, etc.) but not *both* the words, which decreased their predictive accuracy across these semantic relationships. Although we acknowledge that these relation classifications are post hoc, these preliminary analyses suggest that the lack of hierarchical structure and different relationships within semantic representations derived from DSMs may be an important predictor of how well DSMs perform across semantic tasks.

4.4. *Model accuracy in predicting guesser responses*

To estimate the accuracy of each semantic model in correctly predicting the guesses of the guesser, we obtained the topmost two guesses predicted by each semantic model based on cosine similarity between the vector for each word on the board and the vector for the clue

Table 7
Guesser performance scores

Model	Average Guess Score (%; *SE*)
USF-S	47.24 (0.92)
USF-PPMI	49.87 (1.04)
USF-RW	48.93 (1.01)
SWOW-S	58.00 (0.95)
SWOW-PPMI	59.33 (0.87)
SWOW-RW	60.33 (0.88)
GloVe	39.69 (0.78)
word2vec	41.11 (0.73)

produced by the speaker. Predictions were scored as a 0 (for no matches with the guesser's responses), 1 (for a single word match), or 2 (for both words matching) for each semantic model. The maximum total score for a given model across all trials per participant dyad could therefore be 2 (guesses)*30 (word pairs) = 60. We calculated the proportion of correct predictions for each of the semantic models for a given participant dyad and then averaged these scores across all participants to yield an *average guess score* per model ranging from 0 to 1.

As shown in Table 7, prediction accuracy was generally higher in the guesser task, compared to the speaker task. This was expected, given that the task of the guesser was constrained by both the 20-word board and also the speaker's clue. Despite high accuracy overall, there were significant differences across the models. LME analyses revealed that the SWOW-RW and SWOW-PMI models outperformed all other models in predicting guesser responses (ps $< .05$). The difference between SWOW-RW and SWOW-PPMI was marginal ($p = .136$). Next, the USF-RW and USF-PPMI models performed similarly ($p = .164$) but outperformed the USF-S model, as well as word2vec and GloVe (ps $< .05$). Finally, word2vec outperformed GloVe in predicting guesser responses ($p = .034$).

4.5. *Effect of semantic similarity on guesser accuracy*

In addition to examining model predictions for explicit responses of the guesser, we also examined whether the *accuracy* of the guesser in the game itself (i.e., correctly guessing the word pair based on the clue in the first attempt) was predicted by the average semantic similarity between the first clue and the two words. For example, for the word pair *exam-algebra*, and the first clue *math*, we calculated the cosine similarity between *math* and *exam*, as well as between *math* and *algebra*, in each of the semantic models. These two estimates were then averaged to obtain the average similarity of the first clue (i.e., *math*) to the word pair (i.e., *exam-algebra*). This similarity was then used to predict the accuracy of the guesser in the first attempt. As shown in Fig. 3, the greater similarity between the first clue and the word pair predicted greater guesser accuracy, suggesting that clues that were closer in semantic space to the original word pairs produced more accurate responses from the guesser.

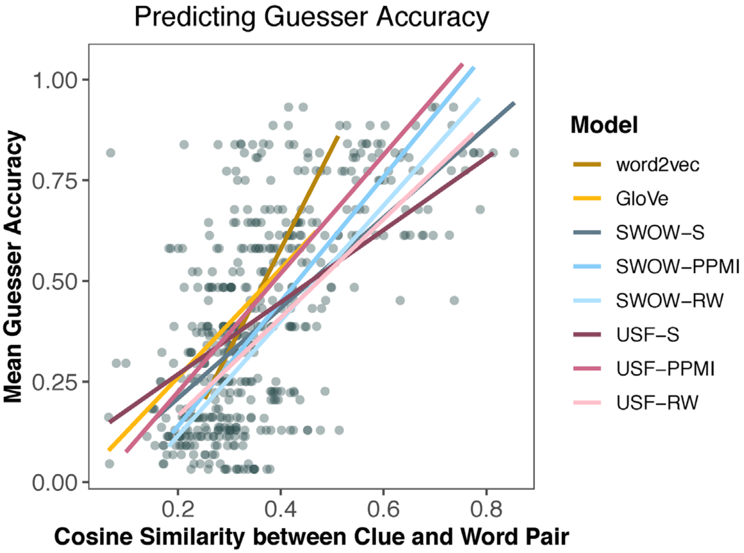


Fig 3. Mean accuracy in the guesser task as a function of average cosine similarity between the first clue and the word pair.

Table 8

Explained variance in models predicting guesser accuracy

Model	***R***^{2} **(Fixed [CI]/Total)**
English Lexicon Project (ELP) variables	1.86 [0.22, 2.76]/30.95
USF-S	4.32 [1.30, 6.46]/29.43
USF-PPMI	5.96 [2.57, 8.77]/28.30
USF-RW	7.62 [3.76, 10.89]/28.85
SWOW-S	6.78 [3.48, 9.82]/28.53
SWOW-PPMI	7.40 [3.90, 10.59]/28.16
SWOW-RW	**8.35 [4.51, 11.83]/28.22**
GloVe	5.38 [2.49, 7.54]/32.78
word2vec	5.63 [2.50, 7.97]/31.98

CI indicates bootstrapped 95% confidence interval for empirically obtained fixed R^2.

LME analyses revealed that similarities from all models significantly predicted guesser accuracy (*p*s $<$ .001), after controlling for item-level lexical variables, that is, word length, concreteness, and frequency for the first clue as well as the word pair, which were obtained from the English Lexicon Project (Balota et al., 2007). Table 8 displays the total explained variance in predicting guesser accuracy across the different models. As shown, the associative models again explained significantly more variance than the DSMs, as indicated by the bootstrapped confidence intervals around the R^2 estimates, and the SWOW-RW model explained the most variance in this task.

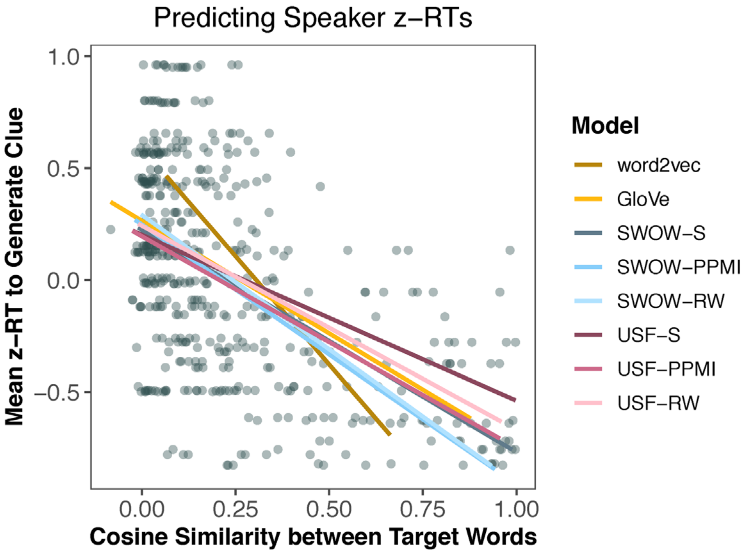


Fig 4. Mean standardized response time (z-RT) to generate first clue as a function of cosine similarity between the target words within different associative and distributional models.

4.6. Effect of semantic similarity on speaker's z-RTs

We also examined the extent to which different models accounted for response latencies in the Connector game. To control for individual differences in RTs (see Faust, Balota, Spieler, & Ferraro, 1999), RTs to generate the clue in the first attempt were standardized in the following manner. RTs below 250 ms and over 120,000 ms (i.e., 2 min) were first excluded, followed by the exclusion of all RTs that fell above or below 3 standard deviations of each participant's mean RT. This process excluded 2.93% of the total trials. Next, RTs were standardized within participants, to produce standardized RTs (z-RTs), which were used for all analyses. We evaluated the extent to which cosine similarity between the words predicted z-RTs to generate the clue in the first attempt. As shown in Fig. 4, the semantic similarity was negatively correlated with z-RTs, that is, the farther the two words were apart in the semantic space (i.e., lower cosine similarity), the longer it took participants to generate the first clue.

LME analyses revealed that semantic similarity estimates from all models significantly predicted z-RTs to generate the first clue. As shown in Table 9, the SWOW-RW model was again the best model in predicting z-RTs and consistently outperformed the DSMs as well as the USF model, based on bootstrapped confidence interval estimates.

4.6.1. Effect of semantic similarity on guesser's z-RTs

We also examined whether z-RTs to generate the first guesses by the guesser were predicted by the average distance between the first clue and the word pairs.

As shown in Fig. 5, cosine similarity between the first clue and word pair was negatively

Table 9
Explained variance in models predicting speaker standardized response times (z-RTs) to generate clue

Model	R^2 **(Fixed [CI]/Total) (%)**
ELP	2.36 [1.07, 3.01]/22.31
USF-S	10.09 [7.09, 12.82]/21.54
USF-PPMI	9.58 [7.44, 11.45]/21.72
USF-RW	8.98/ [6.93, 10.76]/21.56
SWOW-S	11.16 [9.00,13.05]/22.70
SWOW-PPMI	10.53 [8.46, 12.34]/22.76
SWOW-RW	**11.69 [9.51, 13.59]/22.68**
GloVe	6.76 [4.85, 8.39]/22.64
word2vec	8.68 [6.49, 10.55]/22.24

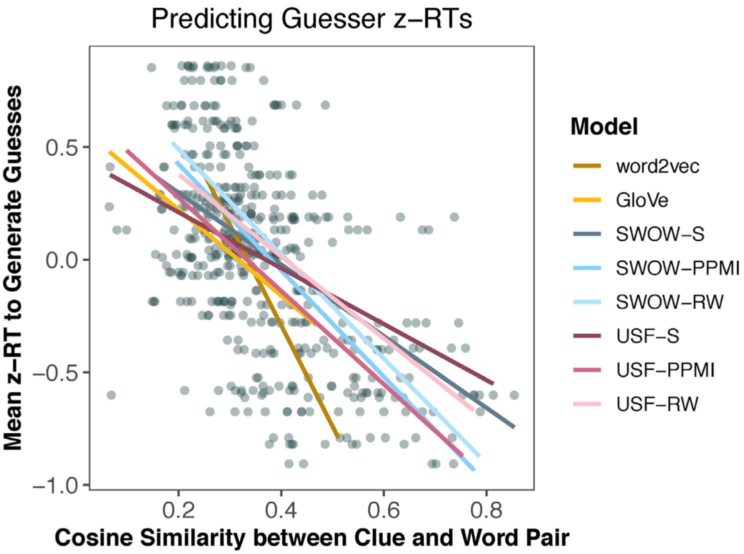


Fig 5. Mean z-RT to produce first guess as a function of cosine similarity between the first clue and word pair across different models.

correlated with guesser z-RTs, indicating that when the first clue was farther from the word pair (i.e., lower cosine similarity), guessers took longer to make their guesses. This is also consistent with the overall positive correlation of speaker and guesser latencies ($r = .21$, $p <$ .001), suggesting that the time course of identifying the word pairs for the guesser was related to the clue generation process for the speaker. LME analyses predicting z-RTs to guess the word pairs in the first attempt revealed that the average similarity between the first clue and the word pairs significantly predicted z-RTs for all models (ps $<$ .05), after controlling for item-level variables (see Fig. 5. As shown in Table 10, the SWOW-RW model again explained the most variance in z-RTs, although variance explained in guesser z-RTs was low overall.

Table 10
Explained variance in models predicting guesser z-RTs to guess word pairs

Model	R^2 **(Fixed [CI]/Total)**
ELP variables	2.21 [0.69, 3.17]/20.09
USF-S	4.48 [1.88, 6.40]/19.71
USF-PPMI	4.05 [1.51, 5.97]]/18.50
USF-RW	4.98 [2.12, 7.28]/18.57
SWOW-S	5.69 [3.02, 7.97]/18.54
SWOW-PPMI	6.14 [3.40, 8.67]/17.71
SWOW-RW	**7.48 [4.51, 10.50]/17.41**
GloVe	3.27 [1.25, 4.58]/19.50
word2vec	4.48 [2.30, 6.17]/18.80

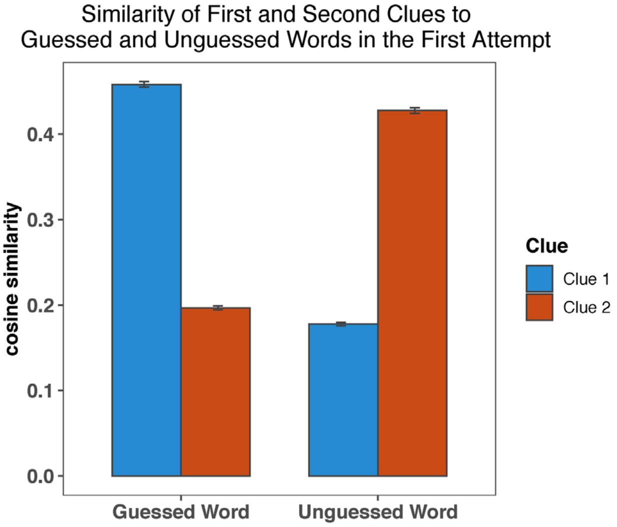


Fig 6. Similarity of the first and second clues to the individual words before and after the guesser's first attempt. Error bars indicate standard errors.

4.6.2. Influence of interactions on second attempt performance

To examine whether patterns of performance differed across the attempts, we evaluated whether the guesser's likelihood of succeeding in the second attempt was dependent on the nature of clues provided by the speaker and whether the speaker changed their strategy based on the guesser's first attempt. It is important to note here the goal of this section was to explicitly identify critical *interactive* patterns in the game. Therefore, although model-based results have been reported, in these analyses we emphasize the cooperative influences captured within Connector.

First, we examined whether the guesser's first response changed the extent to which speakers focused on the individual words within the word pairs, for trials on which the Guesser correctly identified one of the words. As shown in Fig. 6, in the first attempt, when Speakers had no information from the Guesser, first clues (Clue 1) were closer to whichever word was ultimately identified by the Guesser ("Guessed Word" in Fig.6), based on cosine similarity between the clue and individual words (collapsed across all semantic models). However, after the first guess, there was a significant shift in the produced clues. If the guesser correctly identified only one of the words in the first guess and therefore incorrectly guessed the other word ("Unguessed Word" in Fig. 6), second clues (Clue 2) were now closer to the unguessed word compared to the guessed word ($p < .001$). For example, for the word pair *exam-algebra*, if the first clue was *math*, and the guesser's response was *algebra-pen*, then the second clue provided by the speaker (e.g., *testing*) was closer to the unguessed word (e.g., *exam*), compared to the guessed word (e.g., *algebra*). Therefore, speakers changed their search strategy after the first attempt and chose clues that steered guessers toward the unguessed word. This shift in clue similarity toward unguessed words was predicted by cosine similarities between the clue and words in all the semantic models (*p*s < .05).

We also examined whether this shift in speaker strategy influenced the accuracy of the guesser for trials on which the guesser correctly identified one of the words. Specifically, LME analyses predicted guesser's accuracy for the second guess attempt for a particular word pair (e.g., *exam-algebra*, referred to as Word1 and Word2), with the following predictors: (a) the similarity of the second clue provided by the speaker (e.g., *testing*, referred to as Clue2) to Word1 (e.g., *testing-exam*, referred to as Clue2-Word1), and the (b) distance of Clue2 to Word2 (e.g., *testing-algebra*, referred to as Clue2-Word2). We also included the average distance of the first clue to the two words (e.g., the average distance between *math-exam* and *math-algebra*) as a covariate, to control for the effects of Clue1 on the guesser's accuracy in the second attempt. Further, given that Clue2 may also depend on the initial guesses themselves, we separated the LME analyses by whether the guesser correctly guessed Word1, correctly guessed Word2, or guessed neither Word1 nor Word2.

As shown in Fig. 7, when one word was guessed correctly but the other word was not guessed in the first attempt, the guesser had higher accuracy in the second attempt if the speaker provided second clues that were closer to the unguessed word (i.e., cosine similarity between the second clue and unguessed word was high). These effects were predicted by all the semantic models (*p*s < .05), although the SWOW-based PPMI and RW models generally outperformed the DSMs and USF-based models in predicting guesser accuracy in the second attempt. Finally, the third attempt followed similar patterns, although these analyses were underpowered due to fewer data points (over 88% of the word pairs were guessed by the second attempt by over half of the participants) and are therefore not reported. Overall, these effects indicate that the speaker was adjusting the type of clues they provided based on the guesser's initial responses, and the degree of this adjustment influenced subsequent guesser performance, suggesting a cooperative influence between speaker and guesser in the game. This is particularly interesting since the guesser did not know on incorrect trials, which, if either, of the two words, were guessed correctly. These results suggest an element of social interaction within the Connector game, and also provide evidence that two-player word games may be particularly suited to examine contextual influences on semantic search processes due to their relatively unconstrained nature as well as the potential to study the interaction of distinct semantic memory systems.

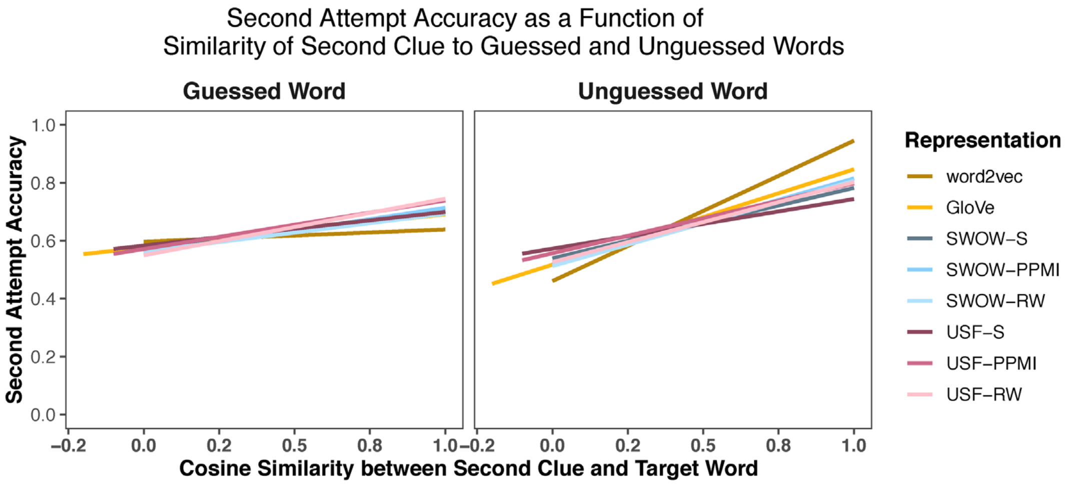


Fig 7. Guesser's second attempt accuracy predicted by cosine similarity between second clue and guessed and unguessed target words.

5. General discussion

The present study investigated semantic search and retrieval processes in a novel cooperative word game, Connector, and also evaluated the extent to which estimates of semantic similarity derived from associative and distributional models explained performance in the game. Although prior work has examined search processes within word games, past research has only studied constrained memory search processes. For example, in the MindPaths game (Marrs et al., 2017), players found paths from one word to another through forced choice at each step, which may have restricted the actual search process. The Connector game introduced in the current study overcomes these limitations and taps into more unconstrained search processes operating on semantic memory representations, by having the speaker freely select any clues that are related to word pairs on the board and having the word guesser guess the word pairs based on their own interpretation of the clues. We now discuss some novel contributions and future directions for this work.

5.1. Comparing SWOW and USF norms

The present study compared the predictive power of two different free association datasets (USF and SWOW) in predicting player performance in Connector. Overall, the SWOW-based associative models outperformed the USF-based models, even after controlling for differences in dataset size. Importantly, the *primary* response dataset (SWOW-R1) also outperformed the USF-based models, suggesting that the advantage of SWOW did not solely arise from capturing secondary and tertiary associations. These model differences may reflect the recency of the SWOW norms (which were collected from 2008 to 2021), compared to the USF norms (published in 2004, collected much earlier), as well as differential task demands/instructions.

First, regarding the recency differences, it is possible that the pattern of associations in SWOW is simply more reflective of present-day associations among the general population. Second, although one may expect the processes for generating *primary* associations to be similar in USF versus SWOW, there were indeed differences in task instructions between the two studies. Specifically, participants in the USF study were asked to write down the first word that came to mind that was "meaningfully related or strongly associated to the presented cue word." In contrast, the SWOW study asked participants to respond with the "first three words that came to mind." Given that the USF study urged participants to produce a single meaningful response, and the SWOW study asked participants to produce multiple responses, it is possible that these instructions biased responses (De Deyne et al., 2019), which in turn manifested in the model predictions examined in the current study. Finally, SWOW is based on *fluent* English speakers, covering a more diverse population than USF (which is restricted to *native* English speakers) which may have also influenced some of the differences across the models derived from the two databases. Future work should further examine the extent to which different instructions and the time period of data collection bias free association responses within different norms. Overall, however, the present findings show how the SWOW-based models generally outperform the USF-based models in predicting game performance.

Another important finding from the present work was that the RW-based SWOW model (SWOW-RW) provided a better account of player behavior, compared to the PPMI-based model (SWOW-PPMI) and associative strength-based model (SWOW-S). This suggests that individuals not only use direct associative strength but also use indirect pathways to come up with a response, similar to the metaphor of using spreading activation-based processes (Collins & Loftus, 1975) to explore the semantic space. It is likely in the Connector game that the explicit search across multiple words capitalizes on instances where indirect paths can arrive at an appropriate clue or guess. This finding is consistent with prior work in the literature (De Deyne et al., 2019; Fathan et al., 2018) where RW models have been shown to successfully capture game-based navigation of semantic space. Of course, given that the present work only examined a RW-based process model, future work should investigate whether alternative search mechanisms such as local-global search and optimal foraging (Hills, Jones, & Todd, 2012) may also be at play within unconstrained semantic tasks such as in Connector.

5.2. Comparing associative and distributional models

An important motivation for the present work was to compare the extent to which semantic representations derived from associative and distributional models successfully accounted for performance in the Connector game. Previous studies of game-based paradigms have examined the performance of players against only association-based network models (e.g., Beckage et al., 2012; Fathan et al., 2018) or only DSMs (e.g., Shen et al., 2018). In the present study, we compared several associative models based on USF and SWOW databases as well as two widely used DSMs in the extent to which they account for performance in the Connector game. The SWOW-based associative models performed relatively better than distributional models in this game task. Of course, one might already expect associative models to outperform DSMs since these associative models were constructed from free association norms, although it is important to note that the differences in predictive power between USF norms and DSMs were minimal. Importantly, one may be concerned about shared method variance between the free association task and the task of finding clues or guessing word pairs in the game. This may have contributed to the higher predictive power of the associative models, compared to the DSMs, which were instead trained on large language corpora. However, it is also possible that free association represents unique *conceptual* information that is not contained within *linguistic* corpora-based DSMs, and tasks that tap into such conceptual processing (such as the speaker and guesser tasks in Connector) may benefit from this representational overlap. Therefore, although comparing associative models to DSMs may be problematic (for detailed arguments, see Jones et al., 2015), it is important to understand the nature of the information contained within these representations, after controlling for differences in the representational format itself (see Kumar, Steyvers, Balota, 2021 for a discussion). In the present work, we ensured that associative models and DSMs were compared in the fairest way possible by constructing WAS and ensuring all words were represented within a high-dimensional space across the two classes of models.

The present results highlight how associative models do indeed emphasize semantic relationships not well-represented within the DSMs and indicate that reliance on pure linguistic corpora within the DSMs may not be sufficient to capture the variety of responses produced by participants in the Connector game. Indeed, in addition to the linguistic content of free associations, associative responses also tend to reflect experiences that evoke mental imagery and emotional responses (De Deyne et al., 2021). It is possible that similar representations are activated when speakers and guessers are searching through semantic space within the Connector game, which the associative models tend to capture. DSMs have been criticized for relying solely on linguistic corpora and therefore their inability to capture non-linguistic features of meaning (Barsalou, 2016; De Deyne et al., 2016). Our results also shed light on some additional aspects of meaning (e.g., hierarchical relationships) that may be readily apparent to humans (and are therefore well-represented in the associative models) but are missing from the DSMs. Within this context, associative models may provide an important behavioral baseline or benchmark for comparisons across DSMs and may therefore be useful in assessing the psychological plausibility of different DSMs (for a detailed discussion, see Kumar, 2021). Indeed, the present work highlights systematic differences across two popular DSMs (GloVe and word2vec) in accounting for performance in the game, with GloVe outperforming word2vec in the speaker task. Although error-driven distributional models such as word2vec have been shown to outperform error-free learning models in other psycholinguistic tasks (e.g., Mandera et al., 2017), it is possible that GloVe may be more sensitive to different *types* of semantic relationships due to capturing more global patterns of co-occurrence, compared to word2vec, which is trained to predict words within a local context window. This may have contributed to the better performance of GloVe in the speaker task within the game. However, word2vec and GloVe appeared to perform at similar levels in the guesser task and in accounting for the response latency patterns, suggesting that these models may also share similar mechanisms to some degree (see Levy & Goldberg, 2014). Given that pretrained models were used in this study, future work should look into how different hyperparameters influence the generated semantic representations to better understand the relative performance of different distributional models, as well as explore more advanced language models.

5.3. Temporal signatures of semantic search

Another important contribution of the present work is the potential to examine temporal signatures of search within semantic memory via response latencies. Indeed, the present analyses showed how semantic similarity between words significantly predicted the time taken by participants to generate clues and guesses in Connector. Of course, the speaker task may involve complex search processes that may reflect local and global search biases and even show patterns of optimal foraging (Hills et al., 2012). Indeed, the present work only examined one fairly simple search algorithm for the speaker based on vector "averaging" which may be more effective for semantically closer words than distant concepts (see Kumar, Garg, & Hawkins, 2021), and it is important to investigate alternative search algorithms that may improve model performance in the speaker and guesser tasks. Future work should focus on further understanding these search processes used by the speaker to navigate the semantic space. Response-time data provide important constraints for model discrimination, and a process-driven model of semantic search should be able to accommodate the patterns observed in the current task.

Although the associative and distributional models nicely predicted the speaker z-RTs, the variance explained was relatively low for guesser z-RTs. This may be partly due to obtaining a single estimate of response latency for the guesser responses in the current study, given that there may be differences in how quickly an individual may select the first and second word in the task. For example, it is possible that some clues lead guessers to quickly identify one of the words but think more carefully about the second word. Although the present work cannot speak to these differential search processes (due to paradigm constraints), future work will focus on fully mapping out the search process of the speaker and guesser in the Connector game.

5.4. Cooperative interactions in the Connector

Another important finding in this work was that the speaker and guesser collaborated to successfully arrive at the correct answers in the game. Specifically, we found that the speaker systematically selected clues in the second round that were biased toward the word that the guesser had failed to guess in the first round. This finding is important in two ways. First, it suggests that individuals constrained their search process based on the task context, and chose unbalanced clues (i.e., not equally related to both words) as the task goal changed during the course of the game. Second, the degree of adjustment in the form of the second clue influenced the accuracy of the guesser in the second round, indicating that both players successfully engaged in a form of referential communication during the game. While most studies of search processes operating on semantic representations focus on individuals, language has an inherent social function (Pinker, 2003, p. 27). The present study sheds light on how individuals modify their search processes based on semantically relevant interactions with other individuals (also see Xu & Kemp, 2010). The experimental design of the Connector affords opportunities to examine dyadic interactions via several recent models of pragmatic reasoning and theory-of-mind (e.g., Frank & Goodman, 2012). Indeed, our other work investigates how individuals incorporate the other player's perspective within the communicative context of the game using pragmatic inference (Rational Speech Act models; Goodman & Frank, 2016) models (see Kumar, Garg, & Hawkins, 2021). Additionally, exploring other forms of *social* interactions that may influence search processes in complex semantic tasks such as didactic dialogue, teaching, second language learning, translation, and so forth, is also an important avenue for future research, and word games may be particularly suited to shed light on these issues. For example, the Connector game would appear to be ideally suited to explore the nature of shared semantic representations across friends, spouses, and family members. In this way, the task can be used to further extend the burgeoning field of shared distributed/cognitive representations (e.g., see Abel, Umanath, Wertsch, & Roediger, 2018; Hollan, Hutchins, & Kirsh, 2000). Indeed, the boardgame Codenames has recently been adapted to examine pragmatic language use (Shen et al., 2018), develop an educational game to teach physics to students in diverse school settings (Souza et al., 2018) and teach English vocabulary to students in Indonesia (Octaviana et al., 2019), suggesting that there are several research opportunities within the domain of cooperative word games.

In conclusion, the present study introduced a novel word game, Connector, to study relatively unconstrained search and retrieval processes operating over semantic representations. The results indicated that associative models based on random walk processes can better capture these search processes and interactions, compared to modern distributional models, as reflected in responses and response latencies in this game task. In addition, player performance in this cooperative word game is sensitive to indirect interactions and players alter their search processes based on task demands and goals.

Acknowledgments

We sincerely thank Simon De Deyne, Theo Rhodes, and David Noelle for their extremely insightful feedback and helpful comments on earlier versions of this manuscript. All data, materials, and analysis scripts can be accessed at https://osf.io/haes8/.

Notes

1. In the original Codenames game, multiple players typically play in teams, and the board has words assigned to different teams via colors. Therefore, the speaker (called the spymaster in Codenames) must try to give clues for words that are relevant to their own team while avoiding words that are relevant to the other team. Additionally, there is an "assassin" word that the speakers and guessers must avoid. We simplify this game to a two-person setting and eliminate teams/colors/assasins to ask targeted research questions.
2. Shen et al. did not use the full board but instead asked participants to choose two words from a set of three words and also did not model real-time interactions between players.

Password (Xu & Kemp, 2010) is not based on Codenames and involves speakers providing one-word clues related to a word to guessers who make one-word guesses.

3. The only differences across both experiments were that we used different items and counterbalanced the direction of presentation for word pairs in the second experiment. However, we did not find any influence of word order in any of the analyses. Further, only two analyses (clue prediction and guesser prediction) yielded significant interactions between model estimates and experiment, and these were largely driven by item-level differences (further discussed in the General Discussion). Therefore, in order to provide the most generalizable estimates of model performance across items, we combine results across the two experiments.
4. We thank Simon De Deyne for sharing the code for computing these similarity measures.
5. Initial analyses also compared "network"-based models based on free association norms to DSMs (as in Kumar et al., 2020), but these analyses yielded similar overall patterns. Therefore, the present work focuses on the USF/SWOW-based models based on S, PPMI, and RW measures.
6. We thank Simon De Deyne for sharing the code for computing these association spaces.
7. Although some boards had more powerful distractors than others, we piloted these boards to ensure that there were no obvious competitors to the target word pairs within the boards. Future work will examine the influence of board-level differences on behavioral performance.
8. Although the program only scored an exact string match as the correct response, we adjusted for spelling errors manually after the experiment and scored those responses as correct in all analyses.
9. The r.squaredGLMM function gives marginal and conditional estimates of R^2 for mixed-effects models and a single estimate for linear regression models. Both estimates are reported wherever mixed-effects models are used.
10. We explicitly focus on search processes relevant to the target words in the current paper, although in other work, we investigate the effect of the words on the boards (see Kumar et al. 2021).
11. Partial string matches or clues found within multi-word model predictions were scored as a correct prediction (e.g., *old age* vs. *age*, *mathematics* vs. *math*, etc.). Additionally, plurals or variants of the target words (e.g., *exams*, *drums*, etc.) were removed from model predictions.

References

Abel, M., Umanath, S., Wertsch, J. W., & Roediger, H. L. (2018). Collective memory: How groups remember their past. In M. L. Meade, C. B. Harris, P. Van Bergen, J. Sutton & A. J. Barnier (Eds.), *Collaborative remembering: Theories, research, and applications* (pp. 280–296). Oxford, England: Oxford University Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, *1*, 238–247.

Bartoń, K. (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142.

Beckage, N., Steyvers, M. & Butts, C. (2012). Route choice in individuals—semantic network navigation. Proceedings of the Annual Meeting of the Cognitive Science Society, 34, 108-113.

Campbell, M., Hoane, Jr, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial Intelligence*, *134*(1-2), 57–83.

Canty, A., Ripley, B. (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In A. K. McCallum & S. Roweis (Eds.), *Proceedings of the 25th international conference on machine learning* (pp. 160–167). New York: ACM.

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., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922.

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.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. *Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical papers*, Osaka, Japan (pp. 1861–1870).

Fathan, M. I., Renfro, E. K., Austerweil, J. L., & Beckage, N. M. (2018). Do humans navigate via random walks? Modeling navigation in a semantic word game. *Proceedings of the Annual Meeting of the Cognitive Science Society*, Madison, WI.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*(6), 777–799.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, *31*(3), 59–79.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In Philological Society (Great Britain) (Ed.), *Studies in linguistic analysis*. Oxford, England: Blackwell.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

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

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

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.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174–196.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570–574. https://doi.org/10.1037/a0039248

Jouravlev, O., & McRae, K. (2016). Thematic relatedness production norms for 100 object concepts. *Behavior Research Methods*, *48*(4), 1349–1357.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: Evidence from the Hebrew mental lexicon. *PloS ONE*, *6*(8), e23912.

Kim, A., Ruzmaykin, M., Truong, A., & Summerville, A. (2019). Cooperation and Codenames: Understanding natural language processing via Codenames. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *15*(1), 160–166).

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*, 40–80.

Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2261–2276.

Kumar, A. A., Garg, K., & Hawkins, R. (2021). Contextual flexibility guides communication in a cooperative language game. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 2457–2463.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*.

Kutuzov, A., Fares, M., Oepen, S., & Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling*, Reykjavik, Iceland, (pp. 271–276).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

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.

Marrs, F., Straka, M. J., & Beckage, N. M. (2017). *Human path finding in a semantic word game*. Retrieved from https://wiki.santafe.edu/images/e/e2/Semantic_Networks.pdf

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. Retrieved from https://arxiv.org/abs/1301.3781

Moskvichev, A., & Steyvers, M. (2019). Word Games as milestones for NLP research. *Proceedings of GAMNLP-19*, San Luis Obispo, California. Retrieved from https://webfiles.uci.edu/msteyver/publications/GAMNLP-2019_paper_9.pdf

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Octaviana, I. T., Rahmah, R. E., & Puspitasari, D. (2019). The use of Codenames game to help students in learning vocabulary. *Vision: Journal for Language and Foreign Language Learning*, *8*(2), 101–116.

Patel, A., Sands, A., Callison-Burch, C. & Apidianaki, M. (2018). Magnitude: A fast, efficient universal vector embedding utility package. arXiv preprint arXiv:1810.11190. Retrieved from https://arxiv.org/abs/1810.11190

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1532–1543). Red Hook, NY: Curran.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. *Studies in the Evolution of Language*, *3*, 16–37.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Shen, J. H., Hofer, M., Felbo, B., & Levy, R. (2018). Comparing models of associative meaning: An empirical investigation of reference in simple word games. In *Proceedings of CoNLL 2018*. Retrieved from https://arxiv.org/pdf/1810.03717.pdf

Siew, C. S., Wulff, D. U., Beckage, N. M., & Kenett, Y. N. (2018). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019, 2108423.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*(1), 64–75.

Souza, P. V. S., Morais, L. P., & Girardi, D. (2018). Spies: An educational game. *Physics Education*, *53*(4), 045012.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Decade of behavior. Experimental cognitive psychology and its applications* (pp. 237–249). Washington, DC: American Psychological Association. https://doi.org/10.1037/10895-018

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Thawani, A., Srivastava, B., & Singh, A. (2019). June). SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In A. Rogers, A. Drozd, A. Rumshisky & Y. Goldberg (Eds.), *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP* (pp. 43–51). Minneapolis: Association for Computational Linguistics.

Veyra, J. D. (2016). *Codenames board game gets local edition*. Retrieved from https://news.abs-cbn.com/life/12/14/16/codenames-board-game-gets-local-edition

Wittgenstein, L. (1953). *Philosophical investigations*. London: Macmillan.

Xu, Y., & Kemp, C. (2010). Inference and communication in the game of Password. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (pp. 2514–2522). Red Hook, NY: Curran Associates, Inc.

Yee, E., Lahiri, A., & Kotzor, S. (2017). Fluid semantics: Semantic knowledge is experience-based and dynamic. In A. Lahiri & S. Kotzor (Eds.), *The speech processing lexicon: Neurocognitive and behavioural approaches* (vol. *22*, pp. 236–255). Berlin: Boston De Gruyter Mouton.

Zunjani, F. H., & Olteteanu, A. M. (2019). Towards reframing Codenames for computational modelling and creativity support using associative creativity principles. In *Proceedings of the 2019 on Creativity and Cognition*, San Diego, CA (pp. 407–413).