Contextual flexibility guides communication in a cooperative language game

Abhilasha A. Kumar
Washington University in St. Louis
abhilasha.kumar@wustl.edu

Ketika Garg
University of California, Merced
kgarg@ucmerced.edu

Robert D. Hawkins
Princeton University
rdhawkins@princeton.edu

Abstract

Context-sensitive communication not only requires speakers to choose relevant utterances from alternatives, but also to retrieve and evaluate the relevant utterances from memory in the first place. In this work, we compared different proposals about how underlying semantic representations work together with higher-level selection processes to enable individuals to flexibly utilize context to guide their language use. We examined speaker and guesser performance in a two-player iterative language game based on Codenames, which asks speakers to choose a single 'clue' word that allows their partner to select a pair of target words from a context of distractors. The descriptive analyses indicated that speakers were sensitive to the shared semantic neighborhood of the target word pair and were able to use guesser feedback to shift their clues closer to the unguessed word. We also formulated a series of computational models combining different semantic representations with different selection processes. Model comparisons suggested that a model which integrated contextualized lexical representations based on association networks with a contextualized model of pragmatic reasoning was better able to predict behavior in the game compared to models that lacked context at either the representational or process level. Our findings suggest that flexibility in communication is driven by context-sensitivity at the level of *both* representations and processes.

Keywords: semantic retrieval; memory search; pragmatic inference; contextualized word representations

Introduction

Accumulating evidence suggests that efficient communication requires flexibility across contexts (Sperber & Wilson, 1986; Clark, 1996; Goodman & Frank, 2016). But *where* should contextual flexibility enter into models of communication? One possibility is that flexibility is supported at the *representational* level. For example, individuals may utilize distributional statistics from natural language to distill context-relevant information directly into the structure of their high-dimensional semantic representations (as suggested by recent models of semantic memory; Kumar, 2021) or lexical association networks (as suggested by recent network-based models; De Deyne et al., 2021). Alternatively, context-sensitivity may arise at the *process* level. For example, speakers may use context to prioritize different retrieval cues, or re-weight different utterances post-retrieval based on pragmatic inferences about the communicative goal at hand.

Context-sensitivity likely reflects contributions at both the representational and process levels. Therefore, disentangling these contributions requires an appropriately rich experimental paradigm exposing the richness of our semantic representations. Reference games — where speakers must produce a referring expression that distinguishes a target object from a context of distractors — have been widely used to operationalize context-sensitivity in communication (e.g. Olson, 1970; Dale & Reiter, 1995). These games typically present a visual context, such as an array of images, to evaluate accounts of *grounded* semantic representations and pragmatic reasoning (e.g. Degen et al., 2020). Yet it has been challenging to evaluate theories of subtler associative and distributional relationships *among* different words using these visual contexts. Recently, Kumar, Steyvers, & Balota (2021) introduced a paradigm called Connector, a simplified variant of the board game *Codenames* (Chvátil, 2016), which used large referential contexts of other words rather than images. Rather than referring to a single target, participants must refer to *sets* of targets (see Xu & Kemp, 2010 which examined a game called *Password* cuing a single target word rather than a set). This task therefore require participants to use richer semantic and conceptual relationships between words to guide their language use.

For example, on the trial depicted in Fig.1, the Speaker is presented with the target pair *tiger-lion* and asked to generate a one-word clue that would allow the guesser to select that pair. Their clue "cat" is transmitted to the Guesser, who is then asked to select exactly two words from the 20-item board (e.g., *clever-lion*). If their first attempt is unsuccessful, Speakers have two more attempts to provide two additional clues to the Guesser. Importantly, unlike the variant of Codenames recently explored by Shen et al. (2018), Connector does not place any hard constraints on word choice, allowing us to track natural search and retrieval processes across the entire lexicon. As such, Connector represents an ideal paradigm to elicit rich, context-dependent communication. In this paper, we used these data to evaluate different proposals for how semantic representations and selection processes work together to enable flexible language use.

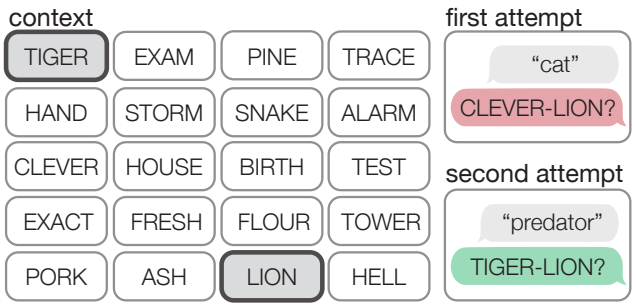

Figure 1: Example trial in Connector. The target pair (*tiger-lion*) is highlighted for the Speaker. Their first clue, "cat," led to an incorrect response. This feedback was used to adjust the clue to "predator," allowing the Guesser to correctly identify the pair.

Table 1: Examples of clues provided by the Speaker

Word pair	Top 3 clues (frequency)
lion-tiger	cat (24), animal (5), feline (4)
exam-algebra	math (22), test (3), school (2)
war-quiet	peace (5), fight (3), ceasefire (2)

Methods

Behavioral Data

We evaluated our models on a dataset of Connector games played by 75 dyads (150 participants). Each game consisted of 30 trials, with a different pair of target words on each trial. These word-pairs were presented in a sequence of 10 blocks, with exactly 3 trials per block. The presentation of the word-pair cue to the speaker was counterbalanced (e.g. between *lion-tiger* and *tiger-lion*) to control for possible salience effects. Each block used a distinct board with different sets of words. The overall sequence of trials was fixed across all pairs, and the set of target word-pairs were chosen to reflect varying levels of difficulty, computed via averaging similarity estimates across different semantic models (see Kumar et al., 2021, for details). We aggregated data across two different experiments, for a total of 60 word pair items and an average of 18 unique clues generated per word pair (SD = 6.89). Participants achieved an overall success rate of 85% across the three attempts, reflecting relatively high accuracy overall. Table 1 provides some examples of clues generated by the Speaker for different target pairs[1].

Candidate Models of Semantic Representation

Communication depends on the underlying semantic representations of words, and different proposals of representational models exist in the literature. We considered 3 different representational proposals in our analyses: two large distributional models, *GloVe* and *BERT*, as well as an associative network model based on the Small World of Words (*SWOW*) dataset. These models include representations for a large vocabulary of 12,218 words[2]. In this section, we introduce each of these models in detail.

GloVe Distributional semantic models (DSMs) assume that individuals extract statistical regularities from natural language to construct semantic representations, which can be inferred from large text corpora. We utilized one such DSM, GloVe (Pennington, Socher, & Manning, 2014)[3]. We obtained 300-dimensional GloVe embeddings from a pre-trained model, trained on a 3 billion-word Wikipedia corpus available from Kutuzov, Fares, Oepen, & Velldal (2017).

BERT Although semantic representations are often assumed to be "non-contextual" (as in GloVe), there are now several modern language models that learn *contextualized* semantic representations. In these models, vector representations for words are learned by attending to not simply word co-occurrence patterns, but also predicting upcoming words within sentential contexts by using positional and syntactic information. Therefore, we also evaluated whether a state-of-the-art contextual word embedding model, BERT (Devlin, Chang, & Lee, 2019) can account for Speaker and Guesser utterances in Connector. To obtain BERT embeddings, we used the BERTModel provided by HuggingFace (Wolf et al., 2019), trained on a $\approx$ 3.8 billion corpus, to obtain 768-dimensional embeddings. Embeddings were obtained by providing each word in the search space to the BERT model via the prompt, "[CLS] word [SEP]", and summing the vectors from the last four hidden layers for each token, as is typically recommended (McCormick & Ryan, 2019). Note that even though these BERT embeddings are not contextualized with respect to the Connector game, the learning mechanisms behind BERT and GloVe considerably differ. Therefore, the present analyses evaluated how these de-contextualized BERT embeddings compare to GloVe, and an associative network model.

Small World of Words (SWOW) It is possible that distributional information from text corpora is insufficient to account for flexible language use. Indeed, significant recent work has shown that associative models, typically based on free association norms, often outperform DSMs in semantic tasks (De Deyne et al., 2019). Therefore, we also evaluated whether an associative representation model, based on the SWOW dataset can better account for performance in this task[4]. SWOW embeddings were obtained by converting the raw associative frequencies for the different cues in the SWOW dataset into a 300-dimensional random walk-based word association space (Kumar, Steyvers, & Balota, 2021).

Candidate Models of Speaker Selection Strategies

Representations and selection strategies are inextricably tied to each other: any communicative action is a combination of a specific selection strategy operating over underlying semantic representations. Therefore, in the current paper, we considered all combinations of the representational models described above with different models of selection to evaluate whether individuals consider only the retrieval cues (i.e., target word pair), or take into account possible distractors on the board.

[1] All data and analysis scripts have been made available at https://github.com/hawkrobe/connector_cogsci

[2] To equate these models, we restricted GloVe and BERT to the 12,216 unique cues in SWOW database, supplemented with each of the words on the board and all valid clues (excluding multi-word responses, $< 1\%$ of total trials) produced by speakers in our dataset.

[3] Initial analyses also included another distributional model, word2vec, but we focus on GloVe for simplicity, as it performed better overall. See Kumar et al., under review for additional comparisons.

[4] The SWOW dataset is based on a continued free association task, where participants are given a cue and produce the first 3 words that come to mind, see https://smallworldofwords.org/

All selection models are based on a 12218-word lexicon where the similarity between any two words can be computed as the cosine distance between their vector embeddings in the given representational space:

$$s(w_i, w_j) = \frac{\vec{w}_i \cdot \vec{w}_j}{\|\vec{w}_i\| \|\vec{w}_j\|}. \quad (1)$$

Target-only Speaker Our simplest speaker model prefers clues c that maximize similarity to the two words in the target pair $\{w_1, w_2\}$ while minimizing similarity to all other words $\overline{B} = B - \{w_1, w_2\}$ that are not in the target pair:

$$U_B(c; \{w_1, w_2\}) = \alpha \cdot s(c, w_1) \cdot s(c, w_2) - (1 - \alpha) \cdot \sum_{b \in \overline{B}} s(c, b) \quad (2)$$

where $\alpha \in [0, 1]$ is a parameter controlling the influence of the distractors. For comparison to more sophisticated probabilistic models, we score clues according to a softmax over this utility, where β is a temperature parameter which was fine-tuned to fit the data for each choice of semantic representation. We call the special case where $\alpha = 1$ the *target-only* model. Under this setting, the agent completely ignores the distractors and only concentrates on maximizing similarity to the target pair.

Target+Board Speaker When $\alpha < 1$, there is a non-zero influence of the rest of the board on the speaker's choice. For example, if the speaker is given the target word pair *lion-tiger*, they may initially want to use the clue *dangerous*, since it is semantically related to both target words. However, if they are paying attention to the set of distractors, they may realize that it is also semantically related to the distractor *snake*, and therefore prefer a clue like *cat* that better minimizes the second term of Eq. 2. We parametrically varied α to find the level of distractor influence that best fits the data.

Pragmatic Speaker Finally, it is possible that context-sensitivity emerges through explicit *pragmatic reasoning* about alternatives, which accounts for the Guesser's decision processes. Specifically, we formulated a model of our task within the Rational Speech Act (RSA; Goodman & Frank, 2016) framework, according to which a pragmatic Guesser G_1 recursively reasons about a pragmatic Speaker S_1, who in turn recursively reasons about a literal Guesser G_0. For the literal guesser G_0, we computed the likelihood of selecting any pair of words $\{w_1, w_2\}$ on the board B, given a clue c, using a product semantics:

$$G_{literal}(\{w_1, w_2\}|c, B) \propto \exp\{s(c, w_1) \cdot s(c, w_2)\} \quad (3)$$

Intuitively, a pair $\{w_1, w_2\}$ is only a good candidate if *both* w_1 and w_2 are semantically related to the clue. For the pragmatic Speaker S_1, we computed the probability of selecting every possible word as a tradeoff between its informativity to the literal guesser and its retrieval cost:

$$S_1(c|\{w_1, w_2\}, B) \propto e^{\beta \ln G_{literal}(\{w_1, w_2\}|c, B) - w_c \cdot cost(c)} \quad (4)$$

where *cost(c)* captures biases towards selecting more accessible words from the search space (operationalized via word frequency) and β again captured the temperature of the softmax distribution. We fine-tuned cost weight $w_c \in [0, \infty]$ and $\beta \in [0,]$ parameters separately for the different representational models to maximize the likelihood of the data.

Candidate Process Models of Guesser Flexibility

For the Guesser task of selecting a word-pair from the board given a particular clue, we evaluate whether the representational similarity of different words on the board to the given clue influences Guesser choices. In addition, we also obtain pragmatic predictions via RSA models to evaluate whether Guessers incorporate the Speakers' perspectives into their selection process.

Baseline Guesser For the Guesser task, the baseline model predictions corresponded to G_0 in Eq. 3, where we maximized the product of similarities of the given clue to the different pairs of words on the board.

Pragmatic Guesser The context-sensitive Guesser G_1 considers not only the absolute relatedness of the clue to the words, but also the possible alternative clues that the Speaker could have provided. For example, a pragmatic guesser may reason that if the Speaker had intended to identify the target pair *snake-tiger* they would be more likely to say something like *striped* rather than the given clue; because they didn't say *striped*, they must *not* have intended that pair. To formalize this reasoning, we computed the probability of selecting $\{w_1, w_2\}$ given a clue c and board B as follows:

$$G_1(\{w_1, w_2\}|c, B) \propto S_1(c|\{w_1, w_2\}, B) P(\{w_1, w_2\}) \quad (5)$$

where $P(\{w_1, w_2\})$ is the prior probability of selecting any two given words on the board in the absence of any clue. We assumed a uniform prior over all words on the board.

Behavioral Results

Before fitting these models to our data, we first characterize two basic qualitative patterns of context sensitivity in speakers' choices. First, to what extent were speakers able to find clues that accounted for *both* words in the target pair? Second, when the first clue did not lead to a successful response, were speakers able to adjust their second clue to be sensitive to their partner's errors?

How do speakers integrate words in target pair? Our task requires speakers to choose a single clue from their vocabulary that allows their partner to select a *pair* of targets

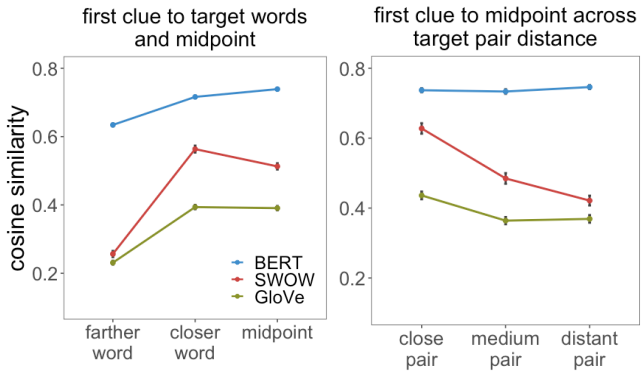

Figure 2: The first clue tended to be closer to one of the target words in all semantic representations. Within the SWOW and GloVe models, the clue was closest to the midpoint of the target pair when the words within the pair were close themselves. Error bars represent 95% confidence intervals.

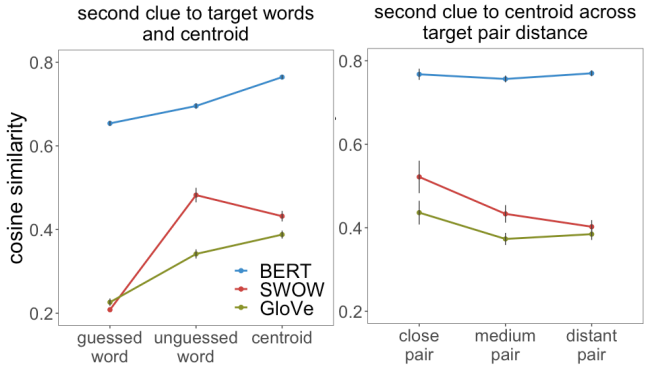

Figure 3: Across all semantic representation models, the second clue was closer to the unguessed word, compared to the guessed word, and was close to the centroid of the target pair when the words were close themselves, within the SWOW and GloVe models. Error bars represent 95% confidence intervals.

on the board. How do speakers simultaneously integrate constraints from both words in the target pair? We hypothesized that speakers aim to select clues are as semantically similar as possible to each component target word without sacrificing similarity to the other; in other words, clues ought to be closer to the semantic "midpoint" of the target pair than to either of the targets considered separately. We tested this hypothesis by computing cosine similarities between the vector embeddings of the clue (e.g. $c_1 =$ *cat*) and the two component words in the target pair (e.g. $w_1 =$ *tiger* and $w_2 =$ *lion*, respectively), as well as to the midpoint of the pair, $(w_1 + w_2)/2$.

Because the order of the two words in the pair was not meaningful, we assigned w_1 to be the further of the two words from the clue and w_2 to be the closer. Next, we fit a linear mixed-effects model with random effects at the participant and word-pair level to evaluate the null hypothesis that the clue was equidistant from the two words (i.e. that $s(c,w_1) = s(c,w_2)$). For all three choices of representational semantics, we found a significant violation of symmetry, suggesting that the first clue was typically closer to one word than the other (p's $< .001$, see Fig. 2). Furthermore, as shown in Fig. 2, we also found that the first clue was closer to the midpoint only in the BERT model ($p<.001$), and in fact showed the opposite effect in the SWOW model ($p < .001$), and did not show a significant effect in the GloVe model ($p = .381$).

To further investigate this effect at the item level, we examined whether the similarity between the words within the word-pair influenced the extent to which first clues were closer to the midpoint. We categorized the word-pairs into three levels: "close" (e.g., *lion-tiger*), "medium"(e.g., *giggle-abnormal*), and "distant"(e.g., *void-couch*), based on averaged cosine similarities between the words across a range of semantic models (see Kumar et al. for details). Next, we tested whether the first clue was significantly closer to the midpoint as a function of semantic distance via mixed effect models. Indeed, as shown in Fig. 2 (right panel), clues were closest to the midpoint when words themselves were "close" in semantic space compared to "medium", although only SWOW ($p<.001$)and GloVe ($p = .013$) models showed this effect. On the other hand, the BERT model was not sensitive to these item-level differences, likely due to overall high cosine similarities which may be indicative of ceiling effects. Overall, however, these analyses suggest that the proximity of the target words within semantic space directed Speaker choices, i.e., when words were semantically distant, it was harder to select clues that were "equidistant" from both words, compared to when words were semantically close and the selection pool itself was optimally constrained for the Speaker to choose clues that were closer to *both* words.

Do Speakers adjust their clues based on Guesser feedback? Next, to understand how guesser feedback affected the speaker's choices, we considered the trials where guessers successfully identified only one of the words. Consistent with Kumar et al.'s findings, when players correctly guessed one of the words, the second clue given by the Speaker was more similar to the unguessed word than to the guessed word ($p < .001$) in all the models. In addition, the second clue was closer to the "centroid", $(w_1 + w_2 + c_1)/3$, compared to either of the target words in the BERT and GloVe models, but showed the opposite effect in the SWOW model (p's $< .001$). When examining this effect at the item level, when the words within the target pair were themselves close, the second clue remained close to the centroid (see Fig. 3) in the SWOW ($p = .008$) and GloVe ($p = .04$) models. Therefore, Speakers attempted to produce clues that were most similar to the global intersection of the two words and the first clue when the words were semantically close, but also switched their proximity to either word within that intersection, based on the Guesser's responses (see Fig. 4).

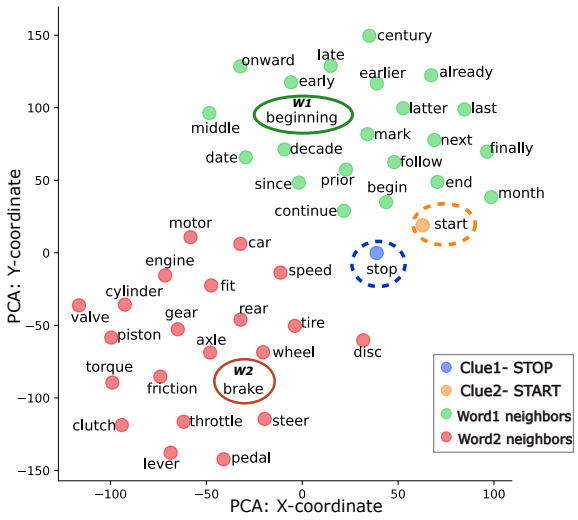

Figure 4: An example illustrating the effect of Guesser responses on Speaker clues using GloVe embeddings. The Speaker selects *stop* as Clue1 from the global intersection of *beginning* and *brake*. When Guesser successfully identifies *brake*, the Speaker selects Clue2 (*start*) that is closer to both the unguessed word (*beginning*), and farther from the guessed word (*brake*).

Model Comparison

Having established these basic patterns of targeted search within semantic space, we next turn to the problem of predicting the *first* clue speakers choose to send, and the *first* pair that guessers select in response. In this section, we conduct a quantitative model comparison evaluating the extent to which different combinations of representational models and process models successfully account for Speaker and Guesser behavior.

Speaker Predictions For the Speaker task, three measures were computed. First, as an overall measure of fit, we computed the *log likelihood* of the data under each model. Second, as a more interpretable measure of absolute performance we computed the *top-5 accuracy*, measuring the proportion of clues in our data that fell within the top 5 predictions produced the model[5]. Table 2 shows some examples of clues correctly predicted by one representation/process model combination but not another.

Finally, as a more interpretable measure of whether each model captured the full distribution, we calculated the *mean rank* of each clue produced by speakers. In other words, each model produced a full ranking over all 12218 words in the vocabulary, and we examined where the clues that were actually produced fell in that distribution. Lower ranks indicate better performance across the entire distribution of Speaker responses.

[5]We used the top-5 criteria because the first few predictions by the models were often the target words (e.g., *jump* or *leap*) or variations of the target words(e.g., *jumping* or *leaping*)

Table 2: Examples of clue predictions

Word-Pair/Modal-Clue	Representation/Process	Prediction
feet-chapel /**kneel**	**SWOW/Target-only**	kneel
	BERT/Target-only	pilgrimage
exam-algebra / **math**	SWOW/Pragmatic	math
	BERT/Target-only	calculus

We found three key patterns in these analyses (see Table 3). First, the associative SWOW-based representations strictly outperformed other representational models in predicting Speaker utterances (p's $< .05$). Second, Speakers mainly prioritized similarity to words within the target pair when retrieving clues, rather than prioritizing distance from distractors, as indicated by higher values of α providing a better overall fit; see Fig. 5. Finally, the pragmatics-driven RSA model, combined with the SWOW representation model, provided the best fit to the data overall. Further, the pragmatic model was also the best-performing model for the BERT model based on log-likelihoods, but this pattern did not hold for the GloVe model.

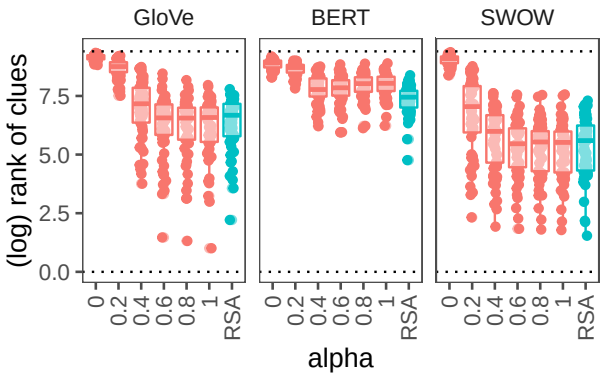

Figure 5: Average (log) rank of empirical clues in full utterance distribution produced by each model. Boxplot represents distribution over 60 wordpair items. Dotted lines represent upper and lower bounds. RSA model shown in blue.

Guesser Predictions For the Guesser task, we similarly obtained the top-5 accuracy, mean ranks, and log likelihood scores for each representational and process-level model. Table 4 displays the predictions scores and log-likelihoods for the literal and pragmatic Guesser models for each of the representational models. As shown, the SWOW-based associative model again performed better than the other representational models in predicting Guesser responses. Log-likelihood scores also indicated that the pragmatics-driven model provided a better fit to the overall data across all models, although the accuracy and rank measures did not appear to show a benefit of incorporating pragmatics-driven information.

Table 3: Model prediction scores for the Speaker's first clues

Representation	Process Model	Optimal Parameters	Top-5 Accuracy (95% CI)	Mean Rank (95% CI)	Log-Likelihood
GloVe	Target-only	$\beta = 21$, $\alpha = 1$	.09 (.05-.13)	798.64 (694-933)	-15773.48
	Target+Board	$\beta = 21$, $\alpha = .9$	.10 (.06-.14)	792.03 (686-917)	-15846.85
	Pragmatic	$\beta = 22$, cost = 0.04	.06 (.03-.09)	876.67 (769-987)	-15990.86
BERT	Target-only	$\beta = 20$, $\alpha = 1$	.03 (.02-.04)	3182.88 (2862-3485)	-18636.95
	Target+Board	$\beta = 20$, $\alpha = .8$	.03 (.02-.05)	2598.26 (2319-2878)	-18752.26
	Pragmatic	$\beta = 30$, cost = 0.03	.02 (.01-.03)	1784.70 (1590-1999)	-17533.20
SWOW	Target-only	$\beta = 23$, $\alpha = 1$	.16 (.11-.22)	336.43 (263-421)	-13204.00
	Target+Board	$\beta = 23$, $\alpha = .9$	.16 (.11-.21)	336.21 (264-414)	-13287.74
	Pragmatic	$\beta = 25$, cost = 0.04	.21 (.15-.26)	361.17 (299-425)	**-12895.74**

Table 4: Model prediction scores for Guesser's first responses

Representation	Process Model	Top-5 Accuracy (95% CI)	Mean Rank (95% CI)	Log-Likelihood
GloVe	Baseline	.17 (.15-.19)	25.29 (23.71-25.32)	-8140.96
	Pragmatic	.13(.11-.13)	26.72 (24.93-28.52)	-10343.96
BERT	Baseline	.09 (.08-.10)	58.60 (56.38-61.05)	-9385.38
	Pragmatic	.09 (.07-.10)	43.98 (41.74-46.19)	-10468.63
SWOW	Baseline	.43 (.41-.45)	9.23 (8.33-10.17)	-10144.16
	Pragmatic	.31 (.29-.33)	20.11 (18.25-21.94)	**-6665.27**

Discussion

Communication is a complex behavior that requires attending to environmental cues as well as initiating search and retrieval processes that operate on underlying knowledge representations to ultimately achieve a specific goal. Contextual flexibility is a key property of efficient communication, that enables speakers and listeners to efficiently convey meaningful information to each other within a shared context. This paper evaluated different representational and process-level models of contextual flexibility, to assess the contribution of retrieval context, representation, and pragmatic information in explaining communicative behavior in a cooperative language game, Connector.

We first descriptively examined the extent to which different representational models can capture Speaker and Guesser behavior in the game. We found that in the face of multiple retrieval cues (i.e., the word pair), Speakers limited their search space to the common neighbors of the two cues. The similarity between the cues also affected the search space, in that greater similarity between the individual words significantly restricted the retrieval context. We also found that the global context defined by the cues changed relative to the clues previously retrieved. Additionally, on trials where Guessers correctly identified one of the words, Speakers produced clues that would guide the Guesser towards the unguessed word. Taken together, these results suggest that Speakers were sensitive to the retrieval cues and produced clues that optimized communication. Furthermore, we found the representational model BERT was least sensitive to these descriptive patterns, likely due to ceiling effects and the lack of finetuning.

Our model comparisons indicated that an associative model (SWOW) combined with a pragmatic search and retrieval model (RSA) best accounted for Speaker and Guesser performance in the game. With respect to representation-level flexibility, it is important to mention here that the associative SWOW model is based on behavioral free association data, and therefore captures conceptual representations that may be activated in an associative task. As such, the Speaker and Guesser tasks are also associative in nature. Therefore, it is possible that the SWOW model provides the best account of the data partly due to shared method variance, in addition to capturing non-linguistic, hierarchical information that is difficult to extract via pure text-based distributional models (see Kumar et al., under review for detailed arguments). In this light, associative models such as the SWOW model may be viewed as an empirical *ceiling* for model comparisons, and one can then evaluate how well models *not* based on behavioral norms compare to this baseline. Indeed, we find that the GloVe model performs significantly better than the BERT model in the Speaker and Guesser tasks. However, it is important to highlight here that the BERT model used in the present work represents an entirely *non-contextual* model, i.e., although BERT is trained to attend to contextual information in text, we did not provide any task-specific context to BERT, but instead used "context-free" BERT embeddings in this work. It is possible that BERT would be able to generate more reasonable predictions when embedded within task-relevant linguistic contexts, and exploring contextualized BERT embeddings within communicative contexts is an avenue for future work.

With respect to process-level contextual flexibility, our analyses indicated that Speakers prioritized the retrieval context of the word-pairs significantly more than the surrounding context of distractors, when generating clues. Furthermore, the pragmatic model-based analyses indicated that both the Speaker and the Guesser benefited from pragmatic information about the communicative context. It is important to mention here that the pragmatics-driven model inherently accounted for the board, and in fact generated predictions that were quite similar to context-sensitive Speaker model that prioritized the word pairs but also incorporate the board to some extent (e.g., $\alpha > 0.8$). In addition, error analyses indicated that when words were relatively dissimilar or difficult (e.g., *communicate-cooking*), Speakers chose clues that were more related to one word than the other in such cases (e.g., *food*) – i.e., Speakers were no longer picking "rational" clues but instead choosing clues purely based on associative information from one of the words. Finally, with respect to the Guesser, although the accuracy and rank metrics did not show a benefit of the pragmatic model, the log-likelihood scores showed that the pragmatic model provided a better fit overall. This may reflect the relatively lower variance in responses produced by the Guesser, as well as the lack of separate fine-tuning parameters for the Guesser, as we prioritized fine-tuning the Speaker parameters which were then directly fed into the pragmatic Guesser model. Exploring independent optimality parameters for the Guesser as well as identifying specific contexts in which players prioritize suboptimal responses and differentially weight the perspective of the other player are avenues for future research in this domain.

Overall, the present findings suggest that players are sensitive to semantic neighborhoods as well as the perspective of the other player in communicative contexts. Therefore, flexibility in communication is driven by sensitivity at multiple levels, i.e., at the representational level in the form of associative information, and at the process level in the form of retrieval context and pragmatic information.

References

Chvátil, V. (2016). Codenames.

Clark, H. H. (1996). *Using language*. New York, NY: Cambridge University Press.

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

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2019). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1).

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

Devlin, J., Chang, M. W., & Lee, K., K.and Toutanova. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*.

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

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin and Review*, *28*, 40–80. doi: https://doi.org/10.3758/s13423-020-01792-x

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). Semantic memory search and retrieval in a novel cooperative word game: A comparison of associative and distributional semantic models. *Cognitive Science*. doi: https://doi.org/10.1111/cogs.13053

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

McCormick, C., & Ryan, N. (2019). Bert word embeddings tutorial. Retrieved from `http://www.mccormickml.com`

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

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

Shen, J. H., Hofer, M., Felbo, B., & Levy, R. (2018). Comparing models of associative meaning: An empirical investigation of reference in simple language games. *Conference on Natural Language Learning*.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press: Cambridge, MA.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, arXiv-1910.*.

Xu, Y., & Kemp, C. (2010). Inference and communication in the game of password. *Advances in Neural Information Processing Systems*, 2514–2522.