SHAPING MEMORIES: THE ROLE OF LANGUAGE AND LINGUISTIC FEATURES OF SPONTANEOUS SPEECH

By

Evgeniia Diachek

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

May 12, 2023

Nashville, Tennessee

Approved:

Sarah Brown-Schmidt, Ph.D.

Ashleigh M. Maxcey, Ph.D.

Sean M. Polyn, Ph.D.

Stephen M. Wilson, Ph.D.

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my mentor Dr. Sarah Brown-Schmidt for the unwavering support and guidance, and a profound belief in my abilities. Thank you for filling these five years with laughter and curiosity. Thank you for teaching me how to be a good researcher, for being proud of me on the good days, and for showing kindness and compassion on the hard days. Thank you, Sarah, for setting a beautiful example of a girl boss. There is no one else I would rather have gone through this journey with.

I must also thank my committee members Dr. Asheigh M. Maxcey, Dr. Sean M. Polyn, and Dr. Stephen M. Wilson for their generous service, unparalleled expertise, and invaluable feedback. Thank you for mentoring me through the years, for putting me in charge of exciting and meaningful projects, for challenging me, and through all of it making me feel like a colleague. Finally, thank you all for a stellar example of mentorship, leadership, and brilliance.

I am also grateful to all the past and current members of the Grad Lab for creating a safe learning environment. To Dr. Duane Watson, Dr. Lisa Fazio, Dr. Meg Saylor as well as Dr. Nick Tippenhauer, soon-to-be Drs. Bethany Gardner, McKay Wright, Raunak Pillai, and many other graduate students for sitting through numberless presentations, for helping me to develop new ideas, for excellent feedback, and interesting questions. Thank you to all Conversation Lab research assistants and lab coordinators for their hard work and tremendous effort. This work would not have been possible without your help.

Thank you to Dr. Ev Fedorenko and Dr. Ted Gibson -- my first research mentors and the people I today consider my American parents for inviting me into their labs and into their home. Thank you for taking a chance on me, for introducing me to the people who have become my

friends and my community. Most importantly, thank you for making me feel like I belong both in science and in the United States.

Special thanks to Dr. David Cole for his service on my qualification exam committee, for being the dream course instructor to TA for, for fascinating statistics conversation, and for bringing meaningful change to our department. I would also like to acknowledge the National Science Foundation, Peabody College, Vanderbilt Graduate School, and Lisa M. Quesenberry Foundation for funding my graduate research.

Спасибо моим родителям, маме и папе, за то, что вы ставите мои мечты и счастье выше своих собственных[1]. Thank you to my big sister, *Ане,* for her unconditional love and for always believing in me, even on the days when I don't believe in myself. To *Сём*, for being the perfect shade of grey. To my best friend Allegra Anderson for being by my side since the very first day of graduate school, for weekly wine nights, and for making Nashville feel like home. Thank you to my partner and the love of my life, Lénie Torregrossa for choosing me every day and for bringing more love than I had ever imagined was humanly possible.

Thank you all for your kindness, love, and support through it all -- I feel so lucky to have crossed paths with you.

[1] To my parents, mom and dad, for putting my dreams and happiness above their own.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER I

Introduction

The relationship between language and thought has been one of the most intriguing questions in the fields of philosophy, neuroscience, and cognitive science and has long been of particular interest in exploring the nature and organizational principles of human cognition. Indeed, on a daily basis humans use language to label entities in the world and to communicate their thoughts and feelings to each other. Yet, the precise impact of linguistic reference on shaping mental representations of concepts, objects, and events remains underspecified. When referring to an object as an "enormous balloon", what kinds of mental representations are being accessed? How does referring vs. not referring to an entity in the world impact the representation of that item or that class of item? And how do linguistic choices such as "enormous balloon" as opposed to "orange balloon" shape the ways in which individuals represent a particular object?

This dissertation aims to further characterize the relationship between language and memory in the human mind. **Using computational, experimental, and observational methods, I examine how language shapes mental representations of concepts, objects, and real-world experiences, as well as subsequent memories for them in three complementary lines of research.** Study 1 examines the role of linguistic labels in accessing and evaluating semantic properties of lexically invoked concepts. Study 2 examines the effect of linguistic reference on shaping mental representations of objects and subsequent memories for them as well as the underlying cognitive mechanisms. Finally, Study 3 utilizes an ecologically valid approach to

examine the linguistic features of spontaneous speech that shape representations and later, memories of real-world experiences.

The Role of Linguistic Labels in Accessing and Evaluating Semantic Properties of Lexically Invoked Concepts

The nature of conceptual representations and their relationship with the corresponding linguistic labels has been a central question in the field of cognitive science for decades (Osgood, 1952; Collins & Quillian, 1969; Tulving, 1972; Collins & Loftus, 1975; Saffran & Schwartz, 1994; McRae et al., 1997; Fodor, 1998; Laurence & Margolis, 1999; Jackendoff, 2002; Cree & McRae, 2003; Barsalou, 2008; Lambon-Ralph & Patterson, 2007; Mahon & Caramazza, 2008; Binder et al., 2009). Technological advances in the late 1990s offered a new powerful computational method of quantifying the meaning of words through ample text corpus data, creating a new class of distributional semantic models. Distributional semantic models, or DSMs, such as latent semantic analysis (LSA; Landauer & Dumais, 1997), word2vec (Mikolov et al., 2013), and Global Vectors (GloVe; Pennington et al., 2014) conceptualize semantic representations as vectors residing in a high-dimensional vector space. These models are based in part on the assumption that the meaning of a word is reflected in the pattern of its usage, namely, that words with similar or related meanings tend to occur in similar contexts (Harris, 1954; Firth, 1957). According to this idea, words like *virus*, *mask*, and *vaccine* tend to occur in proximity to each other (e.g., in the same sentence, paragraph, or document) because the meanings denoted by the words are semantically associated. In contrast, the words *virus* and *flowers* do not tend to occur in similar contexts, suggesting that the meanings associated with the words have little to no semantic association or similarity.

Distributional semantic models have been incorporated into a variety of cognitive models of semantic memory, predicting human performance on a variety of tasks including the TOEFL synonym task (Landauer & Dumais, 1997), word analogies (Mikolov et al., 2013), concept naming (Pennington et al., 2014), free recall (Morton & Polyn, 2016), feature generation (Cutler et al., 2019), the remote associates test (Smith, Huber, & Vul, 2013), the preferential decision making task (Bhatia, 2019; Bhatia, Ritchie, & Zou, 2019), semantic fluency (Hills et al., 2012), and binary semantic classification (Grand et al., 2022). To model behavior in any one of these tasks, the semantic representational structure captured by the distributional semantic models must be integrated with cognitive mechanisms that make use of it. For example, on the TOEFL synonyms task, participants are presented with a target word and several choices. The participant's task is to identify the synonym among the alternatives. In this example, the model's cognitive machinery is relatively simple -- the algorithm calculates the cosine similarity of the target word to each choice word and picks the word with the greatest similarity to the target among the alternatives.

Despite the success of distributional semantic models (DSMs), challenges arise in broadly incorporating them into the models of semantic tasks. While many of the tasks considered above involve evaluating words in terms of their similarity, problems arise with tasks involving the evaluation of specific properties of the words in question, since the dimensions of the semantic space are not necessarily meaningful. In other words, the proximity of the words within the representational space indicates semantic relatedness but not the nature of the relation. For example, such models would perform well on identifying the oddball among words "flower", "garden", and "vehicle". However, it is unclear how they could identify which properties of a vehicle make it the oddball.

These limitations are not true of all models of semantic memory. For example, the graph theoretic semantic models of Collins, Quillian, and Loftus (Collins & Quillian, 1969; Collins & Loftus,1975) overcome this issue by incorporating labeled links that specify the relationship between the properties of concepts. For example, the node *canary* is linked to other nodes *animal, yellow, beak,* and *fly* by the respective links *isa, is, hasa,* and *can*. Similarly, Rumelhart, McClelland, Rogers, and the PDP group (1986) developed connectionist models of semantic knowledge that are explicitly trained to store and retrieve item properties, e.g., if *bird* and *hasa* are activated, *beak* is retrieved. Finally, Smith et al.'s (1974) featural model specifies that concepts have an associated list of features that can be queried to determine properties of the concept. While these classic models offer information about the properties of items, and the relationships between concepts, these relations have been experimenter coded, and we are unaware of any current technology that can generally automate this process. DSMs, on the other hand, offer a substantial advantage in terms of their scale (e.g., 3 million words in word2vec vs. 541 concepts in a norming study by McRae et al., 2005), but lack specificity regarding the nature of the relations between concepts.

In a recent study, Grand et al. (2022) addressed this problem. Using a method similar to Osgood's semantic differential technique (Osgood, 1952; Osgood et al., 1957), Grand and colleagues (2022) collected human ratings evaluating words in terms of a variety of semantic dimensions (e.g., size, danger, gender, intelligence). For example, to evaluate the target word *elephant* on the size dimension, the words *small* and *large* were linked to the extremes of a 5-point scale, and the participant selected which number best went with the target word. They proposed a computational model which used distributional semantic representations to simulate these simple binary decisions about the characteristics of real-world objects on the semantic

dimensions examined with the human ratings. Their model proposes that the cognitive system uses the adjective labels assigned to the two extremes of the semantic dimension to construct a semantic axis in the representational space of the DSM. In other words, to make a *size* judgment, the vector representations of *large* and *small* are retrieved and are used to construct a semantic axis in this representational space. A judgment is made by projecting a given word vector onto the semantic axis and calculating which extreme it is closer to (**Figure 1**). We refer to this as an adjective-composite model of binary semantic classification. Grand et al. (2022) demonstrated the utility and flexibility of this semantic projection model, which was able to capture approximately 0.37 of variability in human ratings.

Figure 1. Schematic representation of the semantic projection. The difference vector (red dotted line) for the dimension of "size" is constructed using the item-composite semantic evaluation method (word2vec embedding), i.e., the two extremes of the vector "big" and "small" are computed as averages of the words in the dataset unanimously judged as "big" or "small" by all participants. The position of a projected word on the difference vector is referred to as dot-product score and is later used as evidence to generate each response probability.

Finally, the focus of the Grand and colleagues' (2020) study was on the geometry of the representational space of the DSMs and its relevance to human semantic judgments, not on the cognitive or decision-making bases of the semantic projection per se, leaving open the question of how such mechanism would be instantiated in the mind. Additionally, Grand et al.'s (2020) approach (also Osgood's semantic differential) is inconsistent with the dominant view of semantics (Jamieson et al., 2018), categorization (e.g., colors), and recognition (e.g., faces; Nosofsky & Palmeri, 2015), all of which favor instance-based theories. To that end, the memory retrieval mechanisms proposed in Hintzman's seminal work with MINERVA 2 (Hintzman, 1984, 1986, 1988), in which a memory store composed of many instance-based traces of past experience can be flexibly probed to reactivate composite representations are more aligned with the contemporary proposals. Furthermore, such machinery was used to great effect in the recently proposed instance theory of semantic memory (ITS; Jamieson et al., 2018), which treats multiple instances of a word's usage in natural language as independent traces in memory. This allows ITS to, among other things, interpret homonyms correctly by flexibly constructing a representation of the word's meaning on the basis of the word's current context. Together, further research is needed to define the cognitive mechanisms of the semantic projection.

The Role of Linguistic Reference in Shaping Mental Representations of Objects

Linguistic labels are useful for accessing conceptual knowledge and performing semantic judgments about objects and classes of objects (e.g., *Are balloons big or small?).* But can linguistic reference simultaneously be used to shape online representations and subsequent memories of familiar objects? Studies with special populations (i.e., congenially blind individuals) have indeed shown that language, beyond its role in knowledge expression (Li &

Gleitman, 2002), can support semantic knowledge acquisition (Kim et al., 2019; Kim et al., 2020; van Paridon et al., 2021). Therefore, it is possible that hearing or producing linguistic reference (e.g., *enormous balloon*) may shape the representations of familiar objects, akin to the process by which linguistic knowledge is continuously updated throughout lifespan (Ryskin et al., 2017).

Studies exploring the effect of language on memory for events offer insight into the nature of the relationship between linguistic reference and object representation. According to some prominent theories, language can alter memories of events by reconstructing memories at *retrieval* and/or by shaping mental representations during *encoding*. For example, Loftus and Palmer (1974) and Loftus (1985) demonstrated that participants judged the speeds of moving cars to be significantly higher when the subsequent test question included the verb "smashed" compared to "contacted." Another example of the influence of language on memory is the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) – a phenomenon where memory performance drops after individuals produce a verbal description of previously studied faces. Schooler & Engstler-Schooler (1990) hypothesized that verbalizing difficult-to-describe stimuli such as faces or colors can create a memory distortion, which interferes with the original memory trace. Together, these findings, among others, demonstrate that language production can shape memory for events and faces even after they are encoded in memory. We now turn focus to how language can shape memories at encoding.

A complimentary line of research provides evidence consistent with the notion that producing language during encoding may enhance subsequent memory for objects and object conjunctions (Szewczuk, 1970; Freund, 1972; Urgolites et al., 2020). In one study (Urgolites et al., 2020), participants studied object-scene combinations and later, had to indicate their binding

memory for them (i.e., *Did this object go together with this scene?)* (**Figure 2**). The results suggested that when participants conducted a rhythmic shadowing or a verbal shadowing task during the study phase, their binding memory was worse compared to when they studied the items passively. Furthermore, the binding memory was significantly worse in the verbal shadowing condition compared to the rhythmic shadowing condition. The authors concluded that reducing access to verbal resources impairs binding memory for objects and speculated that the ability to label object-scene conjunctions (i.e., *banana in the forest*) might be critical for maintaining memory for them. While, consistent with previous work (Szewczuk, 1970; Freund, 1972), these findings suggest that language might be beneficial for memory, an outstanding question remains: would *providing* access to language alleviate the observed memory deficit? Lastly, the cognitive mechanisms by which linguistic labeling could improve binding memory for objects remain to be explored.

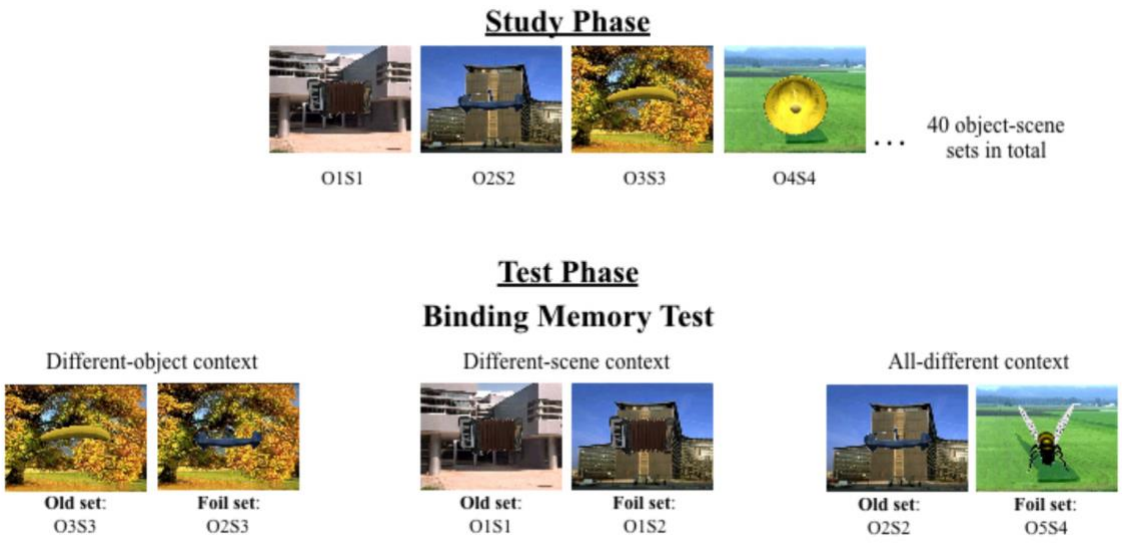


Figure 2. From Urgolites et al. (2020). Schematic of the experimental procedure. During the study phase, participants studied object-scene conjunctions on the screen one at a time. Then, participants completed individual and binding memory tests. In the binding memory test,

participants had to identify object and scene occurred together or not together during the study phase in a 2-alternative choice paradigm.

Studies that directly test the effect of language production on memory for objects show that generating a linguistic label out loud benefits memory above and beyond reading the label or thinking about it silently (Fawcett & Ozubko, 2016; Zormpa et al., 2018). A paradigm commonly used to target this question involves a task where participants view images of objects one at a time (Zormpa et al., 2018). Critically, some images simultaneously have an associated linguistic label displayed on the screen (e.g., a picture of broccoli with the word "broccoli" displayed in the middle). In both the picture+word and picture-only conditions, participants alternate between overtly naming the corresponding linguistic labels of the objects on the screen and thinking about them silently. Finally, participants complete a recognition memory test, where they have to indicate if the probe image is old or new. Zormpa et al. (2018) found a main effect of presentation such that participants were more likely to correctly recognize a probe image if it had been presented without the corresponding linguistic label compared to the picture+word condition. Further, participants were more likely to correctly recognize an image in the out loud vs. silent condition. In sum, these findings indicate that generating a linguistic label improves memory for objects and this effect is amplified when language production occurs out loud vs. silently.

In turn, only few studies examined the effect of processing of linguistic reference on memory for objects. For instance, in Lord and Brown-Schmidt's (2022) study, participants listened to auditory instructions and selected corresponding items on the screen. Importantly, Lord and Brown-Schmidt (2022) manipulated the referential form of the instructions such that

the modification of the noun was either pre-nominal (e.g., *Click on the dotted bag*) or post-nominal (e.g., *Click on the bag with dots*). Later, the authors tested participants' memory for the target (bag with dots), early competitor (dotted bowtie/bag with stripes), late competitor (bstriped bag/bowtie with dots), and non-competitor (ice-cream cone) in a 2-AFC paradigm. The authors found that targets were remembered better than competitors, and in turn, competitors were remembered better than non-competitors indicating that hearing a linguistic reference improves memory for the objects it describes. Importantly, Lord and Brown-Schmidt (2022) found an early > late competitor effect both for post-nominal and pre-nominal expressions, due to a natural and manipulated longer temporal ambiguity in post-nominal phrases and pre-nominal phrases at slow speech rate respectively. Overall, the authors concluded that referential expressions such as "the bag with dots" can create temporal ambiguity and activate the target along with the contextual competitors, all of which are encoded in memory. The significance of these findings is two-fold: it illustrates that referring to an object boosts memory for it, and it shows that the linguistic form of the referential expression matter in what is and what is not encoded in memory.

Finally, we review work examining the effect of various types of linguistic reference on memory in children, which attempted to further illuminate the cognitive mechanisms underlying the mnemonic benefits for object conjunctions associated with language. More specifically, Dessalegn and Landau (2008) demonstrated that certain types of linguistic labels helped to maintain the conjunction of visual features in a delayed matching task in 4-year-olds. Across four experiments, children were able to correctly indicate the location and color of an item after it was no longer on the screen when the objects were studied along with auditory directional labels (i.e., *Where is red? The red is on the left*). In contrast, children's performance dropped

when they viewed the items either without hearing directional labels (e.g., *This is a dax/The red is touching the green*), or along with nonlinguistic attentional cues (e.g., objects flashing). In a different study (Scott & Sera, 2018), children from 5 to 10 years of age completed a nonverbal spatial memory task, where they first encoded and later indicated their memory for the location of a dot inside a circle. Children's memory for the dot location improved when they practiced the relational labels (e.g., *left, right, top, bottom*) before the study. The authors concluded that practicing relational terms prior to study increased children's access to language which was later used to unify different instantiations of objects (i.e., dot and circle) into a single representation. Both studies speculated that language and vision interact to create a unified "hybrid" representation to aid visual memory. However, neither of them specified what this hybrid representation is, or exactly how linguistic labeling contributes to this enhancement.

While Lord and Brown-Schmidt (2022) propose an activation-time account according to which temporal ambiguity in language can activate multiple objects which are later represented in memory, Dessalegn and Landau's (2008) idea of a "hybrid representation" lacks a mechanistic explanation of how it may improve memory for objects. Relatedly, linguistically oriented work probes a similar idea and suggests that certain types of language may evoke conceptual composites of individual objects during language processing. Specifically, in a study by Brown-Schmidt et al. (2005), participants manipulated objects on the table as instructed by the experimenter. The instructions included two sentences with the first sentence indicating participants to place one object on top of the other (e.g., *put the cup on the saucer*). Critically, in the second sentence, the referent was manipulated (e.g., *now put <u>it/that</u> over by the lamp*). The findings revealed that when the experimenter used the pronoun "it", participants interpreted it as referring to the theme noun (i.e., cup), while the pronoun "that" was interpreted as referring to

the composite (i.e., cup and saucer). Brown-Schmidt and colleagues (2005) concluded that the pronoun *that* is used to refer to conceptually complex items or sets of items, and when processed, evokes an interpretation of a composite entity similar to the idea of a "hybrid representation". Though it is still unclear how such language would affect memory for objects.

In conclusion, previous evidence suggests that both language production and comprehension can shape memory for objects and object conjunctions (Dessalegn & Landau, 2008; Scott & Seta, 2016; Zormpa et al., 2018; Urgolites et al., 2020; Lord & Brown-Schmidt, 2022). One of the two primary hypotheses elucidating how processing different types of referential expressions may shape memory for objects is the activation-time account (Lord & Brown-Schmidt, 2022), according to which in the course of language processing, temporal ambiguities may cause activation of multiple meaning candidates, which are encoded and later retrieved from memory. In contrast, the alternative proposal of a "hybrid representation" (Dessalegn & Landau, 2008; Scott & Sera, 2016) is underspecified, lacks a mechanistic explanation, and requires further investigation.

The Role of Linguistic Features of Spontaneous Speech in Shaping Mental Representations of Real-World Experiences

While computational and experimental approaches provide utility in targeting the cognitive underpinnings associated with the effect of language on memory through meticulously controlled experimental designs, they leave questions regarding the generalizability of the findings. In contrast, observational studies that utilize ecologically valid designs and explore natural language use offer insights into the linguistic features of spontaneous speech that might be difficult to manipulate or elicit in laboratory settings. For example, studies of conversational

memory often employ paradigms where participants engage in a conversation as the interaction is recorded. After a delay, participants are then asked to recall the conversation in detail, or to make judgments about utterances that did or did not appear in the original conversation. Findings from these paradigms offer important insights into the nature of conversational memory. First, conversational memory tends to be gist-like with participants recalling the thrust of what was said in the conversation but simultaneously, struggling to correctly distinguish verbatim statements from paraphrases as soon as 5 days after the interaction (Sachs, 1974; Kintsch & Bates, 1977). Second, many studies report a generation benefit, such that memory is much better for what a participant said themselves, compared to what their conversational partner said to them (Slamecka & Graf, 1978; Hjelmquist & Gidlund, 1985; Isaacs, 1990; Miller et al., 1996; Fischer et al. 2015; McKinley et al., 2017; Ross & Sicoly, 1979; Zormpa et al., 2019). Finally, conversational recall is quite limited, with estimates of the percent of idea units that can be accurately recalled after delays of minutes to days being between 5-20% of all ideas expressed in the original conversation (Benoit & Benoit, 1998; Ross & Sicoly, 1979; Samp & Humphreys, 2007; Stafford & Daly, 1984; Stafford, Burggraf, & Sharkey, 1987).

What makes a portion of conversation more likely to be recalled? Studies examining the type of *content* that impacts the memorability of conversation show that participants remember emotionally charged content such as sexually explicit statements better than emotionally neutral content (Pezdek & Prull, 1993). For example, in one study, Pezdek and Prull (1993) presented participants with an audio recording of a conversation between a man and a woman during which the speakers made sexually explicit statements. The authors manipulated the perceived context in which the conversation occurred. In an "incongruent" context, participants were told that the two speakers were in an office. In a "congruent" context, participants were told that the two speakers

were at a singles bar. Finally, the authors tested participants' gist and verbatim memory 5 weeks after hearing the conversation (Experiment 1) or 3 hours after hearing the conversation (Experiment 2). The results indicated that participants recognized and recalled the gist of sexually explicit statements significantly better than non-sexual ones. Further, the incongruent context increased the memorability of the sexually explicit statements, potentially because they violated the expectations of what is considered an appropriate conversation topic in a professional setting.

Other findings suggest that spoken language that has high-interactional value is more likely to be recalled than language with low-interactional value. Keenan and colleagues (1977) recorded a meeting held between faculty members and graduate students. The experimenters then identified statements from the meeting that were either high-interactional value, i.e., conveyed humor or personal criticism (such as a joke), or low-interactional value (e.g., *You put a little morpheme that says you're going to choose the Object as Subject*). Thirty hours later, they tested the attendees' recognition memory of the utterances from the meeting and found that participants were more successful at discriminating verbatim statements made in the conversation from paraphrases for high- than low-interactional content. The authors also recruited a separate sample of participants, played the same recording of the meeting to them, and immediately after, tested their recognition memory for the same statements. Group comparisons revealed that individuals who were physically present at the meeting demonstrated better memory than the individuals who listened to the recording. These results highlight the importance of inter-personal interaction for conversational memory.

Linguistic Features of Conversation and Memory

Some evidence indicates that linguistic features inherent to spontaneous speech may similarly shape memory for language. Here we review several potentially relevant linguistic features of conversation.

Disfluency

An emerging body of research indicates that disfluencies, or hesitations in speech such as *um/uh*, boost memory for spoken language by orienting listener's attention to the upcoming speech stream (Corley et al., 2007; MacGregor et al., 2010; Fraundorf & Watson, 2011; Diachek & Brown-Schmidt, 2022). For example, Fraundorf and Watson (2011) presented participants with audio recordings of passages from the *Alice in Wonderland*, and later tested participants' memory for them. Critically, the recordings were either fluent or contained disfluencies. Participants were more likely to remember the gist of plot points when the passages were disfluent compared to fluent. This disfluency-memory benefit was observed even when disfluency occurred in unpredictable positions in the sentence, consistent with the hypothesis that disfluency orients listeners' attention to speech regardless of its location.

In a different study, Diachek and Brown-Schmidt (2022) observed a disfluency-related memory boost for individual words in pre-recorded sentences. Participants listened to sentences, some of which contained a penultimate disfluency (e.g., "Everyone's got bad habits and mine is biting my um nails"). A subsequent recognition memory test revealed that the odds of correct recognition were 1.45 times higher for sentence-final probe words preceded by disfluency compared to probes from fluent sentences. Follow-up experiments determined that this disfluency-related memory boost occurred with different disfluency types (fillers, pauses, and repetitions), and was short-lived, fading rapidly after the disfluency onset. These findings

support an attentional orienting account of the disfluency-related memory boost, according to which disfluency orients attention to the upcoming linguistic material, improving encoding and subsequent memory for it (Collard et al., 2008; Fraundorf & Watson, 2011).

An important caveat, however, is that Diachek and Brown-Schmidt (2022) found that the disfluency memory benefit was *not* observed when disfluency occurred earlier in the sentence, or when words at earlier sentence position were probed. Yet in unscripted language, naturally produced disfluency occurs in a variety of sentence positions (Butterworth, 1975; Bortfeld et al., 2001; Clark & Fox Tree, 2002). This, then, raises the question as to whether the previously reported disfluency memory boost for individual pre-recorded sentences (Corley et al., 2007; MacGregor et al., 2010; Diachek & Brown-Schmidt, 2022), or pre-recorded stories (Fraundorf & Watson, 2011) would generalize to memory for what was said in natural and unscripted conversation. One reason to think the disfluency boost may be limited, are findings by Toftness and colleagues (2018) who examined the effect of a professors' fluency in recorded instructional videos. While students rated disfluent professors as less effective at teaching compared to the fluent professors, learning outcomes did not differ between the two types of instruction, putting into question the ecological validity of the previously observed mnemonic benefit associated with disfluency (Corley et al., 2007; MacGregor et al., 2010; Fraundorf & Watson, 2011; Diachek & Brown-Schmidt, 2022). Another reason to believe the disfluency boost may be limited are findings by Donahue, Schoepfer, and Lickley (2017) which indicated that recall for passages that contained disfluencies was significantly *worse* compared to fluent passages[2]. In sum, the outstanding question is whether the disfluency-memory benefit generalizes to the unscripted forms of language use in which disfluency is likely to occur.

[2] However, in this study, the materials were manipulated in a way that resembled stuttering, which might be processed by listeners differently than disfluency.

Backchanneling

Another feature of spontaneous speech that, like disfluency is related to attention, is backchanneling. Backchanneling refers to feedback that conversational partners give each other, including brief expressions or phrases (e.g., *uh huh; mmm; right*; *so true*), as well as facial and manual gestures, and actions in shared spaces (Bangerter & Clark, 2003; Fox Tree & Mayer, 2008; Krauss et al., 1977). Backchanneling is prevalent in spontaneous speech with some studies estimating that 8 out of 10 spoken backchannels made in conversation are produced within 1-15 seconds of each other (Oreström, 1983). Conversational partners use backchannels in a variety of ways, including offering evidence of understanding, or to seek clarification (e.g., A: *I could really use a coffee;* B: *coffee?*), among other speech acts (Clark & Schaefer, 1989; Healey et al., 2018; Watzlawick, 1964). Backchannel responses can also help maintain the flow of a discourse (Dittman & Llewellyn, 1967; Oreström, 1983), and communicate emotions and approval (O'Keeffe & Adolphs, 2008). Critically, backchannel responses can also be used to signal continued attention (Fries, 1952; Kendon, 1967; O'Keeffe & Adolphs, 2008; but see Schegloff, 1982 for an argument that backchanneling signals agreement, not attention), and as a result might be predictive of conversational recall.

Yet little is known about the effect of backchannels on representations of conversation. Some initial work examining the cognitive implications of backchannel responses reports that when a speaker says something in conversation, and their partner produces a backchannel response, that this has little to no effect on the speaker's belief that what they said was heard and understood (Brown-Schmidt, 2012; Brown-Schmidt & Fraundorf, 2015). This raises the

intriguing possibility that while people may regularly produce backchannels, that they have little impact on their partner's representation of the conversation.

Discourse Marker "like"

The use of "like" as a discourse marker is prevalent in spontaneous speech with some studies estimating that its frequency exceeds the frequency of the conjunction "and", "you know", "I mean", and "well" (Tagliamonte, 2005; Fox Tree, 2006; Beeching, 2016). While the role of "like" remains debated, some argue that "like" and disfluency are used interchangeably (Valentine, 1991). For example, Fox Tree (2006) found that "like" and disfluency systematically occurred in similar positions in stories, namely, before proper nouns and times. These findings tentatively suggest that, like disfluency, the discourse marker "like" may orient attention to the upcoming speech stream, improving memory for it.

Disagreement

Finally, recall that studies examining *content* memorability reveal a memory benefit for emotionally charged content such as sexually explicit statements (Pezdek & Prull, 1993). Other findings suggest that spoken language with high-interactional value (e.g., humor or personal criticism is more likely to be recalled than language with low-interactional value (e.g., linguistic definitions; Keenan et al., 1977). Given that disagreement might be considered both emotionally charged and of high interactional value, statements that contain disagreement might be recalled better.

Overall, studies examining conversational recall have established that memory for conversation is limited (recalling less than half of what was said) and asymmetric (each party

recalling more of what they said themselves). What factors determine what will be remembered following conversation? We have identified three features of spontaneous speech, namely, disfluency, backchanneling, and the discourse marker “like”, that might improve memory for conversational speech due to the attentional processes (Fries, 1952; Kendon, 1967; Fraundorf & Watson, 2011; Diachek & Brown-Schmidt, 2022). Additionally, studies showing that utterances that have high interactional value tend to be remembered better, suggest that disagreement might be predictive of better recall (Keenan et al., 1977; Pezdek & Prull, 1993). Yet, the effect of these four features on memory for conversation remains to be explored.

CHAPTER II

Study 1. The Role of Linguistic Labels in Accessing and Evaluating Semantic Properties of Lexically Invoked Concepts

One question fundamental to understanding the relationship between language and memory concerns the role of language in the representation of conceptual knowledge. To date, distributional semantic models (DSMs), which construct vector spaces with embedded words, are a proposed framework for understanding the representational structure of human conceptual knowledge. However, unlike the classic semantic models (e.g., graph theory, Collins & Quillian, 1969; Collins & Loftus,1975; connectionist theory, Rumelhart et al., 1986), because DSMs are constructed using the text co-occurrence information only, they lack a mechanism for specifying the properties of concepts, raising questions regarding their utility for a general theory of semantic knowledge. To address this issue, Grand et al. (2022) proposed a computational model of binary semantic classification that used the adjective labels assigned to the two extremes of a semantic dimension to construct a semantic axis in the representational space of the DSM. While Grand's adjective-composite model reliably and accurately predicted human ratings across an array of semantic dimensions, they did not propose how the model would be employed by the cognitive system. The goal of **Study 1** is to contrast the adjective-composite model with an alternative item-composite model in terms of their ability to predict human task performance. This novel computational mechanism proposes that while performing a binary semantic classification task, participants use each adjective label of the judgment (i.e., *big* and *small)* to retrieve a set of items representative of each extreme. Importantly, the item-composite model is

reminiscent of the highly impactful memory retrieval mechanisms MINERVA 2 (Hintzman, 1984, 1986, 1988) and ITS (Jamieson et al., 2018).

Methods

Behavioral Data

The data used to create and evaluate the computational models were collected for a previously published study described in Polyn et al. (2009). Forty-two participants were presented with a series of target words one at a time along with a task cue on a computer screen (**Figure 3**). On size trials, participants indicated whether the referent was big or small (compared to a shoebox) with a keypress. On animacy trials, participants indicated whether the referent was living or nonliving with a keypress. Each target word was presented in the middle of the screen for 3 s. If the participant did not make a response within 3 s, a warning message was displayed and they advanced to the next trial automatically. Each set of 24 target words was followed by a free-recall period which is not examined here. The 24 target words were either all associated with the same task, or a mix of the two tasks. Target words were drawn from a word pool with 1,650 unique words. Each participant completed 4 experimental sessions, each with 12 sets of target words, for a total of 1,152 trials. The final dataset contained 47,520 unique responses (after excluding missed trials).

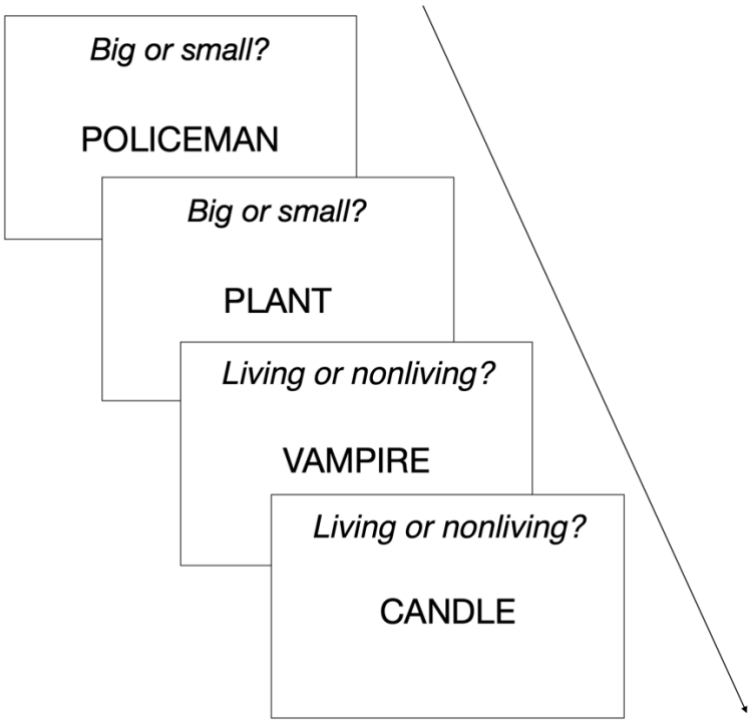


Figure 2. Schematic representation of the binary semantic classification task. Target words were presented one at a time underneath a task cue, and participants indicated their response via keypress. See text for details.

Modeling

We created a likelihood-based modeling framework to simulate and predict human performance on the binary semantic classification task. Each individual model, described in more detail below, was defined in terms of the following subcomponents: i) a *distributional semantic model* that was used to retrieve semantic vectors (GloVe or word2vec); ii) a *semantic evaluation model* that was used to produce an evidence estimate for each choice alternative (single-adjective, adjective-composite, or item-composite); and iii) a *decision model* that was used to convert the evidence into a decision likelihood for each choice alternative, and for the second

decision model, calculate response latency likelihood (logistic or linear ballistic accumulator). This yielded a family of 12 models, which were evaluated against each other.

Distributional Semantic Models. We used two different word embeddings for the semantic vectors for the target words and adjective labels. Specifically, we chose 400-dimensional pretrained word2vec vectors (Mikolov et al., 2010) produced with local context methods, trained on the Google News dataset (approximately 100 billion words). Additionally, we used 300-dimensional pretrained GloVe vectors (Pennington et al., 2014) trained on a combination of Wikipedia 2014 and Giga-word 5 (6 billion tokens).

Semantic Evaluation Algorithm. Three semantic evaluation algorithms were created to calculate evidence (referred to here as dot-product scores) for each choice alternative. Each evaluation algorithm constructs a decision axis in the semantic space for each judgment task (size or animacy). To do this, the algorithm selects one or more representational vectors for each extreme of the continuum. These vectors are used to construct the decision axis as described below.

The *single-adjective model* used a difference vector that was constructed by subtracting the vector for the adjective label "small" from the vector for "big" for the size trials, and subtracting "inanimate" from "animate" for the animacy trials. Following the method of Grand and colleagues (2022), the *adjective-composite model* used a difference vector that was constructed by taking the difference between the two averages of three synonyms, i.e., the difference of {"huge", "big", "large"} and {"small", "little", "tiny"} for the size trials, and the difference of {"animate", "living"} and {"inanimate", "nonliving"} for the animacy trials. Finally, the *item-composite model* used a difference vector that was constructed by averaging the semantic vectors for a set of words that are representative of the extremes on each semantic

dimension. To determine the representative words, we identified the words that had been judged unanimously as big (n = 402, 24% of all unique words) or small (n = 222, 13%) and animate (n = 203, 12%) or inanimate (n = 569, 34%) across all 42 participants (the list of unambiguous words used in the construction of the composite semantic axis can be found in the Appendix).

To derive the evidence for each individual trial, we calculated the dot product between the semantic vector for the target word and the difference vector resulting in one value, which we refer to as a dot-product score.

It is important to note that the difference vector was only created when evaluating the logistic decision model, not the linear ballistic accumulator model (LBA). Because the LBA model implies two competing accumulators for each response alternative, on each trial, we derived two pieces of evidence by calculating two dot product scores between the semantic vector for the target word and the vectors for the two response extremes. For the single-adjective semantic evaluation model, we used vectors for individual words “big”, “small” for the respective accumulators on the size trials, and “animate”, “inanimate” for the animacy trials. The Grand model used the average of multiple synonyms, i.e., “huge”, “big”, “large” or “small”, “little”, “tiny” on the size trials, and “animate”, “living” or “inanimate”, “nonliving” on the animacy trials. The item-composite model used the average of the semantic vectors for all the words unanimously judged as big or small and animate or inanimate to create four respective accumulators.

Decision Models. Two decision models -- logistic transformation and linear ballistic accumulator -- were used to convert the evidence into a decision likelihood for each choice alternative. Each decision model is described in more details below.

Logistic transformation. In the logistic version of the decision model, we generated the predicted responses for a given word using the logistic function. The probability for a given response was calculated using the logistic function using the following equation:

$$f(x) = \frac{1}{1 + e^{-k(x)}} \qquad \text{Equation (1)}$$

where e = the natural logarithm base, x = *dot product score value between the semantic vector for a given word and a difference vector*, i.e., the distance between the difference vector and a given word (determined by the semantic model), k = free parameter indicating the steepness of the curve, and the denominator = the curve's maximum value fixed at 1.

Linear Ballistic Accumulator. In addition to the logistic decision model, we implemented a linear ballistic accumulator (LBA) model (Brown & Heathcote, 2008), which allowed us to evaluate model performance not only in terms of the predicted responses but also reaction times. LBA is a simple model of decision and response time that assumes multiple independent accumulators raising towards a certain decision threshold in a linear and deterministic manner (i.e., noise-free) until the decision is made. In the LBA model, each evidence accumulator begins with a certain amount of evidence reflected at the starting point *k* that increases at a speed determined by the drift rate *d* until it reaches the response threshold *b* (**Figure 4**). The first accumulator to reach the threshold determines the response and the time to reach the threshold, or RT, is calculated as *(b-k)/d*.

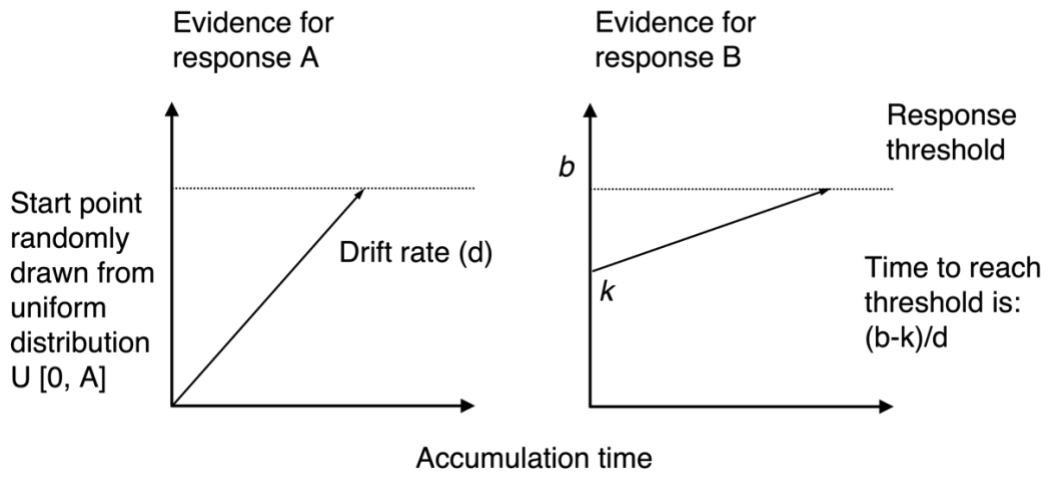


Figure 4. Adapted from Brown and Heathcote (2008). Two-choice version of the LBA. Panel on the left, shows the evidence for Response A, panel on the right - evidence for Response B. Starting values *k* are randomly drawn from a random distribution. The drift rate *d* determines the rate at which the decision process approaches the decision threshold, and is drawn independently for each response from normal distributions with standard deviations *s*. A response is made when the first accumulator reaches the threshold *b*.

Model Evaluation

We evaluated the fitness of each model variant using the maximum likelihood estimation technique. As mentioned earlier, the item-composite model used the words that participants unanimously judged at each extreme in construction of the difference vector. To ensure that the results of the item-composite model were not driven by its performance on these items, we examined each model's performance on the rest of the words.

Maximum Likelihood Estimation. For both the logistic decision model and linear ballistic accumulator, we calculated the likelihood of each model given the observed data by summing the log values for the model's trial-level predictions after excluding the words used to

construct the item-composite vector (this is equivalent to taking the product of the probability values associated with these trials). The probability of a given dataset is the product of the estimated probabilities of each of the experimental events (trials). This number was transformed into a log-probability for model comparison statistics. We used the log-likelihood value to calculate the weighted Aikake's Information Criterion (wAIC) to evaluate the fitness of multiple models relative to each other.

Following Grand et al. (2022), we evaluated the results of the semantic projection using a linear correlation analysis and pairwise order consistency.

Linear Correlation. For each word in the dataset not used in the construction of the item-composite vector, we first calculated the dot-product score and the mean judgment value separately for each task (with the mean of 0 indicating all participants judged the word as small or inanimate, and 1 indicating all participants judged it as big or animate) averaged across the participants. Second, we calculated the Pearson correlation coefficient and the associated p-value between the dot-product score and mean judgment value for the words in this restricted dataset.

Pairwise Order Consistency. Following the method of Grand et al. (2022), we calculated the proportion of two-word combinations in the restricted dataset for which the difference between the human judgment and the dot-product scores was in the same direction, out of all possible two-word combinations. For example, if the words *elephant* was judged on average as larger than the word *mouse* and the dot-product score for *elephant* was larger than for *mouse*, then the *elephant-mouse* word pair would get a score of 1 and 0 otherwise. We repeated that procedure for each possible two-word combination, resulting in $1{,}650^2$ possible word combinations and scores (0 and 1). The final score is the proportion of 1s across all possible two-word combinations.

Results

Maximum Likelihood Estimation – Logistic

Figure 5 reports the fitness of each model variant in terms of the likelihood of generating the observed data. The results indicated that the item-composite semantic evaluation model was most likely to generate the observed data (wAIC for item-composite: 1.0, for adjective-composite and single-adjective: 0.0 each). The results additionally indicated that when summed across all models and tasks, the models that utilized GloVe outperformed the models that utilized word2vec (wAIC for GloVe: 1.0, word2vec: 0.0). The same analysis on the full set of items showed similar results and is reported as a supplemental analysis.

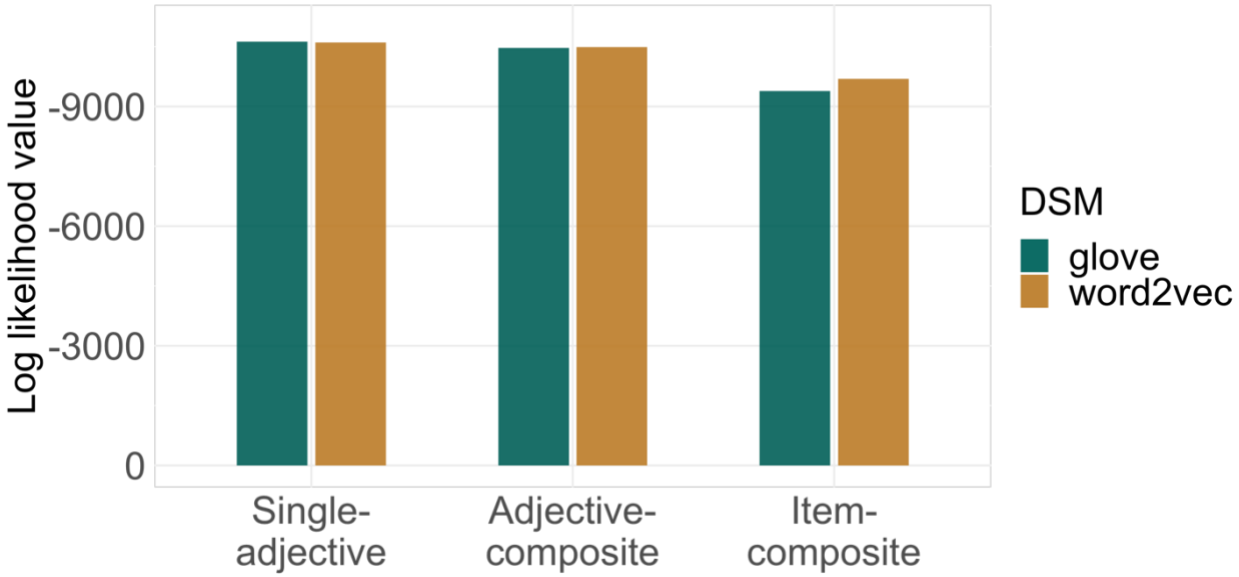


Figure 5. Log likelihood values for the logistic decision model combined across the size and animacy tasks. Different colors represent two distributional semantic models. Values closer to zero correspond to a closer fit.

Correlation Analysis

Using the logistic decision-making model, we carried out a correlation analysis to characterize the degree of correspondence between each model's predicted responses and the observed responses (**Figure 6**). The item-composite model yielded numerically largest and consistently reliable (all *p*s < 0.05) correlations between the predicted and observed values across the two embedding spaces. The item-composite model had the highest correlation of 0.63, followed by the adjective-composite model with a correlation of 0.10, and the single-adjective model with a correlation of -0.02 (all differences between Fisher-transformed correlation coefficients *z*s > 13.71, all ps < 10^{-5}). When averaged across all models and embeddings, the correlation coefficients for the size and animacy tasks were numerically different (means 0.35 and 0.13 for size and animacy tasks respectively). When averaged across all models and tasks, the GloVe and word2vec embeddings produced numerically comparable results with the means of 0.28 and 0.20 respectively. The same analysis on the full set of items showed similar results and is reported as a supplemental analysis.

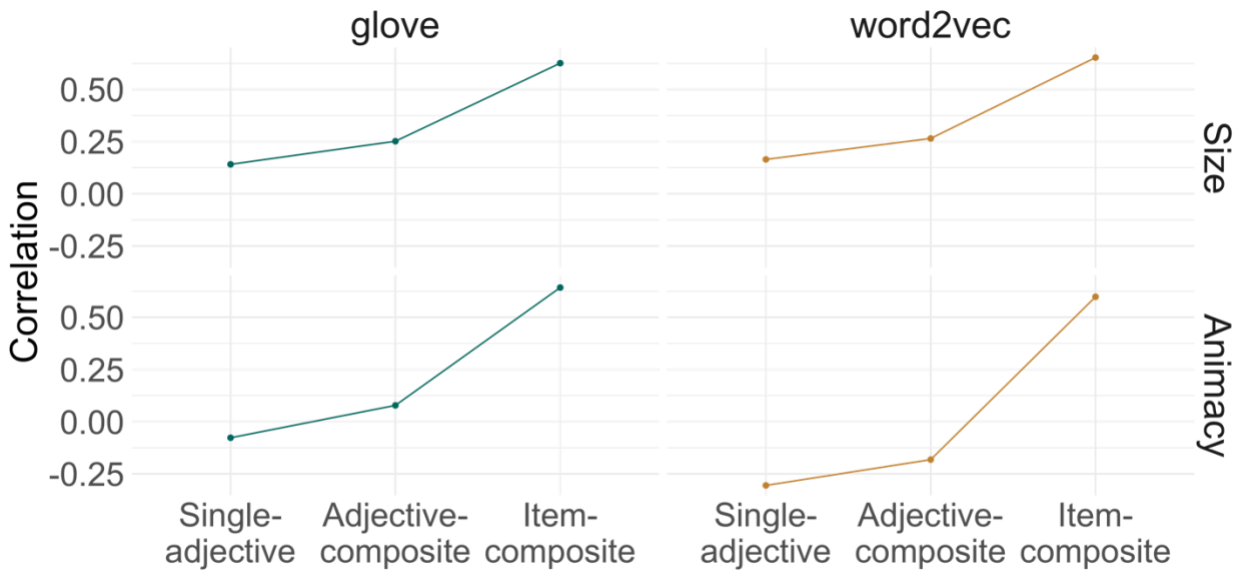


Figure 6. Pearson *r* coefficients between the predicted responses and the mean human judgments for the size task (upper panel) and animacy task (lower panel). Different colors represent two distributional semantic models.

One potential advantage of the item-composite model is that a larger number of word vectors are used to construct the semantic classification model as compared to the other two models. However, a follow-up analysis suggests that the item-composite model performs at a superior level even when the number of words used to make the composite is matched across the different model types. To do this, we carried out a specialized permutation analysis on the trials using the size task. For each permutation, we randomly selected 3 words each from the sets of unanimous big and small words used to construct the item-composite difference axis. We used these words to construct a new difference axis and re-ran the correlation analysis reported above. We repeated this procedure 100 times for both the GloVe and word2vec models to obtain a distribution of correlation values. The mean correlation coefficient across the hundred

permutations was 0.48 for GloVe and 0.52 for word2vec (**Figure 7**). While this value was numerically smaller than for the full item-composite model (means of 0.69 and 0.72 for GloVe and word2vec), it was significantly larger than the correlation values associated with adjective-composite model (0.19 and 0.29 for GloVe and word2vec respectively). In other words, 99 and 97 out of 100 permuted values for GloVe and word2vec respectively exceeded the adjective-composite correlation score. This indicates that the predictive advantage of the item-composite model is due to the semantic identities of the words used to construct the semantic model, rather than the quantity of words.

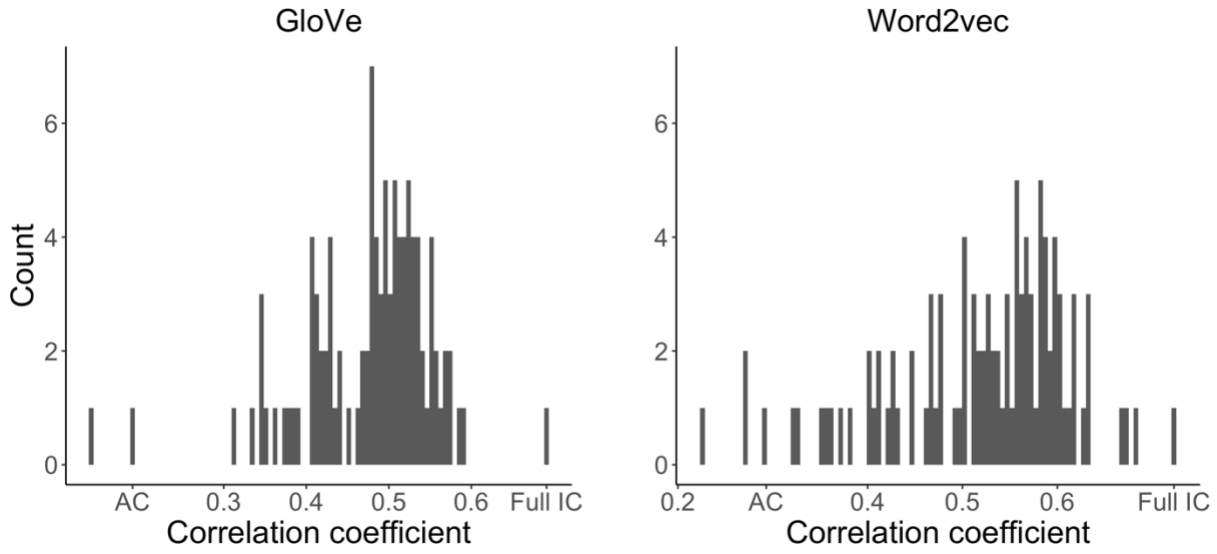


Figure 7. Pearson *r* coefficients between the predicted responses and the mean human judgments for the size task fit using GloVe (left panel) and word2vec (right panel) obtained through the permutation analysis. AC indicates the correlation coefficient calculated for the adjective-composite model, and Full IC indicates the correlation coefficient calculated for the item-composite model used in all prior simulation analyses (i.e., with all unanimously judged items included in the semantic composite).

Pairwise Order Consistency

Following the methods of Grand et al. (2022), we carried out a pairwise order consistency analysis to characterize the degree of correspondence between each model's predicted responses and the observed responses (**Figure 8**). A pairwise order consistency score of 50% indicates chance-level performance. On average, the single-adjective model performed at chance (mean = 49.78%). The adjective-composite model performed significantly above chance (mean = 53.74%). The item-composite model performed substantially better than the other two models (mean = 72.41%). The GloVe and word2vec models performed similarly well with mean pairwise order consistency of 59.96% and 57.33% respectively. The same analysis on the full set of items showed similar results and is reported as a supplemental analysis.

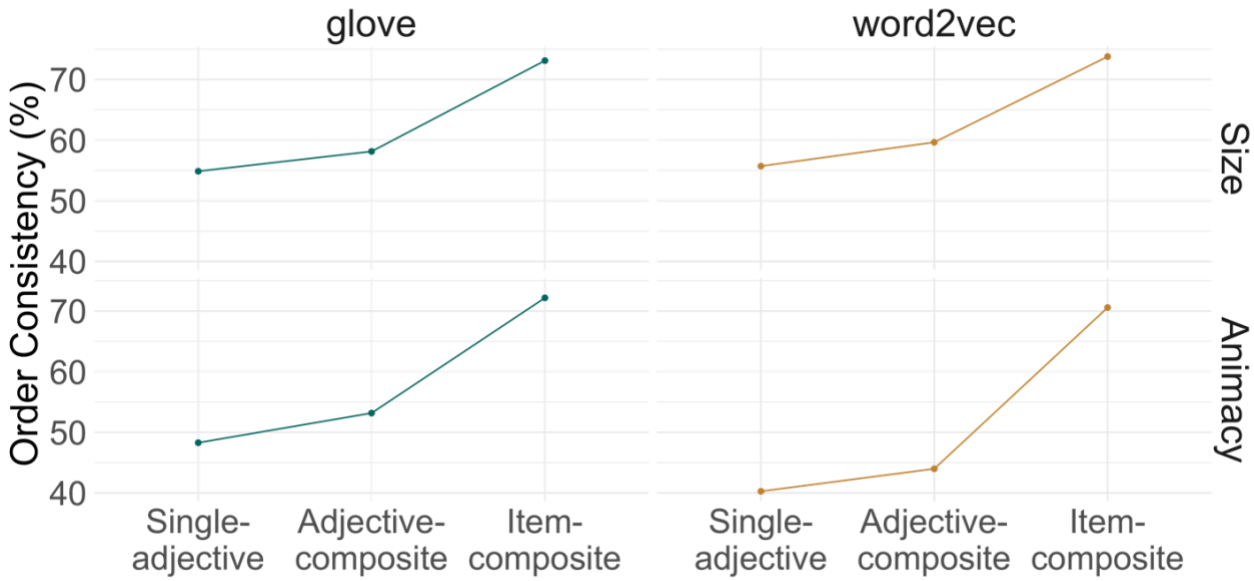


Figure 5. Pairwise order consistency values for the size task (upper panel) and animacy task (lower panel). Different colors represent two distributional semantic models.

Maximum Likelihood Estimation – Linear Ballistic Accumulator

To evaluate if the item-composite framework could be extended to response times, and if its advantage is retained when reaction times are predicted, we additionally implemented the LBA decision model. **Figure 9** reports the fitness of each model variant in terms of the likelihood of generating the observed data. Consistent with the results of the logistic decision model, the item-composite semantic evaluation model was most likely to generate the observed data (wAIC for item-composite: 1.0, for adjective-composite and single-adjective: 0.0 each). One difference emerged with regard to the underlying DSM. Here, word2vec yielded the best fit to the data (wAIC for GloVe: 0.0, word2vec: 1.0). The same analysis on the full set of items showed similar results and is reported as a supplemental analysis.

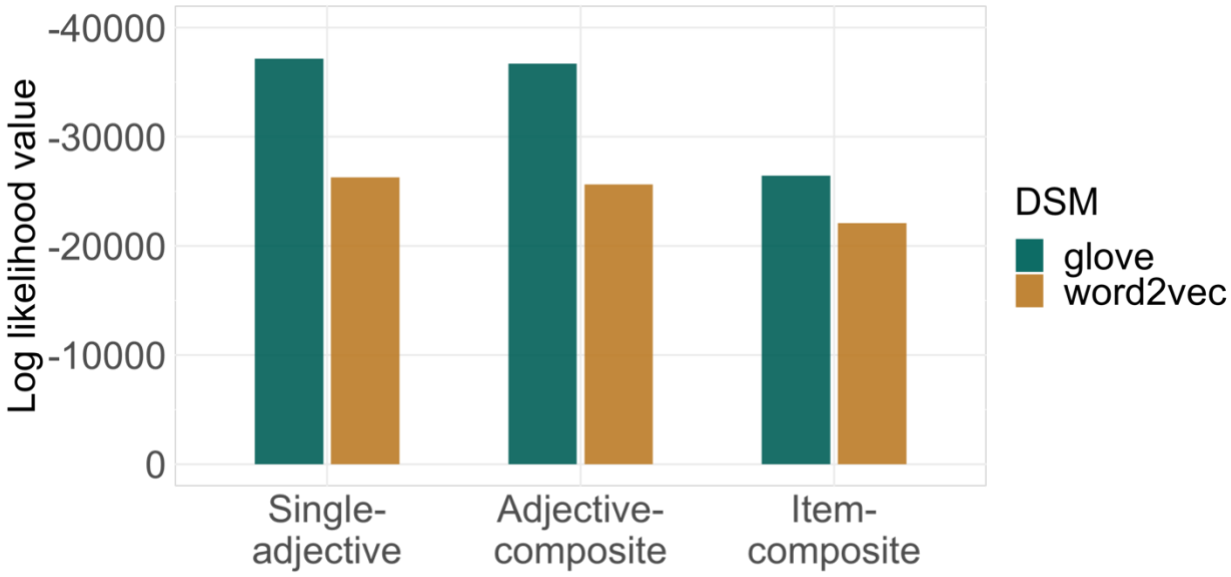


Figure 6. Log likelihood values for the LBA decision model combined across the size and animacy tasks. Different colors represent two distributional semantic models. Values closer to zero correspond to a closer fit.

Our simulations with the linear ballistic accumulator model indicate that the item-composite semantic model captures not only the binary responses on the semantic classification task but also the response times on individual trials. **Figure 10** shows the response time histograms for the observed correct responses as well as the predictions for the item-composite LBA model. The distributions were plotted for the trials that included words with strong dot-product scores (the top 50% of the scores) and weak dot-product scores (the bottom 50% of the scores). One weakness of the item-composite linear ballistic accumulator model is apparent from this analysis. The model successfully predicts that words with strong dot-product scores will have faster response times than the words with weak dot product scores but predicts the effect will be much larger than is actually observed. More specifically, a paired-samples t-test on the *predicted* values indicated that the mean difference between by-participant RTs for the words with the strong and weak dot-product scores for the axis extreme "big/animate" was 648.23 ms ($t_{(41)} = -58.65$, $p < 10^{-15}$) and 521.78 ms ($t_{(41)} = -58.66$, $p < 10^{-15}$) for the axis extreme "small/inanimate", indicating a significantly slower response time for the words with weak dot-product scores. In contrast, a paired-samples t-test on the *observed* values indicated that the mean difference between the RTs for the words with the strong and weak dot-product scores for the axis extreme "big/animate" was only 64 ms ($t_{(41)} = -21.73$, $p < 10^{-15}$) and 30.83 ms ($t_{(41)} = 8.51$, $p < 10^{-9}$) for the axis extreme "small/inanimate". While the item-composite model predicts larger RT differences between the words with the strong dot-product scores and weak dot-product scores that observed, it performs better than the adjective-composite model. For comparison, the adjective-composite model estimated an even larger by-participant mean RT difference between the words with strong and weak dot-product scores of 1,110.70 ms ($t_{(41)} = -76.46$, $p < 10^{-15}$) for

the axis extreme “big/animate” and 1,084.58 ($t_{(41)}$ = -66.54, $p < 10^{-15}$) for the axis extreme “small/inanimate”.

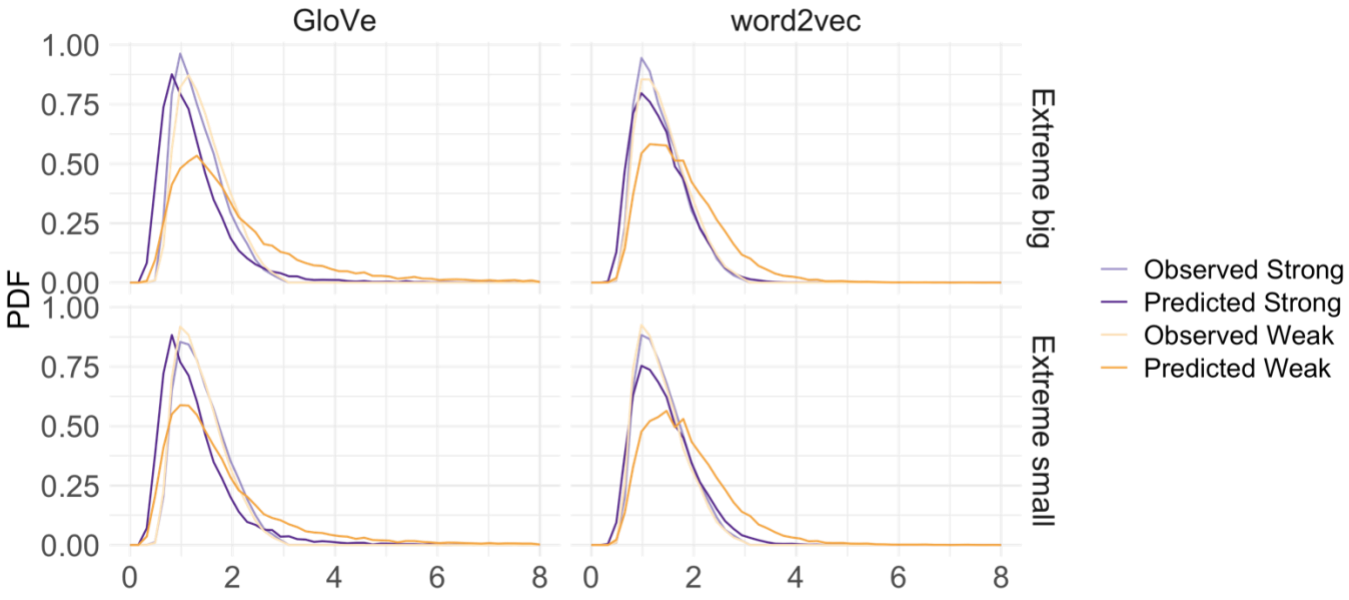


Figure 7. Probability density function for the observed and predicted reaction times according to the LBA model for the item-composite semantic evaluation model fit with the word2vec (on the right) and GloVe (on the left) embeddings. The top panel corresponds to the RTs for the response “big” or “animate”, the bottom panel corresponds to the RTs for the response “small” or “inanimate”. Generally, the item-composite model predicts the RT difference between strong and weak words will be bigger than what is observed.

Discussion

Semantic memory -- a component of long-term memory -- stores information about the world and the things in it. This memory system has been a central component of theories of memory for decades, yet many open questions remain regarding how humans store and

manipulate these rich representations. Distributional semantic models (DSMs) have offered insight into the nature of human semantic memory, as they have been used both as a tool to understand behavioral data and as theories of representation of semantic knowledge (Landauer & Dumais, 1997; Lund and Burgess, 1996; Jones and Mewhort, 2007). These models offer a way to represent semantic spaces and can be combined with the cognitive mechanisms of decision-making to characterize human semantic categorization behavior.

In the current study, we combined the principles of the multiple-trace theory of memory (MINERVA 2, Hintzman, 1986), the instance theory of semantic knowledge (ITS, Jamieson et al., 2018) and the methods from Grand and colleagues (2022) to build a computational model of binary semantic classification. ITS proposes that encounters with words are stored as individual traces in episodic memory, and that the semantic meaning of a word can be constructed on-the-fly by retrieving a blend of many memory traces containing independent instances of usage of that word from episodic memory. In our simulation of the binary semantic classification task, we compared two instance-inspired scenarios. In one case, the semantic identity of a set of adjective labels is retrieved (as in Grand et al., 2022). In the other, the semantic identity of representative items is retrieved. In developing our novel semantic evaluation model, we additionally assessed two cognitive models of decision making. The first model used a logistic function to simulate the likelihood of each choice decision. The second model incorporated linear ballistic accumulators (Brown & Heathcote, 2008) to simulate both responses and response times as a race between the accumulators representing the two extremes of the decision axis.

Our findings demonstrate that the item-composite semantic evaluation model provides a better account of human classification responses and response times on a binary semantic classification task, relative to two other models. Importantly, the improved performance of the

item-composite model was not driven by the words that were used to construct the semantic classification model, as the results on a restricted dataset are consistent with the results of the full dataset for each of the three analyses. Additionally, the improved performance of the item-composite model cannot be explained by the fact that more items are used to construct the composite. Using a permutation analysis, we matched the number of words used to construct the difference axes for the adjective-composite model and the item-composite model. This suggests it is the quality of the words used to construct the difference axis, not the quantity, that drove the observed pattern of results. Finally, in a set of supplemental analyses, we reproduced these findings using an independent dataset, which further highlights the generalizability of our model. (For more information on the independent dataset, see Supplemental Materials). Together, these findings suggest that when performing a binary semantic classification task, participants retrieve and blend together items that are representative of the labels defining the two extremes on a given semantic dimension to make a decision.

While the primary goal of the study was to evaluate the different semantic judgment algorithms, we also compared our two distributional semantic models with one another. These comparisons suggested that the two DSMs performed similarly well and were not conclusive regarding the two DSMs relative cognitive utility. While GloVe outperformed word2vec using the logistic model, word2vec outperformed GloVe using the LBA model. It is not clear what aspects of the DSMs are responsible for these discrepancies. Word2vec is a predictive model with hidden layers that learn representations of words through prediction and self-correction, and GloVe is a latent semantic abstraction model which lacks this predictive component. How these architectural differences contribute to the differences in goodness of fit remains an open question. However, both GloVe and word2vec produced generally similar performance, and thus

both offer utility in modeling human semantic decision-making in future studies. This conclusion is consistent with previously reported findings. For example, Pereira et al. (2016) found that word2vec and GloVe produced comparable results in a large study comparing various distributional semantic models on word association, synonyms and analogy problems, and similarity and relatedness judgments.

Our assumption that semantic reasoning is based on an on-the-fly retrieval of individual word instances is broadly consistent with a variety of findings from the study of real-time lexical processing, which show that word meanings are flexible in context, drawing on multiple possible meanings in a context-dependent manner (Eberhard et al., 1995; Metzing & Brennan, 2003). For example, interpretation of referential expressions like "the girl" and "the peanut" is shaped by the properties of the overall discourse they are embedded in, including the referents and their properties. For example, in a context that illustrates the animacy of a cartoon peanut, a sentence like "*The peanut was in love*" is easily processed, but a locally coherent sentence like "*The peanut was salted*" results in confusion (Nieuwland & Van Berkum, 2006). Likewise, an instruction like "*Put the cube inside the can*" given a context with two differently sized cans causes momentary confusion if the cube is small enough to fit in either can, whereas this confusion is lifted if the cube is larger and only will fit in the larger of the two cans (Chambers et al., 2002; also see Chambers et al., 2004; Nieuwland et al., 2007). In addition to clear relevance to the language processing literature, the item-composite model can be conceptualized as an instance-based or an exemplar model consistent with dominant views of categorization (e.g., colors) and recognition (e.g., faces) (Nosofsky & Palmeri, 2015) (for an extended discussion of these connections, see Jamieson et al., 2018).

The extra step of comparing the target with other representative items also offers a potential mechanism for explaining how a relevant comparison class shapes semantic judgment. While we do not address this question in the present work, the set of extreme exemplars that are retrieved may itself be a contextually dependent process; if so, this may explain some of the contextual dependency in how certain linguistic expressions are *interpreted* in rich contexts. In our study, when making a size judgment, participants had a reference point as they were asked to judge a size of an object compared to a shoebox. However, the item-composite model allows for the reference point to shift in different contexts by selecting the representative examples for each semantic category. Consider that when judging the size of an amoeba, the set of extreme exemplars that are retrieved (e.g., virus, cell → cat, dog), should be different than exemplars retrieved when judging the size of Texas (e.g., Seattle, Denali → Spain, Africa). Indeed, it is well-established in the referential processing literature that the real-time interpretation of phrases like *the small glass* is driven by the relevant comparison set in the immediate context (Sedivy et al., 1999; Sedivy, 2003): The adjective "small" evokes a 4 cm tall glass when the context contains a 4 cm and an 8 cm glass, but "small" evokes the 8 cm glass when it is paired with a 12 cm one. Further, these comparison classes are created on the fly, based on multiple cues in the local context. In a context where a listener views 3 drinking glasses (4 cm, 8 cm, 12 cm tall), and the speaker says "*Pick up the small glass*", this sentence is typically interpreted as referring to the smallest glass that the *speaker* can see: if the 4 cm glass is obscured from the speaker's view, the listener interprets "the small glass" to be the 8 cm tall one, rather than the "small" 4 cm glass that the speaker cannot see (Heller et al., 2008; Heller et al., 2016; Ryskin et al., 2015). If the retrieved set of extreme exemplars was shaped by properties of the local context when making a semantic judgment about a target, this could be a way to capture findings like these.

The idea that semantic classification requires an extra step of comparing the target with other items is also supported by a variety of findings from studies of language *production*. For example, adjectives like "small" and "large" tend not to be produced by speakers unless the immediate context contains items that contrast along the size dimension and the speaker has noticed them (Brown-Schmidt & Tanenhaus, 2006; Brown-Schmidt & Konopka 2008; Pechmann, 1989). For example, when naming a butterfly, if the speaker fails to notice a larger one in the scene, they are likely to simply say "butterfly", and if they do notice the larger butterfly, the timing of when the adjective is produced is strongly predicted by the latency of the eye-fixation to the size-contrasting item, with early looks producing prenominal modifiers (e.g., "the small butterfly"), and later looks producing late modifiers (e.g., "the butterfly, uh small one"), unless the speaker is using a language that affords postnominal modification (e.g. "la mariposa pequeña", Brown-Schmidt & Konopka, 2008).

Our study raises a number of questions for future work. The item-composite model utilizes representations constructed from large linguistic corpora. However, these semantic vectors do not have easily interpretable semantic dimensions, which makes it unclear how the relevant words used to construct the axes are retrieved from memory. One possibility is that some perceptual features of concepts can be recovered through the co-occurrence statistics alone. Previous research has shown that individuals who lack certain sensory experiences – for example, congenitally blind individuals – possess detailed semantic knowledge about perceptual features of various objects. For example, van Paridon et al. (2021) demonstrated that congenitally blind people, despite the lack of visual perceptual experience, formed associations between colors and adjectives (e.g., blue is cold, red is hot) that were similar to the intuitions of sighted people. Similarly, Kim and colleagues (2019) compared blind and sighted people's

knowledge of the appearance of common animals. The authors found that individuals who were blind inferred features of animal appearance from taxonomy and habitat properties (e.g., because sharks live in the water, they must have scaley skin like other fish). These results indicate that knowledge of animal appearance (even if incorrect) can be acquired through inference from language, rather than through memorization of facts directly specifying those properties. Together these findings indicate that in the absence of direct perceptual experience, distributional information obtained from linguistic input can serve as a source of semantic knowledge. An open question is how and when certain concepts might be tagged for various semantic features.

Finally, our study does not answer the question of whether the difference vectors used for the semantic projection are a part of the individual's representational knowledge or if they get constructed on-the-fly to meet specific task demands. According to the instance-based theory (Jamieson et al., 2018), a representation of word meaning can be constructed on-the-fly in a highly parallel, probe-driven retrieval process. Following Jamieson et al. (2018), we speculate that composite representations used in our models might be constructed during performance and not necessarily constitute a part of the participant's core semantic representation. This account would also explain the flexibility of human semantic knowledge. Previous studies indicate that humans are capable of rapidly and flexibly reconfiguring their semantic knowledge to meet various task demands. A good example of such conceptual flexibility is ad-hoc categories (Barsalou, 1983), such as "things to sell at a garage sale" or "things that can fall on one's head". While these features are unlikely to be part of a person's core semantic knowledge, participants can nevertheless perform such a classification rapidly, suggesting that they can quickly construct a representation of a category they have never encountered before.

CHAPTER III

Study 2. The Role of Linguistic Reference in Shaping Mental Representations of Objects

Study 1 examined the type of conceptual information encoded through linguistic labels and the mechanisms by which humans can access and manipulate it. **Study 2** probed how linguistic reference can shape mental representations of objects and subsequent memories for them.

Study 2A. The Effect of Linguistic Form on Memory for Objects and Object Pairings

Previous research exploring the relationship between language and object representation indicates that producing or processing linguistic reference at encoding benefits subsequent memory for objects (Dessalegn & Landau, 2008; Scott & Sera, 2016; Zormpa et al., 2018; Lord & Brown-Schmidt, 2022) while restricting access to language via verbal shadowing results in mnemonic deficits for object-scene conjunctions (Urgolites et al., 2020). Further, Lord and Brown-Schmidt (2022) demonstrated that the linguistic form of the referential expression matters in what is encoded in memory at the time that language is processed and proposed an activation-time account of the observed effect. An alternative hypothesis exploring the effect of different types of linguistic reference on memory for objects is the idea of a "hybrid representation" which is adopted within developmental research (Dessalegn & Landau, 2008; Scott & Sera, 2016). A notable shortcoming of this proposal is the lack of a mechanistic explanation, namely, what constitutes a "hybrid representation", how it is built, or why it would benefit memory for objects.

Simultaneously, linguistic work demonstrates that certain types of language, specifically, the demonstrative pronoun "that" evokes a composite interpretation (i.e., referring to conceptually complex entities) during language processing (Brown-Schmidt et al., 2005), reminiscent of the concept of a "hybrid representation". The goal of **Study 2A** is to elucidate the effect of the demonstrative pronoun "that" on the representation and subsequent binding memory for object pairs as well as to characterize the cognitive mechanisms underlying any potential mnemonic benefits associated with such language.

Methods

Participants

Following the method of Urgolites et al. (2020), we recruited 120 people on Prolific. The experiment took on average 30 minutes to complete. Participants were compensated $0.70 for each 5 minutes of their time. One person was excluded because their data was not recorded properly. Consistent with the pre-registered exclusion criteria, we excluded one additional participant because they provided response "new" on more than 98% of trials on the individual memory test, indicating either poor memory or failure to understand/perform the task, leaving 118 people in the final analysis. All participants reported themselves as native or fluent speakers of English.

Procedure

The experiment included two phases. In Phase 1, participants completed 96 trials on which they saw 9 objects on the screen in a 3 x 3 layout (**Figure 11**). Importantly, participants

simultaneously heard an audio description of two objects on the screen. First, participants were instructed to place one object on top of the other (e.g., *put the shirt on the sheep*) because "on top" has previously been found to create a more intuitive composite (versus other expressions e.g., *next to*, see Brown-Schmidt et al., 2005). Then, participants were instructed to place the two objects in the box. Importantly, in the critical "that" condition, the two objects were referred to using the demonstrative pronoun "that" (e.g., *now put <u>that</u> in the box*). In the first comparison condition, the two objects were referred to using a conjoined NP with both referents (e.g., *now <u>put the shirt and the sheep</u> in the box*). Finally, in the second comparison condition, the two objects were referred to using the pronoun "them" (e.g., *now put <u>them</u> in the box*). Both comparison conditions utilized the language that highlights the individuality of the two objects – in contrast to the critical "that" reference, with the "them" condition additionally controlling for the length of the referential expression. In Phase 2, participants completed two blocked recognition memory tests, on which half of the studied items were tested for binding memory, and the other half – for individual memory. For the binding memory test, participants saw 24 old pairs of objects (manipulated together at study) and 24 new pairs (had appeared at study but not on the same trial) presented horizontally side-by-side. Participants were instructed to click "Together" if they thought the object pair went together during the study phase, and "Not together" if they didn't. Finally, for the individual memory test, participants saw 48 old and 48 new individual objects. Participants were instructed to click "old" if they recognized the object from the study phase, and "new" if not. The order of the test presentation was counterbalanced across participants. The order of the trial presentation was randomized.

Study phase

"Put the shirt on the sheep. Now put that / them / the shirt and the sheep in the box"











OLD **NEW**

Binding memory test

Did this pair go together?


TOGETHER NOT TOGETHER

Did this pair go together?


TOGETHER NOT TOGETHER

Individual memory test

Is this object old or new?

OLD NEW

Is this object old or new?

OLD NEW

Figure 8. Schematic of the experimental procedure. During the study phase, participants heard an audio recording instructing them to first, place one object on top of the other, and then, to place them both in the box. In the critical language condition, the two objects were referred to using the demonstrative pronoun "that". In two control conditions, the audio recording referenced the objects either with the pronoun "them" or a conjoined noun phrase with both referents.

Materials

The materials included 912 photos of real-world objects in a white background (**Figure 11**). The images were randomly selected from the Dinolab Database Website of 1000 unique objects (link: https://mariamh.shinyapps.io/dinolabobjects/). We created three experimental lists that counterbalanced the images across three experimental conditions using a modified Latin square design. Each list contained 192 critical objects at study, 64 in each of the three conditions, and 672 foils. An additional 48 new objects were drawn from the same database as the old objects and were presented at the individual memory test. Because of the large number of items, new items were not counterbalanced with all 48 new items occurring in each of the three experimental lists. Each participant was randomly assigned to a single experimental list. The auditory stimuli were recorded by a female research assistant with a North American accent of English at a natural speaking rate.

Predictions

If the language that is typically used to refer to conceptual composites (i.e., pronoun *that*; Brown-Schmidt et al., 2005) binds the referenced items into a composite, we predict that for the

binding memory test but not for the individual memory test, pronoun "that" will result in statistically better memory than both the pronoun "them" and a conjoined noun phrase (NP) with both referents (e.g., *the shirt and the sheep*). If, on the other hand, linguistic forms like "that" invoke a conceptual composite in the moment as language is being interpreted, without affecting how this information is encoded in memory, we predict that binding memory will be similar across the three language conditions i.e., the pronoun "that", pronoun "them", and a conjoined NP.

Results

We used a signal-detection theoretic mixed-effects analysis (Wright, Horry, & Skagerberg, 2009) for the response data. We fit two logistic mixed effect regression models to the participants' old-new responses; one on the individual item memory test, and the other -- on the binding memory test. The individual memory model included item status (whether the item was actually old vs. new) as a factor, and then for old items, the condition as a predictor. These fixed effects were coded using weighted Helmert contrasts (see Table 1) to test following comparisons:

Old vs. new – actually old probes vs. new probes

Noun vs. pronoun – a conjoined NP with both referents vs. pronouns

That vs. them – pronoun “that” vs. pronoun “them”

The binding memory model included item status (whether the pairing was old or new) as a factor, and then for old items, the condition as a predictor. Similar to the coding used for the individual memory test, these fixed effects were coded using weighted Helmert contrasts (see

above and Table 2). The models included random intercepts by participant and item. Random by-participant and by-item slopes were included if determined so by the model specification package in R -- buildmer (Voeten, 2020), which performs a stepwise elimination based on the change in model's log-likelihood until an optimal model is found and converges successfully.

Individual memory

The results of the individual memory test are reported in **Table 1**. A negative intercept was due to participants' significant bias to say "new" ($b = -0.50$, $z = -5.23$, $p < 10^{-6}$). A significant effect of item type (actually old vs. new) indicated successful recognition of the old images ($b = 3.61$, $z = 20.89$, $p < 10^{-15}$). Non-significant effects of the noun phrase ($b = 0.08$, $z = 1.20$, $p = 0.23$) and pronoun type ($b = -0.02$, $z = -0.39$, $p = 0.70$) indicated that different ways of referring to objects at study did not significantly modulate individual memory for them (**Figure 12**).

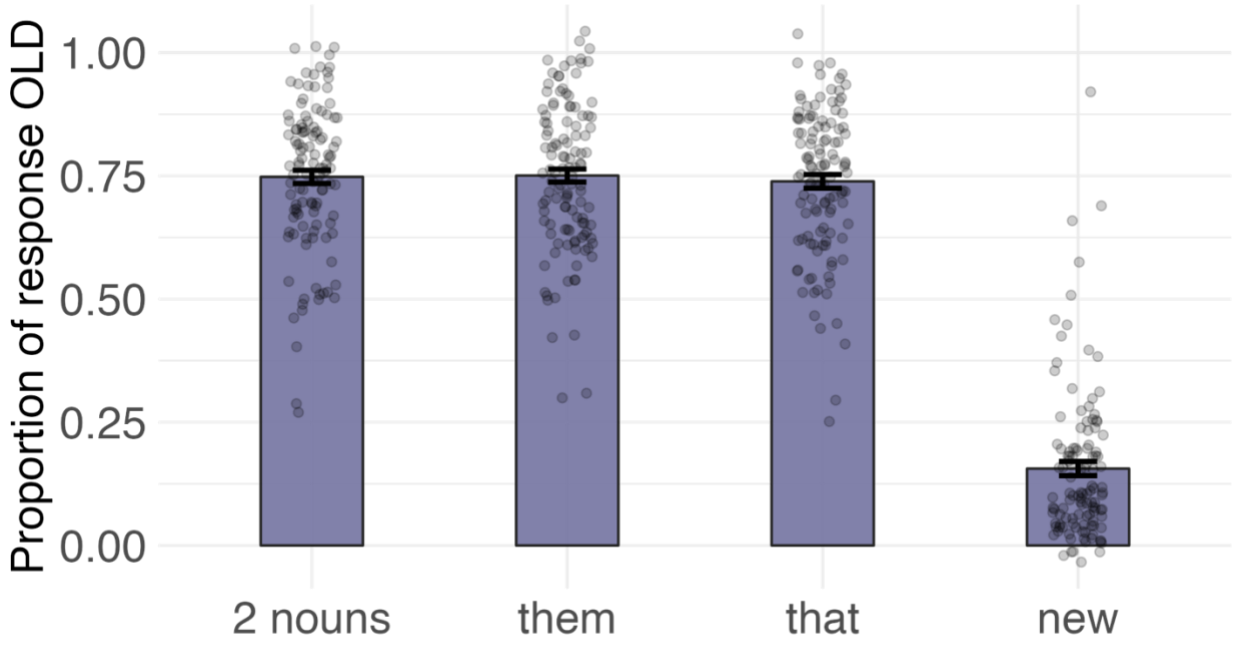


Figure 9. Individual memory results. Error bars represent by-subject SEM. Data points represent mean accuracies for each participant. Means per condition are 75%, 75%, 74%, 16% for the conjoined noun phrase, "them", "that" conditions, and the new items respectively.

Table 1. Individual memory results: Mixed effect model with a binary dependent measure - whether the image was recognized or not on the recognition memory test. Values in bold indicate significant results at an alpha level of .05.

response (0/1) ~ 1 + new.vs.old + that.vs.them + noun.vs.pronoun +

(1 + new.vs.old + that.vs.them | participant ID) + (1 + noun.vs.pronoun | item ID)

Fixed effects	***Estimate***	***SE***	***z-value***	***p-value***
Intercept	-0.50	0.10	-5.23	$< 10^{-6}$
new.vs.old (new = 0.5 each, old = -0.5)	3.61	0.17	20.89	$< 10^{-15}$
that.vs.them **(them = 0.5, that = -0.5, noun, new = 0 each)**	0.08	0.06	1.20	0.23
noun.vs.pronoun (nouns = -0.58333333, pronoun = 0.41666667 each,	-0.02	0.06	-0.39	0.70

new = -0.08333333)

Random effects	*Variance*	*SD*	*Correlations*	
Item ID	0.52	0.72		
noun.vs.pronoun	0.04	0.19	-0.53	
Participant ID	0.71	0.84		
new.vs.old	1.91	1.38	-0.54	
that.vs.them	0.07	0.26	-0.68	0.67

Number of observations: 22848; groups: item ID: 192, participant ID: 118

Binding memory

The results of the binding memory test are reported in **Table 2**. A non-significant negative intercept was due to similar rates of "together" and "not together" responses (b = -0.01, z = -0.03, p = 0.98). A significant effect of item type (actually together vs. not together) indicated successful recognition of the old image pairs (b = 0.92, z = 7.08, $p < 10^{-11}$). A significant effect of the noun phrase (b = -0.18, z = -2.09, p = 0.04) indicated that referring to objects using a conjoined noun phrase resulted in better binding memory compared to using either pronoun "that" or "them". Finally, there was no significant difference in binding memory between the two pronoun types (b = -0.04, z = -0.36, p = 0.72) (**Figure 13**).

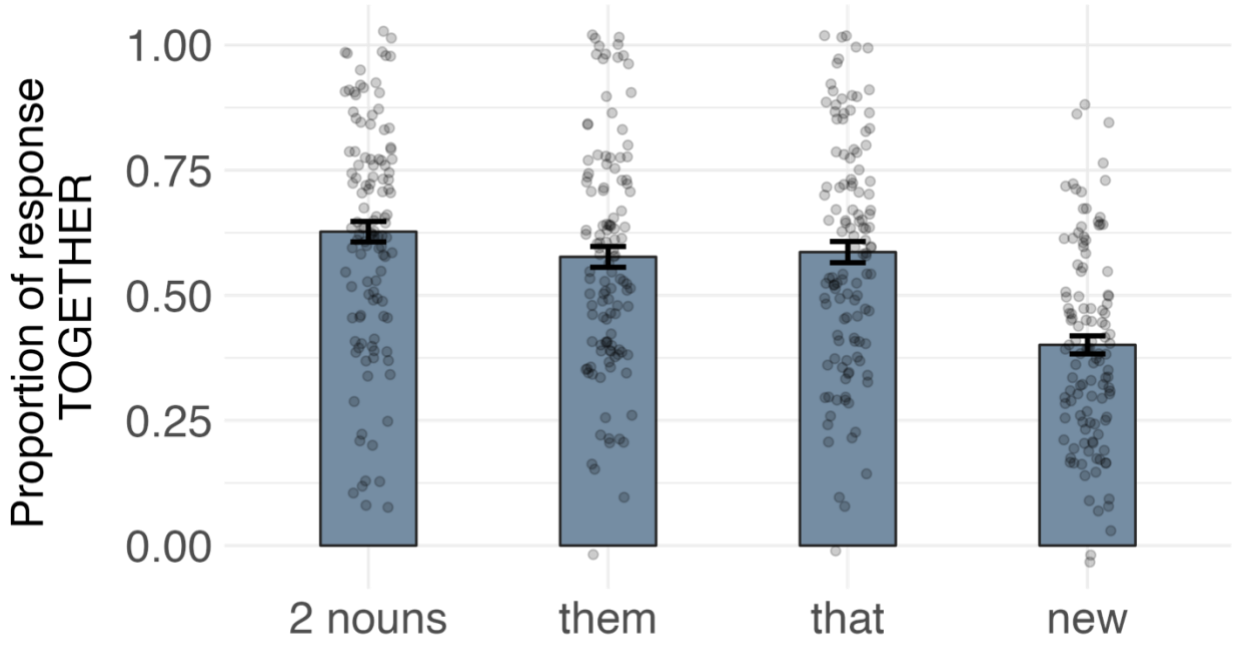


Figure 10. Binding memory results. Error bars represent by-subject SEM. Data points represent mean accuracies for each participant. Means per condition are 63%, 58%, 59%, 40% for the conjoined noun phrase, "them", "that" conditions, and the new items respectively.

Table 2. Binding memory results: Mixed effect model with a binary dependent measure - whether the image was recognized or not on the recognition memory test. Values in bold indicate significant results at an alpha level of .05.

response (0/1) ~ 1 + new.vs.old + noun.vs.pronoun + that.vs.them +

(1 + new.vs.old + noun.vs.pronoun + that.vs.them | participant ID) +

(1 + noun.vs.pronoun | item ID)

Fixed effects	***Estimate***	***SE***	***z-value***	***p-value***
Intercept	-0.01	0.09	-0.12	0.90
new.vs.old (new = 0.5 each, old = -0.5)	0.96	0.16	6.16	$< 10^{-9}$
noun.vs.pronoun (nouns = -0.58333333, pronoun = 0.41666667 each, new = -0.08333333)	-0.20	0.10	-2.02	0.04

that.vs.them (them = 0.5, that = -0.5, noun, new = 0 each)	-0.04	0.10	-0.41	0.69

Random effects	***Variance***	***SD***	***Correlations***		
Participant ID	0.47	0.69			
new.vs.old	0.61	0.78	-0.19		
noun.vs.pronoun	0.04	0.19	-0.25	0.02	
that.vs.them	0.02	0.12	-0.54	0.27	-0.67
Item ID	0.25	0.50			
noun.vs.pronoun	0.04	0.19	-0.87		

Number of observations: 5712, groups: item ID: 85, participant ID: 118

Discussion

The current experiment explored the effect of referential expression on individual and binding memory for distinct objects. More specifically, we tested if the language that is typically used to refer to conceptual composites (i.e., pronoun *that*; Brown-Schmidt et al., 2005) can alter the representation of two objects by binding them into a conceptual composite and subsequently, improving binding memory for them. The results indicated that while participants could successfully recognize old images, the individual memory for objects was not differentially modulated by the three types of linguistic reference. Critically, binding memory for objects was better when the items were labeled using a conjoined noun phrase (i.e., *the shirt and the sheep*) compared to either pronoun "them" or "that". One possibility is that repetition (i.e., hearing the linguistic labels twice) could drive the observed mnemonic benefit. However, it is not clear why repetition would aid binding but not individual memory making it an unlikely explanation of the observed pattern of results. Another possible explanation is that the binding memory test was more difficult than the individual memory test, so the repetition of the referents was particularly helpful on a more challenging task. Even though we could not statistically compare the

performance on the individual and binding memory tests, the fact that the mean endorsement rates for actually old items were numerically higher for the individual memory test (74.6%) compared to the binding memory test (60%) generally supports this idea. Together, our findings do not support the hypothesis that the demonstrative pronoun “that” invokes a conceptual composite in the moment during language processing and by doing so, changes the representation of the objects it refers to.

Study 2B. The Effect of Linguistic Form on Memory for Distinct Object Features

Another type of language that is routinely used to describe objects is modified noun phrases (e.g., *enormous balloon*). Language processing literature shows that noun phrases that are semantically rich (i.e., modified) confer processing benefits and subsequently, are more accessible in memory than unmodified noun phrases (Hofmeister, 2011; Karimi & Ferreira, 2016; Troyer et al., 2016). Further, stimuli that contain more information such as modifiers are associated with better memory possibly due to the longer encoding times, the distinctiveness of the encoded representation, more effortful/elaborate encoding, or the existence of multiple retrieval cues (Craik & Tulving, 1975; Fisher & Craik, 1980; Bradshaw & Anderson, 1982; O'Brien & Myers, 1985; Marks, 1987; Gallo et al., 2008). Consistent with this idea, Yoon and colleagues (2016; 2021) found that using modification improves memory for objects compared to referring to them using nouns only. In particular, the authors recruited pairs of participants to complete a referential communication task, during which participants described objects on the screen, and later, completed a surprise memory test. Importantly, the critical manipulation was designed to elicit participants' use of adjective modifiers when describing the objects on the screen. The results indicated that when participants used an adjective to describe the target image (e.g., *dotted sock*) their memory for that item was better compared to when described using only a noun (e.g., *sock*) or using a locative phrase (e.g., *the top left one*). Yoon et al. (2016) hypothesized that more elaborate encoding of the objects associated with the use of adjective modifiers might facilitate object recognition. While the mnemonic benefit associated with the use of adjective modifiers is well-established, if and how modification shapes various *aspects* of object representation remains poorly characterized in the literature.

This question is particularly relevant since previous research indicated that different object features are represented and forgotten independently from each other (Brady et al., 2013). In their study, Brady and colleagues (2013) instructed participants to study a list of items that varied on multiple dimensions, such as object's state (e.g., *full glass / half-empty glass*) and exemplar (e.g., *beer glass / brandy glass*) (**Figure 14**). Later, they tested participants' recognition memory in a short-delay and long-delay task in a 4-AFC paradigm. The authors found that the memory for object states and object exemplars decayed at different rates suggesting that various visual features of objects are not represented as bound units and are stored and forgotten independently from each other. If so, the independent representations of object features might be modulated independently by certain linguistic forms like modifiers which can be used to highlight individual features of an object.

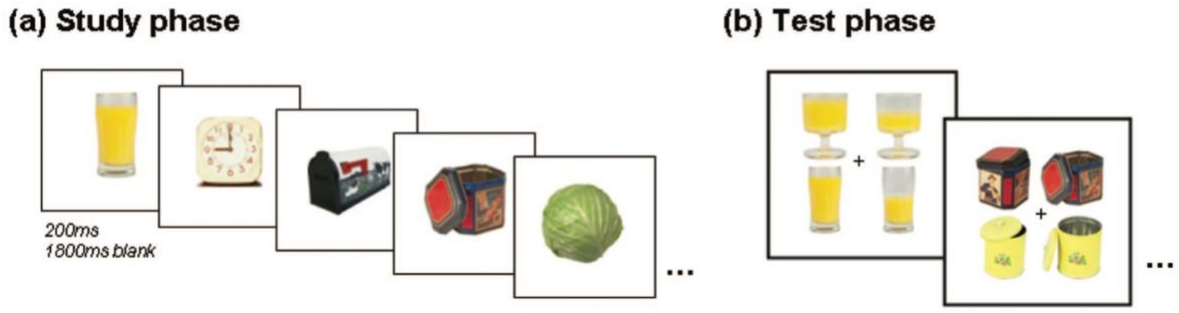


Figure 11. From Brady et al. (2013). Schematic of the experimental procedure. During the study phase, participants made size judgments for items presented on the screen one at a time. At test, participants had to identify an object that appeared during the study phase in a 4-alternative choice paradigm. Note that participants completed the test phase in the short-delay conditions (immediately after the study phase) or in the long-delay condition (three days after the study phase).

In the current study, we probe how adjective modifiers affect the representations of independent object features. We entertain three alternative hypotheses with respect to the underlying cognitive mechanisms at play. According to one hypothesis, lexicalizing one feature of an object (e.g., sleeve length, *sleeveless shirt*) may highlight it, strengthening its representation, and consequently, making it more memorable. In return, the representation of the other, unlexicalized feature (e.g., overall length, *cropped top*), becomes weaker and consequently, less memorable. According to another hypothesis, a more memorable lexicalized feature of an object might serve as a retrieval cue for the unlexicalized feature, boosting subsequent memory for it. Finally, it is possible that lexicalizing object features does not impact how these features are represented in the mind.

One reason to believe that the use of adjective modifiers might reduce memory for the unmodified feature is findings by Lupyan (2008). Across six experiments, Lupyan (2008) probed if category label generation augments the representation of everyday objects. To investigate this question, Lupyan (2008) presented participants with images of chairs and lamps (Experiment 1, 3-6) or chairs and tables (Experiment 2) and asked them to make preference judgments (e.g., like or dislike) or classification judgments (e.g., *Is it a chair or a lamp/table?)*. Later, participants' recognition memory was tested. In an old/new paradigm, the foil images of chairs and tables/lamps were selected from the same furniture catalog as the targets and differed from the target images on shape and/or color. The critical findings indicated *worse* recognition memory for the items that were classified compared to the items for which participants indicated their preference. The proposed mechanism is a representational shift account, according to which, overtly classifying objects using linguistic labels co-activates top-down category features

simultaneously with the bottom-up features of an exemplar. Consequently, when presented with a probe, participants retrieve a representation that is a combination of the previously observed features and the top-down category features, ultimately, decreasing the familiarity of the old probe and resulting in lower recognition performance for the classified items. Following this logic, if labeling an object as "sleeveless shirt" results in a representational shift such that the feature "sleeveless" becomes more salient, we might observe worse recognition of the unnamed feature (i.e., overall dress length).

Methods

Participants

We recruited 120 participants on Prolific. The experiment took on average 20 minutes to complete. Participants were compensated $0.70 for each 5 minutes of their time. Four people were excluded because their data was not recorded properly. Consistent with the pre-registered exclusion criteria, two additional people were excluded because they provided response "new" on more than 95% of test trials, indicating either poor memory or failure to understand/perform the task, leaving 114 people in the final analysis. All participants reported themselves as native or fluent speakers of English.

Procedure

The experiment included two phases (**Figure 15**). In the first phase, participants studied 104 pictures of various pieces of clothing while listening to the audio descriptions of them. An additional 104 trials were included as foils, on which participants had to click on animals, half of

which were also described using one noun (e.g., *click on the giraffe*) and the other half -- a modifier and a noun (e.g., *click on the left-facing giraffe*). Immediately after completing Phase 1, participants started Phase 2, which was a surprise recognition memory test and during which they viewed 208 (104 old + 104 new) images of clothes presented in a random order one probe at a time. Before starting the second phase, participants were instructed to look for subtle differences between the clothing items and were given an example (**Figure 16**) which was not a part of the experimental stimuli.

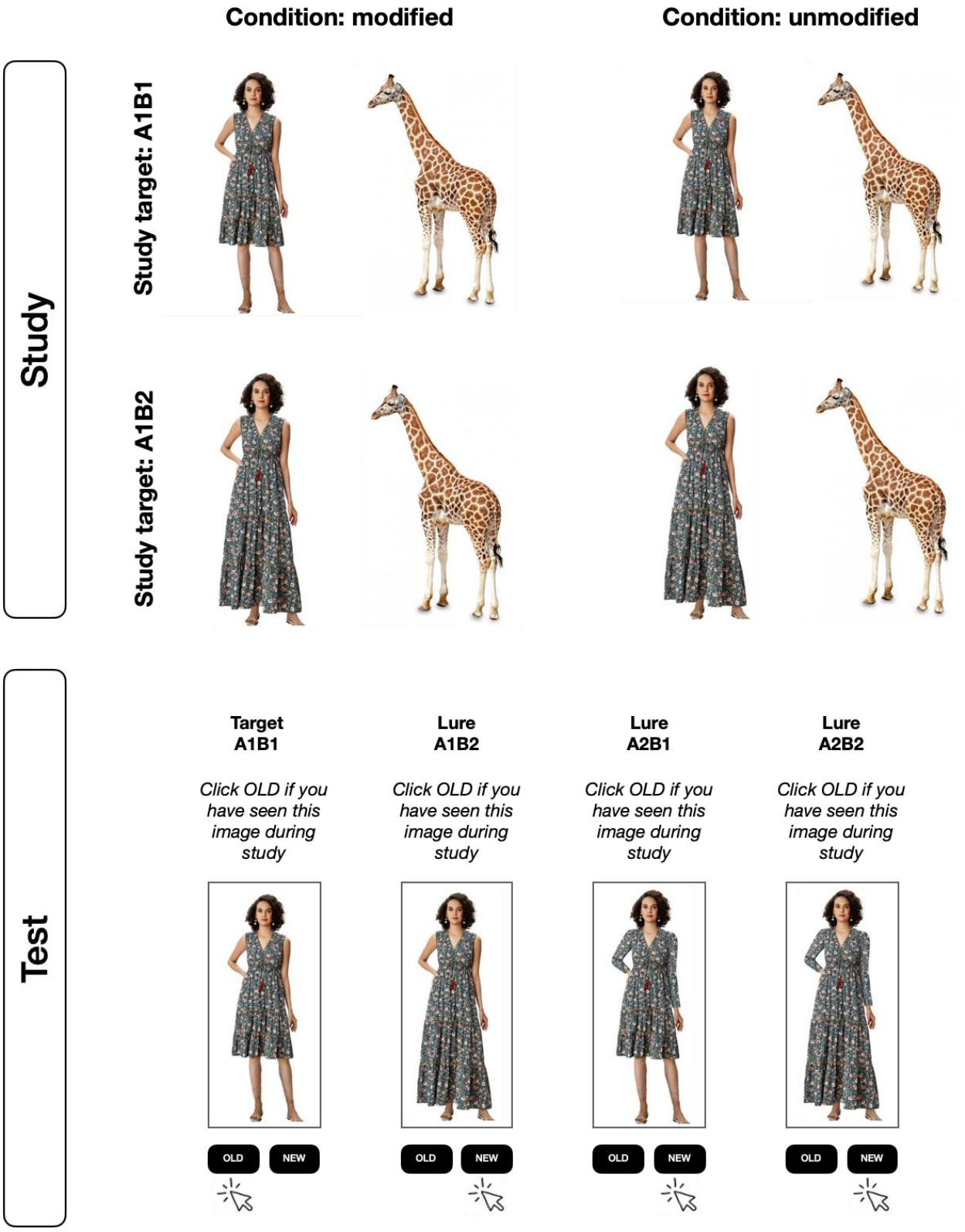


Figure 12. Schematic of the experimental procedure. During study, participants were instructed

to click on the target image (a1b1/a1b2) which was presented in either a modified (e.g., *click on the sleeveless dress*) or unmodified condition (e.g., *click on the dress*). During test, participants completed a recognition memory test during which they saw one image that either matched the test image or mismatched it on one or both features (a1b1/ a1b2/ a2b1/ a2b2).

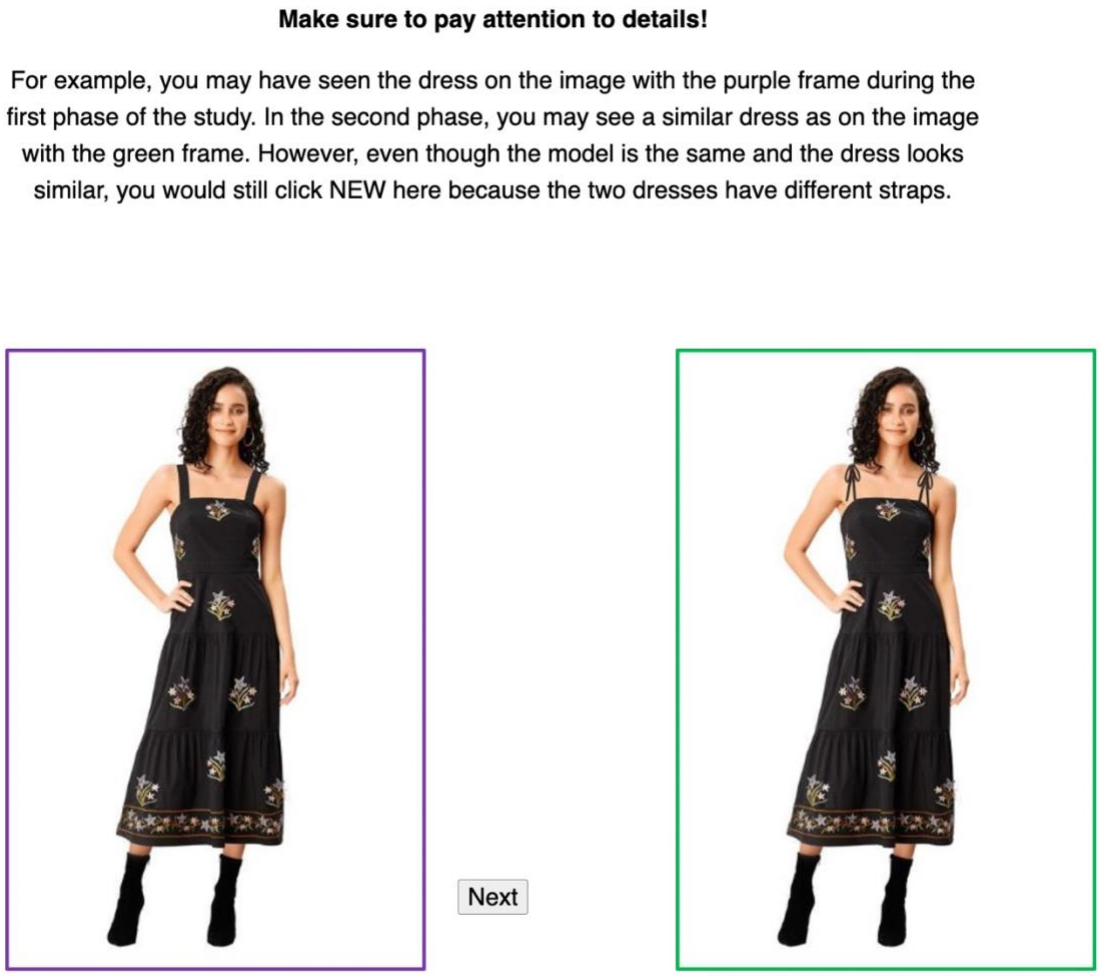


Figure 13. Instructions presented to participations prior to beginning Phase 2 of the experiment -- recognition memory test. Participants were instructed to pay attention to details and were provided with an example of two similar but not identical images.

Materials

The materials included 104 unique images of clothing items. The images were selected from the online customizable clothing shop (link: https://www.eshakti.com/). The final set of items included 7 button-ups, 10 cardigans, 10 dresses, 5 dusters, 7 jackets, 10 jumpers, 5 shirts, 10 sweaters, 20 tops, and 20 tunics. Each item varied on two features, A and B, (e.g., sleeve length and overall length), resulting in 4 versions for each item: a1b1 (e.g., sleeveless knee-long dress), a1b2 (e.g., sleeveless ankle-long dress), a2b1 (e.g., knee-long dress with long sleeves), or a2b2 (e.g., ankle-long dress with long sleeves), and 416 images total (**Figure 15**). For each unique item, we manipulated: a) which version of the target image was presented as study (a1b1/a1b2), b) if it was modified or unmodified, and c) which version of the target image was presented at test (a1b1/a2b1/a1b2/a2b2). If the item was presented in the unmodified (noun) condition, the image was described using only a noun (e.g., *click on the dress*), while in the modified condition, the image was described using an adjective and a noun (e.g., *click on the sleeveless dress*). Because of the large number of items and multiple manipulations, we did not manipulate which feature was named (i.e., A or B). In other words, for the items that were presented in the modifier condition, only Feature A was named. Which feature was considered A or B was selected at random ensuring that across all items, various clothing features were named roughly equally. At test, all 104 unique items were presented in one of the 4 test conditions (a1b1/a2b1/a1b2/a2b2). Only 25% of the test items were truly "old" (i.e., matched on both features) while the rest were "new". The 416 items were distributed across 16 experimental lists. Each list contained 104 unique items at study, 52 in the noun condition and another 52 in the modifier condition. All items were counterbalanced across the conditions following a modified Latin square design such that each participant only saw a unique item once. The auditory stimuli

were recorded by a female research assistant with a North American accent of English at a natural speaking rate.

Results

Model 1

We fit a logistic mixed effects model with participants' response on the recognition memory test as a binary dependent variable (0 - NEW/1 - OLD) and modifier condition (modified/unmodified), Feature A (same vs. different), Feature B (same vs. different), and their interactions as predictors. All predictors were effects coded. The model included random intercepts by-participant and item. Random by-participant and by-item slopes were included if determined so by the model specification package in R -- buildmer (Voeten, 2020), which performs a stepwise elimination based on the change in model's log-likelihood until an optimal model is found and converges successfully.

The results of this analysis are reported in **Table 3.** A negative intercept was due to participants' significant bias to say "new" ($b = -0.41$, $z = -4.9$, $p = < 10^{-6}$). A significant effect of the modifier condition ($b = 0.16$, $z = 2.04$, $p = 0.04$) indicated that overall, when the actual target for an item set had been modified, the associated image at test (regardless of whether it was old or new) was more likely to generate a response "old". A significant Feature A effect ($b = 0.19$, $z = 3.31$, $p = 0.0009$) indicated that participants were more likely to say "old" when Feature A at test matched Feature A at study. Similarly, a significant Feature B effect ($b = 0.11$, $z = 2.01$, $p = 0.04$) indicated that participants were more likely to say "old" when Feature B at test matched Feature B at study even though Feature B was never named. All interactions were non-

significant indicating that naming Feature A did not boost memory for it significantly better than for Feature B. Additionally, the lack of the Feature A * Feature B interaction indicated that memory for these properties was independent of each other (**Figure 17**).

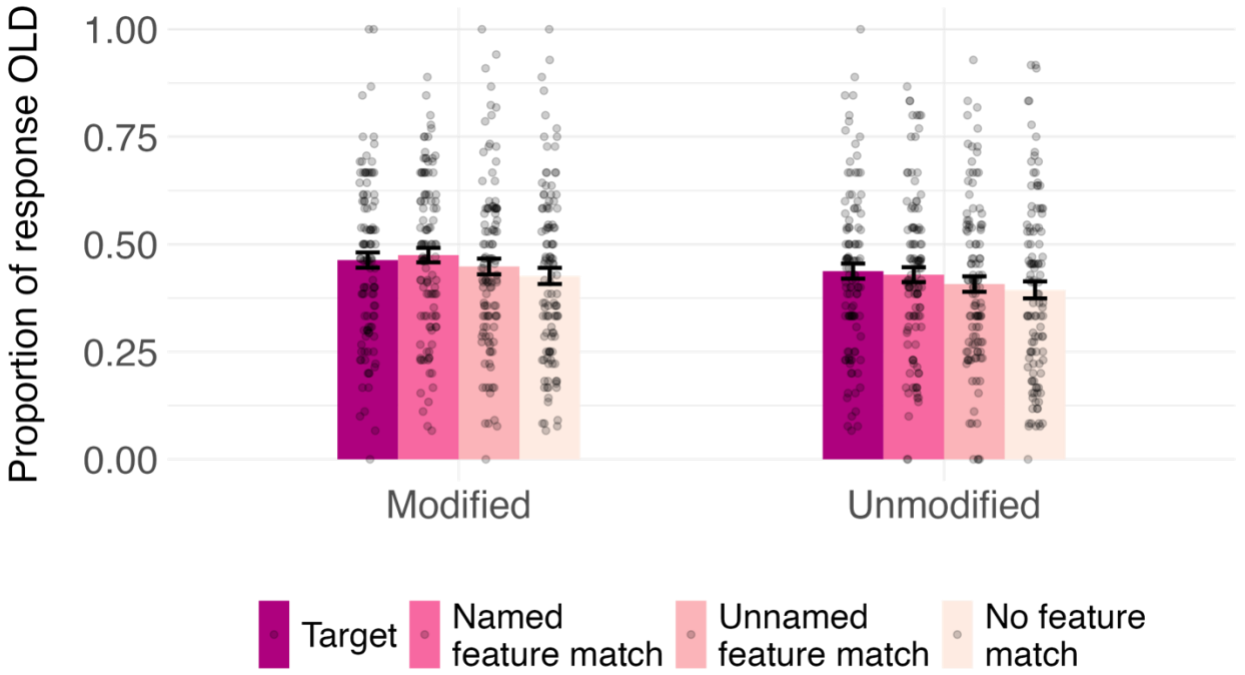


Figure 14. Recognition memory results. Error bars represent by-participant SEMs. Data points represent mean proportions for each participant.

Table 3. Memory results for Model 1: Mixed effect model with a binary dependent measure - whether the image was recognized or not on the recognition memory test. Values in bold indicate significant results at an alpha level of .05.

*response (0/1) ~ 1 + Condition * Feature A match *Feature B match +*

(1 | participant ID) + (1 | item ID)

Fixed effects	***Estimate***	***SE***	***z-value***	***p-value***
Intercept	**-0.41**	**0.08**	**-4.91**	**$<10^{-6}$**

Condition (noun = -0.5, modifier = 0.5)	**0.16**	**0.08**	**2.04**	**0.04**
Feature A match (Feature A match = 0.5, Feature B match = -0.5)	**0.19**	**0.06**	**3.31**	**< 0.05**
Feature B match (Feature B match = 0.5, Feature A match = -0.5)	**0.11**	**0.06**	**2.01**	**0.04**
Condition * Feature A match	0.03	0.11	0.26	0.79
Condition * Feature B match	0.00	0.11	-0.03	0.98
Feature A match * Feature B match	-0.10	0.08	-1.24	0.21
Condition * Feature A match * Feature B match	-0.04	0.16	-0.25	0.80

Random effects	***Variance***	***SD***
Participant ID	0.40	0.63
Item ID	0.20	0.45

Number of observations:11856, groups: item ID: 104, participant ID: 114

Model 2

Model 1 indicated that both match on Feature A and Feature B independently increased the likelihood of an "old" response. Although the model parameter estimates for Feature A and B were numerically different (b_A =0.19 vs. b_B = 0.11), Model 1 did not allow us to statistically compare them. As a result, Model 2 aimed to test the difference in recognition memory between the items that were matched on Feature A vs. Feature B. The model selection procedure was identical to that used for Model 1. Modified condition was effects coded (unmodified = -0.5, modified = 0.5). Critically, we used weighted Helmert contrast coding scheme for a predictor indicating a match/mismatch between study and test with the following comparisons:

1 – target vs. the rest (i.e., match on Feature A and B vs. the rest)

2 – match on one feature vs. mismatch on both

3 – match on Feature A vs. match on Feature B

The results for Model 2 are reported in **Table 4**. The findings from Model 2 replicated the findings from Model 1 such that participants demonstrated a bias to say "new", and the

modified imaged were more likely to generate an "old" response compared to unmodified targets. Importantly, participants were more likely to endorse the target vs. the rest of the images (Contrast 1: b = 0.15, z = 3.09, p = 0.00). Additionally, participants were more likely to say "old" if the image was matched on at least one feature vs. mismatched on both (Contrast 2: b = -0.13, z = -2.51, p = 0.01). Finally, there was no significant difference between the test images that were matched on Feature A or Feature B (Contrast 3: b = 0.07, z = 1.29, p = 0.20). Interactions of these Helmert contrasts with modifier condition were not significant.

Table 4. Memory results for Model 2: Mixed effect model with a binary dependent measure - whether the image was recognized or not on the recognition memory test. Values in bold indicate significant results at an alpha level of .05.

response (0/1) ~ 1 + Condition + Contrast 1 + Contrast 2 + Contrast 3 + Condition:Contrast 1 + Condition:Contrast 2 + Condition:Contrast 3 + (1 + Contrast 1 | participant ID) + (1 + Contrast 1 + Condition + Condition:Contrast 1 | item ID)

Fixed effects	***Estimate***	***SE***	***z-value***	***p-value***
Intercept	**-0.29**	**0.08**	**-3.74**	**0.00**
Condition **(noun = -0.5, modifier = 0.5)**	**0.17**	**0.04**	**3.88**	**0.00**
Contrast 1 **(target = -0.7490722, else = 0.2509278 each)**	**0.15**	**0.05**	**3.09**	**0.00**
Contrast 2 **(Feature A match = 0.4992830634, Feature B match = -0.500716936, else = -0.0007169366 each)**	**-0.13**	**0.05**	**-2.51**	**0.01**
Contrast 3 (A and B mismatch = -0.6247048, Feature A or B match = 0.3752952 each, target = -0.1247048)	0.07	0.06	1.28	0.20
Condition * Contrast 1	0.04	0.11	0.34	0.73
Condition * Contrast 2	0.02	0.10	0.18	0.86
Condition * Contrast 3	0.01	0.10	0.13	0.90

Random effects	***Variance***	***SD***	***Correlation***
Participant ID	0.40	0.64	
Contrast 1	0.04	0.21	0.09
Item ID	0.20	0.45	

Contrast 1	0.01	0.11	-0.06		
Condition	0.03	0.18	-0.15	-0.41	
Contrast 1 * Condition	0.18	0.42	-0.14	0.98	-0.56

Number of observations: 11856; groups: item ID: 104, participant ID: 114

Discussion

The current study examined the effect of adjective modifiers on memory for distinct object features. Overall, we found that the use of adjective modifiers increased endorsement rates for all items consistent with previous reports (Hofmeister, 2011; Troyer et al., 2016; Yoon et al., 2016; Yoon et al., 2021). Additionally, at test, participants were more likely to categorize an image as "old" if it matched the study image on both or at least one feature. Critically, we found that memory for two distinct object features was independent of one another (also Brady et al., 2013). Finally, the use of adjective modifiers did not enhance (or reduce) endorsement rates for the named feature above and beyond the unnamed feature. These preliminary findings are partially consistent with the hypothesis that lexicalizing a feature of an object does not impact the representation of the non-lexicalized feature. However, a single null result only provides inconclusive evidence and further experimentation is required to either confirm or disconfirm the null hypothesis. The interpretation of the findings is further complicated by the noteworthy limitations of the current experiment. In particular, while participants were able to distinguish between targets and non-targets, overall discriminability was low. For the items that were actually old, participants said "old" 45.44 % of the time. Similarly, for the items that were actually new, participants said "old" 43.01% of the time, together indicating that participants failed to detect the subtle differences between the images unless they were unmatched on both features compared to the study target.

Study 2. General Discussion

Study 2 aimed to evaluate how different types of linguistic reference can impact individual and binding memory for objects and distinct object features. The results of Study 2A indicated that the demonstrative pronoun "that" does not change the representation of individual objects by binding them into a composite above and beyond using a conjoined noun phrase or a personal pronoun "them". Specifically, participants were more likely to recognize an object or object pair as "old" when the linguistic labels that referred to them were repeated twice as evidenced by the higher recognition rate for the 2-noun condition compared to the pronoun conditions (e.g., *the sheep and the shirt* vs. *them/that*). These findings are at odds with the idea of a "hybrid representation" – which argues for a composite of visual and linguistic information (Dessalegn & Landau, 2008; Scott & Sera, 2018) – at least, in the context of the demonstrative pronoun "that". Why then did Dessalegn and Landau (2008) find a memory boost for the location and color of shapes in 4-year-olds, which Scott and Sera (2018) later replicated? Recall that the reported mnemonic benefit in both studies was specific to the directional labels (e.g., *on the left/on the top*) but not other relational labels (e.g., *the green is next to the red*), neutral labels (e.g., *this is dax*), or nonlinguistic attentional cues (e.g., objects flashing). On one hand, relational language may be special for memory, even more so that the language that is used to refer to conceptual entities, namely, the pronoun "that" (Brown-Schmidt et al., 2005). On the other hand, if the effect of relational language on memory is limited to children who develop relational knowledge and acquire relational terms around the age of 4 or 5 (Gentner et al., 2011), it is possible that learning new terms like "left" and "right" may involve larger attentional resources compared to learned words and flashing objects.

The findings of Study 2B further revealed that the use of adjective modifiers when referencing objects facilitates its recognition compared to when referred to using nouns only in line with previous work (Hofmeister, 2011; Troyer et al., 2016; Yoon et al., 2016; Yoon et al., 2021). While we found evidence suggesting that distinct object features are indeed remembered independently from one another (see also Brady et al., 2013), lexicalizing one feature of an object did not impact recognition of the unlexicalized feature. However, further experimentation is needed to narrow down the effect of referential modification on memory for both the lexicalized and unlexicalized features. If any follow-up experiment should find evidence consistent with the hypothesis that memory for both object features is boosted following linguistic reference, such findings would be at odds with the previously reported mnemonic deficits following linguistic labeling (Schooler & Engstler-Schooler, 1990; Lupyan, 2008). This discrepancy then could be attributed to the differences in experimental designs. Specifically, both Lupyan (2008) and Schooler and Engstler-Schooler (1990) utilized generation (vs. processing) tasks with participants either categorizing a visual image as *chair* or *table/lamp* or verbally describing faces after they were encoded in memory. Yet, Zormpa and colleagues (2018) found better memory for objects when participants generated linguistic labels for them during study. Taking into consideration the differences between these paradigms, such a pattern of findings would raise an intriguing possibility that accessing language during encoding benefits memory, whereas using language at retrieval creates memory interference, weakening the original memory trace, and impairing subsequent memory.

Finally, it is worth mentioning that both studies had notable limitations. In Study 2A, contrary to our predictions, we found that repeating linguistic labels resulted in better binding but not individual memory when compared to both pronoun conditions. While it was outside of the

scope of the current work to probe the reason why this was the case, we speculate that the differences might have stemmed from the variability in the difficulty levels of the two memory tests. More specifically, the mean endorsement rates were ~75% and 60% for the individual and bindings tests respectively. Further work needs to control for the difficulty levels to adequately compare participants' performance on both tests. Similarly, in Study 2B, the mean hit rate across participants was only 0.45, indicating that the task was difficult. To address this concern, follow-up experiments might consider utilizing a 2-AFC paradigm which might highlight the subtle differences between the clothing items. Additionally, Study 2B only had 35% power to detect the effect associated with Contrast 3 from Model 2 targeting the difference between the items that matched on Feature A vs. the items that matched on Feature B.

CHAPTER IV

Study 3. The Role of Linguistic Features of Spontaneous Speech in Shaping Mental Representations of Real-World Experiences

So far, Studies 1 and 2 have focused on the role of linguistic reference in accessing and retrieving representations of concepts and objects. However, human language is complex, encompassing multiple domains extending beyond the lexicon. Therefore, in **Study 3**, we probed the impact of various linguistic features inherent to spoken language on memory for real-world experiences, as well as the underlying cognitive mechanisms by which these features operate. Specifically, we focused on memory for conversation and employed an ecologically valid paradigm to explore the relationship between language and memory as it occurs in real life.

Previous research on conversational recall shows that the amount of conversation that can be recalled after a delay is limited (Benoit & Benoit, 1998; Ross & Sicoly, 1979; Stafford & Daly, 1984) and biased in favor of one's own contributions (Slamecka & Graf, 1978; McKinley et al., 2017; Ross & Sicoly, 1979). In targeting the aspects of conversational interaction that shape subsequent recall, we reviewed the literature predicting memory for pre-recorded materials. In particular, studies with pre-recorded sentences and stories indicated that disfluency (Diachek & Brown-Schmidt, 2022; Fraundorf & Watson, 2011) improves memory for linguistic input by orienting attention to speech. Yet, given the previous unsuccessful attempts to replicate the disfluency-driven memory benefit for stories (Donahue et al., 2017), and the narrow scope of the effect reported for sentences (Diachek & Brown-Schmidt, 2022), it remains possible that the disfluency memory benefit would not generalize onto unscripted speech. Further, similar to

disfluency, backchanneling has been associated with heightened attention and might consequently, improve memory for language. Finally, studies examining the content that is more likely to be remembered indicated that statements that express disagreement might be more memorable due to their high interactional value (Keenan et al., 1977). Taken together, we examine how four features inherent to spontaneous speech, namely, disfluency, backchanneling, "like", and disagreement shape memory for natural language.

Note that originally, the study additionally aimed to investigate the effect of mood on recall and included a pre-registered mood manipulation (see Method). However, the mood manipulation did not last the duration of the experiment, did not significantly impact recall, and is not reported here.

Methods

Participants

Following the methods from Stafford & Daly (1984) and Benoit & Benoit (1988) on conversational memory (i.e., 20 dyads per cell), and ensuring that we reach sufficient power, consistent with our pre-registration, we recruited 120 participants through the Vanderbilt University participant pool (i.e., 30 dyads per cell in a 2x2 design). 84 participants received partial course credit; the remaining 36 participants received a $20 Amazon gift card. The recording from one conversation was lost leaving 59 conversational dyads (118 participants) in the final analysis.

Procedure

At the beginning of the experiment, participants watched a mood induction video. Each participant was seated at an individual computer in the same experimental room and was offered a separate set of headphones. The computers were situated back-to-back so that each person could not see the screen of the other participant. The experimenter remained in the room for the duration of the mood induction procedure. Each mood induction video lasted 1 minute and 33 seconds and included 20 images from the International Affective Picture System (IAPS; Lang et al., 2008) accompanied by classical music. Previous research indicated that classical music enhances affective experiences and as a result, has been recommended to be presented along with the images to increase the effectiveness of the mood induction procedure (Zhang et al., 2014). Participants were randomly assigned to one of four conditions created by the crossing of mood congruence (congruent and incongruent mood) and mood (happy or sad). In the happy condition a positive mood was induced by presenting a series of happy images (IAPS items: 1440, 1441, 1460, 1710, 1750, 1811, 1920, 2080, 2154, 2209, 2340, 5210, 5760, 5825, 5830, 5833, 5910, 7502, 8190, 8501) with a mean valence of 8 (SD = 0.20, min = 7.62, max = 8.34) as the participant listened to Beethoven, Symphony no.6 (3rd mvt) (Baumgartner et al., 2006). Participants in the sad mood condition watched a series of unhappy images (IAPS items: 2205, 2691, 3550, 6300, 6563, 7380, 9163, 9220, 9280, 9295, 9332, 9560, 9600, 9623, 9800, 9810, 9830, 9910, 9921, 9940), with a mean valence of 2.30 (SD = 0.39, min = 1.62, max = 3.04) while listening to Samuel Barber, Adagio for Strings (Baumgartner et al., 2006).

After the mood induction procedure, participants completed a Positive and Negative Affect Survey (PANAS; Watson et al., 1988) via Survey Monkey to assess their positive and negative affect as a manipulation check (**Figure 18**). Next, participants were asked to change their chairs to sit facing each other and to engage in an unscripted conversation for 15 minutes.

We offered participants 3 conversation prompts ("Living in Nashville", "Favorite TV show", and "Favorite artist") to get the conversation going, but the conversation was unscripted and participants were free to talk about anything. All conversations were recorded on the computer using the PRAAT software. For the duration of the conversation, the experimenter left the room but returned for the next phase of the experiment. After 15 minutes of conversation, participants completed a distraction task which included watching a 20-minute informational video about outer space. The use of a video as a filled delay task is a common procedure in the conversational memory literature (e.g., Samp & Humphries, 2007; Stafford, Waldron, & Infield, 1989; Stafford & Daly, 1984).

Next, participants returned to their computers and completed PANAS for the second time to assess affect at the time of completing the free recall task. The second responses from 2 participants for PANAS were lost and thus, were not included in the analyses. Finally, participants completed a surprise written free recall task. They were instructed to type as much as they could recall about what was said in the conversation, word for word, including the source of each utterance (i.e., who said what), and using a separate line for each speaker. Participants saw a short example from a made-up conversation of how the recall should be typed. The experimenter instructed participants that they had 30 minutes to complete the task to ensure that they would write everything they remember without trying to leave early. However, if participants indicated to the experimenter that they could not remember any more details from the conversation before the 30-minute mark, they were allowed to leave. Additionally, the experimenter instructed participants not to consult each other about the past conversation and stayed in the room with the participants until they finished the task. Together, the entire experiment took approximately 90 minutes to complete.

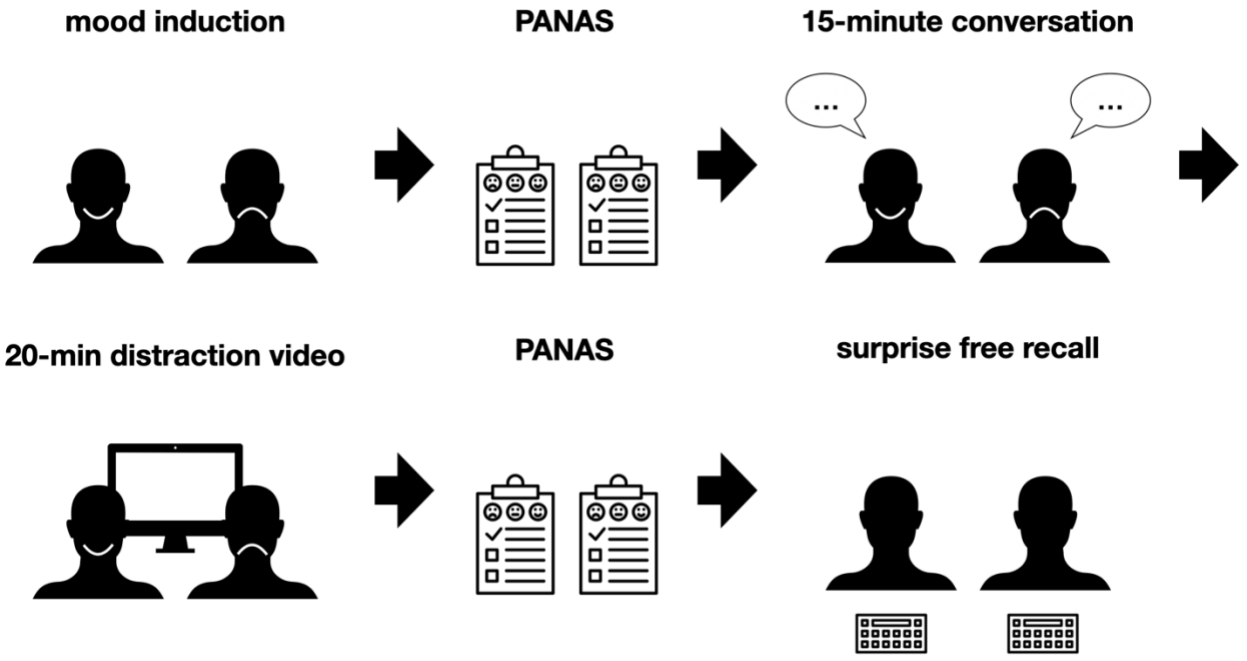


Figure 15. Schematic of the experimental procedure.

Conversation analysis

Audio recordings of each conversation were transcribed word-by-word, with utterances numbered and labeled for who said what. Then, the utterances were broken down into idea units (Ross & Sicoly, 1979; Stafford & Daly, 1984; Stafford et al., 1987; 1989). An idea unit corresponds to "the smallest unit of meaning that has informational or affective value; it represents the gist of each thought expressed by the interactants" (Stafford et al., 1989, p.600) and typically, includes one subject and one verb (for more details on the coding scheme, see project's associated OSF page).

When transcribing the data, we also coded the serial position of each idea unit in the conversation as serial order effects in recall are well-documented in the literature (Deese & Kaufman, 1957; Murdock, 1962). While this literature primarily uses lists of words as stimuli

(which have a natural ordered structure), we hypothesized that serial order effects might generalize to conversational recall as well, with idea units at the beginning and end of the conversation recalled better than idea units in the middle. Lastly, we coded each idea unit in the conversation for a series of linguistic features (see **Table 5** for examples), as follows:

Disfluency. An idea unit was coded as disfluent if it contained fillers (um/uh), pauses, disfluent repetitions, rephrases, or restarts. Disfluent utterances were coded as a 1 if a disfluency was present regardless of its position within the idea unit, and a score of 0 if not. If a disfluency occurred in-between two idea units, we coded the second of the two idea units as disfluent since our hypothesis specified that disfluency orients attention towards what is about to be said. In addition, for disfluent idea units, we also coded whether there were multiple disfluencies present or only a single disfluency present in the idea unit.

"like". Each idea unit was assigned score of 1 if the discourse marker "like" was present, and 0 if not. "like" was not considered a disfluency as we were interested in whether "like" and disfluency shared functional similarities with respect to memory.

Backchannel. We hypothesized that backchanneling is indicative of the listener's heightened attention to spoken language. We coded an idea unit as backchanneled (1) if it was followed by a backchannel, and 0 – if not. Thus, if one participant said "I'm a huge fan of Beyoncé" and the other participant said "cool", the IU "I'm a huge fan of Beyoncé " would be coded as back-channeled (1). Backchannels included exclamations such as "yeah", "okay", "cool", "that's dope", "right", "really", "wow". We considered short responses like this to be a backchannel if: a) it contained no more than 3 words; b) indicated attention/following the conversation; and c) could be removed without any loss of information from the conversation

(i.e., there was no follow-up on it). If a backchannel interrupted an utterance, only the idea unit that directly preceded the backchannel was coded as backchanneled.

Disagreement. Idea units that expressed disagreement with a previous statement were assigned a score of 1 for disagreement, and 0 – if no disagreement was expressed. In the following example, Speaker B's utterance was coded as containing disagreement since it directly contradicted Speaker A's statement:

Speaker A: "I didn't know it was gonna be that warm in Nashville."

Speaker B: "It wasn't really that warm last year.

Table 5. Example of conversation coding for idea units and four linguistic features.

IU#	PID	Idea unit (IU)	DISFLUENCY	MULTIPLE DISFLUENCIES	BACKCHANNEL	DISAGREEMENT	LIKE
167	9	Um one of the grad	**1**	0	0	0	0
168	9	Well post-doctoral scholars um	**1**	0	0	0	0
169	9	went to grad school	0	NA	0	0	0
170	9	In Rochester	0	NA	**1**	0	0
	10	Ah					
171	9	And she really liked it	0	NA	0	0	0
172	10	they apparently have a really good linguistics program	0	NA	0	0	0
173	10	It makes sense	0	NA	0	0	0
174	10	because there is a huge Deaf community	0	NA	**1**	0	0
	9	Oh					
175	10	Yeah, it's like one of the biggest Deaf communities	0	NA	0	0	**1**
176	10	in the country	0	NA	0	0	0
177	10	is in Rochester	0	NA	0	0	0
178	10	so they probably have some relation	0	NA	0	0	0
179	10	to Linguistics	0	NA	0	0	0

180	9	Mhm, that's really cool	0	NA	0	0	0
181	9	And that's not too far	0	NA	0	0	0
182	9	from um New York City	**1**	0	0	0	0
183	9	right?	0	NA	0	0	0
184	10	Ehh, it's pretty far	0	NA	0	**1**	0
	9	Really?					
185	10	6 hours	0	NA	**1**	0	0
	9	Oh ok					

Note. **IU#** denotes the serial order of idea units within the conversation. **PID** is participant ID.

Recall analysis

Each written recall was broken down into idea units following the same guidelines used to code the conversations. To quantify memory for conversation, we matched the idea units in the recalls to the idea units in the conversation if they conveyed the same gist (**Table 6**). If participants recalled something that was not in the original conversation, that idea unit was assigned an "NA". The accuracy of the recalls was not coded. After the initial coding was complete, a second coder independently coded 10% of randomly selected recalls (n=12) masked to the original coding. An inter-coder reliability coefficient was calculated as the proportion of recalled idea units for which both coders were in agreement as to which specific conversation idea unit it corresponded to. The ICC was equal to 0.79 (SD = 0.05) indicating substantial agreement.

Conversational recall data, PANAS data, along with the associated analyses scripts as well as the details on the conversation coding scheme are available on project's associated OSF page.

Table 6. Example of recall coding (participant's spelling preserved).

Recaller PID	Recall IU	Conversation IU	IU# from the conversation

9	One of the postdocs	[167: um one of the grad] well post-doctoral scholars	168
9	in the lab	---	NA
9	I work in	---	NA
9	down the hall	---	NA
9	went to grad school	went to grad school um	169
9	at University of Rochester	in Rochester	170
9	They have a really good linguistics program there	They apparently have a really good linguistics program	172
9	That makes sense	It makes sense	173
9	Because there's a really big deaf population there	Because there is a huge Deaf community	174
9	I'm guessing you aren't too far	That's not too far	181
9	from new york city?	from New York City	182
9	I am kind of far	It's pretty far	184

Results

Descriptive statistics

Across 59 conversations, interlocutors produced 29,943 idea units. On average, there were 507 idea units (SD = 94.52, min = 325, max = 764) per conversation. The interlocutors contributed to the conversation roughly equally, with the participants who spoke more in each dyad producing on average 58.52% of all idea units in the conversation (SD = 6%, min = 50%, max = 74.02%). Each written recall included an average of 138.38 idea units (SD = 56.97, min = 29, max = 267). On average, participants recalled 27.89% of the original conversations (SD = 11.66%, min = 5.94%, max = 56%), including 29.80% (SD = 13.10%, min = 4.80%, max = 59.86%) what they said, and 26.53% (SD = 11.51%, min = 5.94%, max = 57.98%) of what their conversational partner said to them (**Figure 19**).

In terms of linguistic features, 4,870 (16.26%) of the idea units were disfluent, and 6,225 (20.79%) contained the discourse marker “like”. Of the 4,870 idea units that contained

disfluency, 1,166 (24%) contained multiple disfluencies, and 4,231 (14.13%) idea units were backchanneled. Only 49 (0.16%) idea units contained disagreement.

Memory predictors

A logistic mixed effects regression model of whether (1) or not (0) each idea unit in the conversation was recalled by each participant included fixed effects of participant role (whether the current participant produced (1) or listened to (0) that idea unit), given well-documented generation benefits in conversational memory (Ross & Sicoly, 1979, *inter alia*). Each linguistic feature (fluency, backchanneling, disagreement, "like") was coded as 1 when the feature was present and 0 when absent. The interaction between each feature with the production variable were included as predictors. In addition, conversation length as measured by the number of idea units in the conversation (centered and scaled), serial order of the idea unit (centered and scaled), and the quadratic function for idea unit serial order were included in the model as control variables. Random effects included by-subject and by-dyad intercepts, and random slopes were included in the model if the model converged with them (see details on model selection procedure above). Analyses were performed in R (R Core Team, 2021) through the RStudio interface (RStudio Team, 2020) using the "lme4" (Bates et al., 2015) package.

The model (**Table 7**) contained a negative intercept ($b = -1.43$, $z = -18.03$, $p = < 10^{-15}$) such that idea units were more likely to *not* be recalled than recalled. The linear term for idea unit serial order was negative ($b = -0.57$, $z = -6.89$, $p = <10^{-11}$), indicating that idea units at the beginning of the conversation were more likely to be recalled than those at the end; the remaining control variables were not significant.

Participants were more likely to recall idea units they produced vs. heard (b = 0.18, z = 4.04, p = $<10^{-4}$), consistent with a generation effect. Consistent with our prediction, disfluent idea units were more likely to be recalled than fluent ones (b = 0.25, z = 6.74, p < 10^{-10}). Notably, the interaction between fluency and production was not significant (b = 0.01, z = 1.12, p = 0.09) indicating that both listening to and producing disfluency boosted recall[3]. Similarly, backchanneled idea units were more likely to be recalled (b = 0.10, z = 2.34, p = 0.02), an effect that did not interact with production (b = -0.06, z = -1.07, p= 0.28). Finally, idea units that expressed disagreement (b = 0.36, z = 1.12, p = 0.26) or that contained "like" (b = 0.07, z = 1.93, p = 0.05) were *not* significantly more likely to be recalled (**Figure 20**).

Table 7. Memory results for Model 1: Mixed effect model with a binary dependent measure - whether the idea unit was recalled or not on the free recall test. Values in bold indicate significant results at an alpha level of .05.

*recalled (0/1) ~ 1 + fluency * produced + produced * backchanneling + produced * disagreement + produced * like + IU number + IU number ^2 + conversation length + (1 + IU number + IU number ^2 + produced | conversation ID : participant ID) + (1 | conversation ID)*

Fixed Effects	***Estimate***	***SE***	***z-value***	***p-value***
Intercept	**-1.43**	**0.08**	**-18.03**	**< 10^{-15}**
IU number	**-0.57**	**0.08**	**-6.89**	**< 10^{-11}**
IU number ^2	0.01	0.06	0.09	0.93
Conversation length	-0.10	0.06	-1.56	0.12
Produced (produced = 1, listened = 0)	**0.18**	**0.04**	**4.04**	**< 10^{-4}**
Disfluency (disfluent = 1, fluent = 0)	**0.25**	**0.04**	**6.74**	**< 10^{-10}**
Backchanneling (backchanneled = 1, not backchanneled = 0)	**0.10**	**0.04**	**2.34**	**0.02**

[3] A post-hoc analysis indicated that the disfluency effect was of a similar magnitude for listeners, b = 0.24, and speakers, b = 0.28.

Like ("like" present = 1, "like" absent = 0)	0.07	0.04	1.93	0.05
Disagree (disagree = 1, agree = 0)	0.36	0.32	1.12	0.26
Produced*backchanneling	-0.06	0.06	-1.07	0.28
Produced*disfluency	0.01	0.05	0.12	0.90
Produced*like	0.03	0.05	0.51	0.61
Produced*disagree	0.02	0.45	0.05	0.96

Random Effects	***Variance***	***SD***	***Correlations***		
Conversation ID (intercept)	0.66	0.81			
IU Number ^2	0.44	0.66	-0.15		
IU Number	0.76	0.87	0.45	0.75	
Produced	0.14	0.38	0.03	-0.02	0.03
Participant ID (intercept)	0.00	0.00			

Number of observations: 59886[4]; groups: conversation_id: 59, participant_id: 118

The beneficial effect of disfluency on conversational memory raised the question of whether disfluency has cumulative effects. Thus, an exploratory analysis including the number of the disfluencies in each idea unit (**Table 8**) revealed that the memory boost for disfluent utterances was similar regardless of the number of disfluencies (b = 0.02, z = 0.41, p = 0.69).

Table 8. Memory results for Model 2: Mixed effect model with a binary dependent measure - whether the idea unit was recalled or not on the free recall test. Values in bold indicate significant results at an alpha level of .05.

recalled (0/1) ~ 1 + disfluency present + multiple disfluencies+ (1 | conversation ID) + (1 | conversation ID: participant ID)

Fixed Effects		***Estimate***	***SE***	***z value***	***p-value***
	Intercept	**-1.04**	**0.07**	**-14.79**	**$< 10^{-15}$**

[4] Note that the number of observations is twice as many as the total number of IUs across the conversations because the dependent variable is an IU for each participant in the dyad.

Disfluency present (fluent = -0.1630097, one disfluency = 0.8369903, multiple disfluencies = 0.8369903)	**0.26**	**0.03**	**9.18**	**$< 10^{-15}$**
Multiple disfluencies (fluent = 0.04233043, one disfluency = -0.45766957, multiple disfluencies = 0.54233043)	0.02	0.05	0.41	0.69

Random Effects	***Variance***	***SD***
Conversation ID (intercept)	0.25	0.50
Participant ID (intercept)	0.17	0.41

Number of observations: 59886; groups: conversation_id: 59, participant_id: 118

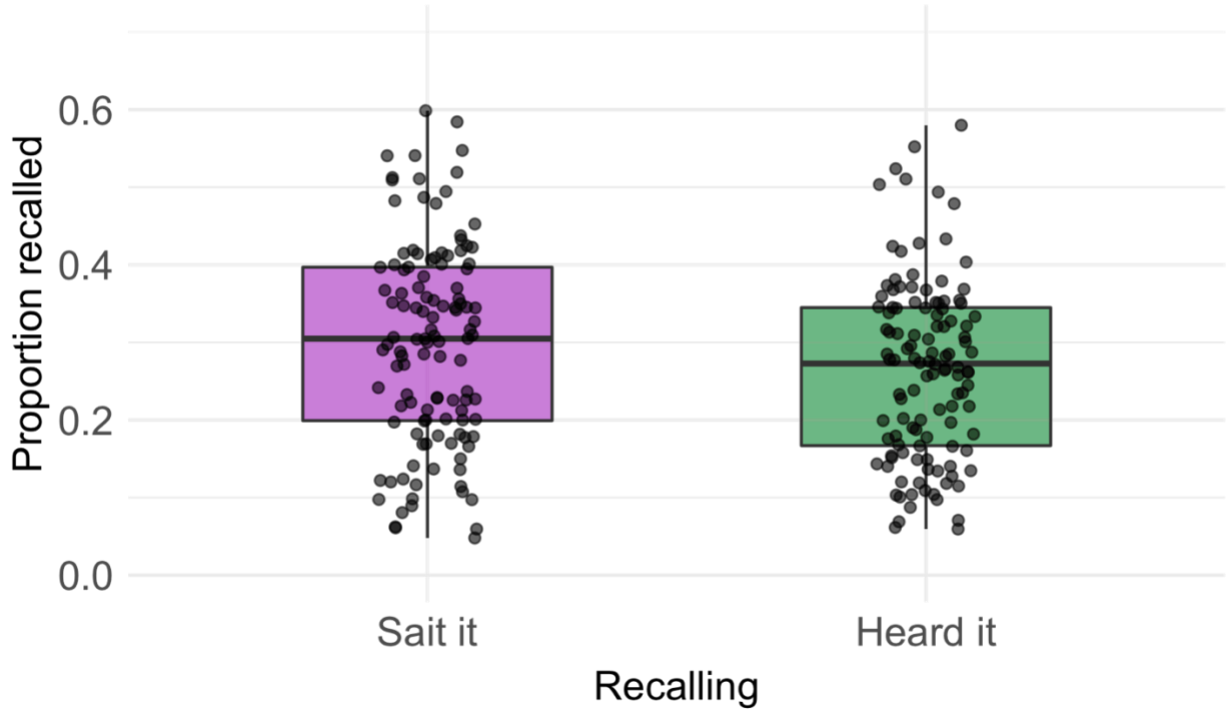
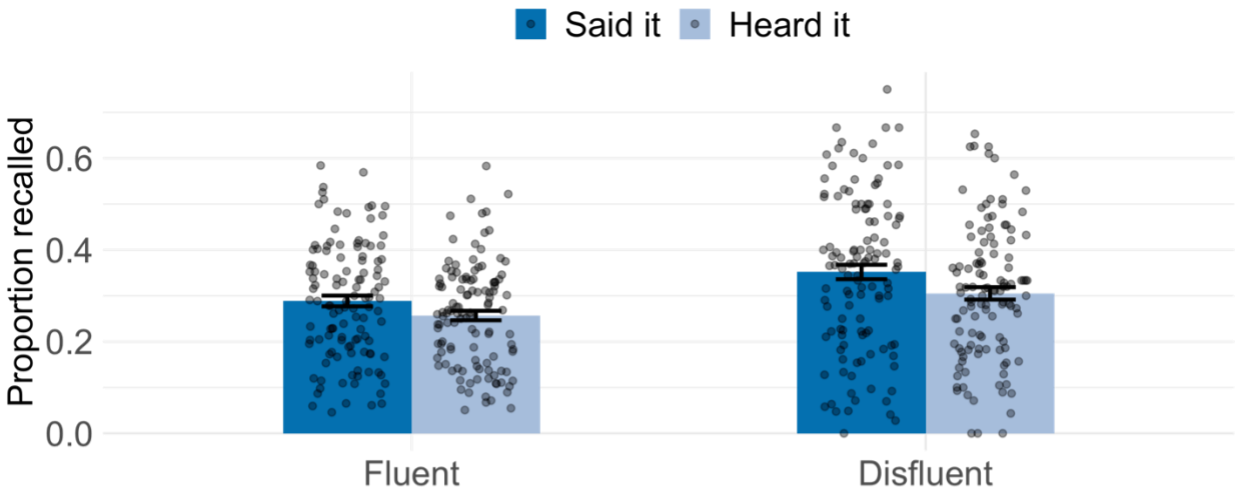
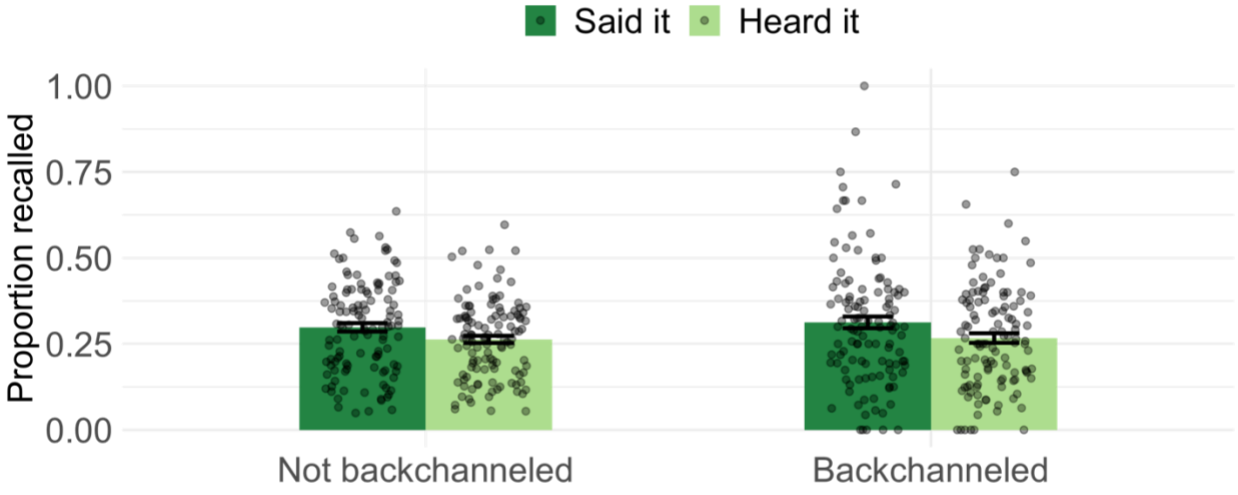


Figure 16. Proportion of idea units from the original conversation that were recalled as a function of whether participants said it or heard their interlocutor say it.


Said it
Heard it
Proportion recalled
0.6
0.4
0.2
0.0
Fluent
Disfluent


Said it
Heard it
Proportion recalled
1.00
0.75
0.50
0.25
0.00
Not backchanneled
Backchanneled

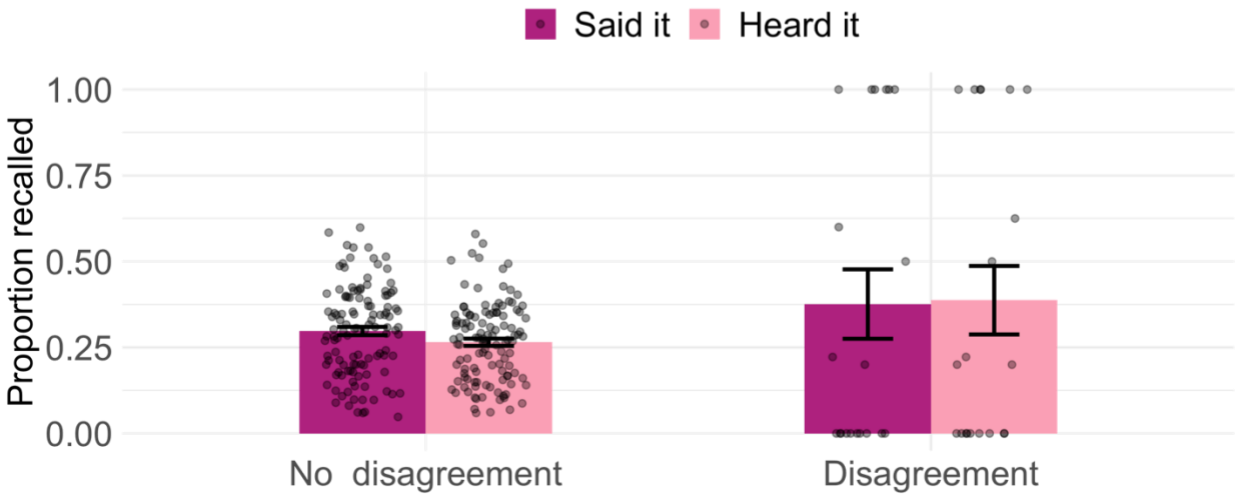

Said it
Heard it
Proportion recalled
1.00
0.75
0.50
0.25
0.00
No disagreement
Disagreement

Figure 17. Proportion of idea units from the original conversation that were recalled as a function of fluency (Panel A), backchanneling (Panel B), and disagreement (Panel C).

Discussion

Conversations can be impactful, important parts of life, yet prior research indicates that our ability to recall the details of prior conversations is limited (Hjelmquist & Gidlund, 1985; Isaacs, 1990; Miller et al., 1996; Fischer et al., 2015; McKinley et al., 2017; Ross & Sicoly, 1979; Zormpa et al., 2019). The aim of this study was to understand the factors that shape conversational memory. In addition to our pre-registered analyses of the effect of mood on conversational memory, we used detailed coding of the linguistic features of interactional talk to attempt to predict what was going to be recalled later. Overall, our findings replicate prior evidence that conversational memory is limited, and egocentrically biased. Dyads conversed for 15 minutes, and then after a 20-minute filled delay, completed a surprise free recall task. Participants recalled on average 28% of the idea units expressed in the original conversation, including 29% of their own and 26% of their partner's ideas. With these findings in hand, we can then turn to what predicts successful conversational recall.

While analyses relating the linguistic features of conversation to memory were not a part of our pre-registration, new discoveries in the lab led us to code four linguistic features prior to conducting any statistical analyses. These findings indicated that disfluency orients listeners' attention to the upcoming linguistic material, improving its encoding and subsequent recognition (Diachek & Brown-Schmidt, 2022; also Collard et al., 2008). The findings further revealed that the disfluency-related mnemonic benefit was short-lived and position dependent, only occurring

when the critical test word was sentence-final and immediately preceded by disfluency. While this and other work observes a disfluency-memory boost for scripted materials (Corley et al., 2007; MacGregor et al., 2010; Fraundorf & Watson, 2011), it was not clear if the disfluency-related memory boost would generalize to unscripted spontaneous speech. Further, consistent with this and prior work indicating that some linguistic features orient attention towards unfolding speech stream (Fraundorf & Watson, 2011), and improve subsequent memory for it, we hypothesized that backchanneling may serve as an indicator of listener's heightened attention, and as a result, might be predictive of better memory for conversation. Consequently, we coded the unscripted conversations for disfluency, backchannels and two other linguistic features, the discourse marker "like" and disagreement. The results of this analysis indicated that both disfluency and backchanneling but not disagreement or the discourse marker "like" increased the likelihood of that idea unit being recalled later.

Our finding that disfluent idea units were more likely to be recalled offers key new insights into the disfluency-memory boost in memory for spontaneous speech. Multiple prior findings suggest that the reason disfluency boosts the *listener's* memory for speech is that the disfluency orients their attention (Corley & Hartsuiker, 2011; Fraundorf & Watson, 2011; Diachek & Brown-Schmidt, 2022). The fact that disfluency boosted memory for *both* speaker and listener speaker raises interesting questions about the nature of the effect.

One explanation is that speakers produce disfluency in cases where they have focused attention on what will be said, and it is this focused attention on what they will say that both results in the disfluency and boosts their own memory. If this explanation is correct, it suggests that it should be possible to identify features of what is *about* to be said (either the message, meaning, phrasing, or word choices) that predict both the speaker's use of disfluency, and their

tendency to later remember what was eventually said. This account points to the intriguing possibility that the reason disfluency orients a listener's attention to the speech stream, thereby boosting memory, is this tendency for disfluency to be associated with to-be-uttered information that grabs the speaker's attention. This line of argumentation is consistent with the argument that production constraints shape distributional patterns in language, thereby shaping comprehension processes (MacDonald, 2013). While testing this account will likely require experimental manipulations of what is (likely to be) said, and thus outside the scope of the present research, we see this as a fruitful line of future inquiry.

Alternatively, it is possible that the cognitive mechanisms underlying the disfluency-related memory boost are different for speakers and listeners. Prior research demonstrated that disfluencies often stem from difficulties in language production (Clark, 1996; Fraundorf & Watson, 2014) and consequently, might be linked to increased cognitive demand, rather than attentional focus per se. Since effortful material is more likely to be recalled later (Tyler et al., 1979), the downstream consequence of producing disfluency might be better memory for what is about to be said. Lastly, we observed that the mnemonic benefit associated with disfluency was not additive, in that listening to multiple disfluencies in an utterance does not significantly improve memory over listening to just one. Given that attentional recourses are not limitless (Kahneman, 1973), there might be a cap on the amount of attention that can be allocated or oriented towards speech upon hearing a disfluency.

In addition to disfluency, another common feature of interactive conversation is back-channel responses (Clark & Schaefer, 1989; Clark & Krych, 2004; Roque & Traum, 2008). The present findings report for the first time that backchannels are associated with better memory for what was said, with the utterances that are backchanneled being more likely to be recalled later

compared to the utterances not followed by back-channels. Given that this effect was present for both the person who produced the backchannel and the person who heard it, we speculate that the mechanism, like disfluency, may be related to attentional orienting towards the speech stream.

Finally, we also coded two further linguistic features, disagreement, and the use of "like", neither of which significantly impacted memory for what was said. Given prior findings that utterances with high interactional value, such as jokes, tend to be remembered better (Keenan et al., 1977), we hypothesized that disagreements – which can also be highly interactive – may be particularly well remembered. The lack of an effect for disagreements might be due to insufficient power to detect this effect. Recall that roughly only 0.2% of utterances communicated disagreement compared to 16% of statements that were disfluent or 14% that were backchanneled. It is possible that because all participants were strangers, they felt reluctant to express their disagreement with their conversational partners. Another possibility is the conversational topics discusses were not polarizing enough.

We hypothesized that the word "like" might also influence conversational memory as, like disfluent filler words (*um, uh…*), "like" can be used by speakers when collecting their thoughts and preparing articulation. The use of "like" as a discourse marker is prevalent in spontaneous speech with some studies estimating that its frequency exceeds the frequency of the conjunction "and" (Tagliamonte, 2005). "like" is also more common than other related forms, including "you know", "I mean", and "well" (Beeching, 2016). While "like" shares some functional similarities with disfluency, some consider it to be qualitatively different from other disfluency types in that "like" does not constitute a natural pause (Croucher, 2004a). In addition, Croucher (2004b) found gender differences in the use of "like" in spontaneous speech while no

such differences were observed for disfluencies “um” and “uh”, further pointing to functional differences between “like” and disfluencies. Finally, studies of deception in both laboratory and real-world settings show that “um” is *less* common with deceptive than non-deceptive speech, whereas no such differences are present for the use of “like” (Arciuli, Mallard, & Villar, 2010; Villar, Arciuli, Mallard, 2012), again pointing to functional differences between “like” and disfluency. While “like” was more common that disagreements (20.79% of all idea units included “like” vs. 16.26% -- disfluency), “like” was not associated with a significant memory boost, pointing to functional differences in the impact that “like” has on memory for conversation.

CHAPTER V

General Discussion

Summary

Linguistic labels are considered a crucial part of a concept's representation. Yet, how humans access and evaluate semantic properties of lexically invoked concepts remains debated. Distributional semantic models, which construct vector spaces with embedded words, offer valuable insight for understanding the representational structure of human semantic knowledge. Unlike some classic semantic models, distributional semantic models lack a mechanism for specifying the properties of concepts, which raises questions regarding their utility for a general theory of semantic knowledge. In **Study 1**, we developed a computational model of a binary semantic classification task, in which participants judged target words for the referent's size or animacy. We created a family of models, evaluating multiple distributional semantic models and mechanisms for performing the classification. The most successful model constructed two composite representations for each extreme of the decision axis (e.g., one averaging together representations of characteristically big things, and another of characteristically small things). Next, the target item was compared to each composite representation, allowing the model to classify more than 1500 words with human-range performance and to predict response times. We proposed that when making a decision on a binary semantic classification task, humans use task prompts to retrieve instances representative of the extremes on that semantic dimension and

compare the probe to those instances. This proposal is consistent with the principles of the instance theory of semantic memory.

In **Study 2**, we focused on the role of linguistic reference in shaping mental representations of objects and object pairs as well as their memories. Previous research points to the beneficial effect of processing of some linguistic referential expressions on memory for objects. To further elucidate the relationship between linguistic reference and object representation, in **Study 2A**, we explored the effect of the demonstrative pronoun "that" on binding memory for objects as it had been implicated to evoke conceptual composites potentially akin to the idea of a "hybrid representation" of visual and linguistic information (Dessalegn & Landau, 2008; Scott & Sera, 2016). To probe this question, in Study 2A, participants viewed and manipulated images of familiar objects. Critically, the instructions were presented via pre-recorded audio in three linguistic conditions. In the first condition, the two critical objects were referred to using a conjoined noun phrase (e.g., *the shirt and the sheep*). In two other conditions, the objects were named either using the pronoun "them" or the pronoun "that". Contrary to our predictions, the results of the binding memory test indicated that participants were more likely to recognize a pair of objects as "old" when they had been referred to using two nouns (e.g., *the shirt and the sheep*). Importantly, we did not observe any differences between the conditions for individual memory. In **Study 2B,** we further probed the effect of language on the representation of objects by shifting focus toward (modified) noun phrases. While previous research established that modification improves memory for objects compared to describing them using nouns only, how referential modification shapes different features of objects remains largely unknown. In our paradigm, participants studied images of clothing items that varied on two distinct features (e.g., sleeve length and overall length). Simultaneously, the images were labeled using pre-recoded

audio descriptions highlighting one feature of an item or neither (e.g., *sleeveless dress* vs. *dress*). Immediately after, participants completed a recognition memory test. We found that using an adjective modifier increased the proportion of response "old" for that clothing item compared to the noun only condition. Yet, the endorsement rates were not significantly different for the lexicalized compared to the unlexicalized feature suggesting that modification did not differentially modulate recognition of distinct object features.

Language, as it occurs in real-world settings, is complex and various linguistic levels beyond the lexicon might shape subsequent memory. In **Study 3**, we examined the linguistic features of spontaneous speech that determine what is and is not recalled following an unscripted conversation. In an empirical study of 59 dyadic conversations (118 participants), we examined the effects of various features of the conversation - disfluency, backchannels, "like", and disagreements, on memory for what was said. While we replicated prior findings that memory was better for what was said than what was heard, our pre-registered predictions regarding mood were not supported, a finding which may relate to changes in mood throughout the conversation. Critically, consistent with the previous findings of a disfluency-related memory boost for pre-recorded sentences and passages, we identified two linguistic features that did promote recall. Extending prior findings with scripted and pre-recorded materials, disfluency (*um/uh*) promoted conversational recall, as did backchannelling (*ok, yeah*). Interestingly, the disfluency-related memory boost was similar for both the speaker and the listener and was observed regardless of the number of disfluencies in the utterance. These findings point to the intriguing possibility that the reason disfluency orients a listener's attention to the speech stream, thereby boosting memory, is that the speaker, too, has focused attention on what will be said. In sum, we report

that linguistic features of spontaneous speech as they occur in unscripted conversation are predictive of what is and is not recalled.

Conclusion

To conclude, this dissertation contributes to the literature characterizing the relationship between language and memory. More specifically, the three studies examined the effect of language on memory and the associated cognitive mechanisms by focusing on the ways in which individuals access and manipulate the information encoded within linguistic labels for concepts (Study 1), the effect of different types of linguistic reference on binding memory for objects and distinct object features (Study 2), and finally, the effect of spoken language features such as disfluency and backchanneling on memory for real-world experiences (Study 3). Taken together, our findings point to an existence of a nuanced relationship between language and memory with many facets of language shaping representations and consequent memories of concepts, objects, and real-world experiences. We hope that this work will deepen our understanding of language and memory, and human cognition more broadly, and will inspire new lines of research.

REFERENCES

Arciuli, J., Mallard, D., & Villar, G. (2010). "Um, I can tell you're lying": Linguistic markers of deception versus truth-telling in speech. *Applied Psycholinguistics*, *31*(3), 397-411.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive science*, *27*(2), 195-225.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2014). Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal. *Journal of Memory and Language*, *68*(3), 1–43.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, *11*(3), 211-227.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Baumgartner, T., Esslen, M., & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International Journal of Psychophysiology*, *60*(1), 34–43.

Benoit, P. J., & Benoit, W. L. (1988). Conversational memory employing cued and free recall. *Communication Studies*, *39*(1), 18-27.

Bhatia, S. (2019). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 627.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31-36.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, *19*(12), 2767-2796.

Bortfeld, H., Leon, Ê. S. D., Brennan, S. E., Bloom, J. E., & Schober, Ë. M. F. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 165-174.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2013). Real-world objects are not represented as bound units: independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, *142*(3), 791.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62-89.

Brown-Schmidt, S., & Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *Journal of Memory and Language*, *84*, 49-74.

Brown-Schmidt, S., & Konopka, A. E. (2008). Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*, *109*(2), 274-280.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592-609.

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, *53*(2), 292–313.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153-178.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic research*, *4*(1), 75-87.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of experimental psychology: Learning, memory, and cognition*, *30*(3), 687.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of memory and language*, *47*(1), 30-49.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of memory and language*, *50*(1), 62-81.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, *13*(2), 259-294.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention Orienting Effects of Hesitations in Speech: Evidence From ERPs. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(3), 696–702.

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

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247.

Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS ONE*, *6*(5).

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, *104*(3), 268.

Cree, G. S., & McRae, K. (2003). Analyzing the Factors Underlying the Structure and Computation of the Meaning of Chipmunk, Cherry, Chisel, Cheese, and Cello (and many Other Such Concrete Nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201.

Croucher, S. M. (2004). I uh know what like you are saying: An analysis of discourse markers in limited preparation events. *National Forensics Journal*, *41*(4), 38-52.

Croucher, S. M. (2004). Like, you know, what I'm saying: A study of discourse marker frequency in extemporaneous and impromptu speaking. *National Forensic Journal*, *22*(2), 38-47.

Cutler, R. A., Duff, M. C., & Polyn, S. M. (2019). Searching for semantic knowledge: a vector space semantic analysis of the feature generation task. *Frontiers in human neuroscience*, *13*, 341.

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. In *Psychology of learning and motivation* (Vol. 59, pp. 43-95). Academic Press.

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology*, *54*(3), 180–187.

Dessalegn, B., & Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. *Psychological Science*, *19*(2), 189–195.

Diachek, E., & Brown-Schmidt, S. (2022). The effect of disfluency on memory for what was said. *Journal of experimental psychology. Learning, memory, and cognition*, 10.1037/xlm0001156. Advance online publication.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, *24*(6), 409-436.

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 99.

Firth, J. R. (2020). Papers in linguistics, 1934-1951.

Fischer, N. M., Schult, J. C., & Steffens, M. C. (2015). Source and destination memory in face-to-face interaction: A multinomial modeling approach. *Journal of Experimental Psychology: Applied*, *21*(2), 195.

Fisher, R. P., & Craik, F. I. (1980). The effects of elaboration on recognition memory. *Memory & Cognition*, *8*(5), 400-404.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.

Fox Tree, J. E. F., & Mayer, S. A. (2008). Overhearing single and multiple perspectives. *Discourse Processes*, *45*(2), 160–179.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175.

Fraundorf, S. H., & Watson, D. G. (2014). Alice's adventures in um-derland: Psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, *29*(9), 1083–1096.

Freund, R. D. (1972). *Verbal and non-verbal processes in picture recognition.*

Fries, C. C. (1952). *The structure of English*. Longmans.

Gallo, D. A., Meadow, N. G., Johnson, E. L., & Foster, K. T. (2008). Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language*, *58*(4), 1095-1111.

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child development*, *82*(4), 1173-1188.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 1-13.

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162.

Healey, P. G., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running repairs: Coordinating meaning in dialogue. *Topics in cognitive science*, *10*(2), 367-388.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831-836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104-120.

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

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96-101.

Hintzman, D. L. (1986). " Schema abstraction" in a multiple-trace memory model. *Psychological review*, *93*(4), 411.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528.

Hjelmquist, E., & Gidlund, Å. (1985). Free recall of conversations. *Text-Interdisciplinary Journal for the Study of Discourse*, *5*(3), 169-186.

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and cognitive processes*, *26*(3), 376-405.

Isaacs, E. A. (1990). *Mutual memory for conversation* (Doctoral dissertation, Stanford University).

Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*(2), 119-136.

Jamieson, R. K., Johns, B. T., Taler, V., & Jones, M. N. (2022). The importance of formal modelling for the development of cognitive theory. *Bilingualism: Language and Cognition*, *25*(2), 218–219.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063, pp. 218-226). Englewood Cliffs, NJ: Prentice-Hall.

Karimi, H., & Ferreira, F. (2016). Informativity renders a referent more accessible: Evidence from eyetracking. *Psychonomic bulletin & review*, 23, 507-525.

Keenan, J. M., MacWhinney, B., & Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of verbal learning and verbal behavior*, *16*(5), 549-560.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63.

Kim, J. S., Aheimer, B., Manrara, V. M., & Bedny, M. (2020). Shared understanding of color among congenitally blind and sighted adults.

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *116*(23), 11213-11222.

Kintsch, W., & Bates, E. (1977). Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(2), 150.

Krauss, R. M., Garlock, C. M., Bricker, P. D., & McMahon, L. E. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of personality and social psychology*, *35*(7), 523.

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.

Lang, P., & Bradley, M. M. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, *29*, 70-73.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. *Concepts: core readings*, *3*, 81.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83(*3), 265-294.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 397.

Lord, K., & Brown-Schmidt, S. (2022). Temporary ambiguity and memory for the context of spoken language. *Psychonomic Bulletin & Review*, *29*(4), 1440-1450.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*(2), 203–208.

Lupyan, G. (2008). From chair to" chair": a representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, *4*, 226.

MacDonald, M. C., & Christiansen, M. H. (2002). *Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996).*

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, *111*(1), 36–45.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Disfluent silent pauses in speech have consequences for listeners. *Neuropsychologia*, *48*(14), 3982–3992.

Mahon, B. Z., & Caramazza, A. (2003). Constraining questions about the organization and representation of conceptual knowledge. *Cognitive Neuropsychology*, *20*(3–6), 433–450.

Marks, W. (1987). Retrieval constraints on associative elaborations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 301.

McKinley, G. L., Brown-Schmidt, S., & Benjamin, A. S. (2017). Memory for conversation and the development of common ground. *Memory & cognition*, *45*(8), 1281-1294.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547-559.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the Nature and Scope of Featural Representations of Word Meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201-213.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Miller, J. B., deWinstanley, P., & Carey, P. (1996). Memory for conversation. *Memory*, *4*(6), 615-632.

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of memory and language*, *86*, 119-140.

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482–488.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098-1111.

Nieuwland, M. S., Otten, M., & Van Berkum, J. J. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *Journal of cognitive neuroscience*, *19*(2), 228-236.

Nosofsky, R. M., Palmeri, T. J., & Nosofsky, R. (2015). An exemplar-based random-walk model of categorization and recognition. *Oxford library of psychology. The Oxford handbook of computational and mathematical psychology*, 142-164.

O'Brien, E. J., & Myers, J. L. (1985). When comprehension difficulty improves memory for text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 12.

O'Keeffe, A., & Adolphs, S. (2008). *Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse*.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3).

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.

Pechmann, T. (1989). Incremental speech production and referential overspecification.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. Cognitive Neuropsychology, 33(3–4), 175–190.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 1-13.

Pezdek, K., & Prull, M. (1993). Fallacies in memory for conversations: Reflections on Clarence Thomas, Anita Hill, and the Like. *Applied Cognitive Psychology*, *7*(4), 299-310.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological review*, *116*(1), 129.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). Task context and organization in free recall. *Neuropsychologia*, *47*(11), 2158–2163.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Roque, A., & Traum, D. (2008, June). Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 54-63).

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of personality and social psychology*, *37*(3), 322.

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations (MIT Press, Cambridge, Massachusetts, 1986).

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*(5), 898.

Ryskin, R. A., Qi, Z., Duff, M. C., & Brown-Schmidt, S. (2017). Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 781.

Sachs, J. S. (1974). Memory in reading and listening to discourse. *Memory & Cognition*, *2*(1), 95-100.

Samp, J. A., & Humphreys, L. R. (2007). "I said what?" Partner familiarity, resistance, and the accuracy of conversational recall. *Communication Monographs*, *74*(4), 561-581.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing Discourse: Text and Talk*, *71*, 71–93.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22*(1), 36–71.

Scott, N. M., & Sera, M. D. (2018). Language unifies relational coding: The roles of label acquisition and accessibility in making flexible relational judgments. *Journal of Memory and Language*, *101*, 136–152.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, *32*(1), 3-23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109-147.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of Experimental Psychology: Human Learning and Memory, 4(6), 592–604.

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

Stafford, L., & Daly, J. A. (1984). Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversations. *Human Communication Research*, *10*(3), 379-402.

Stafford, L., Burggraf, C. S., & Sharkey, W. F. (1987). Conversational memory: The effects of time, recall, mode, and memory expectancies on remembrances of natural conversations. *Human Communication Research*, *14*(2), 203-229.

Stafford, L., Waldron, V. R., & Infield, L. L. (1989). Actor-observer differences in conversational memory. *Human Communication Research*, *15*(4), 590-611.

Szewczuk, W. (1970). *The role of verbalization in the retention of nonverbal material*. Państwowe Wydaw. Naukowe.

Tagliamonte, S. (2005). So who? Like how? Just what?: Discourse markers in the conversations of young Canadians. *Journal of Pragmatics*, *37*(11), 1896-1915.

Toftness, A. R., Carpenter, S. K., Lauber, S., & Mickes, L. (2018). The limited effects of prequestions on learning from authentic lecture videos. *Journal of Applied Research in Memory and Cognition*, *7*(3), 370–378.

Troyer, M., Hofmeister, P., & Kutas, M. (2016). Elaboration over a discourse facilitates retrieval in sentence processing. *Frontiers in psychology*, 7, 374.

Tulving, E. (1972). 12. Episodic and Semantic Memory. *Organization of memory/Eds E. Tulving, W. Donaldson, NY: Academic Press*, 381-403.

Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 607.

Urgolites, Z., Brady, T., & Wood, J. (2020). *Verbal interference suppresses object-scene binding in visual long-term memory*.

van Paridon, J. Liu, Q., & Lupyan, G. (2021). How do blind people know that blue is cold? Distributional semantics encode color-adjective associations.

Villar, G., Arciuli, J., & Mallard, D. (2012). Use of "um" in the deceptive speech of a convicted murderer. *Applied Psycholinguistics*, *33*(1), 83-95.

Voeten, C.C. (2020). Package “buildmer”: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. Retrieved from https://cran.r-project.org/package=buildmer

Wang, Y., & Gennari, S. P. (2019). How language and event recall can shape memory for time. *Cognitive Psychology*, *108*, 1–21.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063.

Watzlawick, P. (1964). *An Anthology of Human Communication...* Science and Behavior Books.

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*(2), 257–267.

Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, 154, 102-117.

Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2021). Referential form and memory for the discourse history. *Cognitive science*, *45*(4), e12964.

Zhang, X., Yu, H. W., & Barrett, L. F. (2014). How does this make you feel? A comparison of four affect induction procedures. *Frontiers in Psychology*, *5*(JUL), 1–10.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340-352.

APPENDICES

List of unambiguous words used in the construction of the composite semantic evaluation model.

Big	Small	Animate	Inanimate
acrobat	almond	acrobat	acid
actor	ant	actress	aircraft
actress	apple	adolescent	airport
adolescent	aspirin	adult	album
adult	bacon	alligator	alley
africa	bait	antelope	ambulance
agent	bandage	ape	anchor
aircraft	bead	apple	antenna
airplane	bean	architect	apartment
airport	bee	artist	application
alley	berry	assistant	apron
ambulance	bible	astronaut	article
ancestor	bluejay	athlete	ashtray
antelope	bracelet	audience	atlas
antler	broccoli	author	attic
apartment	bruise	ballerina	automobile
ape	bubble	bartender	award
arena	buckle	bear	badge
army	bug	beaver	bag
artist	butter	beggar	balcony
asia	butterfly	biologist	ball
assistant	button	bird	balloon
astronaut	camera	boy	ballot
atmosphere	candle	boyfriend	bandage
attorney	card	brother	barn
audience	cardinal	bull	baseball
aunt	carrot	burglar	basement
author	cashew	butcher	basket
automobile	caterpillar	butler	basketball
baker	cent	butterfly	bassinet
ballerina	chalk	camel	bath
bandit	charcoal	canary	bathroom
bank	checkers	candidate	bathtub
banker	cheddar	captain	battery
barn	cheek	carpenter	bay
bartender	chemical	cat	beach
bay	cherry	cheerleader	bedroom
beach	chip	chef	beer
bedroom	chocolate	child	belt
beggar	cinnamon	chimpanzee	bench

bicycle	clove	climber	beverage
bike	coal	cobra	bicycle
biologist	cocktail	colonel	bill
bison	coin	comedian	biscuit
blackboard	coleslaw	companion	blackboard
blockade	collar	consumer	blanket
boat	compass	cousin	blockade
body	cookie	cow	blueprint
booth	cork	cowboy	board
boss	cotton	creature	boat
boy	cream	cricket	bolt
boyfriend	crumb	criminal	bomb
bridge	crystal	crocodile	book
brother	cue	crow	boot
brunette	cuff	customer	booth
buffalo	cup	dad	bottle
building	daisy	dancer	bouillon
bull	dandruff	deer	boulder
bully	diamond	dentist	boulevard
bureau	diaper	dictator	bowl
bus	dice	doctor	box
camel	dime	driver	bracelet
canoe	dollar	eagle	brake
canvas	doorbell	electrician	brandy
canyon	dough	elephant	brick
capital	drug	elk	bridge
captive	dust	emperor	brook
car	ear	employee	broom
caravan	earring	employer	brush
carnival	egg	farmer	buckle
carpenter	electron	father	buggy
carriage	envelope	fireman	building
cashier	eyelash	fish	bulb
castle	feather	flower	bulletin
cathedral	fig	friend	bun
cattle	finger	frog	bureau
ceiling	fingernail	gentleman	bus
cellar	fish	girl	button
champion	fist	goose	cabin
chapel	flask	gorilla	cafe
chauffeur	flea	grasshopper	cafeteria
cheerleader	flower	guest	cage
chef	fly	gymnast	cake
chemist	foot	hawk	calculator
chief	fragrance	hen	calendar
church	freckle	hornet	camera

citizen	fries	horse	can
clerk	frost	hostess	canal
cliff	garlic	hound	candle
climber	gem	husband	cane
closet	gene	infant	cannon
coach	germ	instructor	canoe
college	gin	inventor	canvas
colonel	glasses	kid	cap
comet	grape	lady	cape
commander	gum	leader	caravan
community	hand	lion	card
computer	heel	lover	carpet
concert	honey	mailman	cart
conductor	jar	man	carton
consumer	jello	manager	cash
contractor	jewel	mayor	casket
convent	key	miner	castle
cook	kitten	mob	cathedral
cooler	label	mongoose	cave
cop	lace	monk	cellar
copier	leaf	monkey	cello
corporation	lemon	moth	cemetery
couch	lens	mother	cent
country	lime	mouse	chain
county	lint	mule	chalk
cow	lipstick	navigator	chamber
cowboy	lizard	nephew	champagne
criminal	lock	niece	charcoal
critic	lollipop	nun	check
cupboard	loop	nurse	checkers
curtain	magnet	octopus	chime
cyclone	mascara	officer	chimney
dad	match	otter	chisel
dam	mint	outlaw	church
daughter	mitten	owl	cigar
dentist	molecule	ox	cigarette
department	money	oyster	cinnamon
designer	mosquito	parent	clay
detective	moss	parrot	cliff
dictator	moth	partner	clippers
dinosaur	mouse	patient	closet
dishwasher	mouth	pedestrian	clothes
diver	nail	pelican	coal
doctor	napkin	penguin	cobweb
dolphin	necklace	person	coffin
donkey	needle	philosopher	coin

door	nitrogen	pig	coleslaw
dorm	nose	pirate	cologne
dragon	note	plumber	column
driver	novel	poet	compass
dryer	nucleus	politician	computer
dungeon	nut	pony	cone
earth	ointment	preacher	contract
editor	olive	president	convent
egypt	ornament	priest	cookbook
electrician	peanut	prince	cookie
elephant	pear	princess	cooler
elk	pearl	prisoner	copier
emperor	pedal	producer	cord
empire	pen	professional	cottage
employee	penny	puppy	couch
employer	pill	quail	court
engineer	pimple	queen	cracker
escalator	pin	rabbit	crater
europe	plaque	raccoon	crayon
factory	pocket	referee	crevice
family	poison	reptile	crown
farm	popcorn	robber	crutch
farmer	proton	roommate	cube
father	prune	rooster	cuff
field	puck	rose	cup
fighter	quarter	runner	cupboard
fleet	raisin	sailor	curb
florida	rat	salesman	cushion
forest	razor	salmon	custard
fort	ribbon	scallop	cyclone
fountain	ring	secretary	cylinder
france	salt	sergeant	dagger
freeway	sand	serpent	dam
friend	sapphire	shark	dart
furniture	saucer	sheep	dashboard
galaxy	sausage	shepherd	deck
gang	screw	sibling	denim
gangster	seed	sister	deodorant
garage	shoe	snake	desk
garden	shoelace	son	dessert
general	shrimp	spider	detergent
gentleman	signature	spouse	diagram
giraffe	slime	stewardess	dial
girl	slug	stranger	diamond
gorilla	snack	student	diary
governor	soap	surgeon	dice

graduate	sock	swan	dime
grave	spice	swimmer	diner
groom	spider	teacher	dinner
guard	sponge	teenager	diploma
guardian	spool	termite	disc
gym	staple	thief	dish
gymnast	straw	toad	dock
haystack	strawberry	tortoise	doll
helicopter	string	tourist	dollar
herd	syringe	traitor	doorbell
hero	tack	turtle	dough
highway	tag	typist	drawer
hiker	tangerine	uncle	dress
horse	tart	victor	drink
hospital	tea	visitor	driveway
house	thermometer	waiter	drug
human	thimble	waitress	dryer
hurricane	thorn	walrus	dune
iceberg	thumb	warrior	dungeon
igloo	tick	whale	dustpan
inmate	ticket	winner	earring
instructor	toad	witness	elevator
inventor	toast	wolf	encyclopedia
island	toe	woman	engine
jeep	tomato	zebra	eraser
jet	toothbrush		escalator
judge	toothpaste		essay
jungle	trigger		explosion
jupiter	tulip		factory
kangaroo	turnip		feast
keeper	tweezers		feather
king	twig		fiddle
kitchen	virus		fireplace
lady	vitamin		flag
landscape	wallet		flannel
lawn	wasp		flashlight
lawyer	wax		flask
leader	wick		fleet
leopard	wire		floor
lieutenant	worm		flour
limousine	wound		fort
lion	wrench		fossil
lodge	wrist		fragrance
london	yolk		freeway
lounge			fudge
lover			funeral

magician			fur
man			furniture
manager			gallon
mansion			garage
mars			garbage
mattress			gauze
meteor			gavel
microwave			gin
military			glacier
mister			glass
moat			glasses
mob			glue
monster			gold
moon			gown
moose			grave
mother			gravel
motorcycle			grease
mountain			grill
museum			ground
neptune			hail
newsstand			hammer
nun			hammock
ocean			hamper
office			handbag
officer			handcuffs
opera			hanger
orchestra			hatchet
outdoors			haystack
owner			heater
painter			helmet
palace			hoe
parent			hood
paris			hook
partner			hoop
party			horizon
passenger			hospital
path			hurricane
patient			hut
patriot			igloo
pavement			incense
pedestrian			inn
people			iron
person			item
philosopher			jacket
piano			jar
picnic			jeans

pirate			jello
planet			jelly
playground			jewel
plumber			journal
pluto			jug
police			keg
politician			kettle
pony			key
pool			keyboard
pope			kitchen
prairie			kite
preacher			kleenex
predator			knapsack
president			knife
priest			knob
primate			knot
prince			labyrinth
prison			lace
producer			lamp
professor			lash
pub			letter
publisher			lightning
queen			linen
radiator			lint
raft			literature
railroad			lock
ram			lodge
rebel			lollipop
receptionist			lounge
referee			luggage
refrigerator			lunch
reindeer			macaroni
resort			magazine
restaurant			magnet
river			mailbox
road			mall
robber			marble
robot			marker
roof			mask
room			mat
roommate			match
runner			mattress
sailor			mayonnaise
salesman			medal
saturn			medication
scientist			medicine

seashore			meteor
senate			microphone
senator			microscope
servant			mirror
shark			missile
shed			mitten
sheep			monument
shelter			moon
shepherd			mop
sheriff			motel
ship			motor
shore			motorcycle
shrine			mug
sibling			nail
sister			napkin
skeleton			needle
slope			net
society			newspaper
soldier			newsstand
spouse			nickel
stable			nicotine
stairs			nightgown
stallion			nitrogen
statue			notebook
store			oboe
stranger			office
stream			ointment
street			ornament
student			outfit
submarine			oval
suburb			oven
sun			pad
supermarket			paddle
supervisor			pail
suspect			paint
sword			painting
tank			palace
tavern			pan
taxi			pants
teacher			paper
team			parcel
technician			passage
temple			pasta
territory			path
tiger			patio
toilet			pavement

tornado tower town tractor traitor tree tribe tricycle trombone tunnel umpire uncle unicorn universe university van vehicle venus villain visitor volcano volunteer waiter waitress wall walrus warehouse warrior waterfall well whale wife winner wolf woman worker world yacht yard zoo			pedal pen pencil penny pepper perfume periscope phone pick pill pipe pistol pit pitchfork plaid plaster plate plaza pliers pocket pocketbook poison polyester pool port portrait pot pottery powder pub puck pudding pump puzzle quill racket radiator radio raft rag railroad rake razor receipt recipe record

			refrigerator relish report restaurant rifle ring road robe rock rocket roof room roost ruby rum saddle saloon salt sand sandwich sapphire saturn saucer scale scalpel scissors scotch screen screw screwdriver scribble sculpture seat shack shampoo shears shed shelf ship shirt shoe shoelace shop shortcake shovel shutter

			sickle sidewalk siding sign signature sink sketch ski skyscraper slacks sleeve slime sliver slope snack snorkel soap sock sofa spatula spit spoon stage stairs stake stamp stapler step stereo stethoscope sticker stocking stone stool stove straw street string submarine suit suite sunrise sunset supermarket supper survey

			swing switch table tack tag tank tape taxi teapot telephone telescope temple thermometer thimble tie tile toilet tool toothbrush toothpaste torch towel toy tractor train trash tray tread treasure treat trench triangle tricycle trophy truck trumpet tub tunnel twine typewriter umbrella underwear uniform vacuum van vehicle

			velvet vent venus vinegar viola violin volleyball wagon wall wallet wand wardrobe wave wax well wheel whip whistle wick windshield xerox yacht yarn