Toward Forgetting-Sensitive Referring Expression Generation for Integrated Robot Architectures

Tom Williams TWILLIAMS@MINES.EDU
Torin Johnson TAJOHNSON@MINES.EDU
Will Culpepper WCULPEPPER@MINES.EDU
Kellyn Larson KELLYNRLARSON@GMAIL.COM
MIRRORLab, Colorado School of Mines, Golden CO, USA

Abstract

To engage in human-like dialogue, robots require the ability to describe the objects, locations, and people in their environment, a capability known as "Referring Expression Generation." As speakers repeatedly refer to similar objects, they tend to re-use properties from previous descriptions, in part to help the listener, and in part due to cognitive availability of those properties in working memory (WM). Because different theories of working memory "forgetting" necessarily lead to differences in cognitive availability, we hypothesize that they will similarly result in generation of different referring expressions. To design effective intelligent agents, it is thus necessary to determine how different models of forgetting may be differentially effective at producing natural human-like referring expressions. In this work, we computationalize two candidate models of working memory forgetting within a robot cognitive architecture, and demonstrate how they lead to cognitive availability-based differences in generated referring expressions.

1. Introduction

Effective human-robot interaction requires human-like natural language and dialogue capabilities that are sensitive to robots' embodied nature and inherently situated context (Mavridis, 2015; Tellex et al., 2020). In this paper we explore the role that models of Working Memory can play in enabling such capabilities in integrated robot architectures. While Working Memory has long been understood to be a core feature of human cognition, and thus a central component of cognitive architectures, recent evidence from psychology suggests a conception of working memory that is subtly different from what is implemented in most cognitive architectures. Specifically, while most models of working memory in computational cognitive architectures maintain a single central working memory store, Converging evidence from different communities suggests that humans have different resource limitations for different types of information. Moreover, recent psychological evidence suggests that Working Memory may be a limited resource pool, with resources consumed on the basis of the number and type of features retained. This suggests that *forgetting* in Working should be modeled in cognitive architectures as a matter of systematic removal (on the basis of decay or interference) of entity *features* with sensitivity to the resource limitations imposed for the specific *type* of information represented by that feature. Of course robot cognition need not directly mirror human cognition, and indeed robots have both unique knowledge representational needs and increased flexibility in how resource limitations are implemented. In this work, we present a robot architecture in which (1) independent resource pools are maintained for different robot-oriented types of information; (2)

WM resources are maintained at the feature level rather than entity level; and (3) both interference- and decay-based forgetting procedures may be used. This architecture is flexibly configurable both in terms of what type of forgetting procedure is used, and how that model is parameterized. For robot designers, this choice of parameterization may be made in part on the basis of facilitation of interaction. In this paper we specifically consider how the use of different models of forgetting within this architecture lead to different information being retained in working memory, which in turn leads to different referring expressions being generated by the robot, which in turn can produce *interactive alignment* effects purely through Working Memory dynamics. While in future work it will be important to identify exactly which parameterizations lead to selection of referring expressions that are optimal for effective human-robot interaction and teaming, in this work we take the critical first step of demonstrating, as a proof-of-concept, that decay- and interference-based forgetting mechanisms can be flexibly used within this architecture, and that those policies do indeed produce different natural language generation behavior.

2. Referring

2.1 Models of Referring Expression Generation

"Referring" has been referred to as the "fruit fly" of language due to the amount of research it has attracted (Van Deemter, 2016; Gundel & Abbott, 2019). In this work, we focus specifically on Referring Expression Generation (REG) (Reiter & Dale, 1997) in which a speaker must choose words or phrases that will allow the listener to uniquely identify the speaker's intended referent. REG includes two constituent sub-problems (Gatt & Krahmer, 2018): referring form selection and referential content determination. While referring form selection (in which the speaker decides whether to use a definite, indefinite, or pronominal form (Poesio et al., 2004; McCoy & Strube, 1999) (see also Pal et al., 2020)) has attracted relatively little attention, referential content determination is one of the most well-explored sub-problems within Natural Language Generation, in part due to the logical nature of the problem that enables it to be studied in isolation, to the point where "REG" is typically used to refer to the referential content determination phase alone. In this section we will briefly define and describe the general strategies that have been taken in the computational modelling of referential content determination; for a more complete account we recommend the recent book by Van Deemter (2016), which provides a comprehensive account of work on this problem.

Referential content determination, typically employed when generating definite descriptions, is the process by which a speaker seeks to determine a set of constraints on known objects that if communicated will distinguish the target referent from other candidate referents in the speaker and listener's shared environment. These constraints most often include attributes of the target referent, but can also include relationships that hold between the target and other salient entities that can serve as "anchors", as well as attributes of those anchors themselves (Dale & Haddock, 1991).

Three referential content determination models have been particularly influential: the *Full Brevity Algorithm* (Dale, 1989), in which the speaker selects the description of minimum length, in order to straightforwardly satisfy Grice's Maxim of Quantity (Grice, 1975); the *Greedy Algorithm*, in which the speaker incrementally adds to their description whatever property rules out the largest number of distractors(Dale, 1992); and the *Incremental Algorithm (IA)*, in which the speaker incrementally adds properties to their description in order of *preference* so long as they help to rule out distractors (Dale & Reiter, 1995). A key aspect of the IA is its ability to *overspecify* through its inclusion of properties that are not strictly needed from a logical perspective to single out the target referent, but are nevertheless included due to being highly preferred; a behavior also observed in human speakers (Engelhardt et al., 2006).

Because the IA's behavior is highly sensitive to preference ordering (Gatt et al., 2007), there has been substantial research seeking to determine what properties are in general psycholinguistically preferred over others (Belke & Meyer, 2002), or to automatically learn optimal preference orderings (Koolen et al., 2012). As highlighted by Goudbeek & Krahmer (2012), however, this focus on a uniform concept of "preference" obscures a much more complex story that relates to fundamental debates over the extent to which speakers leverage listener knowledge during sentence production. A notion of "preference" as encoded in the IA could be egocentrically grounded, with speakers "prefer" concepts that are easy for themselves to assess or cognitively available to themselves (Keysar et al., 1998); it could be allocentrically grounded, with speakers intentionally seeking to facilitate the listener's ease of processing (Janarthanam & Lemon, 2009); or a hybrid model could be used, in which egocentric and allocentric processes compete (Bard et al., 2004), with egocentrism vs. allocentrism "winning out" on the basis of factors such as cognitive load (Fukumura & van Gompel, 2012). These approaches, which require accounting for listener knowledge to be slow and intentional, stand in contrast to memory-oriented account of referential content determination in which such accounting can naturally occur as a result of priming.

2.2 Memory-Oriented Models of Referring Expression Generation

Pickering & Garrod (2004)'s *Interactive Alignment* model of dialogue suggests that dialogue is a highly negotiated process (see also Clark & Wilkes-Gibbs (1986)) in which priming mechanisms lead interlocutors to influence each others' linguistic choices at the phonetic, lexical, syntactic and semantic levels, through mutual activation of phonetic, lexical, syntactic, and semantic structures and mental representations, as in the case of lexical entrainment (Brennan & Clark, 1996).

While there has been extensive evidence for lexical and syntactic priming, research on semantic or conceptual priming in dialogue has only relatively recently become a target of substantial investigation (Gatt et al., 2011). A theory of dialogue including semantic or conceptual priming would suggest that the properties or attributes that speakers choose to highlight in their referring expressions (e.g., when a speaker chooses to refer to an object as "the large red ball" rather than "the sphere") should be due in part to these priming effects. And in fact, as demonstrated by Goudbeek & Krahmer (2010), speakers can in fact be influenced through priming to use attributes in their referring expressions that would otherwise have been dispreferred.

These findings have motivated dual-route computational models of dialogue (Gatt et al., 2011; Goudbeek & Krahmer, 2011) in which the properties used for referring expression selection are made on the basis of interaction between two parallel processes, each of which is periodically called upon to provide attributes of the target referent to be placed into a WM buffer that is consulted when properties are needed for RE generation (at which point selected properties are removed from that buffer). The first of these processes is a priming-based procedure in which incoming descriptions trigger changes in activation within a spreading activation model, and properties are selected if they are the highest-activation properties (above a certain threshold) for the target referent. The second of these processes is a preference-based procedure in which a set of properties is generated by a classic REG algorithm (cp. Gatt & Krahmer, 2018) such as the Incremental Algorithm, in which properties are incrementally selected according to a pre-established preference ordering designed or learned to reflect frequency of use, ease of cognitive or perceptual assessability, or some other informative metric (Dale & Reiter, 1995).

One advantage of this type of dual process model is that it accounts for audience design effects (in which speaker utterances at least appear to be intentionally crafted for ease-of-comprehension) within an egocentric framework, by demonstrating how priming influences on WM can themselves account for listener-friendly referring expressions. That is, if a concept is primed by an interlocutor's

utterance, a speaker will be more likely to use that concept in their own references simply because it is in WM, with the side effect that that property will then be easy to process by the interlocutor responsible for its' inclusion in WM in the first place (Vogels et al., 2015). Moreover, this phenomenon aligns well with evidence suggesting that despite the prevalence of lexical entrainment and alignment effects, people are actually slow to explicitly take others' perspectives into account (Bard et al., 2000; Fukumura & van Gompel, 2012; Gann & Barr, 2014).

Another advantage of this type of dual process model is its alignment with overarching dual-process theories of cognition (e.g., Kahneman, 2011; Evans & Stanovich, 2013; Sloman, 1996; Oberauer, 2009): the first priming-driven process for populating WM, grounded in semantic activation, can be viewed as a reflexive System One process, whereas the second preference-driven process leveraging the Incremental Algorithm can be viewed as a deliberative System Two process. Of course in the model under discussion the two routes do not truly compete with each other or operate on different time courses, but are instead essentially sampled between; however, it is straightforward to imagine how the two processes used in this type of model could be instead deployed in parallel.

One major disadvantage of this type of model, however, is that its focus with respect to WM is entirely on retrieval (i.e., how priming and preference-based considerations impact what information is retrieved from long-term memory into WM), and fails to satisfactorily account for maintenance within WM. Within Gatt et al. (2011)'s model, as soon as a property stored in WM is used in a description, it is removed from WM so that that space is available for another property to be considered. This behavior seems counter-intuitive, as it ensures that representations are removed from WM at just the time when it is established to be important and useful, which should be a cue to retain said representations rather than dispose of them.

Moreover, this model is surprisingly organized from the perspective of models of WM such as Cowan (2001)'s, in which WM is comprised of the set of all activated representations, of which a small subset (e.g., three or four) are maintained in the focus of attention. In Gatt et al. (2011)'s model, in contrast, activated representations are used as just one source populating WM, and decaying activation within the spreading activation network results in representations losing activated status without also being removed from WM. This suggests that the WM buffer within Gatt et al. (2011)'s model may in fact be better conceptualized as a model of the focus of attention (an interpretation also justified by the two-item size limitation of their WM buffer) than as a model of WM.

A final complication for this model is its speaker-blind handling of priming. Specifically, within Gatt et al. (2011)'s model a speaker's utterances are only primed by their interlocutor's utterances, when in fact the choices a speaker makes should also impact the choices they themselves make in the future (Shintel & Keysar, 2007), either due to Gricean notions of cooperativity (Grice, 1975) or, as we argue, because a speaker's decision to refer to a referent using a particular attribute should make that attribute more cognitively available to themselves in the immediate future.

These concerns are addressed by our previously proposed model of robotic short-term memory (Williams et al., 2018b), in which speakers rely on the contents of WM for initial attribute selection and, if their selected referring expression is not fully discriminating, select additional properties using a variant of the IA. While this does not align with dual-process models of cognition, it does account for both encoding and maintenance of WM, and provides a potentially more cognitively plausible account of REG with respect to WM dynamics. One shortcoming shared by both models, however, is neither the dual-process model of Gatt et al. (2011) nor our WM-focused model appropriately account for when and how information is removed from WM over time, or how this impacts REG.

Different theories of WM "forgetting" necessarily lead to predicted differences in cognitive availability. Accordingly, these different models of forgetting should similarly predict cognitive availability-based differences in the properties selected during REG. To design effective intelligent agents, it is

thus necessary to determine how different models of forgetting may be differentially effective at producing natural human-like referring expressions.

In this work, we first computationalize two candidate models of WM forgetting within a robot cognitive architecture. Next, we propose a model of REG that is sensitive to the WM dynamics of encoding, retrieval, maintenance, *and forgetting*, discuss the particulars of deploying this type of model within an integrated robot architecture, where WM resources are divided by domain (i.e., people, locations, and objects) rather than by modality (i.e., visual vs. verbal). Finally, we provide a proof-of-concept demonstration of two parametrizations of our model into an integrated robot cognitive architecture, and demonstrate how these different parametrizations lead to cognitive availability-based differences in generated referring expressions.

3. Models of Forgetting in Working Memory

Models of forgetting in Working Memory are typically divided into two broad categories (Reitman, 1971; Jonides et al., 2008; Ricker et al., 2016): decay-based models, and interference-based models.

3.1 Decay-Based Models

Decay-based models of WM (Brown, 1958; Atkinson & Shiffrin, 1968), that time plays a causal role in WM forgetting, with a representation's inclusion in WM made on the basis of a level of activation that gradually decays over time if the represented information is not used or rehearsed. Accordingly, in such models, a piece of information is "forgotten" with respect to WM if it falls below some threshold of activation due to disuse. This model of forgetting is intuitively appealing due to the clear evidence that memory performance decreases over time (Brown, 1958; **?**; Ricker et al., 2016).

Computational Advantages and Limitations: As with Gatt et al., spreading activation networks can be used to elegantly model how activation of representations impacts the rise and fall of activation of semantically related pieces of information. One disadvantage of this approach, however, is that activation levels need to be continuously re-computed for each knowledge representation in memory. While this may be an accurate representation of actual cognitive processing, artificial cognitive systems do not enjoy the massively parallelized architectures enjoyed by biological cognitive systems, meaning that this approach may face severe scaling limitations in practice.

Computational Model: To allow for straightforward comparison with other models of forgetting, we define a simple model of decay that operates on individual representations outside the context of a semantic activation network. We begin by representing WM as a set $WM = Q_0, \ldots, Q_n$, where Q_i is a mental representation of a single entity, represented as a queue of properties currently activated for that entity. Next, we define an encoding procedure that specifies how the representations in WM are created and manipulated on the basis of referring expressions generated either by the agent or its' interlocutors. As shown in Alg. 1, this procedure operates by considering each property included in the referring expression, and updating the queue used to represent the entity being described, by placing that property in the back of the queue, or by moving the property to the back of the queue if it's already included in the representation. Note that this procedure can be used either after each utterance is heard (in which case the representation is updated based on all properties used to describe the entity) or incrementally (in which case the representation is updated after each property is heard). If used incrementally, then forgetting procedures may be interleaved with representation updates. Finally, we define a model of decay that operates on these representations. As shown in

Alg. 2, this procedure operates by removing the property at the front of queue Q at fixed intervals defined by decay parameter δ.

Algorithm 1 Per-entity encoding model

```
1: procedure ENCODE(R, P, Q)
2:     R: the object being described
3:     P: the set of properties being used by the speaker or hearer to describe R
4:     Q: set of per-entity property queues.
5:     for all p ∈ P do
6:         Q[R] = Q[R] \ p
7:         push(Q[R], p)
8:     end for
9: end procedure
```

Algorithm 2 Per-entity decay model.

```
procedure PERIODICDECAY(Q, δ)
    Q: the per-entity property queue
    δ: decay parameter
    repeat                                          ▷ Every δ seconds
        pop(Q)
    until Q = ∅
end procedure
```

3.2 Interference-Based Models

In contrast, interference-based models (Waugh & Norman, 1965) argue that WM is a fixed-size buffer in which a piece of information is "forgotten" with respect to WM if it is removed (due to, e.g., being the least recently used representation in WM memory) to make room for some new representation.

Interference-Based Models have been popular for nearly as long as decay-based models (Keppel & Underwood, 1962) due to observations that the evidence used as evidence for "decay" can just as easily be used as evidence for forgetting due to intra-representational interference, as longer periods of time directly correlate with higher frequencies of interfering events (Lewandowsky & Oberauer, 2015; Oberauer & Lewandowsky, 2008), and because tasks with varying temporal lengths but consistent overall levels of interference have been shown to yield similar rates of forgetting, thus failing to support time-based decay (Oberauer & Lewandowsky, 2008). Recent work has trended towards interference-based accounts of forgetting, with a number of further debates and competing models opening up within this broad theoretical ground.

First, there is debate as to whether interference *alone* is sufficient to explain forgetting, or whether time-based decay still plays some role in conjunction with interference. Recent work suggests that in fact these two models may be differentially employed for different types of representations, with phonological representations forgotten due to interference and non-phonological representations forgotten due to a combination of interference and time-based decay (Ricker et al., 2016).

Second, within interference-based models, there exist competing models based on reasons for displacement. In particular, while theories of pure displacement (Waugh & Norman, 1965) posit that incoming representations replace maintained representations on the basis of frequency or recency of use, or on the basis of random chance (similar to caching strategies from computing systems

research (Press et al., 2014)), theories of retroactive interference instead posit that replacements are made on the basis of semantic similarity, with representations "forgotten" if they are too similar to incoming representations (Wickelgren, 1965; Lewis, 1996).

Third, within both varieties of interference-based models, there has been recent debate on the structure and organization of the capacity-limited medium of WM. Ma et al. (2014) contrasts four such models: (1) slot models, in which a fixed number of slots are available for storing coherent representations (Miller, 1956; Cowan, 2001); (2) resource models, in which a fixed amount of representational medium can be shared between an unbounded number of representations (with storage of additional features in one representation resulting in fewer feature-representing medium being available for other representations) (Wilken & Ma, 2004); (3) discrete-representation models, in which a fixed number of feature "quanta" are available to distribute across representations (Zhang & Luck, 2008); and (4) variable-precision models, in which working memory precision is statistically distributed (Fougnie et al., 2012).

Computational Advantages and Limitations: One advantage of interference-based models for artificial cognitive agents is decreased computational expense, as only a fixed number of entities or features must be maintained in WM, and WM need not be updated at continuous intervals if no new stimuli are processed. Rather, WM only needs to be updated when (1) new representations are encoded into WM, or (2) existing representations are manipulated. Another advantage of this approach is its conceptual alignment with the process of *caching* from computer science, which means that caching mechanisms from computing systems research, such as *least-recently-used* and *least-frequently-used* caching policies, can be straightforwardly leveraged, with prior work providing substantial information about their theoretical properties and guarantees. In fact, recent work has explored precisely how caching strategies from computer science can be used for this purpose (Press et al., 2014)). Within the interference-based family of models, slot-based and discrete-representation models are likely the most easy to computationalize due to the ephemeral and undiscretized nature of "representational medium."

Computational Model: To model interference-based forgetting, we use the same WM representation and encoding procedure as used to model decay-based forgetting, and propose a new model designed specifically for robotic knowledge representation. This model can be characterized as a per-entity discrete-representation displacement model. As shown in Alg. 3, this procedure operates by removing properties at the front of queue Q whenever the size of Q is greater than some size limitation imposed by parameter α. This model is characterized as discrete-representation because resource limitations are imposed at the level of discrete features rather than holistic representations. It is characterized as a displacement model because features are replaced on the basis of a Least-Recently Used (LRU) caching strategy (Knuth, 1997) rather than on the basis of semantic similarity, due to the pragmatic difficulty of assessing the similarity of different categories of properties without mandating the use of architectural features such as well-specified conceptual ontologies (cp. Tenorth & Beetz, 2009; Lemaignan et al., 2010), which may not be available in all robot architectures. This model is characterized as per-entity because resource limitations are imposed locally (i.e., for each entity) rather than globally (i.e., shared across all entities). While this is obviously not a cognitively plausible characteristic, it was selected, as a starting point, to reduce the need for coordination across architectural components and to facilitate improved performance (as entity representations need not compete with each other). However, if desired, it would be straightforward to extend this approach to allow for imposition of global resource constraints.

Algorithm 3 Per-entity discrete representation displacement model.

```
procedure DISPLACEMENT(Q, α)
    Q: the per-entity property queue
    α: maximum buffer size
    loop
        if |Q| > α then
            pop(Q)
        end if
    end loop
end procedure
```

4. Models of Working Memory in Integrated Robot Architectures

Memory modeling has long been a core component of cognitive architectures, due to its central role in cognition (Baxter et al., 2011). The relative attention paid to WM has, however, varied widely between cognitive architectures. ARCADIA, for example, has placed far more emphasis on attention than on memory (Bridewell & Bello, 2016). While ARCADIA does include a visual short term memory component, it is treated as a "potentially infinite storehouse," with consideration of resource constraints left to future work (Bridewell & Bello, 2016).

ACT-R and Soar, in contrast, do place larger emphases on WM. ACT-R did not originally have explicit WM buffers, instead implicitly representing WM as the subset of LTM with activation above some particular level (Anderson et al., 1996), with "forgetting" thus modeled through activation decay (cp. Cowan, 2001). In more recent incarnations of ACT-R, a very small short-term buffer is maintained, with contents retrieved from LTM on the basis of both base-level activation (on the basis of recency and frequency) and informative cues. Similarly, Soar Laird (2012) has long emphasized the role of WM, due to its central focus on problem solving through continuous manipulation of WM (Rosenbloom et al., 1991). And while Soar did not initially represent WM resource limitations (Young & Lewis, 1999), it has also by now long included decay-based mechanisms on at least a subset of WM contents (Chong, 2003; Nuxoll et al., 2004), as well as base-activation and cue-based retrieval methods such as those mentioned above Jones et al. (2016). While the models may be intended *primarily* as models of human cognition, flaws and all, rather than as systems for enabling effective task performance through whatever means necessary regardless of cognitive plausibility, a significant body of work has well demonstrated the utility of these architectures for cognitive robotics (Laird et al., 2012, 2017; Kurup & Lebiere, 2012) and human-robot interaction Trafton et al. (2013).

There has also been significant work in robotics seeking to use insights from psychological research on WM to better inform the design of select components of larger robot cognitive architectures that do not necessarily aspire towards cognitive plausibility. For example, a diverse body of researchers has collaborated on the development and use of the WM Toolkit (Phillips & Noelle, 2005; Gordon & Hall, 2006; Kawamura et al., 2008; Persiani et al., 2018); a software toolkit that maintains pointers to a fixed number of chunks containing arbitrary information. At each timestep, this toolkit proposes a new set of chunks, and then uses neural networks to select a subset of these chunks to retain. This model has been primarily used for enabling *cognitive control*, in which the link between robot perception and action is modulated by a learned memory management policy.

Models of WM have also been leveraged within the field of Human-Robot Interaction. Broz et al. (2012), for example, specifically model the episodic buffer sub-component of WM (Baddeley, 2000). Baxter et al. (2013) leverages models of WM to better enable non-linguistic social interaction

through alignment, similar to our approach in this work. Researchers have also leveraged models of WM to facilitate communication. For example, Hawes et al. (2007) leverage a model of WM within the CoSy architecture, with concessions made to accommodate the realistic needs of integrated robot architectures, in which specialized representations are stored and manipulated within distributed, parallelized, domain-specific components (see also Williams & Scheutz, 2016). Similarly, in our own work within the DIARC architecture (Scheutz et al., 2019), we have demonstrated (as we will further discuss in this paper) the use of distributed WM buffers associated with such architectural components (Williams et al., 2018b), as well as hierarchical models of common ground (Williams & Scheutz, 2019) jointly inspired by models of WM (e.g. Cowan, 2001) and models of *givenness* from natural language pragmatics (e.g. Gundel et al., 1993). In the next section we propose a new architecture that builds on this previous work of ours to allow for flexible selection (and comparison between) between different models of forgetting.

5. Proposed Architecture

Our forgetting models were integrated into the Distributed, Integrated, Affect, Reflection, Cognition (DIARC) Robot Cognitive architecture: a component-based architecture designed with a focus on robust spoken language understanding and generation (Scheutz et al., 2019). DIARC's Mnemonic and Linguistic components integrate via a *consultant framework* in which different architectural components (e.g., vision, mapping, social cognition) serve as heterogeneous knowledge sources that comprise a distributed model of Long Term Memory (Williams, 2017; Williams & Scheutz, 2016).

These consultants are used by DIARC's natural language pipeline for the purposes of reference resolution and REG. In recent work we have extended this framework to produce a new *Short-Term Memory Augmented* consultant framework in which consultants additionally maintain, for some subset of the entities for which they are responsible, a short term memory buffer of properties that have recently been used by interlocutors to refer to those entities. In this work, we build upon that STM (Short Term Memory)-Augmented Consultant Framework through the introduction of a new architectural component, the WM Manager, which is responsible for implementing the two forgetting strategies introduced in the previous section.

Our model aligns with two key psychological insights. First, converging evidence from different communities suggests that humans have different resource limitations for different types of information (Wickens, 2008), due to either decreased interference between disparate representations (Oberauer et al., 2012) or the use of independent domain-specific resource pools (Baddeley, 1992; Logie & Logie, 1995). Our approach takes a robot-oriented perspective on this second hypothesis, with our use of the WM-Augmented consultant framework resulting in independent resource pools maintained for different types of entities (e.g., objects vs. locations vs. people) rather than for different modalities (e.g., visual vs. auditory) or different codes of processing (e.g., spatial vs. verbal).

Second, while early models of WM suggested that WM resource limitations are bounded to a limited number of chunks (Miller, 1956), more recent models instead suggest that the size of WM is affected by the complexity of those chunks (Mathy & Feldman, 2012), and that maintaining multiple features of a single entity may detract from the total number of maintainable entities, and accordingly, the number of features maintainable for other entities (Alvarez & Cavanagh, 2004; Oberauer et al., 2016; Taylor et al., 2017). Our approach, again, takes a robot-oriented perspective on these models, maintaining WM resources at the feature-level rather than entity-level, while enabling additional flexibility that may not be reflected in human cognition. Specifically, instead of enforcing global resource limits, we allow for flexible selection between decay-based and interference-based (i.e., resource-limited) memory management models, as well as for simultaneous employment of both models, in order to model joint impacts of interference and decay as discussed by Ricker et al.

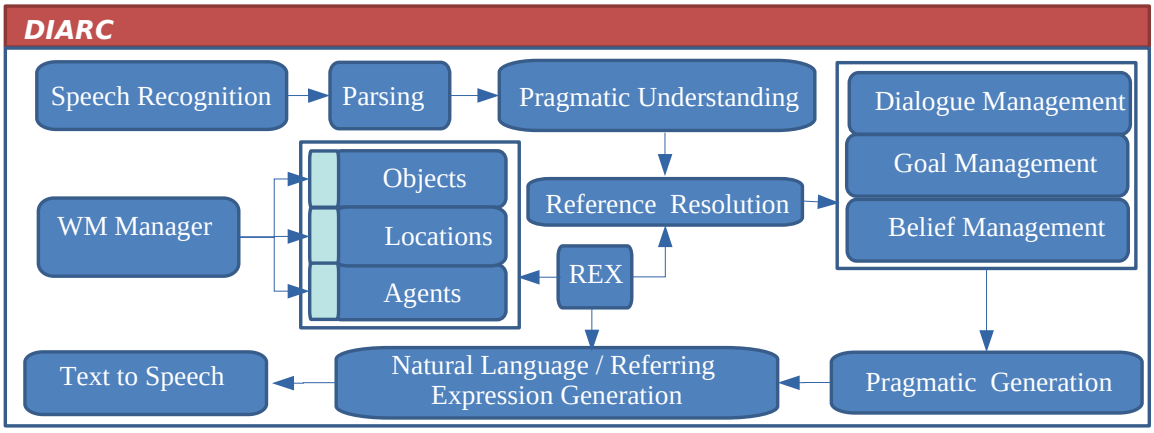

Figure 1. Architecture Diagram

(2016). Moreover, while we currently focus on local (per-entity) feature-based resource limitations, our system is designed to allow for global resource limitations (cp. Just & Carpenter, 1992; Ma et al., 2014) in future work, due to our use of a global WM MANAGER Component.

Using this architecture, our different forgetting models can differentially affect REG without any direct interaction with the REG process. Rather, the WM MANAGER simply interfaces with the WM buffers maintained by each distributed consultant, which then implicitly impacts the properties available to the SD-PIA algorithm that we use for REG. This algorithm takes a lazy approach to REG in which the speaker initially attempts to rely only on the properties that are currently cached in WM, and only if this is insufficient to distinguish the target from distractors does the speaker retrieve additional properties from long-term memory.

Incorporating the WM MANAGER into DIARC yields the configuration shown in Fig. 1. When a teammate speaks to a robot running this configuration, text is recognized and then semantically parsed by the TLDL Parser (Scheutz et al., 2017), after which intention inference is performed by the Pragmatic Understanding Component (Williams et al., 2015), whose results are provided to the Reference Resolution module of the Referential Executive (REX), which leverages the GROWLER algorithm (Williams et al., 2018a) (see also (Williams et al., 2016; Williams & Scheutz, 2019)), which searches through Givenness Hierarchy (GH) (Gundel et al., 1993) informed data structures (Focus, Activated, Familiar) representing a hierarchically organized cache of relevant entity representations. This is important to highlight due to its relation to the WM buffers described in this work. While the WM buffers described in this work serve as a model of the robot's own WM, the Referential Executive's data structures can instead be viewed as either second-order theory-of-mind data structures or as a form of hierarchical common ground.

When particular entities are mentioned by the robot's interlocutor or by the robot itself, pointers to the full representations for these entities (located in the robot's distributed long-term memory) are placed into the Referential Executive's GH-informed data structures, and the properties used to describe them are placed into the robot's distributed WM buffers. In addition, properties are placed into WM whenever the robot affirms that those properties hold through an LTM query. Critically, this can happen when considering other entities. For example, if property P is used during reference resolution, then P will be placed into WM for any distractors for which P is affirmed to hold, before being ruled out for other reasons. Similarly, if R is the target referent during REG, and if property P holds for R and is considered for inclusion in the generated RE, then for any distractors for which

P also holds, P will be added to those distractors' WM buffers at the point where it is affirmed that P cannot be used to rule out those distractors because it holds for them as well. Once Reference Resolution is completed, if the robot decides to respond to the human utterance, it does so through a process that is largely the inverse of language understanding, including a dedicated REG module, which uses the properties cached in WM, along with properties retrievable from Long-Term Memory, to translate this intention into text (Williams & Scheutz, 2017; Williams et al., 2018b).

6. Experimental Setup and Results

Turn	Human's Turn		Robot's Turn			
	Face	Description	Face	Description		
				Decay	Interference	No WM
1	1	H_L, G_M, C_L, G_Y	1	H_L, C_L, G_Y	H_L, C_L, G_Y	H_L, C_L, G_Y
2	2	G_F, G_Y	5	H_D, G_M	H_D, G_M	H_D, G_M
3	3	G_F, C_H, G_N	2	G_F, G_Y	G_F, G_Y	G_F, G_Y
4	4	H_L, H_S, C_L, G_N	1	H_S, C_L, G_Y	H_L, C_L, G_Y	H_L, C_L, G_Y
5	5	H_D, H_S, C_T, G_M	3	G_F, C_H	G_F, C_H	G_F, C_H
6	6	H_S, C_H, G_Y	1	C_L, G_Y	H_S, C_L, G_Y	H_L, C_L, G_Y

Table 1. 6 Face Case Study. Column contents denote properties used to describe target under each forgetting model, listed in the conssitent order below for easy comparison rather than in order selected. Properties: H_L = LIGHT-HAIR(X), H_D = DARK-HAIR(X), H_S = SHORT-HAIR(X), G_M = MALE(X), G_F = FEMALE(X), C_T = T-SHIRT(X), C_L = LAB-COAT(X), C_H = HOODIE(X), G_Y = GLASSES(X), G_N = NO-GLASSES(X).

We analyzed the proposed architecture by assessing two claims: (1) The proposed architecture demonstrates interactive alignment effects purely through WM dynamics; and (2) the proposed architecture demonstrates different referring behaviors when different models of forgetting are selected. These claims were assessed within the context of a "Guess Who"-style game in which partners take turns describing candidate referents (assigned from a set of 16 faces). On each player's turn, they are assigned a referent, and must describe that referent using a referring expression they believe will allow their interlocutor to successfully identify it (a process of REG). The other player must then process their interlocutor's referring expression and identify which candidate referent they believe to be their interlocutor's target (a process of Reference Resolution).

Ideally, we would have assessed our claims in a setting in which a robot played this reference game with a naive human subject. This proved to be impossible due to the COVID-19 global pandemic. Instead, we present a case study in which a series of three six-round reference games are played between robot agents and a single human agent. All three games followed the same predetermined order of candidate referents and used the same pre-determined human utterances. The robot's responses were generated autonomously, with the robot in each of the three games using a different model of forgetting. In the first game, the robot uses our decay-based model of forgetting with $\delta = 10$; in the second game, the robot uses our interference-based model of forgetting with $\alpha = 2$; in the third game, the robot did retain any properties in short-term memory at all.

The referring behavior under each model of forgetting is shown in Tab. 1. As shown in this table, the three examined models perform similarly in initial dialogue turns, but increasingly diverge over time. To help explain the observed differences in robot behavior, we examine specifically turn 6, in which the robot refers to Face 1 for the third time. This face could ostensibly be referred to using four properties: LIGHT-HAIR(X), SHORT HAIR(X), LAB-COAT(X), and GLASSES(X).

The architectural configuration that did not maintain representations in WM (No WM) operated according to the DIST-PIA algorithm (Williams & Scheutz, 2017), which is a version of the Incremental Algorithm that is sensitive to uncertainty and that allows for distributed sources of knowledge. This algorithm first considers the highly preferred property LIGHT-HAIR(X), which is selected because it applies to Face 1 while ruling out distractors. Next, it considers SHORT-HAIR(X), which it ignores because while it applies to Face 1, the faces with short hair are a subset of those with light hair, and thus SHORT-HAIR(X) is not additionally discriminative. Next, the algorithm considers LAB-COAT(X), which it selects because it applies to Face 1 and rules out further distractors. Finally, to complete disambiguation, the algorithm considers and selects GLASSES(X).

In contrast, the configuration that used the decay model had the following properties in WM: {LAB-COAT(X), SHORT-HAIR(X), GLASSES(X)} (ordered from least-recently used to most-recently used[1]). The algorithm starts by considering the properties stored in WM, beginning with LAB-COAT(X), which is selected because it applies to Face 1 while ruling out distractors. Next, it considers SHORT-HAIR(X), which is ignored because the set of entities with short hair is a subset of those wearing lab coats, and thus this is not additionally discriminative. Next, it considers GLASSES(X), which it selects because it applies to Face 1 and rules out distractors. In fact, {LAB-COAT(X), GLASSES(X)} is fully discriminating for Face 1, so no further properties are needed.

Finally, the configuration that used the interference model had the following properties in WM: {SHORT-HAIR(X) GLASSES(X)}. This is easy to see as those properties were recently used in the Human's description of Face 6, and thus would have been considered for Face 1 when ruling it out during reference resolution. The algorithm thus starts by considering both of these properties, which are both selected because they apply to Face 1 and rule out distractors. However, because these are not sufficient for full disambiguation, the algorithm must also retrieve another property from LTM, i.e., LAB-COAT(X), which allows for completion of disambiguation.

The differences in behavior demonstrated in this simple example validate both our claims. First, the proposed architecture's ability to demonstrate interactive alignment effects purely through WM dynamics is demonstrated by the systems' tendency to re-use properties originating from its interlocutor. Second, this example clearly demonstrates that the proposed architecture demonstrates different referring behaviors when different models of forgetting are selected.

7. Conclusions and Future Work

We have presented a flexible set of forgetting mechanisms for integrated cognitive architectures, and conducted a preliminary, proof-of-concept demonstration of these mechanisms, showing that they lead to different referring expressions being generated due to differences in cognitive availability between different properties. The next step of our work will be to fully explore the implications of different parametrizations of each of our presented mechanism, as well as the combined use of these mechanisms, on REG, and whether the referring expressions generated under different parametrizations are comparatively more or less natural, human-like, or effective, which would present obvious benefits for interactive, intelligent robots. In addition, the perspective taken in this paper may also yield insights and benefits for cognitive science more broadly. Specifically, we argue that our perspective may suggest alternative interpretations of the role of cognitive load on attribute choice. In Goudbeek & Krahmer (2011)'s work building on Gatt et al. (2011)'s model, they suggest that when speakers are under high cognitive load, they rely less on previously primed attributes, and are thus more likely to rely on their stable preference orderings. Their explanation for this finding

1. Future work should consider other algorithmic configurations, such as having properties within WM considered in the reverse order, or according to the preference ordering specified by the target referent's consultant.

is that a decrease in available WM capacity leads to an inability to retrieve dialogue context into WM. We suggest an alternative explanation: cognitive load leads to decreased priming not because priming-activated representations cannot be retrieved into WM, but because those representations are less likely to be in WM in the first place. An additional promising direction for future work will thus be to compare the ability of the model presented in this paper to those presented by Goudbeek & Krahmer (2011) with respect to modeling of REG under cognitive load.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*, 106–111.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive psychology*, *30*, 221–256.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*.

Baddeley, A. (1992). Working memory. *Science*, *255*, 556–559.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trend. Cog. Sci.*

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Jour. Memory and Language*.

Bard, E. G., Aylett, M. P., Trueswell, J., & Tanenhaus, M. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. *Approaches to studying world-situated language use*.

Baxter, P. E., de Greeff, J., & Belpaeme, T. (2013). Cognitive architecture for human–robot interaction: Towards behavioural alignment. *Biologically Inspired Cognitive Architectures*, *6*, 30–39.

Baxter, P. E., Wood, R., Morse, A., & Belpaeme, T. (2011). Memory-centred architectures: Perspectives on human-level cognitive competencies. *2011 AAAI Fall Symposium Series*.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different "decisions. *European Journal of Cognitive Psychology*, *14*, 237–266.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*.

Bridewell, W., & Bello, P. (2016). A theory of attention for cognitive systems. *Adv. Cognitive Sys.*.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*, 12–21.

Broz, F., Nehaniv, C. L., Kose-Bagci, H., & Dautenhahn, K. (2012). Interaction histories and short term memory: Enactive development of turn-taking behaviors in a childlike humanoid robot. *arXiv preprint arXiv:1202.5600*.

Chong, R. (2003). The addition of an activation and decay mechanism to the soar architecture. *Proc. of the 5th Intl. Conf. on Cognitive Modeling* (pp. 45–50).

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*, 87–114.

Dale, R. (1989). Cooking up referring expressions. *Proc. Assoc. Computational Linguistics*.

Dale, R. (1992). *Generating referring expressions: Constructing descriptions in a domain of objects and processes.*. The MIT Press.

Dale, R., & Haddock, N. J. (1991). Generating referring expressions involving relations. *Fifth Conference of the European Chapter of the Association for Computational Linguistics*.

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

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*, 554–573.
Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*, 223–241.
Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature communications*, *3*, 1–8.
Fukumura, K., & van Gompel, R. P. (2012). Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? *Cognitive Science*, *36*, 1289–1311.
Gann, T. M., & Barr, D. J. (2014). Speaking from experience: Audience design as expert performance. *Language, Cognition and Neuroscience*, *29*, 744–760.
Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Attribute preference and priming in reference production: Experimental evidence and computational modeling. *Proc. CogSci*.
Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.
Gatt, A., Van Der Sluis, I., & Van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. *Proc. Eur. WS on Natural Language Generation*.
Gordon, S., & Hall, J. (2006). System integration with working memory management for robotic behavior learning. *Proc. 5th Int. Conf. Development and Learning*. Citeseer.
Goudbeek, M., & Krahmer, E. (2010). Preferences versus adaptation during referring expression generation. *Proceedings of the ACL 2010 Conference Short Papers*.
Goudbeek, M., & Krahmer, E. (2011). Referring under load: Disentangling preference-based and alignment-based content selection processes in referring expression generation. *Proc. PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference*.
Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science*, *4*.
Grice, H. P. (1975). Logic and conversation. In *Speech acts*, 41–58. Brill.
Gundel, J., & Abbott, B. (2019). *The oxford handbook of reference*. Oxford University Press.
Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307.
Hawes, N., Sloman, A., Wyatt, J., Zillich, M., Jacobsson, H., Kruijff, G.-J. M., Brenner, M., Berginc, G., & Skocaj, D. (2007). Towards an integrated robot with multiple cognitive functions. *AAAI*.
Janarthanam, S., & Lemon, O. (2009). Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. *Proc. Eur. WS on natural language generation*.
Jones, S. J., Wandzel, A. R., & Laird, J. E. (2016). Efficient computation of spreading activation using lazy evaluation. *Ann Arbor*, *1001*, 48109–2121.
Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol.*, *59*, 193–224.
Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*, 122.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kawamura, K., Gordon, S. M., Ratanaswasd, P., Erdemir, E., & Hall, J. F. (2008). Implementation of cognitive control for a humanoid robot. *International Journal of Humanoid Robotics*, *5*.
Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of verbal learning and verbal behavior*, *1*, 153–161.
Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current directions in psychological science*.
Knuth, D. E. (1997). *The art of computer programming*, volume 3. Pearson Education.

Koolen, R., Krahmer, E., & Theune, M. (2012). Learning preferences for referring expression generation: effects of domain, language and algorithm. *Proc. INLG*.

Kurup, U., & Lebiere, C. (2012). What can cognitive architectures do for robotics? *Biologically Inspired Cognitive Architectures*, *2*, 88–99.

Laird, J. E. (2012). *The soar cognitive architecture*. MIT press.

Laird, J. E., Kinkade, K. R., Mohan, S., & Xu, J. Z. (2012). Cognitive robotics using the Soar cognitive architecture. *Workshops at the twenty-sixth AAAI conference on artificial intelligence*.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*, 13–26.

Lemaignan, S., Ros, R., Mösenlechner, L., Alami, R., & Beetz, M. (2010). Oro, a knowledge management platform for cognitive architectures in robotics. *Proc. IEEE/RSJ IROS*.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the nonexistent problem of decay. *Psychological Review*, *122*, 674.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of psycholinguistic research*, *25*, 93–115.

Logie, R. H., & Logie, R. (1995). *Visuo-spatial working memory*. Psychology Press.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nat. Neuro.*.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, *122*, 346–362.

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, *63*, 22–35.

McCoy, K. F., & Strube, M. (1999). Generating anaphoric expressions: pronoun or definite description? *The Relation of Discourse/Dialogue Structure and Reference*.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*, 81.

Nuxoll, A., Laird, J. E., & James, M. (2004). Comprehensive working memory activation in Soar. *International Conference on Cognitive Modeling* (pp. 226–230).

Oberauer, K. (2009). Design for a working memory. *Psychology of learning and motivation*, *51*.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*, 758.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological review*, *115*, 544.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic bulletin & review*, *19*, 779–819.

Pal, P., Zhu, L., Golden-Lasher, A., Swaminathan, A., & Williams, T. (2020). Givenness hierarchy theoretic cognitive status filtering. *Proc. CogSci*.

Persiani, M., Franchi, A. M., & Gini, G. (2018). A working memory model improves cognitive control in agents and robots. *Cognitive Systems Research*, *51*, 1–13.

Phillips, J. L., & Noelle, D. C. (2005). A biologically inspired working memory framework for robots. *Proc. Int'l Symp. on Robot and Human Interactive Communication, 2005.*.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*, 169–190.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational linguistics*, *30*, 309–363.

Press, A., Pacer, M., Griffiths, T., & Christian, B. (2014). Caching algorithms and rational models of memory. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*, 57–87.

Reitman, J. S. (1971). Mechanisms of forgetting in short-term memory.

Ricker, T. J., Vergauwe, E., & Cowan, N. (2016). Decay theory of immediate memory: From brown (1958) to today (2014). *The Quarterly Journal of Experimental Psychology*, *69*.

Rosenbloom, P. S., Laird, J. E., Newell, A., & McCarl, R. (1991). A preliminary analysis of the soar architecture as a basis for general intelligence. *Artificial Intelligence*, *47*, 289–325.

Scheutz, M., Krause, E., Oosterveld, B., Frasca, T., & Platt, R. (2017). Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. *Proc. AAMAS*.

Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., & Frasca, T. (2019). An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cog. Arch.*

Shintel, H., & Keysar, B. (2007). You said it before and you'll say it again: Expectations of consistency in communication. *Jour. Exp. Psych.: Learning, Memory, and Cognition*, *33*.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psych. bulletin*, *119*, 3.

Taylor, R., Thomson, H., Sutton, D., & Donkin, C. (2017). Does working memory have a single capacity limit? *Journal of Memory and Language*, *93*, 67–81.

Tellex, S., Gopalan, N., Kress-Gazit, H., & Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, *3*.

Tenorth, M., & Beetz, M. (2009). Knowrob—knowledge processing for autonomous personal robots. *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello II, F. P., Khemlani, S. S., & Schultz, A. C. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, *2*, 30–55.

Van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT.

Vogels, J., Krahmer, E., & Maes, A. (2015). How cognitive load influences speakers' choice of referring expressions. *Cognitive science*, *39*, 1396–1418.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological review*, *72*, 89.

Wickelgren, W. A. (1965). Acoustic similarity and retroactive interference in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *4*.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, *50*, 449–455.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of vision*.

Williams, T. (2017). A consultant framework for natural language processing in integrated robot architectures. *IEEE Intelligent Informatics Bulletin*.

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. *Proc. HRI*.

Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond command-based instructions: Extending robotic natural language interaction capabilities. *Proc. AAAI*.

Williams, T., Krause, E., Oosterveld, B., & Scheutz, M. (2018a). Towards givenness and relevance-theoretic open world reference resolution. *RSS Workshop on Models and Representations for Natural Human-Robot Communication*.

Williams, T., & Scheutz, M. (2016). A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. *Proc. AAAI*.

Williams, T., & Scheutz, M. (2017). Referring expression generation under uncertainty: Algorithm and evaluation framework. *Proc. 10th Int'l Conf. on Natural Language Generation (INLG)*.

Williams, T., & Scheutz, M. (2019). Reference in robotics: A givenness hierarchy theoretic approach. In *The oxford handbook of reference*.

Williams, T., Thielstrom, R., Krause, E., Oosterveld, B., & Scheutz, M. (2018b). Augmenting robot knowledge consultants with distributed short term memory. *Proc. Int'l Conf. on Social Robotics*.

Young, R. M., & Lewis, R. L. (1999). The Soar cognitive architecture and human working memory. *Models of working memory: Mechanisms of active maintenance and executive control*.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–235.