Referring to the recently seen: reference and perceptual memory in situated dialogue

John D. Kelleher
ADAPT Research Centre
ICE Research Institute
Technological University Dublin
john.d.kelleher@dit.ie

Simon Dobnik
CLASP and FLOV
University of Gotenburg, Sweden
simon.dobnik@gu.se

Abstract

From theoretical linguistic and cognitive perspectives, situated dialogue systems are interesting as they provide ideal test-beds for investigating the interaction between language and perception. To date, however much of the work on situated dialogue has focused resolving anaphoric or exophoric references. This paper opens up the question of how perceptual memory and linguistic references interact, and the challenges that this poses to computational models of perceptually grounded dialogue.

1 Introduction

Situated language is spoken from a particular point of view within a shared perceptual context (Byron, 2003). In an era where we are witnessing a proliferation of sensors that enable computer systems to *perceive* the world, effective computational models of situated dialogue have a growing number of practical applications, consider applications in human-robot interaction in personal assistants, driverless car interfaces that allow interaction with a passenger in language, and so on. From a more fundamental science perspective, computational models of situated dialogue provide a test-bed for theories of cognition and language, in particular those dealing with the binding/fusion of language and perception in interactive settings involving human conversational partners and an ever-changing environment.

The history of computational models of situated dialogue can be traced back to systems in the 1970's such as SHRDLU which enabled a user to control a robot arm to move objects around a simple simulated blocks micro-world (Winograd, 1973). Since these early beginnings there has been consistent research on computational models of the interface between language and vision, examples of such research spanning the decades include (McKevitt, 1995; Kelleher et al., 2000; Kelleher, 2003; Gorniak and Roy, 2004; Kelleher and Kruijff, 2005a; Kruijff et al., 2006a; Dobnik, 2009; Tellex, 2010; Sjöö, 2011; Kelleher, 2011; Hawes et al., 2012; Dobnik and Kelleher, 2016; Schütte et al., 2017; Larsson, 2018). A commonality across many of these systems is that they have a primary focus on grounding[1] the references within a single utterance against the current perceptual context. For example, many of these systems are concerned with grounding spatial references.[2] Some of these systems do maintain a model of the evolving linguistic discourse. However, many of these systems assume a fixed view of the world, and hence the question of how to store perceptions of entities that have not yet been mentioned does not arise as the necessary perceptual information relating to these entities is always present through direct perception of the situation. Consequently, these systems have no perceptual memory, and so cannot handle reference to entities that have been

[1] In the sense of Harnad (1990) rather than Clark et al. (1991)

[2] Herskovits (1986) provides an excellent overview of the challenges posed by spatial language. Many computational models of spatial language are based on the spatial template concept (Logan and Sadler, 1996); see Gapp (1995a), Kelleher and Kruijff (2005b), Costello and Kelleher (2006), and Kelleher and Costello (2009) for examples of spatial template based computational models of the semantics of topological prepositions, and Gapp (1995b), Kelleher and van Genabith (2006), and Brenner et al. (2007) for computational models of projective prepositions. More recently models based on the concept of an attentional vector sum (Regier and Carlson, 2001; Kelleher et al., 2011), and the functional geometric framework (Coventry and Garrod, 2004) have been proposed. Another stream of research on spatial language deals with the question of frame of reference modelling and ambiguity (Carlson-Radvansky and Logan, 1997; Kelleher and Costello, 2005; Dobnik et al., 2014, 2015; Schultheis and Carlson, 2017)

perceived but are no longer visible. Within this context, this paper highlights the challenges posed to computational models of situated dialogue in designing models that are capable of resolving references to previously perceived entities.

Paper structure: Section 2 frames the paper's focus on reference, and highlights the role that memory plays in reference within dialogue; Section 3 overviews some of the main cognitive theories and models of human memory; Section 4 expands the focus to include models of reference in situated dialogue, including models of data fusion from multiple modalities; Section 5 compares two different approaches to designing computational data structures of perceptual memory (one approach is discrete/local/episodic in nature, the other is an evolving monolithic model of context); Section 6 concludes the paper, where we argue that a blend of these approaches is necessary to do justice to the richness and complexity of situated dialogue.

2 Reference in Dialogue

Referring expressions can take a variety of surface forms, including: definite descriptions ("the red chair", indefinites ("a chair"), pronouns ("it"), demonstratives ("that"). The form of referring expression used by a speaker signals their belief with respect to the status the referent occupies within the hearer's set of beliefs (Ariel, 1988; Gundel et al., 1993). For example, a pronominal reference signals that the intended referent has a high degree of salience within the hearer's current mental model of the discourse context.

The term "mutual knowledge" describes a set of mutually shared propositions that a particular set of things are in the joint focus of attention of the interlocutors, and hence are available as referents within the discourse (McCawley, 1993). In a situated dialogue, an interlocutor may consider an entity to be available as a potential referent: (i) they consider it to be part of the cultural or biographical knowledge they share with their dialogue partner, or (i) it is in the shared perception of the situation the dialogue occurs within.

The term *discourse context* (DC) is often used in linguistically focused research on dialogue to describe the set of entities available for reference due to the fact that they have previously been mentioned in the dialogue:

> "*The DC has traditionally been thought of as a discourse history, and most computational processes accumulate items into this set only using linguistic events as input*" (Byron, 2003, pg. 3).

In this paper, we will often distinguish between the mutual knowledge set and the discourse context, where the mutual knowledge set contains the set of entities that are available for reference but which have not been mentioned previously in the discourse, and the discourse context being a record of the entities that have been mentioned previously. Given this distinction between mutual knowledge and the discourse context, the process of resolving a referring expression can be characterized as follows: a referring expression in an utterance introduces a representation into the semantics of that utterance and this representation must be bound to an entity in the mutual knowledge set (in the case of evoking or exophoric references) or in the discourse context (in the case of anaphoric references) for the utterance to be resolved.

This process of resolving a referring expression against the mutual knowledge set or the discourse context means that we can distinguish at least three types of referring expressions based on the information source they draw their referent from (as opposed to their surface form), namely: *evoking*, *exophoric* and *anaphoric* references. An *evoking* reference refers to an entity that is known to the interpreter through their conceptual knowledge but which has not previously been mentioned in the dialogue. Consequently, the referent of an evoking reference is found in the mutual knowledge set, and the process of resolving this reference introduces a representation of the referent into the discourse context. An *exophoric* reference denotes an entity that is known to the interpreter through their perception of the situation of the dialogue but which has not previously been mentioned in the dialog. Similar to an evoking reference, the process of resolving an exophoric reference introduces a representation of the referent into the discourse context. An *anaphoric* reference refers to an entity that has already been mentioned in the dialogue and hence a representation of its referent is already in the discourse context. Figure 1 illustrates the relationships between the data structures and categories of reference described above.

All of these forms of reference draw upon human memory. Mutual knowledge and the maintenance of a discourse context are both 'stored' in memory. Therefore in order for a computational

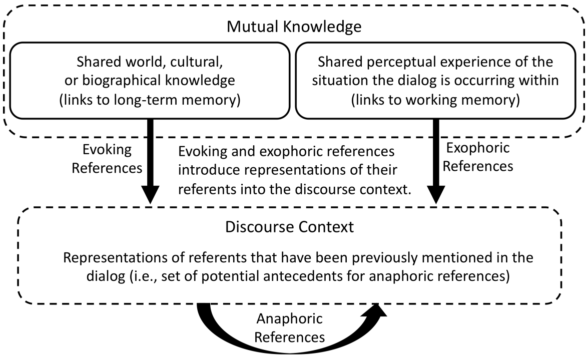
Mutual Knowledge
Shared world, cultural, or biographical knowledge (links to long-term memory)
Shared perceptual experience of the situation the dialog is occurring within (links to working memory)
Evoking References
Evoking and exophoric references introduce representations of their referents into the discourse context.
Exophoric References
Discourse Context
Representations of referents that have been previously mentioned in the dialog (i.e., set of potential antecedents for anaphoric references)
Anaphoric References

Figure 1: The relationship between mutual knowledge, the discourse context, and evoking, exophoric, and anaphoric references.

system to be able to resolve exophoric references it must include, and maintain, data structures that represent the memory component that maintains the mutual knowledge element of shared perceptual experience. To inform the design of this memory data structure in the next section we will review cognitive theories of memory.

3 Cognitive Theories of Memory

Cognitive psychology[3] distinguishes between a number of different types of memory including:

sensory memory which persists for several hundred milliseconds and is modality specific

working memory which persists for up to thirty seconds and has limited capacity

long-term memory which persists from thirty minutes up to the end of a person's lifetime, and has potentially unlimited capacity.

Figure 2 illustrates the (Atkinson and Shiffrin, 1968) model of how these different types of memory interact. External inputs are initially stored in modality specific sensory memory buffers. There is an attentional filter between these sensory specific memories and working memory. Information that is attended to passes through to working memory, and unattended information is lost. Information in the working memory that is frequently rehearsed is transferred to long-term memory and may be retrieved later. Information in working memory that is not rehearsed is displaced as new information arrives.

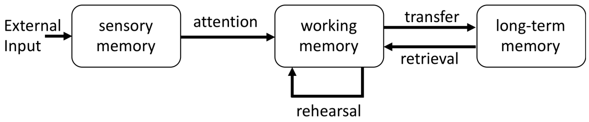

Figure 2: Atkinson and Shiffrin's Multi-store Model of Memory, based on a figure from `https://en.wikipedia.org/wiki/Atkinson?Shiffrin_memory_model`

Evoking references draw on long-term memory and exophoric references draw on working memory.[4] Furthermore, it is reasonable that the discourse context model should be considered a part of working memory. These observations point to a partial mapping between components of Figure 1 and Figure 2. Working memory is where the part of mutual knowledge that is based on perception of the situation and also the discourse context model are stored and maintained; whereas, long-term memory is where the information used to resolve evoking references is stored. The mapping indicates that working memory is at the centre of handing exophoric references.

According to Baddeley (2002) working memory has four major systems, see Figure 3, these are:

central executive is modality independent and is responsible for supervising the integration of information, directing attention, and coordinating the other systems

phonological loop holds speech based information and can maintain this information over short periods by continuous rehearsal

visual-spatial sketchpad stores visual and spatial information and can construct visual images and mental maps

episodic buffer a limited capacity buffer that temporarily stores and integrates information from the phonological loop and the visuo-spatial sketchpad, and can also link to long-term memory, and perhaps other modules dedicated to smell, taste, and so on. The information sources that the episodic buffer draws upon use different encoding schemes, however the episodic buffer integrates these

[3] See, for example Eysenck and Keane (2013).

[4] Exophoric references can also affect the attention filter between sensory memory and working memory, see Dobnik and Kelleher (2016) for more discussion on this point.

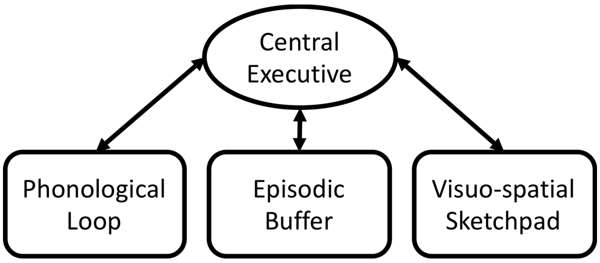

Figure 3: Baddeley's Model of Working Memory, figure inspired by Figure 3 of (Baddeley, 2002)

> disparate encodings into a unitary representation of chronologically ordered episodes.

4 Grounding Language in Vision

Grosz (1977) highlighted that attention processes can affect how references are resolved during a dialogue. In particular, Grosz observed the interaction between the shared focus of attention and the use of exophoric definite descriptions. Specifically, if an object is in the mutual focus of attention it can be denoted by means of a definite description even though other entities fulfilling the description are present in the mutual knowledge set. Grosz and Sidner (1986) extended this work and developed a focus stack model of global discourse attentional state. Other models of global discourse structure and processing have since been proposed, for example Hobbs (1985); Mann and Thompson (1987); Kempson (1988); Kempson et al. (2000); Asher and Lascarides (2003); Kamp et al. (2011). However, whichever model of global discourse structure is assumed the question of how the focus of attention and reference interact within a local discourse context must also be addressed, and a number of approaches to this question have been proposed, for example Alshawi (1987), Hajicová (1993), Lappin and Leass (1994), and Grosz et al. (1995).[5] However, none of these models explicitly accommodate multimodal contexts.

Harnad (1990) addresses the question of grounding language in perception. More recently, Coradeschi and Saffiotti (2003) has addressed this in terms of the symbol anchoring framework, Roy (2005) has proposed semantic schemas, and Kruijff et al. (2006b) proposed an ontology-based mediation between content in different modalities. Generally, these works focus on exophoric references but assume that the referent is still perceptually available. An interesting, and understudied, category of reference are exophoric references to entities that are not perceptually available at the time of the reference. For example, consider an entity that was seen by two interlocutors just prior to either of them referring to it, but which is no longer visible to either of them, perhaps because they (or it) has changed location. The fact that the entity is no longer accessible through direct perception highlights the need for a memory of perception to be maintained to handle these references, and we will refer to these types of exophoric references as references to perceptual memories. These types of references are interesting for two reasons. First, in general, (as noted above) to date exophoric references have been studied under the assumption that the referent is still perceptually available to the interlocutors'. Second, enabling a computational model to handle exophoric referents to entities that are no longer perceptually available requires the design of a perceptual memory data structure. This perceptual memory data structure stores the mutual knowledge information related to the interlocutors shared perceptual experience of the situation (see Section 2). Furthermore, this perceptual memory data can be understood as part of working memory (see Section 3).

5 Perceptual memory

The design of a perceptual memory data-structure opens up a number of significant research questions, for example: should all entities that are perceived be entered into this data structure or is there a filtering process (e.g. an attentional filter); once an entity enters the perceptual memory is it there indefinitely or can it be removed (forgotten); how does the perceptual memory interact with the linguistic discourse history (are they separate); how is the perceptual memory structured, for example, is it episodic or monolithic, does it have a chronological order; and so on.

There are examples of computational models that can function as perceptual memories in the literature. For example, in Robotics there is a long tradition of research on the problem known as Simultaneous Localisation and Mapping (SLAM), Thrun et al. (2005) provides an introduction and overview of SLAM research. SLAM algorithms integrate sensor information received over a period of time as a robot moves around an environment

[5] See (Kruijff-Korbayová and Hajicová, 1997) for a comparison of these approaches.

into a single map representation. Once constructed this map enables a robot to navigate through the environment without colliding with fixed obstacles, such as walls. However, at least in the standard versions of SLAM these maps have no semantic information about what things are, rather the focus is on mapping there are things. So, in some ways, SLAM models can be understood as akin to the visuo-spatial scratchpad in Baddeley's model of working memory. Although undoubtably useful for robot navigation, SLAM models, and the encodings they use, are not designed to facilitate linguistic reference. For this, we need a model that integrates both visuo-spatial information and linguistic information, something akin to the episodic buffer in Baddeley's model.

5.1 A Local/Episodic Architecture

The LIVE system (Kelleher et al., 2005), is a candidate architecture for this episodic buffer module. The LIVE system is designed as a natural language interface to a virtual town, similar in spirit to Winograd's SHRDLU system discussed earlier. A distinctive characteristic of the LIVE system, is that the user is able to move around the environment, and the system has a perceptual memory module that enables the user to refer to off-screen objects that have been seen recently. The LIVE system uses a false colouring visual salience algorithm to process each frame (visual scene) generated as the user moved through the virtual environment (Kelleher and van Genabith, 2003, 2004), there are 28 such frames generated per second. This visual salience algorithm identifies each object instance visible in a frame, and associates a normalised visual salience score to each object, based on its size and location within the frame. For each object in a scene the system also retrieves the object type (e.g. house, tree, etc.) and colour information from the scene graph. Consequently, for each frame a list of the visible objects along with their type and colour information and a salience score is created. This frame information is then used to populate a data structure, known as a reference domain. There is a separate reference domain created for each frame. In a sense a reference domain can be understood as a representation of the perceptual information in a frame that is designed to facilitate the grounding of exophoric references.

A reference domain is composed of a number of lists, known as partitions, and the elements of each partition is ordered, in descending order, by their visual salience. The function of these partitions is to predict the different ways a user may refer to an object in the scene. Every reference domain contains a general *object* partition which lists all the objects in the scene ordered by their salience, there is also a partition for each object type in the scene (e.g., if there are trees visible in a frame then the corresponding reference domain includes a tree partition listing all the trees visible ordered by their salience), and for each object colour (e.g., if there are red objects in the scene then there is a red partition listing all the red objects ordered by colour). The set of potential partitions that could be included in a reference domain is huge, for example there could be a partition for red houses, or green trees, and other combinations of features. In the design of the LIVE system the decision was taken to limit the initial set of partitions to categories that are reasonably likely to be preattentively available, namely, object, type, and colour. Partitions modelling more complex criteria may be created within a reference domain in response to a linguistic utterances, the reasoning being that the act of a referring expression specifying a set of selection restrictions draws attention to the set of objects fulfilling the criteria and therefore creating a partition to explicitly model this set is cognitively plausible at this point. The feature structure below illustrates the reference domain for the frame shown in Figure 4.

$$\begin{bmatrix}
\text{p1} & \begin{bmatrix} \text{criterion} & \text{‘object’} \\ \text{elements} & \begin{bmatrix}\text{H1,1.0; H3,0.2;H2,0.1}\end{bmatrix} \end{bmatrix} \\
\text{p2} & \begin{bmatrix} \text{criterion} & \text{‘house’} \\ \text{elements} & \begin{bmatrix}\text{H1,1.0; H3,0.2;H2,0.1}\end{bmatrix} \end{bmatrix} \\
\text{p3} & \begin{bmatrix} \text{criterion} & \text{‘red’} \\ \text{elements} & \begin{bmatrix}\text{H1,1.0}\end{bmatrix} \end{bmatrix} \\
\text{p4} & \begin{bmatrix} \text{criterion} & \text{‘blue’} \\ \text{elements} & \begin{bmatrix}\text{H3,0.2}\end{bmatrix} \end{bmatrix} \\
\text{p4} & \begin{bmatrix} \text{criterion} & \text{‘green’} \\ \text{elements} & \begin{bmatrix}\text{H2,0.1}\end{bmatrix} \end{bmatrix}
\end{bmatrix}$$

The LIVE system stores these reference domains in a chronologically ordered data structure with a capacity to hold 3,000 reference domains and using a first-in-first-out policy; i.e., when the data structure is full the oldest reference domain

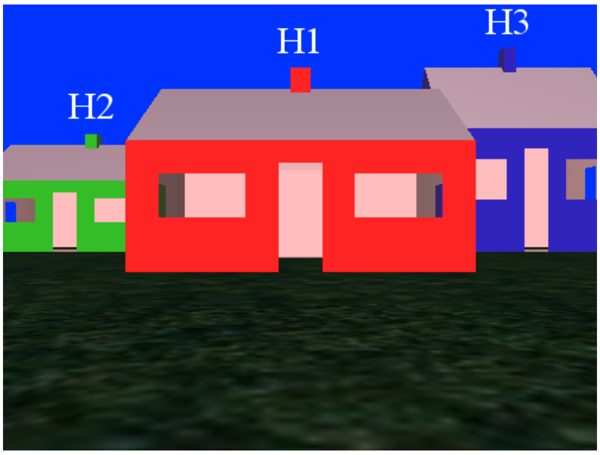

Figure 4: A frame from the LIVE System. Note: the H1, H2, and H3 labels were added to the image to help readers cross-reference with the reference domain feature structure listed in the paper.

is deleted to make space for the new reference domain. This gives the system a perceptual memory of $\frac{3{,}000}{28} = 108$ seconds.

The LIVE system also maintains a discourse context model. This model is similar in structure to the perceptual memory, it consists of up to 3,000 chronologically ordered reference domain data structures and uses a first-in-first-out policy when the buffer is full. New reference domains are added to this discourse context model as a result of resolving a referring expression. The LIVE system defines different algorithms for resolving referring different forms (i.e. surface forms) of references (i.e, there are separate resolution algorithms for demonstratives, indefinite, definite, pronominal, one anaphora, and other anaphora references). The high-level processing of all of these algorithms is: (i) select a reference domain from either the perceptual memory or the discourse context that contains at least one representation of entity whose features match the selection restrictions in the reference (the selection process also considers the recency and internal structure of the reference domain), (ii) make a copy of the selected reference domain, (iii) restructure the reference domain (potentially by adding new partitions) to mark the entity selected as the reference, and (iv) add the restructured reference domain to the head of the discourse context list. The restructuring and augmentation of reference domains in response to a referring expression is dependent on the selection restrictions specified in the reference and is designed to facilitate the processing of potential subsequent anaphoric references.

In summary, the LIVE system maintains a separate perceptual memory and discourse context model, although both of these data structures have similar internal structures (chronologically ordered lists of reference domains). The structure of these components is somewhat similar to the episodic buffer in Baddeley's model: limited capacity, chronologically ordered, and integrating visual perceptual information with semantic information. Furthermore, the similarity in the encodings in the perceptual memory and discourse context model facilitates reference resolution, which entails copying, restructuring, and inserting of a reference domain. Indeed, the approach to resolving a reference taken by the LIVE system can be understood as searching memory for a suitable episodic memory, using this episode as local context within which the reference is resolved, and updating the episode to mark the fact that the reference has occurred. Such a model is capable of handling exophoric references to entities that were recently seen but are no longer on-screen. However, using a reference domain representation of a frame/episode as defining the (local) context for a reference makes it extremely difficult to handle references to refer to two or more entities that never appeared in the same frame. Handling these forms of references requires the system to be able to integrate multiple reference domains, and this is non-trivial; e.g., it is not clear how salience scores from different frames, and hence different times, should be updated during this merger.

5.2 A Global/Monolithic Architecture

An approach to the design of a perceptual memory, that naturally answers the question of how to integrate information from perceptions received across distinct times, is to use an evolving global structure where all referents are stored in a single data structure that is continuously updated to reflect the current state.

Koller et al. (2004) describes an interface for playing textual computer games, based on description logics and theorem proving. This model does not have a visual component, instead the information relating to the physical environment of the game world is provided via textual descriptions. However, the game world is never fully observable, and therefore a player's knowledge of the game world increases as they move through the game. The context model proposed in this work is based on Description Logics, and uses a data structure known as the *T-Box* to encode axioms related to concepts and roles (in a sense the ontology of the world), and another data structure known as the *A-Box* to encode the entities (instances of concepts) in the world. Interestingly, the system maintains two A-Box data structures: (i) the game A-Box representing the full current game world state, and (ii) the player's A-Box representing what the player knows about the game world (this A-Box is typically a sub-part of the world A-Box). As the player moves through the game environment and explores new locations new instances are added to the player's A-Box. As a result, the player's A-Box represents a perceptual memory of what they have experienced in the world. Entities in the player's A-Box are marked with the property of *here* when they share the same location as the player (i.e., the player and the entity are both in the same room in the world), *visible* if the entity is deemed to be currently visible to the player, and *accessible* if the player can currently manipulate the entity. Consequently, the system has the ability to distinguish between entities that are currently visible and entities that are known about but which are not visible. However, the design of the reference resolution algorithms used by the system presupposes that: *players will typically only refer to objects which they can "see" in the virtual environment, as modelled by the concept 'visible'* (Koller et al., 2004, page. 12). This assumption allows the resolution algorithm to ignore entities in the world which are known to the player (and, hence are in the player's A-Box) but which are not currently visible when resolving a referring expression. This assumption means that the system cannot handle exophoric references to recently seen entities that are no longer visible, as they are deliberately excluded from the context used to resolve references. It should be noted that this is not a simple assumption to remove from the system. The system has no model of perceptual salience (although it does have a model of linguistic salience). As a result it must use this strict visible/invisible criterion to exclude potential distractor entities (that are in the model of the player's knowledge of the world but which are not currently in the perceptual focus), which if not excluded would make a reference appear unspecified and ambiguous to the system.

Kelleher (2006) is another natural language interface to a virtual world. It is similar to (Kelleher et al., 2005) in that it uses the same visual salience algorithm to analysis the visual frames the user sees as they navigate through the environment. However, the data structure used to store perceptual memories and discourse structure is very different. This system maintains a single global context model throughout a user's session. Once an entity has been rendered on screen a representation of that entity is introduced in this global context model. There is only ever a single representation of an entity in the global context model. This representation of an entity stores the physical information of the entity (e.g., *type*, *colour*, *size*, and so on) and also stores a visual salience and a linguistic salience score for the entity. The visual salience score is updated after each frame is processed. The visual salience of an entity that is not in the current frame is halved when the frame is processed. As a result the visual salience of an entity drops off once it goes out of (visual) focus (i.e., off-screen), and continues to reduce the longer out of focus it remains. The linguistic salience scoring is based on the assumption that entities that have been mentioned recently are more salient than entities that have not. The particular function used to calculate and update the linguistic salience scores is in the spirit of Centering Theory (Grosz et al., 1995) and is similar to the model proposed by (Krahmer and Theune, 2002). The linguistic salience of an entity is updated after each utterance has been processed. The linguistic salience of any entity not mentioned in an utterance is halved when the utterance is processed. Consequently, similar to the visual salience of an entity, the linguistic salience of an entity drops once it leaves the (linguistic) focus, and continues to drop the longer out of focus it remains. As the above description indicates the representation of an entity in the global context model is a relatively complex feature structure. However, the structure of the global context model itself is minimal, it is simply an unordered set of these entity representations. The fact that the linguistic and visual salience scores are updated based on recency of being visible or mention means that the context model does not need to explicitly model recency.

Reference resolution in this system is done by calculating an integrated salience score for each entity in the context model, and then selecting the entity with the highest integrated score as the referent. The integrated salience score of an entity is recalculated each time a referring expression is processed. The integrated salience score is calculated in three steps: (i) a reference relative visual salience score is calculated by scaling the standard visual salience score to reflect the fit of the entity with the selection restrictions specified in the expression (e.g., in the simplest case the reference relative visual salience score is set to zero if the entity is of the wrong type to be the referent of the reference); (ii) a reference relative linguistic salience score is calculated in a similar way to the reference relative visual salience score; and (iii) the integrated salience score is calculated as a weighted sum of the reference relative visual and linguistic salience scores, where the weighting is dependent on the form of the expression (e.g., for pronominal references the system weights linguistic salience more then visual salience).

The fact that this monolithic global context model does not encode an episodic (frame based) structure means that the integration of information from different scenes is straightforward. As a result, this system can handle references to entities that do not appear on screen together. However, this flexibility is at a cost. The loss of the episodic chronological order means that a system using this context model would not be able to handle exophoric references based on chronology (such as *the first blue house we saw*), or co-occurrence within a local temporal context (such as *the car that was in front of the house when the man fell*).

6 Discussion

The two approaches to perceptual memory described in Sections 5.1 and 5.2 are exemplars at opposing ends of a design spectrum: one focuses on identifying a local context and resolving the reference within that context, the other on creating and continuously evolving a global context model. These approaches have complementary strengths and weaknesses. Consequently, it is likely that a blend of these approaches is necessary. This is not surprising as there are many examples in language processing[6] where there is a need to be able to switch from a local focus to a global perspective, and back again, as the context requires.

[6] Switching between local and global representations, similar to the challenge of modelling long-distance dependencies in sequential data (Mahalunkar and Kelleher, 2018)

References

Hiyan Alshawi. 1987. *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge, UK.

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes1. In *Psychology of Learning and Motivation*, volume 2, pages 89–195. Elsevier.

Alan D Baddeley. 2002. Is working memory still working? *European Psychologist*, 7(2):85.

Michael Brenner, Nick Hawes, John D. Kelleher, and Jeremy L. Wyatt. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2072–2077. AAAI.

Donna Byron. 2003. Understanding referring expressions in situated language: Some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*, Hokkaido University.

Laura Carlson-Radvansky and Gordan D. Logan. 1997. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37:411–437.

Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*, 13(1991):127–149.

Silvia Coradeschi and Alessandro Saffiotti. 2003. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96.

Fintan Costello and John D. Kelleher. 2006. Spatial prepositions in context: The semantics of *Near* in the presense of distractor objects. In *Proceedings of the 3rd ACL-Sigsem Workshop on Prepositions*, pages 1–8.

Kenny R. Coventry and Simon Garrod. 2004. *Saying, Seeing and Acting. The Psychological Semantics of Spatial Prepositions*. Taylor & Francis, New York, NY, USA.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, UK. 289 pages.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (DialWatt)*, pages 43–52, Edinburgh.

Michael W Eysenck and Mark T Keane. 2013. *Cognitive psychology: A student's handbook*, 5th edition edition. Psychology press, New York, NY, USA.

Klaus P. Gapp. 1995a. An empirically validated model for computing spatial relations. In *The 19th German Conference on AI*, pages 245–256.

K.P. Gapp. 1995b. Angle, distance, shape, and their relationship to projective relations. In *The 17th Conference of the Cognitive Science Society*.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Barbara Grosz. 1977. *The Representation and Use of Focus in Dialogue Understanding*. Ph.D. thesis, Standford, University.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling local coherence of discourse. *Computational Linguistics*, 21(2):203–255.

Barbara Grosz and Candy Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expression in discourse. *Language*, 69:274–307.

Eva Hajicová. 1993. *Issues of Sentence Structure and Discourse Patterns*, volume 2 of *Theoretical and Computational Linguistics*. Charles University Press, Prague, Czech Republic.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Nick Hawes, Matthew Klenk, Kate Lockwood, Graham S Horn, and John D. Kelleher. 2012. Towards a cognitive system that can recognize spatial regions based on context. In *Proceedings of the 26th Naitional Conference on Artificial Intelligence*.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Jerry Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information.

Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2011. Discourse representation theory. In *Handbook of Philosophical Logic*, pages 125–394. Springer, Dordrecht.

John Kelleher and Josef van Genabith. 2003. A false colouring real time visual saliency algorithm for reference resolution in simulated 3-d environments. In *Proceedings of the Conference on Artifical Intelligence and Cognitive Science*, pages 95–100.

John D. Kelleher. 2003. *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. Ph.D. thesis, Dublin City University.

John D. Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1-2):21–35.

John D Kelleher. 2011. Visual salience and the other one. In *Salience. Multidisciplinary Perspectives on Its Function in Discourse. Mouton de Gruyer*, number 227 in Trends in Linguistics. Studies and Monographs., pages 205–228. de Gruyter, Berlin/New York.

John D. Kelleher and Fintan Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-Sigsem Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*.

John D. Kelleher, Fintan Costello, and Josef van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and lingusitic discourse context. *Artificial Intelligence*, 167(1-2):62–102.

John D. Kelleher and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

John D. Kelleher, Tom Doris, Quamir Hussain, and Sean ONuallain. 2000. Sonas: Multimodal, multiuser interaction with a modelled environment. In *Spatial Cognition - Foundation and Applications*, pages 171–185. John Benjamins Publishing, Amsterdam.

John D. Kelleher and Josef van Genabith. 2004. Visual salience and reference resolution in simulated 3d environments. *AI Review*, 21(3-4):253–267.

John D. Kelleher and Josef van Genabith. 2006. A computational model of the referential semantics of projective prepositions. In *Syntax and Semantics of Prepositions*. Kluwer.

John D. Kelleher and Geert-Jan M. Kruijff. 2005a. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

John D. Kelleher and Geert-Jan M. Kruijff. 2005b. A context-dependent model of proximity in physically situated environments. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*.

John D. Kelleher, Robert J. Ross, Colm Sloan, and Brian Mac Namee. 2011. The effect of occlusion on the semantics of projective spatial terms: A case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108.

Ruth Kempson. 1988. *Mental representations: The interface between language and reality*. Cambridge University Press, Cambridge, UK.

Ruth Kempson, Wilfried Meyer-Viol, and Dov M Gabbay. 2000. *Dynamic syntax: The flow of language understanding*. Wiley-Blackwell, Oxford, UK.

Alexander Koller, Ralph Debusmann, Malte Gabsdil, and Kristina Striegnitz. 2004. Put my galakmid coin into the dispenser and kick it: Computational linguistics and theorem proving in a computer game. *Journal of Logic, Language and Information*, 13(2):187–206.

Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In Rodger Kibble Kees van Deemter, editor, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CLSI Publications, Stanford University in Palo Alto, California, US.

Geert-Jan Kruijff, John D. Kelleher, Gregor Berginc, and Alex Leonardis. 2006a. Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pages 343–344. ACM.

Geert-Jan M. Kruijff, John D. Kelleher, and Nick Hawes. 2006b. Information fusion for visual reference resolution in dynamic situated dialogue. In *Proceedings of Perception and Interactive Technologies*, volume 4021 of *LNCS*, pages 117 – 128.

Ivana Kruijff-Korbayová and Eva Hajicová. 1997. Topics and centers: A comparison of the salience-based approach and the centering theory. *Prague Bulletin of Mathematical Linguistics*, 67:25–50.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in Cognitive Science*, 10(2):389–408.

Gordan D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In *Language and Space*, pages 493–529. MIT Press, Cambridge, MA, USA.

Abhijit Mahalunkar and John D Kelleher. 2018. Using regular languages to explore the representational capacity of recurrent neural architectures. In *International Conference on Artificial Neural Networks*, pages 189–198. Springer.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 83–96. Springer, Dordrecht.

James D. McCawley. 1993. *Everything That Linguists Have Always Wanted To Know About Logic*(but were ashamed to ask)*, 2nd edition. University of Chicago Press, Chicago.

Paul McKevitt, editor. 1995. *Integration of Natural Language and Vision Processing (Vols. I-IV)*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Terry Regier and Laura Carlson. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Holger Schultheis and Laura A Carlson. 2017. Mechanisms of reference frame selection in spatial term use: computational and empirical studies. *Cognitive Science*, 41(2):276–325.

Niels Schütte, Brian Mac Namee, and John D. Kelleher. 2017. Robot perception errors and human resolution strategies in situated human–robot dialogue. *Advanced Robotics*, 31(5):243–257.

Kristoffer Sjöö. 2011. *Functional understanding of space: Representing spatial knowledge using concepts grounded in an agent's purpose*. Ph.D. thesis, KTH Royal Institute of Technology.

Stefanie Tellex. 2010. *Natural Language and Spatial Reasoning*. Ph.D. thesis, Massachusetts Institute of Technology.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 2005. *Probabilistic Robotics*. MIT Press, Cambridge, MA, USA.

Terry Winograd. 1973. A procedural model of language understanding. In R.C. Schank and K.M. Colby, editors, *Computer Models of Thought and Language*, pages 152–186. W. H. Freeman and Company, New York, NY, USA.