Utilizing Visual Attention for Cross-Modal Coreference Interpretation

Donna Byron, Thomas Mampilly, Vinay Sharma, and Tianfang Xu

The Ohio State University,
Department of Computer Science and Engineering,
2015 Neil Ave, Columbus, Ohio, 43210, USA
{dbyron, mampilly, sharmav, xut}@cse.ohio-state.edu

Abstract. In this paper, we describe an exploratory study to develop a model of visual attention that could aid automatic interpretation of exophors in situated dialog. The model is intended to support the reference resolution needs of embodied conversational agents, such as graphical avatars and robotic collaborators. The model tracks the attentional state of one dialog participant as it is represented by his visual input stream, taking into account the recency, exposure time, and visual distinctness of each viewed item. The model correctly predicts the correct referent of 52% of referring expressions produced by speakers in human-human dialog while they were collaborating on a task in a virtual world. This accuracy is comparable with reference resolution based on calculating linguistic salience for the same data.

1 Introduction

A challenging goal in computational linguistics is understanding all of the ways context modulates the meaning of linguistic forms. One contextual effect that has been observed across multiple experimental disciplines is the use of ambiguous referring expressions for entities that are salient in the context. Speakers use underspecified nominal expressions, especially pronouns such as *this* and *he* but also common noun phrases (NP) such as *the button*, freely in discourse, relying on the addressee's ability to understand which button or person is being referred to. This preference for certain entities, given a prior context, will be called *salience* in this paper. Salience corresponds to a prediction or expectation that a certain entity will be the topic of an utterance. Estimating the relative salience of each entity in the universe of discourse is an important task in computational models of referring behavior - both in producing felicitous noun phrases and also in interpreting connected discourse. The long-term objective of our research program is to create robust, accurate algorithms for reference interpretation in automated agents. This task is impossible without a firm understanding of contextual effects on referring behavior.

It has been well-established in the computational linguistics literature that discourse history can be interrogated to estimate the salience of entities in a

A. Dey et al. (Eds.): CONTEXT 2005, LNAI 3554, pp. 83–96, 2005.

sentence one is trying to interpret. However, recent technology improvements create opportunities for human-computer conversations in which several contextual factors in addition to the discourse history are in play at the same time, each impacting entity salience in different ways. The goal of our project is to create conversational software agents that can carry on a *situated* conversation with a human partner. For the purposes of this paper, situated language will be defined as language having these properties:

Immersion. The conversation takes place within a 3D setting that is perceptually available to the conversational partners. The partners can speak to each other face to face within the setting.

Mobility. Both conversational partners are at liberty to move about in the world, independently of each other, to gather information or change their perceptual perspective of the world.

These characteristics distinguish situated language from other interaction paradigms. The bulk of reference processing algorithms in computational linguistics have been developed using data collected in traditional experimental settings where conversational partners are explicitly prevented from exploiting extra-linguistic contextual clues, such as gesture and gaze [13]. However, in the current study, we examine situated language between two human partners, using new data generated in our lab. This allows us to investigate the interplay between the discourse context and the conversational setting and its effect on the interpretation of referring expressions in a visually-rich domain.

The primary focus of the present work is to develop a model of visual attention that can be used to interpret *exophors*, references to items in the discourse setting. Similar to the way that an anaphor constitutes a repeated mention of an item introduced into the context by the linguistic history, an exophor is a repeated mention of an item already introduced into the context by the physical world, in other words, a cross-modal coreference. Our hypothesis is that the world that is visually perceptible to the conversational partners will be likely to shape the content of their discussion, especially when they are performing a task involving objects in that world. Moreover, a likely source of denotations for exophors are items that the speaker's attention is directed toward as the utterance is produced. Given these two factors influencing the dialog, our aim is to test a method of tracking one speaker's view of the world over the course of a dialog, and use that information as input in a reference resolution algorithm to interpret ambiguous referring expressions. Our eventual goal is to construct a model that fuses attentional information provided in the visual channel with that provided by the discourse history. In the present work, we perform pencil-and-paper analyses and offline simulations of our model, as a first step in developing the algorithms that will eventually be implemented.

2 Motivation and Background

2.1 Overview

Our work is motivated by the goal of building automated agents or interactive characters that cohabit a virtual space with a human partner. Such agents will be able to discuss the world they are in, as well as collaborate, reason, plan, and perceive the virtual world. In order to design this type of agent, we first need to learn how human beings behave in similar environments.

The data used to inform algorithm development in this study was collected by placing pairs of human partners in a first-person graphical world, rendered by the QuakeII game engine[1]. In the virtual world, the partners collaborate on a treasure hunt task. One person in each pair of players, "the leader", was given the list of tasks. The other player, "the follower", had no prior knowledge of these tasks. This setting forced the players to converse in order to solve their task. The partners communicated through headset-mounted microphones, and an audio recording of their dialog was collected and transcribed. In addition, each player's movement and activity in the virtual world was recorded to video tape. The QuakeII game engine allows the two partners to move about in the world independently and manipulate objects. As he moves about, each player sees a first-person view of the virtual world. We trapped separate recordings from each person's viewpoint.

It is obvious that understanding natural language is important in order to collaborate in this domain, and that the two partners discuss not only items that they see but also items that have been discussed or seen in the past. There is evidence in the literature showing that visual context influences how people organize and interpret the meaning of spoken language [25, 21]. Figure 1 shows a sample dialog fragment from our study[2], which contains instances of both cross-modal and linguistic coreference over a chain of references to the helmet. Before utterance 1, speaker F sees the helmet in the room. The expression *it* in utterance 1 denotes the physical helmet in the world[3]. After this mention, the helmet is repeated several times. This discourse fragment also shows the high concentration of referring expressions in this domain. The partners are unlikely to successfully complete their task without correctly interpreting these phrases.

F: I see it {vn:AH} I see *the helmet*
L: yeah
F: yeah {vn:doo} {vn:ack} and to pick *it* up I do control right
L: yes

Fig. 1. A sample dialog demonstrating both linguistic and cross-modal coreference

[1] www.id.com
[2] The notation {vn:} signifies non-word vocal noise.
[3] Forty-six utterances prior to this point in the dialog, the partners had discussed finding the helmet, so the speaker's use of *it* in this example is partially anaphoric.

The visually-perceived Quake world is not only the source of semantics for the conversational partners, but also impacts their focus of attention. For example, when the partners walk through a door and suddenly have a view of a set of new objects in a room, their attention is fixated on the new objects they have just discovered. Therefore, our computational model must attempt to calculate what items of interest might be discussed next, given a 2D plane of pixels which represents the field of view of one partner. There are a variety of issues that must be addressed in utilizing this visual information as input to language processing.

Video Segmentation and Alignment with Language. One of the most interesting challenges in this domain is that the field of view is an ongoing data stream. This stream must be broken into units in order to compute which items are within the speaker's field of view at each point in the stream. We will call these units *visual context frames*. The frequency at which these frames are captured will affect the sensitivity of any algorithm that uses the resulting data. If the sample frequency is too low, many visual events might be missed. The highest available sample frequency is the frame rate of the video (30Hz).

Gaze Direction. In each video context frame, our system should discern which items the viewer is looking at. An object's proximity to the center of the field of view [18] turned out to be a poor indicator of its visual salience. This is primarily because many of the subjects had difficulty making fine-grained movements using their keyboard controls, so they would sometimes pan just until the object came

Fig. 2. The speaker's view when he said "will you punch *that little button* over there?"

Fig. 3. The view at the word "*cabinets*"

Fig. 4. The view at the word "*them*"

into view and then stop. Figure 2 shows an example. Although the speaker is talking about the button, it is not in the center of his field of view. Also, an item may move out of view by the time the speaker refers to it. For example, Figures 3 and 4 show the speaker panning the scene while saying "There's a couple of cabinets here. While I'm here, let me see if I can open the cabinets and not fall into *them*."

Foreground vs. Background Objects. Each visual context frame contains not only items of interest in the task, but also walls, floors, ceilings, etc. These background items are in view in most frames, therefore a model that simply favors items that have been seen frequently/recently will over-weight such items.

2.2 Computational Linguistics Background

Computational Models of Referring. An automated agent that can collaborate in rich domains such as ours will need sophisticated reference understanding software. For collaborative agents, reference resolution is the module that provides a mapping from the noun phrases spoken by the user to the objects the user intend to denote. For example, "Let's see what's in that room" might be a command to explore a particular room, and the reference resolution module determines which room the system thinks the user meant. For a given referring expression, there are many possible places to search for the referent: the physical context, items previously mentioned in the discourse, mutually known objects from the *ambient context* such as 'the president', etc. Although individual algorithms vary in their details, resolution systems primarily rely on linguistic information such as syntax or semantics or the combination of these two. With the development of multi-modal systems, researchers have begun to incorporate visual information into the resolution process [19, 20, 22]. Most of this work is still very preliminary. For example, Campana et al. [5] propose incorporating eye-tracking into a reference resolution module to take advantage of gaze information, but the idea is yet to be evaluated.

The process of reference resolution is normally modeled as two separate steps. First, the *context management* step prepares a set of possible referents that might be referred to in subsequent discourse. To contain the full list of available referents, a system interacting in situated discourse will need:

- A *Linguistic Context* (LC) that contains a list of possible referents that are introduced by the verbal interaction. This is used to track the attentional state as it is portrayed by the discourse. Generally, the entities added are only those that were mentioned as nominal constituents, however more recent algorithms also add high-order referents such as events and propositions [7, 4]. In systems with a visual interaction component, GUI items are added to the LC [17, 3, 15], because they are considered to be part of the communication process.
- A *Mutual knowledge Context* (MC) that contains a list of the objects assumed to be known to both parties before the conversation begins. For example,

an airline reservation system might initialize its MC with a list of airline companies and cities.

- A *Visual Context* (VC) to track items in the world that the partners have interacted with and might discuss. In systems such as [9, 18] that allow the user to move through a virtual world, items encountered by the users are also added to the context. The VC represents the attentional state of the conversational participants based on their extra-linguistic perception in the world.

Taken together, these lists are meant to represent all of the entities which might be mentioned in subsequent discourse.

Each time the context is updated with new entities, the relative salience of all the items is adjusted. A small subset of LC entities comprise the current linguistic focus of attention. A large variety of techniques exist for calculating the focus or salience ranking from linguistic cues [28, 24, 27, 1]. For example, items encountered or discussed recently carry more salience than items that have not been mentioned recently. The salience update process might also take into account attributes of the visual world. For example, Kaiser et al [16] developed a model of visual salience for an augmented reality application using four factors: *time, stability, visibility*, and *center-proximity*. Time represents persistence: the portion of frames over a certain window in which the object appears. Stability results in a penalty for objects that enter and leave the region multiple times. Visibility represents the amount to which the user's gesture overlaps with the object's visible projection, and center-proximity gives items at the center of the user's gaze or gesture a higher salience ranking.

The second step, *interpreting referring expressions*, is triggered as each referring expression is encountered and the context is searched for a semantically compatible referent. The search may integrates a number of different properties of the expression itself, the local linguistic context in which it appeared, and the context as it existed when the expression was spoken. Information about the expression itself can include lexical semantics, such as the lexical head and agreement features, and also the form of the expression, i.e. whether it was a pronoun, description, or locative adverb, etc. Different NP forms indicate different relationships with the context, and therefore different search procedures are invoked for each form. For example, to interpret a pronoun, the search begins with entities that are judged to be most salient [12, 2, 11].

Information about the local context might include the predication context, for example the expression might have been used to describe the PATIENT of a PUSH() action. This information, combined with a semantic resource that defines basic categories and the semantic restrictions on particular argument positions, can provide a powerful level of discrimination for resolving ambiguous phrases. A wide assortment of search methods for anaphora resolution have been developed, some exploiting syntactic structure or semantic features, others using statistical preferences (see [26] for a recent survey).

2.3 Visual Perception and Context

We aim to determine the topic of an utterance using the visual context of the speaker. Our definition of VC is the set of all objects that have ever been within the speaker's field of view and the timing information associated with their appearances and disappearances. In our study, visual salience is defined as that property of an entity in the VC that makes it the most likely topic of a person's utterance. We believe that a speaker's allocation of visual attention provides a reliable indicator of the relative visual salience of objects in the scene. Hence, factors influencing visual attention can provide valuable information about an object's visual salience, which could then be used to determine the topic of utterance.

Research in visual perception has shown that several factors influence the focus of our attention when presented with complex scenes. A well established theory in visual attention research is that the deployment of visual attention can be "guided" by the result of preattentive visual processing [8, 14]. The preattentive visual processing stage is of interest since it is known to be sensitive to certain features, such as color, orientation, curvature, and size, which form a feature-space in which the objects of our Quake world vary greatly. An object's novelty within this feature-space causes it to "pop-out", making it the focus of attention in later stages of visual processing [23].

Based on the role of the preattentive visual processing stage, we provide a simple measure that quantifies the novelty of an object, and how it changes over time, using a *Uniqueness* (U) parameter. Our definition of visual salience does not require an object to be in the current field view for it to be selected as the topic of the utterance. When an object falls out of view, its saliency is determined not only by its Uniqueness, but also by the amount of time since it was last seen ("recall delay"), and its exposure time before it dropped from view. The positive and negative effects of recall delay and exposure (or presentation) time on visual memory performance [10] is well known, and form the basis of the *Recency* (R) and *Persistence* (P) terms respectively.

3 Visual Salience Algorithm

As a person moves through the virtual world, different entities enter the visual context, and the factors affecting their visual salience need to be updated periodically.

Uniqueness (U): As mentioned earlier, the novelty of an object influences its pop-out, and hence influences the deployment of attention over the scene. The Uniqueness term models an object's novelty based purely on how frequently the object appears within the field of view. However, this model can be enhanced using computer vision based approaches that utilize object features pertinent to the preattentive visual processing stage (Sect. 2.3) to determine novelty. In our current formulation, an entity such as the floor, by virtue of being almost constantly visible within a given period, should be assigned a small U value,

while an uncommon object (e.g., the button in the Quake world (Fig.2)) should get assigned a relatively larger value. Initially all objects are assigned a maximum U value of 1, signifying equal Uniqueness. At each instant, the U value of an entity is penalized by a quantity proportional to the frequency of its occurrence in the field of view over a time window called the Uniqueness Window (T_u),

$$\begin{aligned} U_{i,j} &= U_{i,j-1} - \delta \\ \delta &= k \times \left(\frac{n_{i,j}}{T_u}\right) \end{aligned} \tag{1}$$

where the subscripts i and j represent the object i.d., and the current time instant, respectively. $n_{i,j}$ is the number of times object i was seen between $j - T_u$ and j. The constant of proportionality between the penalizing factor δ and the frequency of occurrence is denoted by k.

An object seen often in the recent past would have a large δ value when computed over a small T_u, and its Uniqueness would thus be heavily penalized. This seems consistent with the phenomenon of object-based Inhibition-of-Return (IoR), which states that "people are slower to return their attention to a recently attended object" [6]. It should be noted that while in our model the U values of all visible objects are penalized, a more faithful model of object-based IoR would penalize only objects that were attended to in the scene.

Recency (R): Once an object drops out of the field of view, we assume that the probability of it being the target of a referring expression decays with time. This relation is analogous to the well known decay of visual memory with increase in recall time [10]. A zero-centered Gaussian is chosen to model R. This profile represents a slow decay in R immediately after an object disappears, followed by a period of rapid decay that leads to an almost constant near-zero value,

$$R_{i,j} = e^{-\left(\frac{t_{i,j}}{\sqrt{2}\sigma}\right)^2}, \quad (t \geq 0) \tag{2}$$

where σ stands for the standard deviation (describe later), and $t_{i,j}$ is the length of the time interval measured from j since object i was last seen. Note that all objects currently visible have a maximum R value of 1.

Persistence (P): Analogous to the effect of presentation time on visual memory, the R values of different objects at each time instant should not only depend on $t_{i,j}$, but also on how long they were visible before disappearing. Persistence is a simple measure of an object's exposure time, computed as the frequency of occurrence of an object within a time interval, called the Persistence Window (T_p),

$$P_{i,j} = \frac{m_{i,j}}{T_p} \tag{3}$$

where $m_{i,j}$ is the number of times an object i was seen between $j - T_p$ and j. Intuitively, T_p should be large enough to allow objects to acquire significant P values, and at the same time small enough to ensure that P values of temporally distant objects fade with time. The dependence of R on P is established by making the value σ in Eq.2 proportional to P, such that

$$\sigma = c \times P_{i,j}$$

where c is a constant.

Visual Salience (S): At any instant, the combination of an object's Uniqueness and Recency determines its visual salience (S). Since neither a novel object seen a long time ago, nor a common one that is currently visible, have high visual salience, we define S such that only objects with large values of U *and* R would get assigned a high visual salience value.

$$S_{i,j} = U_{i,j} \times R_{i,j} \tag{4}$$

Since the range of U and R is [0,1], an $S_{i,j} = 1$ corresponds to maximum visual salience.

4 Our Study

We trained our model using a randomly selected five minute discourse segment involving one pair of participants. We chose a small segment of data in this pilot study due to the time-consuming process of manually annotating the video frames. In order to quantify the relationship between the visual and linguistic contexts without the possible interference of factors such as mutual knowledge, we modeled the visual context of a single participant. Frames of the selected video segment were manually annotated at a fixed time interval with a list of objects visible in that frame. The manual annotation assumed perfect object segmentation, that is, each visible object was given a unique label[4], however locative entities such as rooms were not considered in the frame annotations. The algorithm takes a stream of frames in sequential order to create the VC. For each sample frame, the objects in the visual context were ordered by their visual salience as computed by the algorithm described in the previous section.

4.1 Our Baseline Algorithm

The closest comparable approach to our work presented in the literature is [18], where "centrality" and "size" are used to determine visual salience in a simplistic simulated 3D world. However the assumptions that form the basis of their approach do not hold in our domain, thus preventing a direct comparison. As described in Sect. 2.1, salient objects may not necessarily be in the centre of the field of view. Further, since *all* entities in the perceptible world are potential referents, the size feature is also inappropriate, since background objects like walls, floors, and ceilings will always be assigned high salience. We hence resort to a discourse-based salience model to provide a comparison to the proposed approach.

[4] Background entities such as walls, floors and ceilings were determined to be possible targets of reference and were, therefore, identified and annotated by color and texture.

The discourse-based reference resolution method worked as follows, and was hand-simulated rather than automatically produced.

- The transcripts were manually segmented into units. Each independent clause and each full constituent that was not part of a complete clause was considered an independent unit.
- The LC was updated after each unit by forming a new context frame containing an id for the referent of each base noun phrase in the unit, and adding this frame to the beginning of the LC list.
- The salience of items in each frame was determined using a left-to-right, breadth-first ordering on NP-based arguments in the utterance [26].
- Given an expression to match to a referent, each update frame in the context was searched in order of recency starting from the utterance containing the RE. Items in each context frame were compared to the RE and semantically incompatible items were discarded. For example, the plural pronoun *them* would not match a singular item, and a description such as *the button* only matches buttons.

4.2 Experiments

Table 1 shows the count of test items in the development dataset used to tune our visual attention algorithm, and the agreement between the item the speaker referred to and the most salient item in the visual context, as computed by our model. Items in the VC are rank-ordered using the visual attention model described above in Section 3. The number of test items in this table is small because we eliminated referring expressions that referred to items that were never seen, such as generic entities and propositions, in order not to penalize the algorithm for items it cannot track.

As the table demonstrates, using visual salience alone (Column (a)), we were able to identify which entity the user is speaking about 41.5% of the time. Column (a), titled "Absolute Highest Rank", shows whether the most-salient item, as calculated by our visual salience algorithm, was the correct referent. The next set of columns show the effect of adding lexical semantics as a filter on

Table 1. Performance on different referring forms in the development corpus

		Highest Ranked Absolute VC	Highest Ranked Semantic Match	
			VC	LC
RE Form	Count	(a)	(b)	(c)
A/An N	2	1	2	1
The N	19	7	16	11
This/These N	1	0	0	1
That/Those N	4	3	3	2
This/That/These/Those	5	2	2	2
It/Them/They	10	4	4	6
Total	41	17 (41.5%)	27 (65.9%)	23 (56.1%)

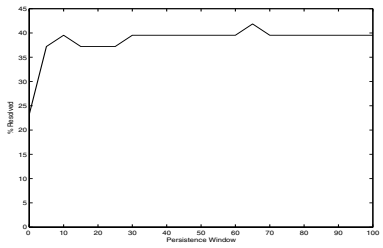

Fig. 5. Accuracy using varying Persistence Windows

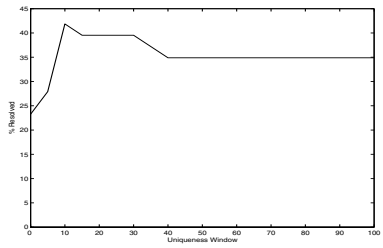

Fig. 6. Accuracy using varying Uniqueness Windows

items in the context. For example, if the speaker says *a button*, the algorithm was considered correct if the most salient button was the correct referent. Column (b) shows the performance of our algorithm with the addition of lexical semantic filtering. As a comparison, Column (c) shows the performance of the discourse-only baseline algorithm, also with lexical semantic filtering.

Effect of Persistence Window and Uniqueness Window. Figures 5 and 6 show the effect of varying the length of the Persistence and Uniqueness Windows on the performance of reference resolution for the development dataset. To account for the possibility of dependence on both parameters, Persistence Window and Uniqueness Window were varied simultaneously. The maximum resolution performance was then determined for each parameter to obtain the optimum lengths of the Persistence and Uniqueness windows in the final algorithm. Figure 5 shows that, for the training dataset, the optimum Persistence window is of length 65 seconds and the optimum Uniqueness window is 10 seconds long. We use these same values for the testing dataset.

Effect of Sampling Frequency. The frequency of update of the visual context can affect not just the performance of reference resolution but the space and time complexity of the resolution algorithm as well. Keeping in mind the goal of enabling real time reference resolution, it is important to minimize the complexity of the algorithm. The performance of the reference resolution model was observed while varying the frequency of visual context updates. The accuracy of reference resolution increases with the frequency of visual context update (0.5Hz=31.70%, 1Hz=41.46%).

We used the development dataset to train these parameters, then tested the algorithm again on two new portions of video, totalling approximately 9 minutes, which used different speakers. The performance is shown in Table 2. The results from our algorithm in this segment are slightly lower compared to that obtained on the development set.

Two examples of correctly resolved noun phrases are shown in Figures 7 and 8. In Figure 7, several objects are in the scene. The Quake logo was correctly chosen for the pronoun *that* in spite of the presence of the table and background objects such as the walls, ceilings and floors. In Figure 8, the correct referent (the button

Table 2. Performance on different referring forms in the test corpus

RE Form	Count	Highest Ranked Absolute VC (a)	Highest Ranked Semantic Match VC (b)	LC (c)
A/An N	1	1	1	0
The N	19	1	9	10
This/These N	8	1	4	2
That/Those N	6	4	4	4
This/That/These/Those	13	6	7	5
It/Them/They	20	8	10	18
Total	67	21 (31.3%)	35 (52.2%)	39 (58.2%)

Fig. 7. Speaker's view when he said "yeah so ***that*** needs to go there"

Fig. 8. Speaker's view at "is there ***a second button*** there"

on the right) associated with the phrase *a second button* was identified by our system. Since this ambiguous expression is the first mention of this object in the discourse, it will never be resolved by the baseline system.

5 Conclusions and Future Work

In this preliminary study we have created a model that assigns a salience measure to each object in the visual context, and automatically updates this value as the speaker moves around in the world and interacts with his surroundings. We use only this salience assignment (along with some minimal assistance from other linguistic analyses, namely semantics) to determine the referents produced by the speaker in two different discourse segments, with promising results. Several aspects of the model were based on findings from related research on visual attention and memory. The results obtained encourage us to believe that the assumptions made were well-founded. Because our visual salience model performs strongly compared to a baseline using linguistically-determined salience, we expect that a fused model, using both sources of evidence, will perform better than currently available methods for tracking the attentional state of a dialog in service of reference resolution.

We intend to develop an automatic process to identify the different objects in the field of view of a user by employing geometric constraints and knowledge of the Quake world. Another stage of the pipeline that needs automation is the segmentation of the transcript into discourse units. With these improvements, we would be able to test our model on larger data sets, and also explore the effect of faster update rates ($>$ 1 Hz). Further, we also intend to incorporate visual characteristics of objects into our model to enable us to better discriminate between objects in the visual context based on their salience.

References

1. Saliha Azzam. Resolving anaphors in embedded sentences. In *Proceedings of the 34^{th} Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 263–269, 1996.
2. Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of ACL '87*, pages 155–162, 1987.
3. Donna K. Byron. Improving discourse management in TRIPS-98. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, 1999.
4. Donna K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of the 40^{th} Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 80–87, 2002.
5. E. Campana, J. Baldridge, J. Dowding, B. A. Hockey, R. W. Remington, and L. S. Stone. Using eye movements to determine referents in a spoken dialogue system. In *Workshop on Perceptive User Interfaces*.
6. S. E. Christ, C. S. McCrae, and R. A. Abrams. Inhibition of return in static and dynamic displays. *Psychonomic Bulletin and Review*, 9(1):80–85, 2002.
7. Miriam Eckert and Michael Strube. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89, 2000.
8. H. E. Egeth, R. A. Virzi, and H. Garbart. Searching for conjunctively defined targets. *J. Exp. Psychol:Human Perception and Performance*, 10:32–39, 1984.
9. Malte Gabsdil, Alexander Koller, and Kristina Striegnitz. Natural Language and Inference in a Computer Game. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING02)*, Taipei, 2002.
10. G. Gillund and R. M. Shiffrin. A retrieval model for both recogintion and recall. *Psychological Review*, 91:1–67, 1984.
11. Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.
12. R. Guindon. Anaphora resolution: short term memory and focusing. In *Proceedings of the 23^{rd} Annual Meeting of the Association for Computational Linguistics (ACL '85)*, pages 218–227, 1985.
13. P. Heeman and J. Allen. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, 1995.
14. J. E. Hoffman. A two-stage model for visual search. *Perception and Psychophysics*, 25:319–327, 1979.

15. M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 376–383, 2002.
16. Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th international conference on Multimodal interfaces (ICMI 2003)*, Vancouver, B.C., Canada, November 2003.
17. Andrew Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas, July 2000.
18. J. Kelleher and J. van Genabith. Exploiting visual salience for the generation of referring expressions. In *Proceedings of the 17th International FLAIRS conference*, 2004.
19. John Kelleher and Josef van Genabith. Dynamically updating and interrelating representations of visual and linguistic discourse. *submitted to Artificial Intelligence*.
20. John Kelleher and Josef van Genabith. Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review*, 21(3):253–267, 2004.
21. Pia Knoeferle, Matthew W. Crocker, Christoph Scheepers, and Martin J. Pickering. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, 2004.
22. Scheutz Matthias, Eberhard Kathleen, and Andronache Virgil. A real-time robotic model of human reference resolution using visual constraints. *Connection Science*, 2004.
23. U. Neisser. *Cognitive Psychology*. Appleton, Century, Crofts, New York, 1967.
24. Candace L. Sidner. Focusing in the comprehension of definite anaphora. In M. Brady and R. Berwick, editors, *Computational Models of Discourse*, pages 363–394. 1983.
25. M.K. Tanenhaus, M.J. Spivey-Knowlton, K.M. Eberhard, and J.E. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634, l995.
26. Joel R. Tetreault. *Empirical Evaluations of Pronoun Resolution*. PhD thesis, University of Rochester, 2004.
27. Marilyn A. Walker. Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264, 1996.
28. Terry Winograd. *Understanding natural language*. New York: Academic Press, 1972.