Between Linguistic Attention and Gaze Fixations in Multimodal Conversational Interfaces

Rui Fang
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
fangrui@cse.msu.edu

Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
jchai@cse.msu.edu

Fernanda Ferreira
Department of Psychology
University of Edinburgh
Edinburgh,UK EH8 9JZ
fernanda.ferreira@ed.ac.uk

ABSTRACT

In multimodal human machine conversation, successfully interpreting human attention is critical. While attention has been studied extensively in linguistic processing and visual processing, it is not clear how linguistic attention is aligned with visual attention in multimodal conversational interfaces. To address this issue, we conducted a preliminary investigation on how attention reflected by linguistic discourse aligns with attention indicated by gaze fixations during human machine conversation. Our empirical findings have shown that more attended entities based on linguistic discourse correspond to higher intensity of gaze fixations. The smoother a linguistic transition is, the less distance between corresponding fixation distributions. These findings provide insight into how language and gaze can be combined to predict attention, which have important implications in many tasks such as word acquisition and object recognition.

Categories and Subject Descriptors

H5.2 [**User Interfaces**]: Natural language

General Terms

Experimentation, Human Factors

Keywords

Linguistic Attention, Gaze Fixation, Multimodal Conversational Interfaces

1. INTRODUCTION

In human machine conversation, understanding human attention is key to the success of communication. Human attention is influenced by an individual's goals, the surrounding, as well as the conversation discourse that is jointly established by the human and the system through turn-taking and grounding. Recognizing attention not only depends on linguistic utterances, conversation discourse, but also non-verbal modalities. While attention has been studied extensively in both language processing [10] and vision processing [12], it is not clear how linguistic attention is aligned with visual attention during human machine conversation. On one hand, linguistic expressions and utterances reflect attention; and on the other hand, directions of gaze also indicate attention [14]. In multimodal conversation where a user can both talk to and look at graphical interfaces, it is not clear how the attention indicated by linguistic discourse (e.g., linguistic attention) is aligned with the attention reflected by gaze fixations. This is an important question since understanding of such alignment will provide insight on how processing of one mode can influence the other in automated computational systems (e.g., automated language processing and vision processing).

To address this issue, we conducted a preliminary investigation on the relationship between linguistic attention and gaze fixations. Here we consider linguistic discourse as a sequence of utterances produced by the user during conversation and visual discourse as the corresponding sequence of gaze streams during human speech production. At each point during conversation, linguistic expressions are used to indicate the focus of attention and tie the utterances together into a coherent discourse. Therefore we capture linguistic attention based on different types of *centers* as specified in Centering Theory [9], a linguistic theory that explains how linguistic expressions are tied together to form a coherent local discourse. Within the visual discourse, we model gaze fixations based on the intensity (i.e., length) of fixations on particular objects on graphical displays. To examine the alignment, we collected a rich set of language and gaze data based on mixed-initiative dialogues. Our empirical findings have shown that more attended entities based on linguistic discourse correspond to higher intensity of gaze fixations. The smoother a linguistic transition is, the less distance between corresponding fixation distributions. These findings provide insight on how language and gaze can be combined to predict attention, which have important implications in many tasks such as word acquisition and object recognition.

In the following sections, we first describe a multimodal conversational system used in our investigation. We then give a brief introduction to the Centering Theory and how it is used to capture linguistic attention and measure the coherence of the discourse. Finally we present the empirical results on the alignment between linguistic attention and gaze fixations.

2. RELATED WORK

The work presented here is motivated by the previous work on eye gaze in human language processing and eye gaze for language interpretation in human machine conversation.

Eye gaze is tightly linked to human language processing in both

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

comprehension and production [26]. This claim is supported by about a decade of psycholinguistic research. In object naming tasks, the onset of a word begins approximately one second after a speaker has looked at the corresponding visual referent [8], and gazes are longer the more difficult the name of the referent is to retrieve [7, 18]. About 100-300 ms after the articulation of the object name begins, the eyes move to the next object relevant to the task [18]. Moreover, objects are fixated in the same order in which they are spoken [1]. Recent work has also demonstrated that eye gazes at objects that are named incorrectly are no different from gazes at objects named accurately [8]. Most of these previous psycholinguistic studies were largely based on simple scene sketches containing only a few simple objects and all objects are relevant to the participant's task [11].

These psycholinguistic findings of time-locking behaviors between eye gaze and linguistic expressions have been applied in computational models that integrate eye gaze and speech. For example, previous work has shown that incorporation of eye gaze improves automated language processing at multiple levels from recognition of spoken hypotheses [6,22], to reference resolution [3, 20], and to automated vocabulary acquisition [17, 23]. Nevertheless, the interactive setting in human machine conversation is much more complex than the settings used in psycholinguistic studies. Recent work has shown that a natural temporal alignment exists between user speech and eye gaze, however, with a large variance compared to those observed in psycholinguistic studies [17]. Gaze fixation intensity serves as an integral role in attention prediction. When combined with visual features, fixation intensity can become even more reliable in predicting user attention [21].

3. MULTIMODAL CONVERSATIONAL SYSTEM

We developed a multimodal conversational system in the domain of treasure hunting. The application is based on a game engine and provides an immersive environment for users to navigate in a 3D castle. This castle has many rooms which contain a total of 115 3D objects. This application allows a user to consult with a remote "expert" (i.e., an artificial system) to find hidden treasures. The expert has some knowledge about the treasures but can not see the castle. The user has to talk to the expert for advice for finding the treasures. A mix-initiative dialogue is enabled to support interactions between a user and the expert. Using this application, we collected conversation data including speech and eye gaze through user studies. During each conversation, the user's speech was recorded, and the user's eye gaze was captured by a Tobii eye tracker. Figure 1 shows a snapshot of our 3D environment.

Figure 1: A snapshot of the 3D environment for the treasure hunting application. Each dot indicates a gaze fixation during speech production, which is invisible during conversation.

Table 1: A segment of conversation between a user and the system.

S_1	What do you see?
U_{1a}	There is *a clock* on *the dresser*.
U_{1b}	*The clock* is round.
U_{1c}	There is *a chair* in front of *the clock*.
S_2	Tell me more about *it*.
U_{2a}	*The chair* is metal and tall.
U_{2b}	And *a wooden chair* next to *it*.
S_3	What else do you see?
U_{3a}	*Two pictures* on *the wall*.
U_{3b}	*A bed* on *the floor*.

Table 1 shows a segment of conversation between the system and a user. S represents a system response and U represents a user utterance. Note that a user turn could comprise multiple utterances (e.g., U_{1a}, U_{1b}, and U_{1c}). This example segment shows a typical coherent discourse in the sense that both system responses and user utterances are not isolated, but rather follow the flow of conversation. The system's responses or user utterances are linked together through the use of linguistic referring expressions (italicized in Table 1), such as pronoun *it* (e.g., *it* in S_2 and U_{2b}) and definite noun phrases (e.g., *the chair* in U_{2a}). Within a local discourse, how expressions are formed and linked together from one utterance to another reflects underlying attention. For example, based on their grammatical roles in U_{1a}, the entity related to *a clock* is more attended than the entity associated with *a dresser*. The clock entity continues to be the attended object for U_{1b}.

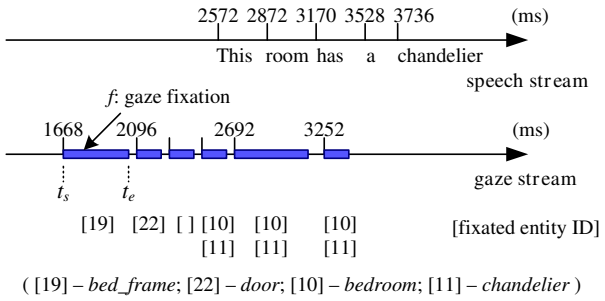

Figure 2: Parallel speech and gaze streams.

Figure 2 shows an excerpt of the collected speech and gaze fixations in one conversation. In the speech stream, each word starts at a particular timestamp. In the gaze stream, each gaze fixation has a starting timestamp t_s and an ending timestamp t_e. Each gaze fixation is also associated with a list of fixated entities (3D objects). An entity e on the graphical display is fixated by a gaze fixation f if the area of e contains fixation points of f. A gaze fixation could fixate on multiple entities due to the overlap of entities on the 3D display. In this case, only the forefront entity (e.g., the entity $chandelier$ in Figure 2) is used in our analysis. For each fixated object associated with a given utterance, the fixation intensity is measured by the length of corresponding fixations in milliseconds. For example, the fixation intensity for the entity $chandelier$ in Figure 2 is the combined length of the three consecutive fixations.

4. LINGUISTIC DISCOURSE

To model the linguistic discourse as shown in Table 1, we use Centering Theory [9]. Centering Theory attempts to relate focus of

attention and choice of referring expressions with the local coherence of a discourse. Centering Theory explicitly models *centers* of utterances, which serve to link adjacent utterances together. The theory aims to explain how centers from one utterance to another utterance form a coherent discourse. Therefore, there are two important notions in Centering Theory. The first is *Linguistic Centers* which reflect the focus of attention in our discussion. The second is *Discourse Transitions* which measure the degree of coherence within a local discourse.

4.1 Linguistic Centers

Given an utterance, three types of centers are identified in Centering Theory:

- **Forward looking centers (CF)**. These are a set of ordered entities mentioned in an utterance which are likely linked to by the succeeding discourse. These centers reflect different degrees of attended entities. There are different schemes to rank entities, but mostly are based on grammatical relations. For example, one ranking scheme indicates that an entity in a subject position is ranked higher than an entity in an object position, which is ranked higher than entities in other positions [9]. We adopted this scheme for our processing of forward looking centers. For example, the utterance U_{1a} in Table 1 has the following ordered forward looking centers corresponding to expressions: {*a clock*, *the dresser*}.
- **Preferred center (CP)**. This is the highest ranked forward looking center. Being highest ranked, the preferred center is most likely to be talked about or attended to in the succeeding utterance to form a coherent discourse. For example, in U_{1a} of Table 1, the preferred center is the referent to the expression *a clock*.
- **Backward looking center (CB)**. This is the highest ranked forward looking center from the preceding utterance that is also realized in the current utterance. This center can be considered to reflect the focus of attention up to the current utterance. For example, in utterance U_{1c} of Table 1, the backward looking center is the referent to the expression *the clock* (instead of *a chair*), which is the highest ranked entity from the previous utterance U_{1b}. Note that since we are interested in conversation, user utterances are influenced by the system's responses. This is particularly the case for identifying the backward looking center for the first user utterance in each user turn. For example, the backward looking center of U_{2a} in Table 1 is the referent to *the chair* since it is the highest ranked entity in S_2 (i.e., the referent to *it*). However, the backward looking center for U_{1a} is undefined since S_1 has no entity specified.

In summary, from the linguistic discourse, the backward looking centers imply the foci from the preceding discourse up to the current utterance and the preferred looking centers reveal the potential focus of attention in succeeding discourse. Therefore, we particularly use preferred centers and backward looking centers in our analysis presented later.

4.2 Discourse Transitions

Based on the centers described above, the Centering Theory provides a mechanism to assess the coherence of a local discourse. The coherence is measured by the movement of the centers from one utterance to another, which is further modeled as transitions. More specifically, three types of transitions are defined across two adjacent utterances: *continuation*, *retaining*, and *shift*. The shift relation is later extended to *smooth shift* and *rough shift* by Brennan et al [2]. These transitions are shown in Table 2. From this table, we can see that two criteria are used to determine the type of transition: whether the backward looking centers of two adjacent utterances are the same (i.e.,$C_b(U_{n+1})$ is the same as $C_b(U_n)$) and whether the backward looking center of the utterance is likely to be the preferred center of the next utterance (i.e., $C_b(U_{n+1})$ is the same as $C_p(U_{n+1})$). The degree of coherence thus is reflected through these transitions. Two utterances are more coherent if they share the same backward looking centers C_b.

Table 2: Four types of transition between two utterances.

	$C_b(U_{n+1}) = C_b(U_n)$ or $C_b(U_n)$ undefined	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Extending Table 2, we characterize linguistic discourse transitions into four types in our investigation:

- **Continue.** The backward looking center reflecting the focus of attention is the same as the preceding discourse and is likely to be continued for the succeeding discourse. For example, the transition between the utterances U_{1a} and U_{1b} (in Table 1) is *Continue*.
- **Retain.** The backward looking center reflecting the focus of attention is the same as the preceding discourse, however there is a tendency to move to a different focus for the succeeding discourse. For example, the transition between the utterances U_{1b} and U_{1c} (in Table 1) is *Retain*, so is the transition between U_{2a} and U_{2b}.
- **Shift.** We combine the smooth-shift and rough-shift together to a single type. Here the focus of attention reflected by the backward looking center is different from the preceding discourse.
- **Switch.** This represents a new type of transition in our dataset such as the transition between utterances U_{3a} and U_{3b}. The backward looking center for each of these utterances is not defined. So the transition is not strictly *Shift* based on the definition. Therefore, we introduce this *Switch* type to account for this new phenomenon where no entities are explicitly shared between the two utterances. This new type represents a complete focus shift from preceding discourse.

It is important to note that the above transitions only characterize the low level local transitions from one utterance to another. For example, *Switch* only represents the change of entities in focus within two utterances. It does not demonstrate the shift of the overall topic at the discourse level.

5. VISUAL DISCOURSE

We model our visual discourse by a sequence of gaze fixations that occurs simultaneously with speech production. More specifically, a visual discourse consists of two components: visual attention characterized by gaze fixations and visual transition.

5.1 Gaze Fixations

The first component relates to visual attention. Visual attention considers an object salient due to many factors including user intention, familiarity, physical characteristics and surrounding context

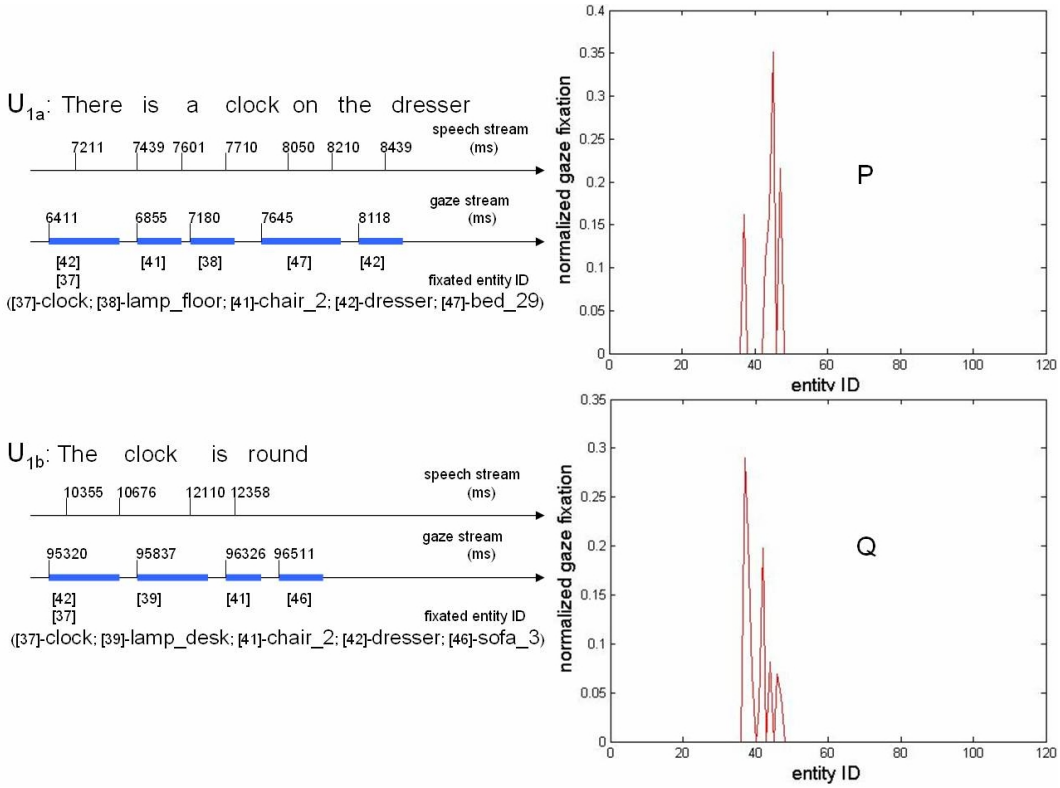

Figure 3: An example of gaze distributions based on fixation intensity.

of objects [12]. Here instead of examining complex visual attention, we simply rely on gaze fixations to serve as a proxy for visual attention since gaze directions indicate attention [14]. Within an utterance, since eye gaze could move to different objects, we use a distribution of gaze intensity on domain entities to represent the overall visual attention during speech production.

For example, Figure 3 shows the distributions of gaze intensity corresponding to the utterances U_{1a} and U_{1b} respectively. In the distribution figures, the X-axis represents the 115 objects modeled by the application. The Y-axis is normalized gaze intensity on each object.

5.2 Visual Transition

Similar to the linguistic discourse, the second component of the visual discourse consists of the transitions between utterances (or gaze streams). There could be different ways to measure the shift of gaze behaviors between utterances. In our current work, we simply measure the distance between the distributions of gaze intensity between two adjacent utterances using Jensen-Shannon Divergence(JSD) [16]. Jensen-Shannon divergence is a symmetrized and smoothed version of the Kullback-Leibler (KL) divergence [15]. It is defined as follows:

Let P and Q be two probability distributions over a random variable x, and let the average of the distribution of P and Q be M, $M = (P + Q)/2$, the JSD between P and Q is defined as:

$$JSD(P, Q) = \frac{D(P\|M) + D(Q\|M)}{2} \quad (1)$$

Where D is the Kullback-Leibler (KL) divergence. For two probabilities distributions p and q, the KL divergence is defined as:

$$D(p\|q) = \sum_x p(x) \times \log \frac{p(x)}{q(x)} \quad (2)$$

Currently, we have 115 objects modeled in the treasure hunting domain. In our case, P and Q are two gaze intensity distributions over the 115 objects as shown in Figure 3.

At each point of interaction, only a partial scene is displayed on the computer. Thus only a few objects on the display have gaze fixations. As a result, the gaze intensity distributions are rather skewed with many zero probability points. JSD provides a good measure to alleviate the problem with these zero probability points.

6. EMPIRICAL RESULTS

We conducted experiments as described in Section 3 to collect data for our investigation. In this section, we give a detailed description of the data used in our analysis and our empirical findings.

6.1 Annotated Data

We manually annotated conversation sessions with eight users in terms of the entities referred to by linguistic expressions. These annotations specify forward looking centers, backward looking centers, and preferred centers for each user utterance given a conversation discourse. The visual attention for each utterance and visual transitions between utterances are automatically captured based on corresponding gaze streams. Table 3 summarizes the data from eight users that were used in our investigation. In general, since a user turn could consist of multiple utterances, the total number of utterances for a given user is usually larger than the total number of turns. There are two situations that affect the number of utterances used in our analysis. First, since the Centering Theory is based on

Table 3: A summary of data from eight users.

User ID	Utterance	Turn	Centers				Transition			
			Type 1	Type 2	Type 3	Type 4	Continue	Retain	Shift	Switch
			notCP & notCB	CP & notCB	notCP & CB	CP & CB				
1	206	129	142	165	7	105	71	5	5	125
2	144	213	37	129	1	82	53	1	4	86
3	145	140	46	147	4	69	46	3	1	95
4	248	179	93	233	9	149	91	5	11	141
5	207	176	55	162	4	124	90	3	8	106
6	159	146	77	156	2	62	44	2	4	109
7	203	127	116	193	10	125	66	8	10	119
8	161	135	61	150	12	88	48	11	10	92

entities, we are only interested in utterances that contain nominals (e.g., nouns and pronouns). Utterances that do not contain nominals were not used in the analysis. Second, some users (e.g., user 2) did not follow the turn-taking behavior, for example, only providing responses after several system's requests. Therefore we see a significantly smaller number of utterances used in our analysis compared to the number of turns.

Given an utterance, every linguistic referring expression corresponds to a forward looking center. Depending on its ranking and the preceding utterance, a forward looking center may also serve as a backward looking center, or a preferred center, or both of them, or neither of them. Therefore, each referring expression uniquely relates to one of the following four types:

- **Type 1**: a forward looking center that is neither a preferred center nor a backward looking center (notCP & notCB).
- **Type 2**: a forward looking center that is a preferred center, but not a backward looking center (CP & notCB).
- **Type 3**: a forward looking center that is a backward looking center, but not a preferred center (notCP & CB).
- **Type 4**: a forward looking center that is a backward looking center and a preferred center (CP & CB).

Table 3 summarizes the number of these four types of centers that appear in user utterances. It is interesting to see that, across all users, a backward looking center is highly likely to also serve as a preferred center simultaneously. This means that an entity related to a backward looking center, being a focus of attention up to current utterance, is also very likely to be continued as focus for the succeeding discourse. Thus the backward looking centers that also serve as the preferred centers (i.e., Type 4) are most salient to reflect attention.

Table 3 also summarizes the number of different types of transitions between the utterances used in our analysis. In our data, across all users, we see fewer occurrences of *Retain* and *Shift* compared to *Continue* and *Switch*.

Based on this dataset, we specifically investigated two questions: (1) whether and how linguistic centers are linked with gaze fixations; and (2) how visual transitions correspond to linguistic transitions. Next we describe empirical results addressing these two questions.

6.2 Linguistic Attention and Gaze Fixations

The explicit modeling of centers in the Centering Theory offers a means to identify salient object or focus of attention from linguistic utterances. The backward looking centers represent the linguistic attention up to the current utterance and the preferred centers represent entities that are most likely attended to in the succeeding utterance. Therefore, one question is whether the linguistic attention related to the preferred centers and backward looking centers correlates with visual attention indicated by gaze fixations. To address this question, we examined the data along three dimensions as discussed next.

6.2.1 *Preferred Centers*

Our first analysis is based on the linguistic attention as indicated by the preferred centers. Since a preferred center is most likely to be attended in the succeeding discourse, our hypothesis is that the preferred center should capture more gaze fixations than the centers that are not preferred. To validate this hypothesis, we compared the average gaze fixation intensity between preferred centers and forward looking centers that are not preferred.

The mean gaze fixation intensity of these two types of centers averaged across all users are shown in Figure 4. Overall, the preferred centers correspond to significantly higher fixation intensity ($t = 4.74$, $DF = 1724.3$[1], $P < 0.001$). The results on individual users are shown in Figure 5. Except for the first three users, the average gaze intensity corresponding to preferred centers is significantly higher than that corresponding to centers that are not preferred for the remaining five users (user 4: $t = 2.49$, $DF = 213.7$, $P = 0.0067$; user 5: $t = 2.33$, $DF = 162.4$, $P = 0.01$; user 6: $t = 2.23$, $DF = 234.5$, $P = 0.0135$; user 7: $t = 2.86$, $DF = 362.0$, $P = 0.0022$; user 8: $t = 3.37$, $DF = 260.2$, $P < 0.001$).

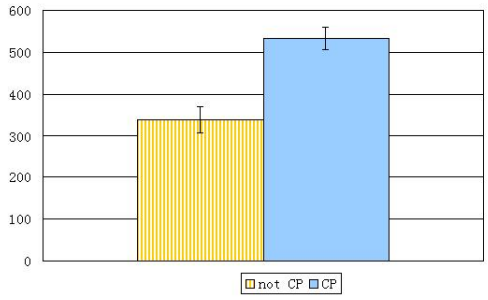

Figure 4: Overall comparison of the mean fixation intensity corresponding to preferred centers and centers that are not preferred.

6.2.2 *Backward Looking Centers*

[1]The degree of freedom is obtained based on Satterthwaite Approximation.

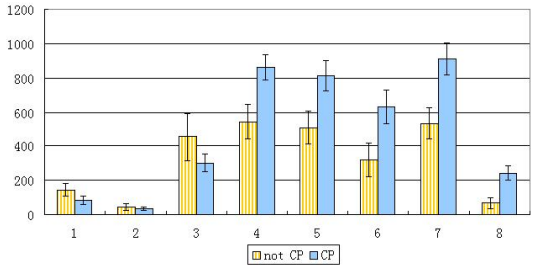

Figure 5: Comparison of the mean fixation intensity corresponding to preferred centers and centers that are not preferred for individual users.

Since a backward looking center carries the attended entity that is highest ranked in the preceding discourse, we hypothesized that it would capture more gaze fixations than the regular forward looking centers which are not backward. To validate this hypothesis, we specifically compared the average gaze fixation intensity for backward looking centers and non-backward looking centers.

The mean gaze fixation intensity of these two type of centers averaged from all users is shown in Figure 6. Overall, the backward looking centers correspond to significantly higher gaze intensity ($t = 8.75$, $DF = 1059.8$, $P < 0.001$).

Results for individual users are shown in Figure 7. Except for user 1 and user 2 who show no significant difference, the mean gaze fixation intensity of backward looking centers is significantly higher than the centers that are not backward looking for the remaining six users (user 2: $t = 2.20$, $DF = 87.1$, $P = 0.0153$; user 4: $t = 3.58$, $DF = 233.4$, $P < 0.001$; user 5: $t = 4.92$, $DF = 143.7$, $P < 0.001$; user 6: $t = 4.00$; $DF = 69.2$; $P < 0.001$; user 7: $t = 4.65$; $DF = 176.7$; $P < 0.001$; user 8: $t = 2.33$; $DF = 126.8$; $P = 0.018$).

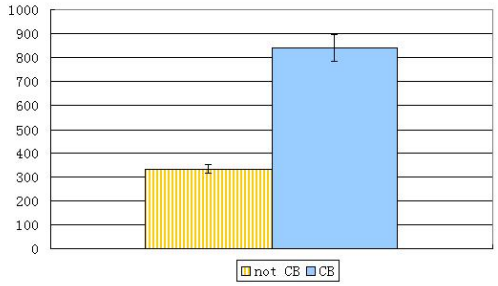

Figure 6: Overall comparison of the mean fixation intensity corresponding to backward looking centers and centers that are not backward looking.

6.2.3 A Combined Analysis

As shown in Section 6.1, a preferred center could simultaneously serve as a backward looking center, or vice versa. In order to understand the interaction between these centers, we further analyzed data based on the four types of combinations described in Table 3.

Overall, as shown in Figure 8, there is a significant difference between the mean fixation intensity among four types of centers (ANOVA, $F_{3,2811} = 40.97$, $p < 0.001$). A follow-up Scheffe pos hoc test has shown that Type 4 (i.e., both backward looking and preferred) centers have a significantly higher mean fixation intensity than Type 1 centers and Type 2 centers. There is no significant

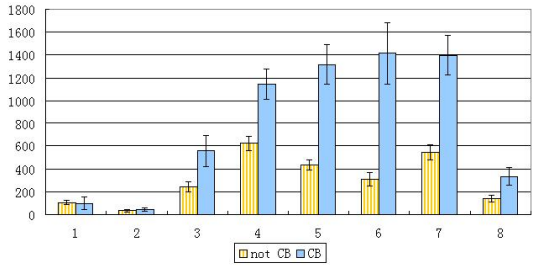

Figure 7: Comparison of the mean fixation intensity corresponding to backward looking centers and centers that are not backward looking for individual users.

difference between Type 1 centers and Type 2 centers. Type 3 has very few occurrences in our dataset (see Table 3). It has not shown statistically significant differences from any other types of centers due to its large variance.

Results for individual users are shown in Figure 9. Again, user 1 and user 2 do not exhibit statistically different behavior among these four types of centers. The average fixation intensity corresponding to these four types of centers is significantly different for the rest of users (user 3: $F(3, 262) = 4.32$, $P = 0.0054$; user 4: $F(3, 480) = 6.36$, $P < 0.001$; user 5: $F(3, 341) = 12.6$, $P < 0.001$; user 6: $F(3, 293) = 13.77$, $P < 0.001$; user 7: $F(3, 440) = 10.97$, $P < 0.001$; user 8: $F(3, 307) = 3.9$, $P = 0.0092$). A Scheffee pos hoc test has shown that, among all the remaining six users, gaze intensity corresponding to Type 1 and Type 2 is significantly less than that with Type 4. There is no significant difference between Type 1 and Type 2. Type 3 again has not shown significant difference from other types within each user.

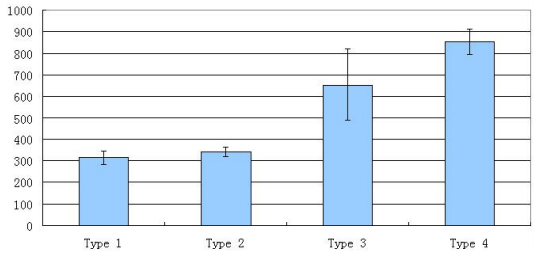

Figure 8: Overall comparison of the mean gaze intensity of the four types of centers.

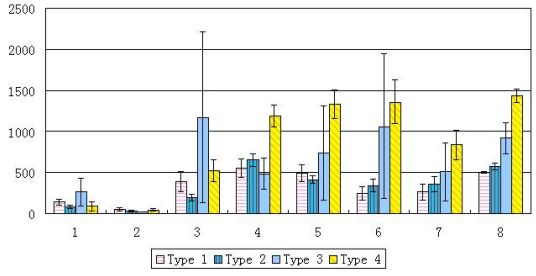

Figure 9: Comparison of the mean gaze intensity of the four types of centers for individual users.

6.3 Linguistic Transitions and Visual Transitions

We further examined how linguistic transitions correspond to visual transitions. We particularly investigated four types of linguistic transitions: *Continue*, *Retain*, *Shift*, and *Switch* (as mentioned in Section 4).

We computed the Jensen-Shannon Divergence (JSD) between the distributions of gaze fixation of two consecutive utterances to represent the corresponding visual transition. We further examined whether different types of linguistic transitions aligned with different degrees of visual transition. The result is shown in Figure 10. Our analysis shows that there is a significant difference among these four types of transitions ($F_{3,1469} = 29.4$, $P < 0.001$) The follow-up Scheffe post hoc test shows no significant difference in visual transition between the transition type *Retain* and *Shift*. The degree of visual transition is significantly higher for the *Switch* type compared to the *Continue* type.

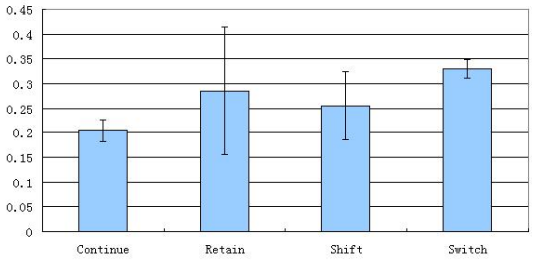

Figure 10: Correspondence between linguistic transitions and visual transitions measured by Jensen-Shannon divergence.

7. DISCUSSION

Our empirical results have shown that five out of the eight users have consistently demonstrated higher fixation intensities associated with linguistically more attention-demanding entities. The remaining three users have somewhat different behaviors. For user 1 and user 2, the gaze intensities corresponding to different centers have no significant difference. A further look at the data indicates that for these users, most gaze streams corresponding to utterances are either not captured or are not "closely coupled". We consider speech and gaze are *closely coupled* if the gaze stream has fixations on at least one object mentioned in the corresponding speech. Although more studies are required to understand the actual differences and what causes these differences, the current observation suggests the possibility of different user behaviors. Therefore, for eye gaze to be useful for language processing, some user modeling may be necessary to first identify closely coupled speech and gaze streams.

Our separate analysis on preferred centers and backward looking centers has shown that preferred centers correspond to higher fixation intensity than non-preferred centers (i.e., Figure 4) and backward looking centers correspond to higher fixation intensity than centers that are not backward looking (i.e., Figure 6). A detailed analysis (i.e., Figure 8) has revealed that backward looking centers are the major contributors to gaze fixations. The higher fixation intensity associated with preferred centers is mainly caused by the preferred centers that are also backward looking. When a center is not backward looking, whether it is preferred or not will not make a significant difference on fixation intensity (see Figure 8). On the other hand, when a center is backward looking, it is most likely also a preferred center and captures higher gaze intensity compared to other types of centers.

These observations can provide insight into how to use gaze fixations to help automated language processing. Previous work has shown that, in multimodal systems involving speech and gestures [4, 5, 13, 24, 25], incorporating gestures enables more robust and stable input interpretation than speech only systems. Fusing two or more information sources can be an effective means of reducing recognition uncertainties, for example through mutual disambiguation [19]. Since speech recognition and language processing still face many challenges, incorporating eye gaze can potentially help identify attention and facilitate interpretation of linguistic expressions. For example, a higher JSD between gaze distributions may indicate a shift of attended objects, which may not be detected if the speech is not correctly recognized and understood. For another example, the alignment between linguistic centers and gaze fixations will provide more reliable mappings between linguistic expressions (particularly nouns) and fixated visual objects. These mappings between words and objects will provide training data for unsupervised learning to enable better vocabulary acquisition for automated language processing and facilitate object recognition for automated vision processing.

8. CONCLUSION

This paper describes a preliminary investigation of how attention reflected by linguistic expressions in a conversation discourse is aligned with attention indicated by eye gaze. We use Centering Theory to model linguistic attention and gaze fixation intensity to measure visual attention. Our empirical findings have shown some interesting alignments between linguistic attention and gaze fixations. Nevertheless, this work is still at the beginning. Many interesting questions remain. Our future work will investigate these questions, for example, examining how the different forms of referring expressions are linked with eye gaze in human machine conversation. We will further incorporate our empirical findings in computational models to improve language processing such as reference resolution and word acquisition.

9. ACKNOWLEDGMENT

This work was supported by IIS-0535112 from the National Science Foundation. We thank Zahar Prasov and Shaolin Qu for their contributions to system development, experiments, and data collection and annotation. We also thank anonymous reviewers for their valuable comments and suggestions.

10. REFERENCES

[1] K. Bock, D. Irwin, and D. Davidson. Putting first things first. In J. M. Henderson and F. Ferreira, editors, *The Interface of Language, Vision, and Action: Eye Movements and the visual world*, pages 249–278. New York: Psychology Press, 2004.

[2] S. E. Brennan, M. W. Friedman, and C. Pollard. A centering approach to pronouns. In *In Proceedings of 25th Annual Meeting of Association for Computational Linguistics*, pages 155–162, 1987.

[3] D. Byron, T. Mampilly, V. Sharma, and T. Xu. Utilizing visual attention for cross-modal coreference interpretation. In *Proceedings of the Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, pages 83–96, 2005.

[4] J. Chai, P. Hong, M. Zhou, and Z. Prasov. Optimization in multimodal interpretation. In *Proceedings of 42nd Annual*

Meeting of Association for Computational Linguistics (ACL), pages 1–8, 2004.

[5] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of ACM Multimedia*, page 31ÍC 40, 1996.

[6] N. J. Cooke and M. Russell. Gaze-contingent asr for spontaneous, conversational speech: an evaluation. In *International Conference in Acoustics, Speech and Signal Processing).*, 2008.

[7] Z. M. Griffin. Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, 82:B1–B14, 2001.

[8] Z. M. Griffin. Why look? reasons for eye movements related to language production. *The Interface of Language, Vision, and Action*, pages 213–247, 2004.

[9] B. J. Grosz, A. K. Joshi, and S. Weinstein. centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225, 1995.

[10] B. J. Grosz and C. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[11] J. Henderson and F. Ferreira. *The Interface of Language, Vision, and Action: Eye Movements and Visual World*. Taylor and Francis, New York, 2004.

[12] L. Itti and C. Koch. Computational modelling of visual attention. *Nat Rev Neurosci*, 2:194–203, 2001.

[13] M. Johnston, S. Bangalore, G. Visireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. Match: An architecture for multimodal dialog systems. In *in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 376–383, 2002.

[14] M. Just and P. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480, 1976.

[15] S. Kullback. The kullback-leibler distance. *The American Statistician*, 41:340–341, 1987.

[16] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.

[17] Y. Liu, J. Y. Chai, and R. Jin. Automated vocabulary acquisition and interpretation in multimodal conversational systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2007.

[18] A. S. Meyer, A. Sleiderink, and W. J. M. Levelt. Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, 66:B25–B33, 1998.

[19] S. L. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *In Proceedings of Conference on Human Factors in Computing Systems: CHI '99*, 1999.

[20] Z. Prasov and J. Y. Chai. What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of ACM 12th International Conference on Intelligent User interfaces (IUI)*, pages 20–29, January 2008.

[21] Z. Prasov, J. Y. Chai, and H. Jeong. Eye gaze in attention prediction in multimodal human machine conversation. In *Proceedings of the AAAI 2007 Spring Symposium on Interaction Challenges for Artificial Assistants*, March 2007.

[22] S. Qu and J. Y. Chai. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *Proceedings of the Conference of the North America Chapter of the Association of Computational Linguistics (NAACL)*, pages 284–291, April 2007.

[23] S. Qu and J. Y. Chai. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2008.

[24] F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. McCullough, and R. Ansari. Multimodal human discourse gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9:171–193, 2002.

[25] R. Sharma, S. Kettevekov, and M. Yeasin. *Integration of Gesture and Speech in Multimodal Interface*. Computer Science Handbook for Displays, 2001.

[26] M. K. Tenenhaus, M. Sivey-Knowlton, E. Eberhard, and J. Sedivy. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268:1632–1634, 1995.