The 14th International Conference on Auditory-Visual Speech Processing
25-26 August 2017, Stockholm, Sweden

Referential Gaze Makes a Difference in Spoken Language Comprehension: Human Speaker vs. Virtual Agent Listener Gaze

Eva M. Nunnemann [1], *Kirsten Bergmann* [1], *Helene Kreysa* [2], *Pia Knoeferle* [3]

[1] CITEC, University of Bielefeld, Germany
[2] Department of General Psychology, Friedrich Schiller University Jena, Germany
[3] Department of German Language and Literature, Humboldt University at Berlin, Germany

enunnemann@techfak.uni-bielefeld.de, kbergman@techfak.uni-bielefeld.de,
helene.kreysa@uni-jena.de, pia.knoeferle@hu-berlin.de

Abstract

An interlocutor's referential gaze is of great importance in face-to-face communication as it facilitates spoken language comprehension. People are also able to exploit virtual agent gaze in interactions. Our study addressed effects of human speaker gaze vs. virtual listener gaze on reaction times, accuracy and eye movements. We manipulated: (1) whether the human speaker - uttering the sentence - was visible, (2) whether the agent listener was present and (3) whether the template following each video matched the scene. Participants saw videos in which a static scene depicting three characters was visible on a screen. We recorded participants ' eye movements as they listened to German SVO sentences describing an interaction between two of these three characters. After each trial a template schematically depicting three characters and their interaction appeared on screen. Participants verified congruence between sentence and template. Participants solved the matching task very well across all conditions. They responded faster to matches than mismatches between sentence and template. Participants were slower when the agent was present. Eye movement results suggest that during the NP2 region participants tended to look at the NP2 referent to a greater extent when the speaker was present compared to the other conditions.

Index Terms: spoken language comprehension, human speaker gaze, virtual agent, listener gaze, eye tracking

1. Introduction

"Gaze is a powerful expressive signal that is used for many purposes, from expressing emotions to regulating human interaction" [1, p.7]. In face-to-face communication these regulating functions of eye gaze comprise the organisation of turn-taking, the request for feedback, as well as a means for emphasizing parts of an utterance. Gaze is also an important signal for the detecting an interlocutor's focus of attention in an interaction (cf. [2, 3, 4]). Already newborn infants are able to detect direct eye contact and at four months of age they can follow the direction of a perceived gaze shift [5]. Adults detect the direction of another's gaze very robustly [6].

Thus, it is not surprising that gaze has become a much investigated topic in research on spoken language comprehension. Studies have examined the beneficial effects of gaze in joint-search tasks [7, 8]. But referential gaze did not only prove to be helpful when the task required collaboration. Passive listeners were also able to rapidly exploit the informativeness of speaker gaze cues [9]. In their experiments a speaker's gaze, which was directed at a target object before it was mentioned, helped participants to disambiguate this target even before it was fully named. Besides that, speaker gaze also had a beneficial effect on the understanding of event roles in a visual world paradigm study by [10]. They showed that people were able to follow a speaker's gaze to a target referent already before she started naming it. Thus, a speaker's gaze helps listeners to anticipate which referent will be mentioned next and to direct their attention towards it.

1.1. The effects of artificial gaze

However, people are not only able to detect and make use of gaze in human-human interaction, but they can exploit artificial gaze cues as well. [11] found that participants established joint-attention with a robotic agent. Although the head and eye movements of the robot they used in their experiments were rather rudimentary, people could still identify the target object, which the robot gazed at, without any difficulties. That the ability to detect a robotic agent's gaze direction in a human-robot collaboration task is robust in human participants was shown by [12]. In one of their experiments people learned to infer the robot's gaze direction quickly only from his head movement in a condition where its eyes were covered by sunglasses.

Even though virtual agents are not perceived as physical entities in the same manner as robots are, a great deal of research has shown that people also react to the gaze of a virtual agent (e.g. [13, 14]). Studies investigated a great variety of individual aspects in gaze behaviour that play a role in human-agent interaction. Among these were gaze aversion and its role in regulating the flow of a conversation [15] or the establishment of rapport in a listening agent [16]. As virtual agents are mainly applied in teaching or learning environments, their affiliative and referential gaze behaviours have recently become prominent topics in research (e.g. [17, 18]).

Consistently with [13], who report that the display of poor or unnatural gaze behaviour in an agent can be worse than no gaze behaviour at all, [19] found that an animated agent who displayed human-like gaze behaviour attracted participants' attention faster than one with either a static gaze or one showing a stepped gaze behaviour, which only consisted of two distinct images. A recent study by [13] showed that virtual agent gaze behaviour can have beneficial effects on learners' recall of study materials. In their experiment Andrist and colleagues (2012) found that participants could remember the taught content better when the virtual teacher, who gave a lesson on history, gazed into the general direction of the learning materials (e.g. a map of ancient China) while talking than when he exclusively looked at the participant.

Overall, the application of human-like gaze behaviour in a virtual agent proved to have beneficial effects for the com-

10.21437/AVSP.2017-4

Figure 1: *The four different conditions in the experiment for the sentence "The waiter congratulates the millionaire".*

munication, such as the facilitating task performance, enhancing learning or enhancing the perception of the agent as being autonomous and natural (e.g.[14, 13, 20]).

1.2. Human speaker vs. agent listener gaze

Although a great variety of research has looked at the effects of either human or virtual agent gaze behaviour in communicative situations (e.g. [9, 14]), to our knowledge none of them has directly contrasted these two types of gaze yet (see [21]). The evidence that people can exploit virtual agent gaze cues when they alone are available remains uncontested. But the question whether people use these artificial gaze cues in exactly the same way as they do with a human interlocutor's gaze cues, is left open. It can only be answered if the two gazes, namely that of a human and of a virtual agent interlocutor are presented at the same time. That means that their gaze information is available simultaneously. Moreover, the acceptance of a gaze cue might also be dependent from whether the interlocutor is a speaker or a listener. Another question - yet unanswered - is how the presence of these different gaze types affects the comprehension of information from a communication situation.

The present experiment investigated these open questions by directly contrasting the speaker gaze of a human with the listener gaze of a virtual agent, which were present on a screen at the same time. Moreover, it looked at the influence of these two referential cues on the comprehension of spoken sentences describing a visually available scene.

2. Experiment

2.1. Method and design

2.1.1. Participants

Thirty-two German native speakers aged between 18 and 30 (mean age 23) took part in the experiment after giving written consent. Their sight was normal or corrected-to-normal. For their participation they were paid 6 Euro. The study was approved by the ethics committee of the University of Bielefeld.

2.1.2. Materials and design

Our experiment used 24 item videos as well as four practice videos. Part of the materials for all these clips came from [10], who had created videos displaying a computer screen with three clearly identifiable static characters placed on a landscape and a human speaker sitting to the right of this screen. The characters for their 24 critical items as well as those for most practice trials came from the online game SecondLife®. The remaining characters originated from clip art programs and were in turn displayed against a neutral white background.

Each of the item videos was accompanied by a grammatically correct, unambiguous German SVO sentence describing a transitive action between the character visible in the middle of the screen (e.g. a waiter) and one of the two outer characters (e.g. the saxophonist and the millionaire; see Figure1). An example sentence is *'Der Kellner beglückwünscht den Millionär'* (The waiter congratulates the millionaire.) In each of the video clips, the speaker is positioned next to the screen at an angle that allowed participants to clearly see her head and eye movements. She always looked at the camera first - smiling at the participant - before she turned towards the screen inspecting each of the three characters in a fixed order. Subsequently, she turned her gaze to the central character, which was always the NP1 referent of the sentence she utters. During the whole utterance, she always looked at the respective character, displaying a gaze shift from the NP1 referent towards the NP2 referent shortly after mentioning the verb.

For our experiment we embedded these "speaker videos" into video clips showing the virtual agent Billie [22] as a listener. In order to produce these videos, we first transcribed the materials from [10] in the transcription software ELAN [23]. This procedure allowed us to extract an exact time course for the speaker gaze for each item. With the data from the transcription we then calculated the time course for Billie's listener gaze behaviour, i.e. the delay with which the agent listener followed the human speaker's gaze towards the referents of the spoken sentences. In this way, we reproduced the speaker's gaze and smile behaviour in the virtual agent Billie, but delayed by 400 ms (this delay was selected based on a pilot test comparing different delays). Furthermore, before the "speaker videos" from [10] appeared on the screen, Billie was already visible and he gazed and smiled at the participant for about 1000 ms. This is a replicated the human speaker's behaviour. Billie's rendered movements were coded in Behavior Markup Language (BML) [24], executed using AsapRealizer [25] and recorded. For the embedding of the "speaker video", we beveled the video at an angle of 40° to make Billie gaze at the speaker as well as the characters depicted in the videos (cf. [10]), while also enabling participants to clearly recognize where the virtual agent was actually looking. To ensure participants could correctly identify Billies locus of gaze, we conducted a pretest on the stimuli.

The design of the experiment included three within-subject factors. The first one is *Speaker Gaze* with the two levels *speaker gaze* and *no speaker gaze*. The second factor is *Agent Gaze*. Correspondingly to *Speaker Gaze* it has the levels *agent gaze* and *no agent gaze*. The third factor comes from the verification task participants solved after each video and captures the *Congruency* between the content of the spoken sentence from the video and a response template (levels *yes* and *no*). The Gaze conditions were distributed over the experiment in a manner that in 50 % of the total videos the human speaker was visible, while in the other 50 % she was obscured (see Figure 1). Also the virtual agent listener was only visible in half of all videos. The overall configuration of interlocutor visibility was distributed in such a way that 25 % of clips showed no interlocutor, in 25% both were present on the screen, in another 25% only the virtual agent was present, and in a final 25% only the human speaker was present. In addition, the referent of the NP2 appeared equally often to the right and to the left of the NP1 referent in the middle of the static scene, which means the human speaker and the agent shifted their gazes equally often to the left and to the right. Also, we balanced for handedness in the response task. Half of the participants had to press the yes button on the CEDRUS box with their right and the no button with their left hand, while the other half pressed the yes button with their left and the no button with their right hand.

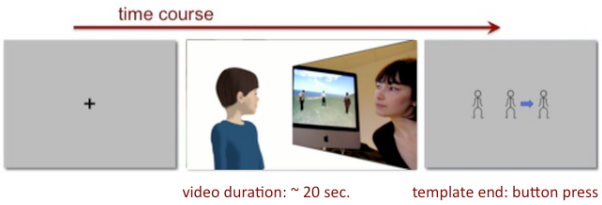

Figure 2: *The time course of a trial from left to right: fixation cross, video clip, verification template.*

2.1.3. Procedure

An Eyelink 1000 desktop head-stabilized tracker (SR Research) monitored participants' eye movements and recorded their response latencies after each video during the first part of the experiment. The stimuli were shown on a computer screen with a resolution of $1680 * 1050$ pixels. We tracked only the right eye, but participants' vision was binocular. We instructed participants to watch the videos closely because there would be a verification task after every video and a memory test in the second part of the experiment. Participants were informed about this memory test right at the beginning of the experiment to ensure they concentrated on the video materials.

All trials followed the same structure (Figure 2). Before each individual trial a fixation cross appeared in the centre of the computer screen, which participants were instructed to fixate. They then watched a video in which the static screen with the three characters and either nobody, one of the two interlocutors (human speaker or agent listener) or both were visible, and heard a sentence describing the scene. This sentence always involved the character in the middle as referent of the NP1 acting upon one of the two outer characters.

After each of these videos, a grey template appeared on the screen depicting the static scene from the video schematically. Three stick men represented the three characters from the previously seen video and a blue arrow depicted the action mentioned between two of them. Participants' task was to decide via button press whether the blue arrow represented the action correctly.For example in the item video for the sentence *Der Kellner beglückwünscht den Millionär* ('The waiter congratulates the millionaire') the waiter was standing in the middle and the millionaire to his right. On the template the arrow pointed from the centered stick man (representing the waiter) to the outer one (in this case the millionaire) on the right. That means, the template depicted the scene from the video correctly and the participant should have pressed the button for yes (see Figure 2).

2.2. Expectations

The human speaker starts shifting her gaze towards the NP2 referent already during the verb region of the uttered sentence. Thus, this visual cue enables participants to identify the character before its mention. Based on extant findings, we would expect listeners to rapidly follow speaker gaze when it is available ([10, 9, 11, 15]).

Alternatively, we might observe hardly any effect of human speaker gaze on listener gaze behavior. Instead, listeners could mostly focus on (and follow) the virtual agents gaze. One reason for such a listener gaze behavior might be that the virtual agent (but not the human speaker) carries a high degree of novelty. Novelty might attract attention, and thus result in listeners gaze following the virtual agent instead of the human interlocutor. The effects of novelty (perhaps coupled with artificiality) could, however, yield a different listener gaze behaviour [26]. If the listeners reject the idea of a virtual agent as an interlocutor of the human speaker, they might chose to focus exclusively on the human speaker and ignore the virtual agent. The fact that the human interlocutor has the role of speaker might enhance this listener gaze response.

In the condition that features both kinds of interlocutors and thus gaze (i.e. human speaker and virtual agent listener gaze) simultaneously, various outcomes are possible. First, only one kind of gaze might be helpful for participants, namely either speaker or agent listener gaze. Moreover, it is also possible that both types of gaze cues in combination turn out to be beneficial for detecting the NP2 referent. This is conceivable, as people do not need to look directly at the interlocutors to detect their gaze direction, but can perceive it peripherally (cf. [10]). Last but not least, two visual cues at the same time could also be distracting. This would manifest itself in participants being slow to look at the target character.

These possible findings might also have consequences for the verification task. Generally, we would expect to find faster response times when the role relations on the template are depicted correctly [10]. In case a gaze cue is helpful, this should have a positive effect on participants ' verification task performance (e.g. [13]), namely that participants are faster when the gaze was visible in the video. If the availability of a certain gaze type or of both in combination is not helpful or even distracting, this might also be visible in slower reaction times for conditions where the gaze was visible.

2.3. Analysis

We computed response times (RT) in the verification task from the onset of the template until participants' button press. In the analysis of the log-transformed RTs we only included accurate trials. The analysis was conducted using linear mixed models with crossed random intercepts and slopes for participants and for items. Following [27] we started out with the most complex converging model and then used backward selection to determine the simplest model with comparable goodness of fit that contained at least all manipulated factors as fixed effects.

For the analyses of the eye movements, two critical time windows were determined in the video. The first of these windows is the *shift time window*, which contains all fixations that started after the human speaker's gaze shift and before the mean onset of the NP2 (approx. 719 ms after shift onset).The second time window is the *NP2 time window* and it comprises all the fixations which started in the first 700 ms after the onset of the NP2. These two time windows were further subdivided into 100ms time bins. We then analysed the log-gaze probability ratio with which participants were likely to fixate the target character (the NP2 referent) over the competitor (the third unmentioned character). In order to analyse this log-gaze probability ratio, we fitted separate linear mixed effects models for participants and items. Moreover, instead of including congruency as a third fixed effect in the models for the *shift* and *NP2 time windows*, time is introduced as factor into the models. This being the main difference to the procedure of model fitting for the RTs, we again followed backward selection [27] to fit the optimal models for the eye movement data. We obtained these optimal models when the removal of a term resulted in a significant decrease of model fit as compared to the next complex one or when the model only contained only main effects.

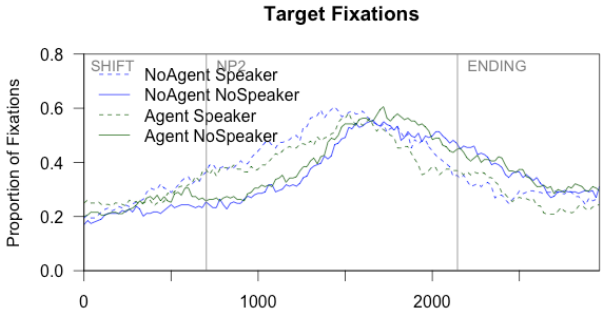

Figure 3: *Time course of participants' fixations to the target character (NP2 referent) in ms from speaker gaze shift. The mean on- and offset of the NP2 are inserted as vertical lines.*

2.4. Results

2.4.1. Accuracy and reaction time results

After each video, participants verified whether the template depicted the visual scene correctly. For 744 out of 768 critical trials participants gave the correct answer (96.7% of cases). There was a maximum of three errors per participant (12.5%) as well as a maximum of three errors per item (9.4%). However, neither human speaker gaze nor virtual agent listener gaze had any effect on accuracy (*p*s $> .88$).

In the final model ($log\_RT \sim agent * speaker * congruency + (1 + agent + congruency|participant) + (1 + agent + condruency|item)$) the factors congruency and agent both significantly affected RTs. Participants responded faster in those cases where the verification template matched the described visual scene from the corresponding video ($p < .001$) than when it did not. Moreover, participants were slower to answer on trials in which agent listener gaze was available ($p < .001$) than when it was not visible.

2.4.2. Eye movement results

Figure 3 shows the time course of participants' fixations on the target character (the NP2 referent) for 3000 ms from the onset of the human speaker's gaze shift as a function of speaker and agent gaze. It illustrates participants' attention to the target character, i.e. the character that was mentioned in the NP2, in all four conditions. The most striking observation here is that participants start fixating the NP2 referent earlier and more in those two conditions where speaker gaze was available. This development starts when the human speaker begins to mention the target character. Agent listener gaze alone does not make a great difference in comparison to the baseline condition in which neither speaker nor agent gaze were available. Participants in those two conditions start looking at the NP2 referent later than in conditions where the speaker was present.

Figure 3 also illustrates the finding that neither human speaker gaze nor virtual agent listener gaze affect participants ' behaviour towards the NP2 referent during the *shift time window*(all *p*s $> .7$). However, the time factor showed a significant effect (*p*s $< .05$) in both the by-participants and by-items models, reflecting a change in participants ' fixation behaviour towards the target over time. Moreover, in the final model for the participant-analysis, there was a trend towards an interaction between time and speaker ($p = .09$).

In the NP2 time window participants looked far more to the target character than to the competitor character, which is evidenced in significant intercepts in both models ($p < .001$). This is not surprising, as the speaker mentions the NP2 referent during this late time window. We also find main effects of speaker and time in the models for participants and items (all *p*s $< .003$). People looked more to the target character when speaker gaze was available as a visual cue than in the conditions where it was absent. The main effect of time indicates an increase in looks to the target character over the time course of this late region. Finding this main effect in the NP2 window is not surprising as the speaker mentions and thereby clearly identifies the NP2 referent here. Again, as in the previous time region, we found an interaction between speaker and time in the model for participants. Here the interaction is significant ($p = .02$) and indicates that participants look more to the target as time passes.

In sum, the only consistently significant effect across both time windows and all analyses was the time window factor, indicating a gradual increase in looks to the target as participants processed the sentence and identified the NP2 referent. More interestingly, virtual agent listener gaze did not have an effect on participants ' gaze behaviour towards the target character at any point. Instead, participants made use of speaker gaze only, although the effect was significant only in the later NP2 region, when the speaker mentioned and thus unambiguously identified the target.

3. Discussion

The present study examined whether the co-presence of human speaker gaze and virtual agent listener gaze had an effect on the comprehension of spoken sentences describing a visual scene. We wanted to assess whether there are any similarities in the exploitation of a human and a virtual interlocutor's gaze or whether one is preferred over the other when both gaze cues are available simultaneously. We looked at these two open questions by tracking participants' eye movements while they watched short video clips showing a human speaker and a virtual agent listener to the sides of a static display with three characters. While watching, participants heard an unambiguous German SVO sentence describing an interaction between two of these characters. After each of the trials participants verified the accuracy between the video clip and a schematic template. Participants' accuracy in the rating task was very high across all conditions. Their response times were faster when the template matched the video content. However, the presence of the virtual agent listener had a negative influence on their response latencies. Neither speaker nor agent gaze turned out to elicit more looks to the target character before its mention. Speaker gaze though, affected participants' visual attention by eliciting more looks to the NP2 referent at an early stage in the NP2 time window in conditions where she was present. These findings show that human speaker gaze is used as a visual cue, which is a replication of findings by [10, 9] and [11]. Although virtual agent gaze did not have any effect upon participants' gaze behaviour, it even had a negative impact on the performance of the verification task. Thus, the agent's presence must have at least been perceived peripherally to affect response times.

4. Acknowledgements

This research was funded by grants from the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University.

5. References

[1] B. J. Lance and S. . Marsella, "The expressive gaze model: Using gaze to express emotion," in *IEEE Computer Graphics and Applications*, vol. 4, no. 30, 2010, pp. 62–73.

[2] M. Argyle and M. Cook, "Gaze and mutual gaze." 1976.

[3] B. Fischer and B. Breitmeyer, "Mechanisms of visual attention revealed by saccadic eye movements," *Neuropsychologia*, vol. 25, no. 1, pp. 73–83, 1987.

[4] W. Steptoe, O. Oyekoya, A. Murgia, R. Wolff, J. Rae, E. Guimaraes, D. Roberts, and A. Steed, "Eye tracking for avatar eye gaze control during object-focused multiparty interaction in immersive collaborative virtual environments," in *Virtual reality conference, 2009*, 2009.

[5] T. Farroni, G. Csibra, F. Simion, and M. H. Johnson, "Eye contact detection in humans from birth," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 14, 2002, pp. 9602–9605.

[6] M. von Gruenau and C. Anston, "The detection of gaze direction: A stare-in-the-crowd effect," *Perception*, vol. 24, pp. 1297–1313, 1995.

[7] S. E. Brennan, X. Chen, C. A. Dickinson, M. B. Neider, and G. J. Zelinsky, "Coordinating cognition: The costs and benefits of shared gaze during collaborative search," *Cognition*, vol. 106, pp. 1465–1477, 2008.

[8] M. Staudte, M. W. Crocker, A. Heloir, and M. Kipp, "The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts," *Cognition*, vol. 133, pp. 317–328, 2014.

[9] J. E. Hanna and S. E. Brennan, "Speakers eye gaze disambiguates referring expressions early during face-to-face conversation," *Journal of Memory and Language*, vol. 57, pp. 596–615, 2007.

[10] H. Kreysa and P. Knoeferle, "Effects of speaker gaze on spoken language comprehension: Task matters," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, L. Carlson, C. Hoelscher, and T. Shipley, Eds. Cognitive Science Society, 2011.

[11] M. Staudte and M. W. Crocker, "Investigating joint attention mechanisms through spoken human-robot interaction," *Cognition*, vol. 120, pp. 268–291, 2011.

[12] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey, "I reach faster when i see you look: gaze effects in human-human and human-robot face-to-face cooperation," *Frontiers in Neurorobotics*, vol. 6, pp. 1–11, May 2012.

[13] S. Andrist, T. Pejsa, B. Mutlu, and M. Gleicher, "Designing effective gaze mechanisms for virtual agents," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2012, pp. 705–714.

[14] S. Raidt, F. Elisei, and G. Bailly, "Face-to-face interaction with a conversational agent: Eye-gaze and deixis," in *International Conference on Autonomous Agents and Multiagent Systems*, 2005.

[15] S. Andrist, B. Mutlu, and M. Gleicher, "Conversational gaze aversion for virtual agents," *Intelligent Virtual Agents*, pp. 249–262, 2013.

[16] N. Wang and J. Gratch, "Don't just stare at me!" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1241–1250.

[17] N. Bee, J. Wagner, E. Andre, T. Vogt, F. Charles, D. Pizzi, and M. Cavazza, "Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010, p. 9.

[18] W. L. Johnson and J. C. Lester, "Face-to-face interaction with pedagogical agents, twenty years later," *International Journal of Artificial Intelligence in Education*, pp. 1–12, 2015.

[19] S. Martinez, R. S. Sloan, A. Szymkowiak, and K. Scott-Brown, "Using virtual agents to cue observer attention," in *CONTENT 2010 : The Second International Conference on Creative Content Technologies*, 2010.

[20] M. Courgeon, G. Rautureau, J.-C. Martin, and O. Grynszpan, "Joint attention simulation using eye-tracking and virtual humans," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 2013.

[21] K. Ruhland, C. Peters, S. Andrist, J. B. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception," *Computer Graphics forum*, vol. 00, no. 0, pp. 1–28, 2015.

[22] K. Bergmann, F. Eyssel, and S. Kopp, "A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time," in *International Conference on Intelligent Virtual Agents*. Springer, 2012, pp. 126–138.

[23] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, vol. 2006, 2006, p. 5th.

[24] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *International Workshop on Intelligent Virtual Agents*. Springer, 2006, pp. 205–217.

[25] H. Van Welbergen, R. Yaghoubzadeh, and S. Kopp, "Asaprealizer 2.0: the next steps in fluent behavior realization for ecas," in *International Conference on Intelligent Virtual Agents*. Springer, 2014, pp. 449–462.

[26] M. Rehm and E. Andre, "Where do they look? gaze behaviour of multiple users interacting with an embodied conversational agent," *Intelligent Virtual Agents*, pp. 241–252, 2005.

[27] P. Knoeferle and H. Kreysa, "Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension?" *Frontiers in Psychology*, vol. 3, pp. 1–15, 2012.