In Dialogue with an Avatar, Syntax Production is Identical Compared to Dialogue with a Human Partner

Evelien Heyselaar (evelien.heyselaar@mpi.nl)
Peter Hagoort (peter.hagoort@mpi.nl)
Katrien Segaert (katrien.segaert@mpi.nl)

Max Planck Institute for Psycholinguistics, Wundtlaan 1
Nijmegen 6525XD, The Netherlands

Abstract

The use of virtual reality (VR) as a methodological tool is becoming increasingly popular in behavioural research due to its seemingly limitless possibilities. This new method has not been used frequently in the field of psycholinguistics, however, possibly due to the assumption that human-computer interaction does not accurately reflect human-human interaction. In the current study we compare participants' language behaviour in a syntactic priming task with human versus avatar partners. Our study shows comparable priming effects between human and avatar partners (Human: 12.3%; Avatar: 12.6% for passive sentences) suggesting that VR is a valid platform for conducting language research and studying dialogue interactions.

Keywords: virtual reality; human-computer interaction; syntactic priming.

Introduction

The use of virtual reality (VR) as a method is becoming increasingly prevalent in behavioural studies in a wide range of fields, including navigation research (Tarr & Warren, 2002) and rehabilitation therapy (Rizzo et al., 2004). However this new trend does not seem to be catching on in the field of psycholinguistics. This may be due to the assumption that humans do not interact with computers in the same way that they interact with other humans, making any behavioural measure of language interaction with a computer-partner ("avatar") ecologically equivocal.

However, research into human-computer interactions has suggested the opposite, at least with regards to desktop modules (Stoyanchev & Stent, 2009a; 2009b). Work by Nass and Moon (2000) has repeatedly shown that humans attribute human-like characteristics to their desktop computer partner, the most unintuitive of these findings being the use of politeness when asked to evaluate the computer and the implementation of social hierarchy. This behaviour was observed even in participants who, during the debrief, agreed that "the computer is not a person and does not warrant human treatment or attribution."

VR is one step up from desktop modules as it offers an immersive 3D world that participants can move in and interact with. However, as it is still a program, VR allows experimental control over parameters that cannot be (as finely) controlled in the real world. What is important for psycholinguists is that VR offers the ability to finely control avatar behavior in parameters that are nearly impossible to control in a confederate, an aspect that is particularly attractive for dialogue research.

There is a rapidly growing interest in interactional aspects of language. Language is increasingly studied in a dialogue context, focusing on, for example, the turn-taking event (Stivers et al., 2009), the role of the dialogue partner (Branigan et al., 2003) and characteristics of the social interaction (Balcetis & Dale, 2012). With this, there is an increasing demand in the field to develop a methodology where these factors can be stringently controlled.

We put VR as a methodology to study language behaviour to the test. In Experiment 1, we established which characteristics make an avatar more human in a rating study. In Experiment 2, we investigated language behaviour in interaction with the most human avatar. We focused on syntactic processing (specifying the syntactic relations between words in the sentence), a core aspect of language production and comprehension. To compare behaviour in a human-human interaction and human-avatar interaction, we used a syntactic priming task.

Priming refers to the phenomenon in which an individual adapts the behavioural characteristics of their conversational partner. This can range from adapting the speech rate of your partner (Giles et al., 1992; Giles & Powesland, 1975) to more complex adaptations such as using the same sentence structure as your partner (Bock, 1986). It has been proposed that speakers align syntactic choices and other linguistic behaviour to increase affiliation with a conversation partner (Giles, Coupland, & Coupland, 1991) and to increase conversation success (Pickering & Garrod, 2004). It is an open question whether you do this to the same extent with an avatar-partner as a human-partner.

Experiment 1

Experiment 1 was conducted to establish which characteristics determine the humanness of an avatar. We asked participants to rate six avatars with different facial combinations on the amount of humanness, familiarity, quality of facial expression, and quality of voice in order to isolate a combination of facial features that causes the avatar to appear as human as possible.

Method

Participants

30 native Dutch speakers (13 male/17 female, M_{age}: 22.5; SD_{age}: 3.1) gave written informed consent prior to the experiment and were monetarily compensated for their participation.

Materials

Avatars The avatar was adapted from a stock avatar produced by WorldViz ("casual15_f_highpoly"). The avatar's appearance suggested that she was a Caucasian female in her mid-twenties, which matched the age of the Dutch speaker who recorded her speech.

The six facial expressions to be tested involved combinations in blink rate, smiling and eyebrow habits (Table 1). Blinks happened once every 1 - 5 seconds. For versions with normal smiling and normal eyebrow habits we explicitly programmed when the avatar would smile and/or raise her eyebrows, such that it would coincide with the content of her speech. For example, the avatar would raise her eyebrows when asking a question, and smile when she was enthusiastic ("*Come, let's play another round!*"). When not speaking, she would smile once every 5 - 10 seconds, and raise her eyebrows once every 1 - 5 seconds such that she would still differ from the no smile/no eyebrow version.

Virtual Environment The virtual environment (VE) was a stock environment produced by WorldViz ("room.wrl") adapted to include a table on which stood a wooden divider. The divider height was such that participants could view the top of the divider and the face of the avatar simultaneously. The table in the VE matched in both dimension and position with a table in the physical world, such that participants could actually touch the "virtual" table.

The experiment was programmed and run using WorldViz's Vizard software. Participants wore an NVIS nVisor SX60 head-mounted display (HMD), which presented the VE at 1280x1024 resolution with a 60 degree monocular field of view. Mounted on the HMD was a set of 8 reflective markers linked to a passive infrared DTrack 2 motion tracking system from ART Tracking, the data from which was used to update the participant's viewpoint as she moved her head. Additionally, a single reflective marker was taped onto the index finger of the participant's dominant hand. This marker was rendered as a white ball in the VE, such that participants knew the position of their finger at all times. Sounds in the VE, including the voice of the avatar, were rendered with a 24-channel WorldViz Ambisonic Auralizer System.

Table 1. Avatar Facial Expressions.

Avatar	Blink Rate	Smiling Habit	Eyebrow Habit
1	No blink	No smile	No movement
2	Slow blink[1]	Random[2]	No movement
3	Slow blink	Continuous	Constantly up
4	Normal	No smile	Random[2]
5	Normal	Normal	Random[2]
6	Normal	Normal	Normal

[1] Duration of a slow blink was 0.5 seconds. Duration of a normal blink was 0.1 seconds.

[2] Random habits occurred once every 3 - 5 seconds.

Procedure and Task

The participants were informed that they would be rating six different avatars. Exposure to each avatar started with the avatar giving a short introduction speech, followed by a short matching card game.

The card game is identical to the one used in Experiment 2 (for more detail see Methods of Experiment 2). But briefly: the participant and the avatar would alternate in describing picture cards to each other. The participant would be presented with six cards from which to freely choose one to describe. After the participants turn, the avatar would select the described card, causing it to be replaced by a novel card. After the avatar's turn, the participant would need to select the card described in order for it to be replaced. The cards consisted of single or paired actors depicting an action, the verb of which would be written underneath the picture.

Between avatar versions, participants removed their headset in order to fill out a pen-and-paper questionnaire. Previous research has shown that if the subject evaluates the avatar in the presence of said avatar, they rate them more favourably (Nass & Moon, 2000).

Results

A significant effect was found of avatar versions on the rating of humanness ($F = 4.970$, $p < .001$), familiarity ($F = 3.065$, $p = .01$) and quality of facial expression ($F = 5.097$, $p < .001$). The voice ratings were not found to be significantly different between avatar versions ($F = 1.418$, $p = .220$), which works as a sanity check as the voice was exactly the same in each version.

A *post hoc* Tukey's HSD test showed that avatars with eyebrow movement (3-6) were rated significantly more human than avatars without eyebrow movement ($p < .05$), whereas smiling habits made no significant difference, a result that is consistent with previous literature (Looser & Wheatley, 2010). Additionally, the avatars with a normal blink rate (4-6) were rated as having a significantly higher quality of facial expression ($p < .05$) but had no impact on humanness rating.

As we were aiming to use the most human avatar in Experiment 2, we drew linear correlations between familiarity and humanness (Figure 1A) and quality of facial expressions and humanness (Figure 1B). For both, two-tailed Pearson's correlations were positive (familiarity: $R^2 = 0.64$, $p < .001$;

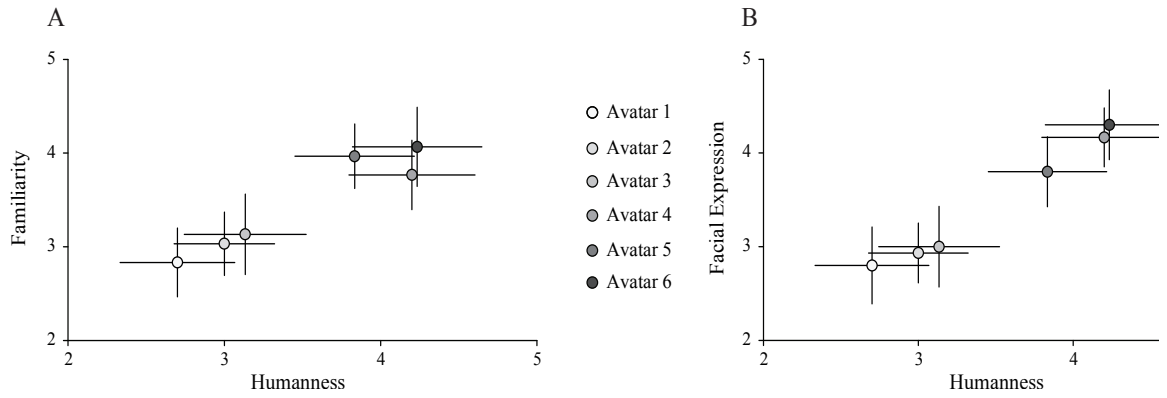

Figure 1. Rating of avatar versions. A. Correlation between the familiarity and humanness ratings, and **B**. correlation between the quality of facial expression and humanness ratings for the six avatars. Avatars with eyebrow movement (3-6) were rated as significantly more human ($p < .001$).

facial expression: $R^2 = 0.58$, $p < .001$) and showed Avatar 6 as being rated the highest in all three conditions. Therefore, Avatar 6 will be used in Experiment 2.

Experiment 2

In this experiment we investigate the language behaviour in interactions with an avatar and human partner. We used syntactic priming as a behavioural measure with which to compare human and avatar conditions.

Methods

Participants

33 native Dutch speakers gave written informed consent prior to the experiment and were monetarily compensated for their participation.

Five subjects were not convinced that the confederate was an ignorant participant or did not believe that the avatar was voice-recognition controlled and were *a priori* not considered part of the data-set. Thus only 28 were included in the analysis (14 male/14 female, M_{age}: 20.6; SD_{age}: 2.4).

Task and Design

Participants conducted a longer version (240 cards) of the task described in Experiment 1. All participants completed the task in VE with an avatar as well as in the physical world with a confederate (order was randomized and counterbalanced across participants).

In each block, the participant was presented with six cards, with the belief that the confederate/avatar had their own spread of six cards behind the divider (Figure 2). The participant and the confederate/avatar would alternate in describing cards to each other. Participants were instructed to describe the picture using one concise sentence (e.g., *The man kisses the woman*). If either member had a card that was described, both members would remove that card from the spread and replace it with a novel card from their deck (in VE this would happen automatically after the subject selected the card). The confederate/avatar description would serve as the prime for the participants' subsequent target description.

The confederate's deck was ordered identically to the participant's deck, so the confederate/participant always had the card described. In the VE block, the avatar was programmed to randomly pick one of the participant's cards to describe.

The confederate's deck of cards showed the stimulus picture but with a full sentence typed underneath, as such the confederate simply needed to read the sentence. 50% of the transitive sentences described the picture in the passive tense, 50% described it in the active tense. In VE, the avatar was programmed to use 50% passives, 50% actives.

Three conditions were included in the analysis: baseline trials (intransitive prime followed by a transitive target), active priming (active prime followed by a transitive target), and passive priming (passive prime followed by a transitive target). To ensure an adequate number of trials in each condition, 2/3 of the cards were transitive and 1/3 were intransitive. *Post-hoc* analysis showed that there was an average of 26.2 (SD: 8.5), 27.8 (SD: 3.9), and 25.9 (SD:

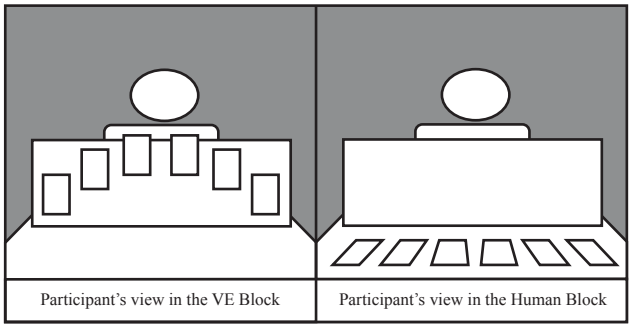

Figure 2. Setup for Experiment 2. Shows the experimental set-up from the view of the participant. The only difference is that in VE the cards were presented at the top of the divider, whereas in the Human block, the cards were laid out on the table.

3.6) trials in the baseline, passive and active conditions respectively in the Human block and 21.0 (SD: 3.9), 25.1 (SD: 3.6), and 24.9 (SD: 4.0) trials in the baseline, passive and active conditions respectively in the VE block.

One subject was discarded as the difference in the proportion of passive prime exposure between the two blocks (Human: 0.40; VE: 0.65) fell two-and-a-half standard deviations outside the mean difference between blocks (Mean: 0.01; SD: 0.09).

Materials

Stimulus Pictures The photos used in this task have been described elsewhere (Segaert et al., 2011) but briefly: our stimulus pictures depicted 40 transitive events such as *kissing, helping* or *strangling* with the agent and patient of this action. Each event was depicted by either one pair of adults or one pair of children. There was one male and one female actor in each picture and each event was depicted with each of the two actors serving as the agent. The position of the agent (left or right) was randomized. These pictures were used to elicit transitive sentences.

Filler pictures were used to elicit intransitive sentences. These fillers depicted events such as running, singing, bowing with one actor. The actor could be any of the actors used in the transitive stimulus pictures.

The verb depicted in each picture was written underneath.

Procedure

Participants were informed that our goal was to compare how experiencing events differed in VE. To ensure that the participants felt that they were communicating with a program and not a programmer, they were told that it worked on voice-recognition, and hence no third party was necessary to operate the program.

Responses were manually coded as active or passive. Target responses were included in the analysis only if 1) both actors and the verb were used correctly and 2) no unnecessary information was included in the description.

Results

We excluded 1.09% (71 out of 6485) of the target responses because they were incorrect (criteria described under *Procedure*).

The responses were analyzed using a mixed-effects logit model in R (R Development Core Team, 2009), the results of which are shown in Table 2. Target responses were coded as 0 for actives and 1 for passives. We used a maximal random-effects structure (Barr er al., 2013): the repeated-measures nature of the data was modeled by including a per-participant and per-item random adjustment to the fixed intercept ("random intercept"). We attempted to include as many per-participant and per-item random adjustments to the fixed effects ("random slopes") until the model failed to converge. The full model included random slopes for *Prime*, *Partner Type*, and *Order* for the per-participant random intercept only. Per-item could not take any random slopes. We used dummy coding with baseline condition as reference level for the effect for prime, and deviation coding for the other factors. Multi-collinearity measures came back as non-significant (VIF <2.5).

Figure 3 summarizes the relative proportion of passive target responses after each prime structure. The fixed effects of the best model fit for these data are summarized in Table 2. The negative estimate for the intercept indicates that in the baseline condition active responses were more frequent than passive responses. Following passive primes, more passive responses were produced compared to baseline ($p < .001$). Following active primes, there was no increase in active responses compared to baseline ($p = .152$). A model with *Partner Type* as an interaction with *Prime Structure* was not significantly better than the current model ($p = 0.32$). In this model, neither active nor passive priming interacted with partner type ($\beta = 0.28$, $p = .37$; $\beta = 0.55$, $p = .13$ respectively) suggesting that the priming effect is the same in the Human and VE block.

As the verbs depicted in the intransitive pictures were not the same as in the transitive pictures, a separate model was created based on data that excluded the baseline condition in order to analyze the effect of verb repetition on target choice. Figure 4 shows the proportion of passive responses with verb repetition and without. Interestingly, there are no trials in which a passive response was produced following an active prime with verb repetition. Therefore, in order to make the model converge, active primes were also removed from the dataset. Table 3 summarizes the fixed effects of the best model fit, indicating an influence of *Verb Repetition* on target structure production ($p < .001$). A model with *Partner Type* as an interaction with *Verb Repetition* was not significantly better than the current model ($p = .99$), suggesting that passive priming is boosted by verb repetition but the influence of verb repetition on target production is the same in the Human and VE block.

Correlation analysis showed a positive correlation ($R^2 = 0.37$) between the priming effects for passives in the Human and VE block. The correlation is significant with both Pearson's r (two-tailed, $p < .001$) and Spearman's rho (to control for the possible influence of outliers; two-tailed, $p = .039$), suggesting that participants primed comparably in each condition (Figure 5).

Discussion

The first experiment showed that the amount of facial expression in the upper face (namely eyebrow movement) has a significant impact on the humanness of a virtual being, whereas other facial features such as smiling habits and blink rate increase the overall realism of the facial expression but has no impact on the perceived humanness of the avatar. Previous studies in which participants rated faces on a desktop computer had suggested this relationship (Looser &

Table 2. Summary of fixed effects in the mixed logit model for the response choices based on prime structure.

Predictor	coefficient	*SE*	*Wald Z*	*p*	
Intercept (baseline)	-3.38	0.31	-10.87	< .001	***
Passive Prime	1.55	0.21	7.41	< .001	***
Active Prime	-0.27	0.19	-1.43	.152	
Partner Type	-0.17	0.20	-0.90	.366	

Note: N = 4019, log-likelihood = -1148.28

Wheatley, 2010) but it had yet to be verified in an immersive environment such as VR.

The results of Experiment 2 show comparable syntactic priming effects when participants interacted with a human partner compared to an avatar partner.

Three findings provide converging evidence that language behaviour was similar when interacting with an avatar and human: i). Syntactic priming effects were found in the VE as well as the Human block and the size of these effects did not differ; ii). In line with the literature, syntactic priming effects showed an inverse preference effect (syntactic priming effects for passives, not for actives (Bock, 1986; Ferreira, 2003)) and a lexical boost (larger syntactic priming effects when the verb between prime and target was repeated (Branigan, Pickering, & Cleland, 2000)) and these again did not differ between the VE and Human block; and iii). Participants' priming effect when interacting with a human was correlated with participants' priming effects when interacting with the avatar.

Our results therefore suggest that humans interacting with an avatar elicit the same language behaviour as if they were interacting with a human partner. We are attributing this finding to the humanness of the avatar. One limitation is that we have not manipulated the humanness of the avatar in Experiment 2. Therefore, we are currently measuring the influences of a low-humanness avatar on language behaviour in order to further support our claim.

Although this study only provides evidence for syntactic production, it suggests the possibility that other behaviours may also be consistent between VE and the real world. This provides a strong argument in favor of the use of VR to investigate interaction behaviour in the field of psycholinguistics.

Table 3. Summary of fixed effects in the mixed logit model for the response choices based on verb repetition.

Predictor	coefficient	*SE*	*Wald Z*	*p*	
Intercept (baseline)	-0.71	0.26	-2.71	.0068	**
Verb Repetition	1.45	0.13	10.83	< .001	***
Partner Type	-0.003	0.16	-0.02	.983	

Note: N = 1400, log-likelihood = -565.72

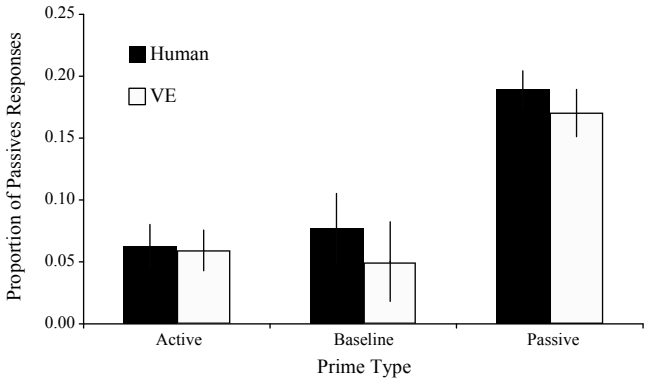

Figure 3. Proportion of passive responses per prime type. There is no significant difference in syntactic priming effects in the Human and VE block. Passive production increases with 12.3% for the Human block and 12.6% for the VE block following a passive prime compared to the baseline condition.

The use of a confederate is a key requirement when studying dialogue, yet it is also a limitation for a variety of reasons. Firstly, it is impossible for the confederate to behave exactly the same (for example, maintaining the same tone and speech rate), thereby causing between-subject variability. Additionally, the use of a human confederate also limits the type of scenarios one can create.

Both of these challenges can be overcome by replacing the confederate with a recording played on a desktop computer. This ensures that speech characteristics are kept consistent across sessions and also allows more finely controlled voice manipulations (such as elongating vowels or decreasing pitch range). Many studies have replicated priming behaviour in participants interacting with a desktop-computer module (Branigan et al., 2010; Branigan et al., 2003; Weatherholtz,

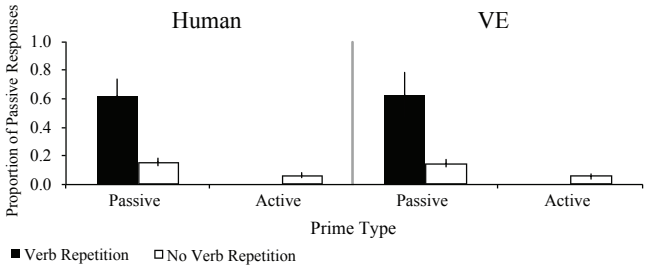

Figure 4. Proportion of passive responses with verb repetition and without verb repetition for each block. The influence of verb repetition on passive target response is the same for Humand and VE block. In both cases, there were no passive responses with verb repetition following an active prime.

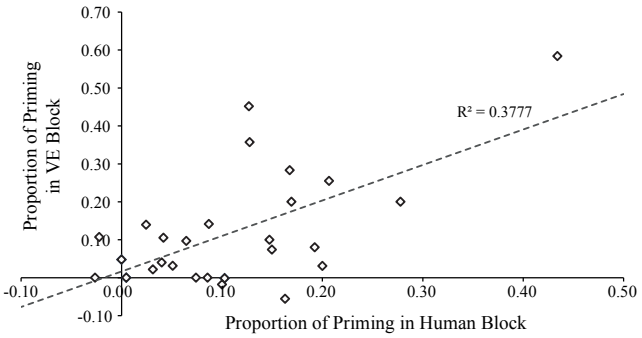

Figure 5. Correlation of passive priming effect between blocks. There is a significant positive correlation between the passive priming effects seen in the Human and VE block.

Campbell-Kibler, & Jaeger, 2012), and therefore desktops have provided a temporary solution.

However, advancements in interaction research are severely hampered due to limitations in the scenarios we can create in the real world or using desktop computers. A key example is the ability to realistically edit the external features of the confederate, such as manipulating facial expressions or even more subtle changes such as varying pupil diameter. Manipulations such as these allow investigations into the social influences on dialogue behaviour, for example does the attractiveness or perceived similarity (Balcetis & Dale, 2012) of your partner affect your word choice. These questions cannot easily be answered using other techniques, and therefore VR provides an important platform on which previously unanswerable questions can now be investigated.

Acknowledgments

We would like to thank our confederate Nadine de Rue for playing the same game 33 times without complaint.

References

Balcetis, E. E., & Dale, R. (2012). An Exploration of Social Modulation of Syntactic Priming. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society.*

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255-278.

Bock, J. K. (1986). Syntactic Persistence in Language Production. *Cognitive Psychology*, 18, 355–387.

Branigan, H. P., Pickering, M. J., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2), B13–25.

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9), 2355–2368.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Nass, C. I. (2003). Syntactic Alignment Between Computers and People : The Role of Belief about Mental States. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*, 186–191.

Casasanto, L. S., Jasmin, K., & Casasanto, D. (1996). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 127–132.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, 48(2), 379–398.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Giles, H., Henwood, K., Coupland, N., Harriman, J., & Coupland, J. (1992). Language attitudes and cognitive mediation. *Human Communication Research*, 18(4), 500–527.

Giles, H., & Powesland, P. F. (1975). *Speech styles and social evaluation.* Academic Press.

Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy. How, when, and where we perceive life in a face. *Psychological Science*, 21(12), 1854–62.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27, 169–225.

Rizzo, A., Schultheis, M., Kerns, K. A., & Mateer, C. (2004). Analysis of assets for virtual reality applications in neuropsychology. *Neuropsychological Rehabilitation*, 14(1-2), 207–239.

Segaert, K., Menenti, L., Weber, K., & Hagoort, P. (2011). A paradox of syntactic priming: why response tendencies show priming for passives, and response latencies show priming for actives. *PloS one*, 6(10).

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., et al. (2009). *Universals and cultural variation in turn-taking in conversation*, 106(26), 10587–10592.

Stoyanchev, S., & Stent, A. (2009a). Lexical and syntactic priming and their impact in deployed spoken dialog systems. *Proceedings of NAACL HLT 2009: Short Papers*, 189-912.

Stoyanchev, S., & Stent, A. (2009b). Concept form adaptation in human-computer dialog. *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, 144-147.

Tarr, M. J., & Warren, W. H. (2002). Virtual reality in behavioral neuroscience and beyond. *Nature Neuroscience Supplement*, 5, 1089–1092.

Weatherholtz, K., Campbell-Kibler, K., & Jaeger, T. F. (2012). Syntactic alignment is mediated by social perception and conflict management. *Architectures and mechanisms for language processing (AMLaP 2012)*.