Bennett CC, et al. (2022) Effects of Cross-Cultural Language Differences on Social Cognition During Human-Agent Interaction in Cooperative Game Environments. *Computer Speech and Language*. 81, pg. 101521. https://doi.org/10.1016/j.csl.2023.101521

Effects of Cross-Cultural Language Differences on Social Cognition During Human-Agent Interaction in Cooperative Game Environments

Casey C. Bennett[1,2,*], Young-Ho Bae[1], Jun Hyung Yoon[1], Yejin Chae[1], Eunseo Yoon[1], Seeun Lee[1], Uijae Ryu[1], Say Young Kim[3], Benjamin Weiss[4]

[1]Department of Intelligence Computing
Hanyang University
Seoul, Korea

[2]Department of Computing & Digital Media
DePaul University
Chicago, IL, USA

[3]Department of English Language & Literature
Hanyang University
Seoul, Korea

[4]Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany

Abstract

A major challenge in human-robot interaction (HRI) is creating the "social fluidity" necessary for humans to perceive the interaction as life-like. During verbal interactions, for instance, the speech content itself is not the only thing that matters. Rather, things like timing, cadence, and manner of speaking are necessary to speak "like a native", yet those attributes vary significantly by language and cultural setting. To that end, we developed a bilingual virtual avatar (Korean and English speaking) capable of autonomous speech during cooperative gameplay with a human participant in a social survival video game. We then ran a series of experiments with 60 participants (30 English speakers and 30 Korean speakers) interacting with the avatar during 30-minute game sessions. The experiments included several conditions, in which we modified the avatar's speech behavior in different ways while collecting multiple types of data (audio-visual recordings, speech data, gameplay data, human perceptions). Results showed significant differences between English and Korean speakers during the experiment. Korean speakers spoke less on average and had more negative speech sentiment, while the English speakers spoke more frequently and had more positive speech sentiment. The avatar was also more likely to interrupt the human's speech in English than Korean, despite having the same design. Furthermore, Korean speakers perceived more social presence when the avatar engaged in more repetitive speech behavior, while English speakers perceived more when the avatar was more "chatty". We suggest that these results likely relate to cultural differences between East Asian cultures and Western cultures in terms of the social norms that govern appropriate social interaction behavior, and discuss the implications for future work on interactive speech agents.

Keywords — human-robot interaction; social cognition; speech system; virtual avatar; language differences; cross-cultural robotics

** Corresponding Author*

1. Introduction

1.1 Background

One of the major goals in the field of human-robot interaction (HRI) is geared toward creating artificial intelligence (AI) that can emulate natural human social behavior to produce better interactive technology. Part of that challenge is creating the "social fluidity" in artificial agents necessary for humans to perceive an interaction as life-like. Such fluidity can be defined in the sense of *social presence*, which is the feeling of being there with a real person (Oh et al., 2018). More broadly, this relates to Dennett's work on triggering an intentional stance in humans toward artificial agents, and the attribution of *agency* to such agents (averse to machines) (Dennett, 1971). However, understanding what might constitute social presence is dependent on the cultural and linguistic setting that the interaction takes place (Lowry et al., 2010). Indeed, many of us have had this experience when speaking a second language which is not our mother tongue. Simply producing the verbal content is not always sufficient ... there is a timing, cadence, and manner of speaking that is necessary to speak "like a native" (Enfield, 2017; Koutsombogera & Vogel, 2019). Past research has even shown there are cognitive differences in bilingual speakers depending on the language they are actively speaking (Kousta et al., 2008).

There are a variety of ways that cultural and linguistic differences might impact HRI (see Section 1.2), but a good starting point is to characterize the effects of those differences on HRI by empirically examining them in detail within particular interaction modalities, e.g. speech behavior. To do so, however, it is necessary to understand how social interaction might work with the *exact* same agent but in different languages or cultures. Moreover, such differences are often context-specific, e.g. idiomatic language (Gwóźdź, 2019; Chahboun et al., 2021), implying that contrived lab experiments often lack the necessary context to truly elicit the subtle differences that occur *in vivo*. Rather, we need goal-oriented environments that require free-form yet context-specific social interaction (Sabanovic et al., 2006; Jung & Hinds, 2018). Cooperative games are one such possible environment, which have been used in various HRI research studies in the past few years (Correia et al., 2016; Jesso et al., 2020).

In this paper, we sought to explore these questions within such a cooperative game paradigm, where an autonomous virtual avatar agent played a survival video game with human participants while socially interacting with them in different languages (Bennett et al., 2022b). To further test the linguistic effects, we also explored how human perceptions of social presence of the agent were affected by alterations to the avatar's speech behavior across languages.

1.2 HRI Speech Systems

There has been much prior research on HRI speech systems over the past 20 years (Thomasz et al., 2016), as well as a more recent focus on the effects of culture (Lim et al., 2021). Such cultural differences are embedded in the languages we speak, otherwise known as *linguistic relativity* (Deutscher, 2010). Human speech behaviors, including things such as turn-taking cues, interruption speech, backchanneling, and sentiment conveyance, manifest in different ways in different languages, and are notably different in languages and cultures considered more "distant", such as Western and

East Asian cultures. For instance, silence (i.e. social pauses) can often be seen as a strong indicator of communicative breakdown in Western cultures, yet it can represent a very appropriate (even necessary) contribution to dialogue in high-context East Asian cultures (Bruneau, 1973).

In the domain of HRI, there has been research on the effects of turn-taking (Skantze, 2021), interruptions (Fischer et al., 2021), affective communication (Bennett & Sabanovic, 2014), politeness levels (Seok et al., 2022), and communication failures (Honig & Oron-Gilad, 2018), among other aspects, during speech interactions between robots and humans. There has also been research exploring how those aspects affect conversational agents in general regardless of the form factor of the agent, such as personal assistant devices (Doyle et al., 2021) or characters in virtual reality environments (Slater, 2009). Primarily, researchers have focused on human perceptions of the robot or agent as it relates to speech behaviors, whether from the standpoint of likeability and animacy (Bartneck et al., 2009) or more complex notions of social presence (see Section 1.1) (Oh et al., 2018). The latter relates to the *immersion* of an interactive experience, which can be thought of as the dividing line between illusion and reality (Gonzalez-Franco & Lanier, 2017).

While there has been some prior work on second-language learners in HRI (Engwall et al., 2021), there is still limited research on direct comparisons of linguistic differences with fluent speakers using the same robot or virtual avatar platform in different languages, particularly the effects of those differences on social interaction during HRI. However, understanding those effects is critical for addressing some of the questions raised above in Section 1.1. Beyond verbal interaction itself, there is also of course an interplay of speech with non-verbal aspects of communication, which is a topic we return to later in this paper. Suffice it to say, there are a number of open questions and research avenues toward better understanding how HRI speech systems are influenced by language and culture.

1.3 Research Aims

The focus of this study is on whether there are differences across languages in how people interact with a robot or virtual avatar. To test this, we created a bilingual virtual avatar (Korean and English speaking) capable of context-specific autonomous speech during cooperative gameplay in a social survival video game (Bennett et al., 2022b). As mentioned in Section 1.1, cooperative game paradigms are a good research environment for this, since they demand goal-oriented behavior where players must socially interact and cooperate to survive. We then recruited both Korean speakers and English speakers to participate in a series of experiments.

To test the effects of language, we created several conditions where the speech behavior of the avatar was altered in different ways, then compared how that affected human speech behavior between the languages, as well as participant perceptions of social presence of the avatar agent. Our focus here is on the speech system itself, and alterations thereof, rather than non-verbal communication. Although, there are numerous potential research avenues, including the interplay with non-verbal aspects, that could be further explored in future work (see Discussion section). In brief, the goals can be summarized as attempting to understand:

1) Differences across languages in how people verbally interact with a robot or virtual avatar

2) How alterations to artificial speech systems have different effects across languages

3) Whether the above differences relate to human perceptions of a robot or virtual avatar

2. Methods

2.1 Virtual Avatar & Speech System Development

To study the questions described in the Introduction section, we developed a virtual avatar capable of autonomous speech during a cooperative survival game (described in Section 2.2). The virtual avatar and speech system (henceforth referred to as the "Social AI") was the subject of extensive development and testing that included both recording naturalistic human vs human gameplay as well as evaluation of the different developed speech components onboard the avatar, which has been described in detail previously elsewhere (Bennett et al. 2022; Suh et al., 2021, Bennett & Weiss, 2022). The Social AI was capable of hundreds of different speech utterances covering 46 different utterance categories, each related to a particular game situation (e.g. collecting resources, fighting monsters, deciding where to go next) organized as a hierarchy with several levels. Those speech utterances were both self-generated based on internal logic of the Social AI, as well as responses to human player speech via automatic speech recognition (ASR). The speech responses were similar in both English and Korean (i.e. the virtual avatar was bilingual, in essence).

The system was developed through first recording and annotating human vs. human gameplay in the same game environment, in both English and Korean (using native speakers), to produce parallel speech corpora. Subsequently, initial versions of the Social AI were then tested during avatar vs. human gameplay pre-tests to identify missing capabilities, which were then augmented. The speech system was implemented in custom code written in Python, using locally-installed (Windows or Mac) voice packages as part of the Text-to-Speech (TTS) module, with the audio output redirected to an internal "virtual" microphone jack. The ASR component used the Microsoft Azure speech-to-text API for human speech recognition in both English and Korean. The speech output (via the internal microphone jack) was then directed to the Loomie application (https://www.loomielive.com/), where we created a visual avatar capable of moving its lips synchronously with the speech (see gameplay example figure in Section 2.3). The Loomie avatar was also capable of some basic built-in gestures, but we did not attempt to modify those for the current experiments. In total, approximately one year was spent developing the system prior to starting the experiments described below.

2.2 Cooperative Game Environment

In the current study, we utilized a video game called *Don't Starve Together* for our cooperative game environment (https://www.klei.com/games/dont-starve-together), which can be downloaded from online sources such as Steam. The *Don't Starve Together* game is a social survival game where players need to collect resources, make tools, fight monsters, and cooperate with each other to survive longer. Similar to other social survival games (e.g. Minecraft), *Don't Starve Together* requires players to collect specific combinations of resources in order to build things, without which they will be vulnerable to various dangers and likely lose the game via player death, though there are multiple

strategies that can be pursued (i.e. free-form). Moreover, it has cooperative multi-player gameplay modes (used here), which allow the players to cooperate on such tasks to survive. The tasks are under time constraints, however, as the level of danger gradually increases over time. As such, the game represents a free-form yet goal-oriented cooperative gameplay environment.

Along with providing an ideal gameplay environment, the *Don't Starve Together* game is heavily customizable through the use of game modification tools, which allow users to alter the mechanics of the in-game environment and non-player-character (NPC) behaviors through the LUA programming language. For our experiments, a custom "Game Mod" was developed in LUA with two aims in mind. First, we wanted to create "game data writing" functionality, so that we could collect real-time data about the game state at every moment. That included information about player status, inventory, movement, items equipped, attacking/fighting, time of day, and entities in the players' immediate environment (e.g., monsters, structures). Second, we wanted to create customizable scenarios within the game to be able to control the types of interactions between the avatar and human. That included a fixed starting position with various resources immediately available ("advanced start"), providing a constant source of light at that starting position in order to encourage players to return to it periodically to encourage more social interaction ("base camp"), and setting the minimum health level at 10% so that players could not die guaranteeing every experiment game session could last approximately the same amount of time, e.g. 30 minutes ("partial invincibility"). During the experiments, human participants were not informed of those game modifications, however.

The Game Mod ran in parallel to the avatar's speech system (i.e. Social AI), so that the written game data could be used in real-time to make the avatar aware of in-game events which then affected the kinds of speech utterances it produced. In other words, the speech utterances were context-specific. Additionally, the written game data was used later for analysis to try to understand differences in speech patterns across conditions and languages (see Section 2.5).

2.3 Experiments

For the experiments here, we recruited 60 participants, 30 Korean speakers and 30 English speakers, recruited via university message boards. All participants were either native speakers or had advanced proficiency in either English (TOEIC Level B2) or Korean (TOPIK Level C1, aka "level 5"). The genders were balanced, with 28 males and 32 females, with an average age of roughly 23.5 years. The sample was 47% female on the Korean side, and 56% female on the English side. The Korean speakers were all native L1 speakers living in Korea. The English speakers were university exchange students in Korea, primarily from North America and Europe (i.e. a mix of L1 and L2 speakers) that met the proficiency criteria above. There were no exclusion criteria based on game skill level or video game experience. The game used in this study (*Don't Starve Together*) was purposely chosen because it is considered accessible for all ages and skill levels.

Participants were randomly assigned to one of 3 conditions (described in Section 2.4). All participants were provided a brief 5-minute tutorial for how to play the game prior to the start of the experiment, in either Korean or English. The experiments were approved by the Hanyang University IRB (#HYU-2021-138).

The experimental setup involved two computers in two separate rooms, one for the human participant ("player computer") and one for the virtual avatar where its code was run ("confederate computer"), both linked to the same online game server. The player computer was further equipped with an HD camera, headphones, and Blue Snowball microphone for high-quality audio-visual input/output. Each game session involved one human participant and the virtual avatar, engaging in a 30-minute game session on a private server in 2-player cooperative gameplay mode. We set up a Zoom meeting to allow direct audio-visual communication between the human and avatar while playing the game, in a side-by-side configuration. An example of this can be seen in Figure 1. During the game session, the virtual avatar interacted autonomously with the human participant through speech and basic gestures via Zoom, though the in-game character actions were controlled surreptitiously by a human confederate on the confederate computer. The confederate could hear and see the participant through Zoom. However, participants were not informed about the existence of the confederate, and all *external* sound input from the confederate side was shut off from Zoom so only the virtual avatar appeared or spoke in Zoom on the participant's end, so as far as the participant knew they were just playing the game with the avatar.

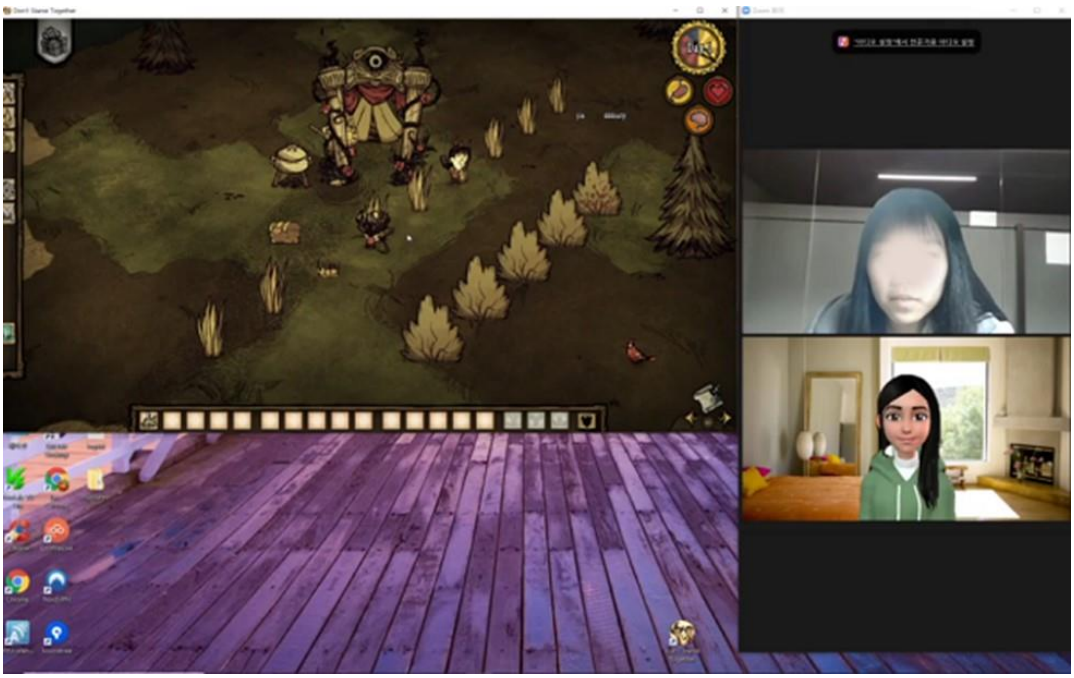

Figure 1: Gameplay example during experiment (human vs avatar)

During each experiment, we collected three main types of data: audio-visual recordings of the gameplay, written game data, and instrument data of human perceptions. We used OBS Studio (https://obsproject.com/) to record the entire computer screen during gameplay, including the game window itself as well as the Zoom window of simultaneous social interactions. That allowed us to later extract the speech from the recordings for both the avatar and human player synced with in-game gameplay events. Additionally we collected written game data via the Game Mod (see Section 2.2), so we could later analyze how different gameplay events influenced the interaction. We also collected several common HRI instruments at the end of each game session, such as the Godspeed scale (Bartneck et al., 2009) for measuring general perceptions of a robot/agent and the Networked Minds instrument (Biocca et al., 2001) for measuring social presence (Oh et al., 2018). The Godspeed

scale was originally developed to understand how perceptions affect the interaction between a robot/agent and a human and has been extensively used throughout the HRI field over the past decade. Meanwhile, the Networked Minds scale originally came from the psychology field and was aimed at understanding human interactions with computers, e.g. chatbots or virtual reality agents, in comparison to social interactions between humans.

2.4 Experimental Conditions

As noted in section 2.3, the experiments described here consisted of 3 conditions that involved different alterations of the virtual avatar's speech behavior. Participants were randomly assigned to one of the three, equally split by language. The **first condition ("Control")** served as our control condition. The speech system was run normally as originally designed (see Section 2.1) without any modification or manipulation.

In the **second condition (H2),** the avatar's speech system was modified so that the avatar was "less chatty" than the other conditions (i.e. talked less). This was accomplished by implementing a priority system for different kinds of utterances, where some utterances had a higher priority level and others had a lower priority level. We then deployed a control mechanism that could be turned up or down like a volume dial so that only utterances above the threshold would be spoken. The result was that the avatar would talk less if the threshold was set higher, and vice versa. For the H2 condition in this study, we set the threshold up high, so that only high priority utterances were produced by the avatar related to critical situations (e.g. imminent danger, monster attacks, starvation). The priority system was developed through a data-driven analysis of human vs human gameplay recordings, described in detail in (Bennett et al., 2022b).

In the **third condition (H3)**, the avatar's speech system was modified to have less "speech awareness" and thus be more likely to repeat things it already just said. This is rooted in the concept of *Social Inhibition of Return* (social IOR), which is based on IOR models from various human sensory functions such as vision (Nafcha et al., 2020). The fundamental idea is that there are mechanisms in the brains of naturally intelligent organisms (including humans) that inhibit us from repeating the same behavior in a short period of time (e.g., 2-3 seconds) in order to maximize task efficiency (e.g., during visual "information foraging") (Klein & MacInnes, 1999). A failure in these mechanisms is thought to play a role in human mental illness, such as obsessive-compulsive disorder. In the context of social IOR, these mechanisms are also important to produce fluid natural behavior, rather than repetitive "robot-like behavior" (Nafcha et al., 2020). In all other experimental conditions, the avatar system had a built-in social IOR mechanism which utilized the top-level utterance categories from the speech hierarchy (7 total) so that the Social AI maintained an internal array to keep track of recently spoken categories, with a "counter" that counted down a certain number of seconds during which any further utterances within that same category were suppressed (though the AI could still make utterances from other categories). This counter was set to 3 seconds, based on prior research on social IOR in humans. However, in this H3 condition here, that social IOR mechanism was turned off. The virtual avatar was free to repeat itself frequently, even sometimes before the human had a chance to respond to the first utterance.

2.5 Analysis Approach

For the analysis of language differences in this paper, we first extracted the speech data from the OBS recordings of the experiments in order to create data for NLP analysis. This entailed using speaker diarization via Google Cloud services to automatically identify avatar and human participant speech in the recorded video of each game session, resulting in output transcripts with timestamps (so they could be synced with in-game gameplay events). It was necessary to perform some post-diarization manual cleanup of those transcripts to ensure accuracy. The speech data was then analyzed in multiple ways, by both language and condition. That analysis entailed various statistical methods (t-tests, ANOVAs, etc.) and data visualizations performed in either Python or R, which are described in the relevant sections in the results below (see Section 3). Unless otherwise noted, the language comparisons were performed with two-tailed independent-samples t-tests, while condition comparisons were performed using one-way ANOVAs. Two-way ANOVAs by language and condition were also performed, but we only detected a significant interaction effect for one particular analysis (interruption frequency), so for the most part those results are omitted here for brevity. A visual flow chart for the overall analysis process is shown in Figure 2.

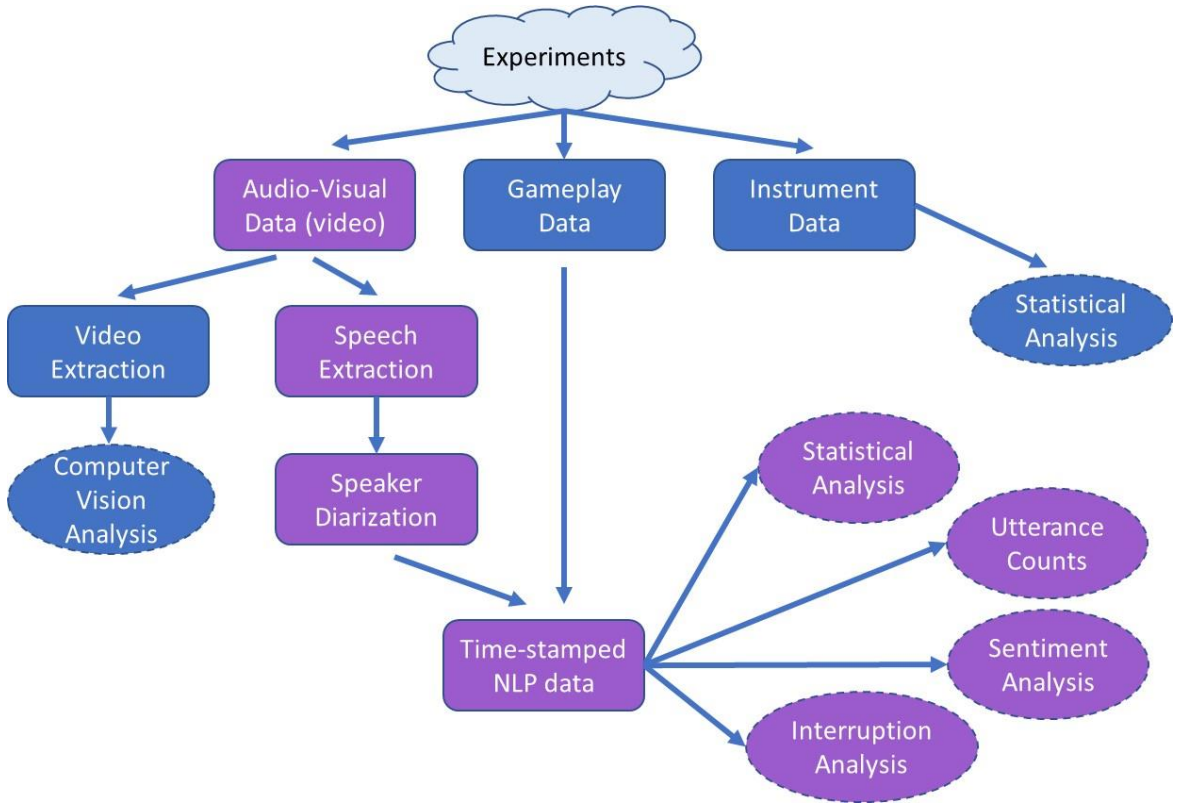

Figure 2: Flow Chart of Overall Analysis Process. Rectangles represent process steps, while circles represent analyses. Since this paper focuses on the speech data (shown in purple), details about the computer vision and gameplay data processing are omitted, for brevity.

To understand the basic frequency of speech, utterance counts were calculated for both the human and avatar. Utterance counts for the avatar were further separated into two categories: self-generated speech and ASR responses to human speech. We also conducted sentiment analysis using lexical

parsing via VADER (Hutto & Gilbert, 2014). For English, VADER was used directly, while a scientifically-validated Vader-like dictionary was used in Korean (Park et al., 2020). We additionally conducted an interruption analysis, looking for places where either the avatar interrupted the human participant by speaking while the human was still speaking (i.e. inter-pausal unit, or IPU), or vice versa the human interrupted the avatar (Skantze, 2021). This was done through manual annotation of the speech transcript data, by identifying places where timestamps of utterances overlapped without any pause between. Finally, we analyzed the instrument data to evaluate whether there were differences in the human perceptions of the avatar social interaction during gameplay.

3. Results

3.1 Utterance Counts

To get an overall sense of the frequency of speech for both the avatar and human participant, we first undertook an analysis of utterance counts. An overall comparison of utterance counts across all experiments can be seen in Table 1, comparing differences between the human and avatar (regardless of language or condition), based on two-tailed independent-sample t-tests. As can be seen in the table, there were significant differences in the speech frequencies of the human and avatar (henceforth indicated as * = 0.05 level, ** = 0.01 level, *** = 0.001 level). This was to be expected, as humans have a greater range of speech capabilities than our current state-of-the-art conversational agents. There were also different types of avatar speech (self-generated vs. ASR response), though that was by design. As shown in Figure 3, roughly 80% of the avatar's speech was self-generated based on contextual gameplay events, while the remaining 20% were ASR-based responses to human speech.

Table 1: Overall Utterance Counts

Categories 1	Category 2	Cat1 Mean (std)	Cat2 Mean (std)	*p*-val	*Sign.*
Human	Avatar	68.88 (61.97)	48.1 (39.21)	0.03040	*
Human	Avatar-self generated	68.88 (61.97)	39.75 (34.19)	0.00190	**
Human	Avatar-ASR	68.88 (61.97)	8.35 (10.65)	0.00000	***
Avatar	Avatar-self generated	48.1 (39.21)	39.75 (34.19)	0.21620	
Avatar	Avatar-ASR	48.1 (39.21)	8.35 (10.65)	0.00000	***
Avatar-self generated	Avatar-ASR	39.75 (34.19)	8.35 (10.65)	0.00000	***

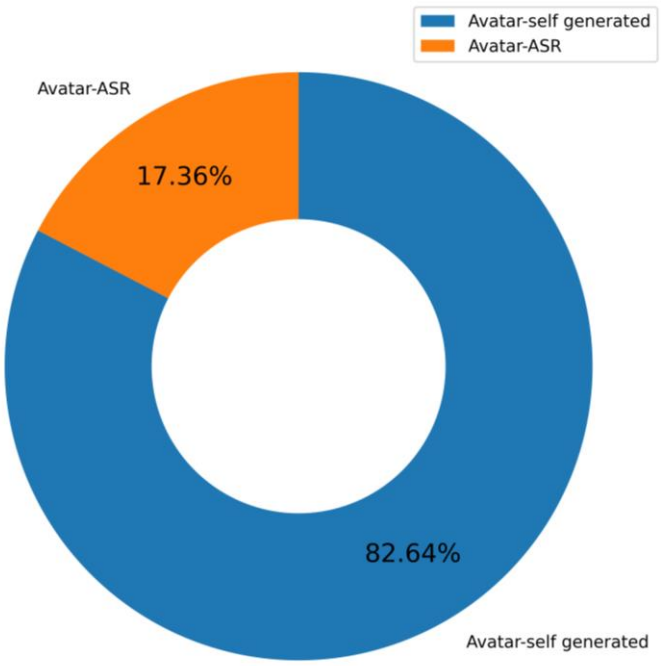

Figure 3: Avatar Speech Patterns, showing the percentage of total utterances that were self-generated versus ASR responses to human speech.

We then analyzed the utterance counts in more detail, by language (Table 2) and by condition (Table 3). As can be seen in Table 2, the utterance counts were significantly different between English speakers and Korean speakers for all types of speech (both human and avatar). Table 3 shows however that the amount of human speech was not significantly different across conditions, but that it was different for the avatar particularly for condition H2, which was by design (in the H2 condition the avatar was deliberately made "less chatty", see Section 2.4). We do note that despite the language differences, the patterns across conditions were quite similar between English and Korean, as can be seen in the visualizations in Figure 4 and Figure 5. We also note that modifications of the avatar's speech behavior across conditions did not seem to impact the utterance counts of the human participant, but we will see that those modifications did impact other aspects of the speech interaction (sentiment, interruptions, etc.) in later Results sections.

Table 2: Utterance Counts, by Language

	English Mean (std)	**Korean Mean (std)**	***p*-val**	***Sign.***
Human	105.38 (69.65)	34.74 (23.09)	0.00001	***
Avatar	71.38 (42.01)	26.32 (19.06)	0.00001	***
Avatar-self generated	56.69 (38.94)	23.9 (18.63)	0.00010	***
Avatar-ASR	14.69 (11.6)	2.42 (4.71)	0.00001	***

Table 3: Utterance Counts, by Condition

	Control Mean (std)	H2 Mean (std)	H3 Mean (std)	*p*-val	*Sign.*
English					
Human	122.67 (54.87)	124.7 (86.77)	70.5 (53.1)	0.10281	
Avatar	99.44 (18.02)	18.4 (12.62)	99.1 (16.77)	0.00000	***
Avatar-self generated	79.33 (10.93)	5.7 (3.34)	87.3 (12.54)	0.00000	***
Avatar-ASR	20.11 (11.24)	12.7 (12.33)	11.8 (10.56)	0.25240	
Korean					
Human	35.1 (18.46)	37.45 (23.84)	31.4 (27.98)	0.87367	
Avatar	32.6 (17.68)	7 (7.77)	41.3 (9.06)	0.00000	***
Avatar-self generated	30.8 (16.87)	3.82 (3.68)	39.1 (7.36)	0.00000	***
Avatar-ASR	1.8 (1.48)	3.18 (7.32)	2.2 (3.33)	0.80341	

What is interesting here is that there were such **stark differences in the utterance counts between the languages, despite using the *exact same* speech hierarchy generated by native speakers as well as using the *exact same* ASR/TTS cloud technology** to detect/generate speech during the experiments. It is possible that was due to the lack of speech interaction from Korean speakers (who on average spoke less frequently), the limitations of current ASR/TTS technology in non-English languages, or perhaps the way the Korean players played the game. To that latter possibility, one potential explanation was that the Korean speakers were not as "adventurous" during gameplay for some reason, which led to encountering less gameplay events that would trigger speech interaction. We return to this topic in the Discussion section.

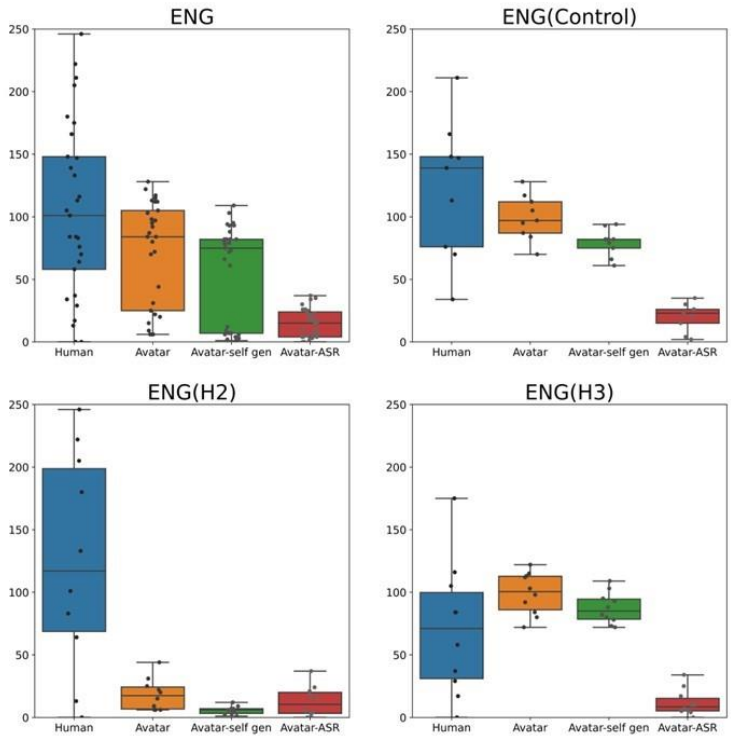

Figure 4: English Utterance Counts, by Condition. The y-axis represents the average number of utterances per participant (and standard deviation).

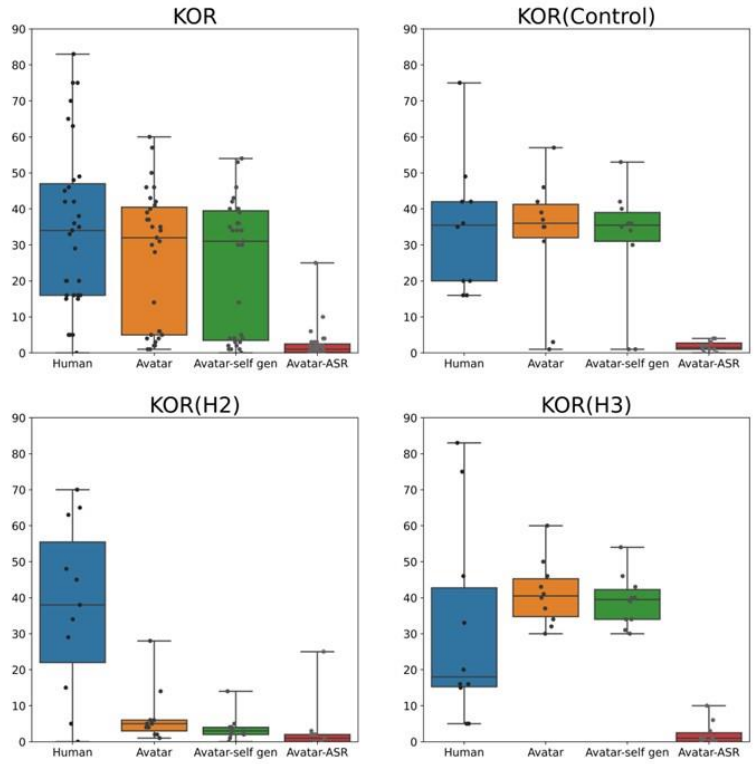

Figure 5: Korean Utterance Counts, by Condition. The y-axis represents the average number of utterances per participant (and standard deviation).

3.2 Sentiment Analysis

We performed a sentiment analysis to evaluate whether there were differences in the affective communication of the speech content between the Korean and English speakers, based on English and Korean versions of VADER (Hutto & Gilbert, 2014; Park et al., 2020). Results can be seen in Figure 6, broken out by human speech and avatar speech for each language. Given the differences in utterance counts between the languages, these are expressed as percentages (rather than raw counts) for a fair comparison.

As can be seen in the two right-side columns in Figure 6, the sentiment percentages for the avatar speech were quite similar between Korean and English. However, for the human participants' speech (left 2 columns), the sentiment was notably different between languages. The "Neutral" category was the same, but Korean speakers had much higher percentage of negative utterances than English speakers (21.1% vs. 12.1%, respectively). Meanwhile the inverse was true for English speakers, who were made more frequent positive utterances than Korean speakers on average (24.8% vs. 14.1%). **In short, the Korean speakers were more negative, while the English speakers were more positive.** Two-tailed independent samples t-test was calculated as p=0.0012 for negative sentiment (std dev: 11.2 KOR, 7.49 ENG, n=60), with positive sentiment having a similar p-value of 0.003 (std dev: 10.6 KOR, 10.7 ENG, n=60). We also analyzed differences across condition, but found none (results omitted here for brevity).

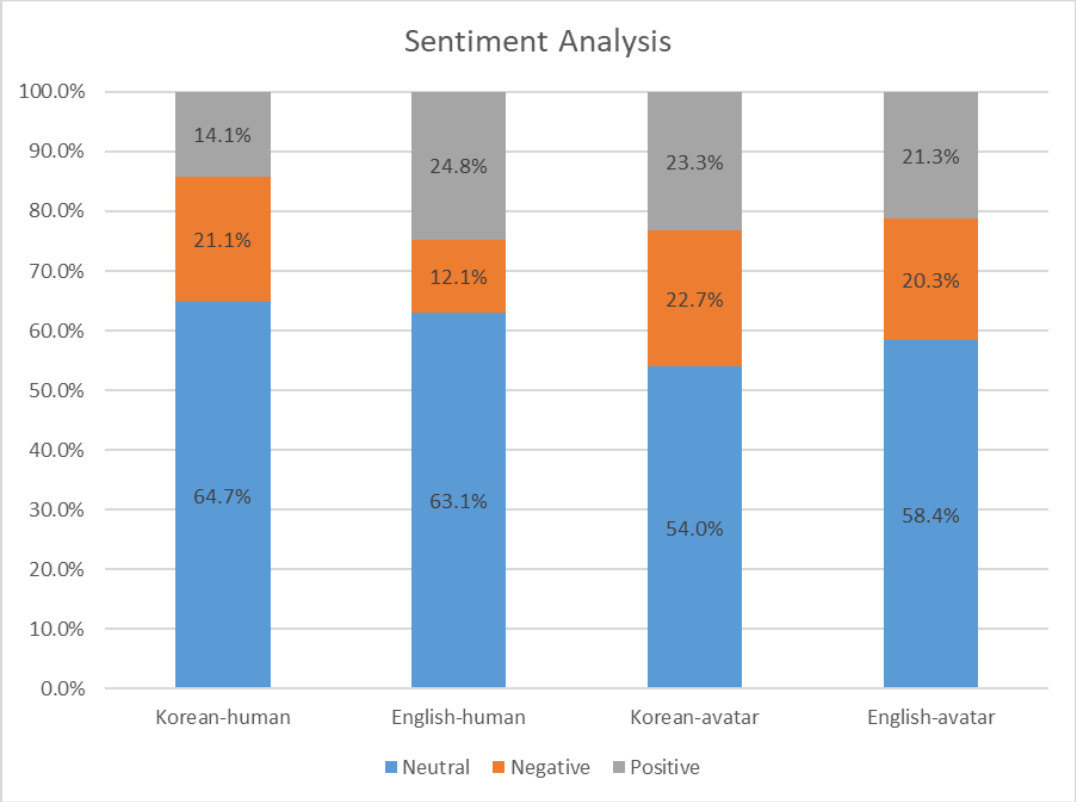

Figure 6: Sentiment Analysis Comparison. The y-axis indicates the percentage of total utterances identified as each sentiment.

The language differences in sentiment could perhaps partially explain the differences in utterance counts between languages (see Section 3.1), whereas if the Korean speakers had a more negative experience they may have been inclined to talk less and/or be less adventurous during gameplay. In general, many of the positive/negative sentiment utterances in both languages related to resources, player status (e.g. hunger, health), and monsters. Adversely, it could also be explained as a cultural difference between Western and East Asian mindsets, either in terms of expectations of how to cooperate during social interaction (e.g. cooperative gameplay) or simply linguistic differences leading to different communication styles (Sanchez-Burks et al., 2003; Imada et al., 2013; Yum, 1988; Liddell & Williams, 2019). Another possible explanation could be technological, i.e. differences in the original English-based VADER and the Korean-version VADER that we used. Though, if that was the case then we would have expected there to also be differences in the sentiment analysis of the avatar's speech, which we did not detect.

3.3 Interruption Analysis

We performed an analysis of speech interruptions to evaluate whether there were differences in the frequency of interruptions by language and by condition, for both the human participant and the avatar. Interruptions were defined as speech that overlapped the other player's speech (IPU) rather than waiting for them to finish speaking, regardless of whether it was accidental or not. More broadly, such interruptions can be viewed as failures of proper *turn-taking* during a social interaction (Skantze, 2021).

Results can be seen in Table 4. Since there were differences in the utterance counts between language and condition (see Section 3.1), the interruption counts were calculated as a percentage of the total utterance count within each category, for fair comparison.

Table 4: Interruption Frequency (interruptions as percentage of total utterance count)

	Avatar Mean (std)	Human Mean (std)	*p*-val	*Sign.*
Language				
English	2.6% (0.61)	1.4% (0.38)	0.00030	***
Korean	1.1% (0.47)	1.4% (1.08)	0.16830	
Condition				
Control	2.5% (0.83)	1.1% (0.33)	0.00010	***
H2	1.0% (0.61)	2.6% (1.61)	0.00010	***
H3	2.1% (0.58)	0.4% (0.20)	0.00010	***
Overall	1.9% (0.39)	1.4% (0.58)	0.02170	*

There are several key takeaways from Table 4. First, **the avatar was much more likely to interrupt in English**, even after controlling for differences in utterance counts. The reason for that is unclear. There were no differences on the human side between languages. We also note that a more detailed analysis showed that the differences for the avatar across conditions were also primarily on the English language side (data not shown for brevity). There was a significant interaction effect (*p*-val=0.025) for the avatar by language and condition, which was the only significant one detected for any analysis in the entire study. As for the conditions, the **human participant was much more likely to interrupt when the avatar spoke less (condition H2)**, and vice versa for the avatar interruptions. One possible interpretation was that this was due to the human participant trying to "fill the empty space" in the conversation when the avatar was spoke less, which inadvertently led to an increased frequency of interruptions by the human. This is something to consider during future design of interactive devices, conversational agents, and other HRI systems.

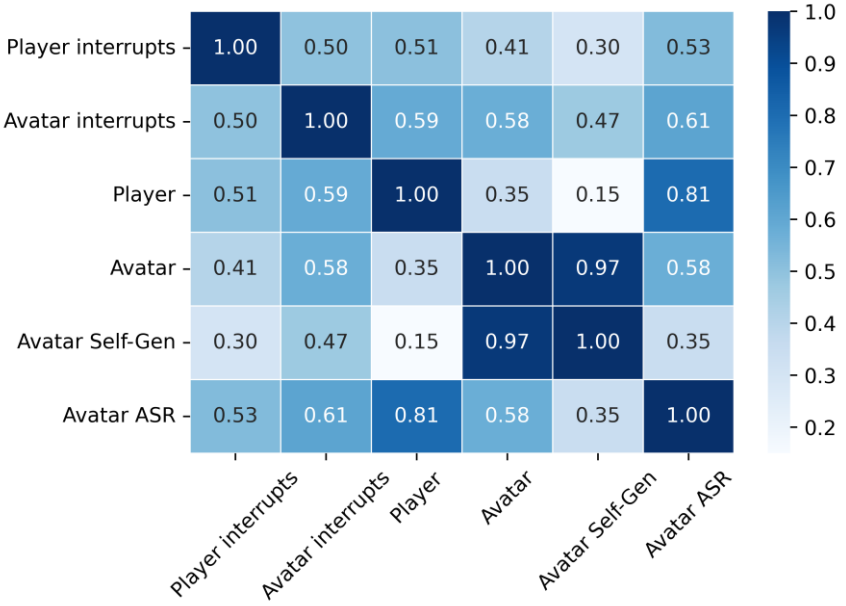

Figure 7: Interruption Frequency Correlations with Utterance Counts (by speaker and speech type)

Another question we had was whether the interruption frequency by the avatar was more related to its self-generated speech or its ASR-based responses to human speech. In other words, were the interruptions due to the avatar making self-generated comments about the gameplay situation or due it trying to converse with the human based on human prompts. To see if there were any clues to this, we conducted a correlation analysis comparing avatar and human interruptions with the utterance counts of different types of speech (human, avatar, avatar self-generated, avatar ASR). The results are shown in Figure 7.

There are several interesting things in that figure, but we would direct the reader's attention to one thing of particular note. **For both human interruptions and avatar interruptions, the frequency was more correlated with the avatar's ASR responses rather than the self-generated speech** (human: 0.53 vs 0.30, avatar: 0.61 vs 0.47). In other words, the interruption patterns did seem to relate more to the ASR-based attempts by the avatar to converse with the human participant, which likely indicates that there need to be other indicators of "turn taking" beyond just the content of the human speech itself during cooperative gameplay in HRI. Those indicators might include acoustic features of speech or possibly non-verbal cues, but they will likely need to be linguistically and/or culturally specific. For instance, rising or falling pitch is used in some languages (but not all) to indicate the end of one speaker's turn (Skantze, 2021).

3.4 Instrument Data Analysis

We also analyzed the instrument data to evaluate whether there were differences in the human perceptions of the avatar social interaction during cooperative gameplay. This was focused on general perceptions of the virtual avatar (using the Godspeed instrument) as well as perceptions of its social presence (using the Networked Minds instrument), which are described in Section 2. Results can be seen in Table 5 (by language) and Table 6 (by condition). To summarize, there were significant differences in general perceptions by language, but not for social presence. **English participants rated the virtual avatar more highly than the Korean speakers**. In contrast, there were significant differences in social presence by condition, but not language. There were no significant interaction effects in any case.

Table 5: Instrument Analysis, by Language

	English	Korean	*p*-val	*Sign.*
Godspeed	3.35	3.03	0.04030	*
NM Self	3.10	3.05	0.66673	
NM Other	3.15	3.03	0.27737	
NM Avg Difference	0.22	0.24	0.65475	

Table 5: Instrument Analysis, by Condition

	Control	H2	H3	*p*-val	*Sign.*
Godspeed	3.44	3.05	3.09	0.07931	
NM Self	3.20	2.88	3.15	0.02268	*
NM Other	3.31	2.89	3.07	0.00740	**
NM Avg Difference	0.23	0.28	0.18	0.23611	

The critical aspect of the Networked Minds instrument however is not so much the raw values, but rather how correlated the ratings of "Self" and "Other" are. Higher correlations indicate a higher degree of social presence, in that the human feels a greater sense of connection with the agent they are interacting with, in terms of actions, intent, emotion, etc. (Biocca et al., 2001). As such, we evaluated the correlations between NM Self and NM Other, by language and condition. The results are shown in Table 7.

Table 7: Networked Minds Correlations (social presence), by Language and Condition

	Control	H2	H3
English	0.904	0.489	0.842
Korean	0.316	0.853	0.835
Overall	0.748	0.717	0.843

The key takeaway from Table 7 is that Korean speakers and English speakers responded very differently to the Control and H2 conditions. **Koreans felt less social presence during the Control condition (when the avatar was less repetitive), while English felt less social presence during H2 (when the avatar was less chatty)**. Both groups reported very similar values for H3, when social IOR was turned off so that the avatar was still spoke more like the Control condition but engaged in more repetitive speech behaviors (see Section 2.4). It is not clear why these differences exist, but we suspect they are related to cultural differences between East Asian cultures and Western cultures in terms of the social norms that govern appropriate social interaction behavior (Masuda et al., 2008; Stankov, 2015; Bennett et al., 2014; Bennett & Weiss, 2022).

4. Discussion

4.1 Summary of Results

The primary goal of this study was to evaluate whether there are differences across languages in how people interact with a robot or virtual avatar in cooperative game environments, where the human and artificial agent must socially interact and cooperate in order to succeed at some goal. To that end, we ran a series of experiments with 60 participants (30 English speakers and 30 Korean speakers) interacting with a bilingual virtual avatar capable of autonomous speech interactions during gameplay (Bennett et al., 2022b). The experiments included several conditions, in which we modified the avatar's speech behavior in different ways. We collected multiple types of data, including audio-visual recordings of interactions between the human and avatar, speech data for both the human and avatar, gameplay data, and instrument data on human perceptions of the virtual avatar agent.

Based on extensive analysis of that data, the main results can be summarized as follows. First, there were notable differences between English and Korean in the amount of spoken utterances during each game session, despite the avatar using the same speech system and technology platform in both languages (with speech content that had been generated by native speakers). Sentiment analysis of the utterances made during the experiments also showed that Korean speakers were more negative, while the English speakers were more positive, which may relate to differences in

communication styles between East Asian and Western cultures. An analysis of speech interruptions showed that the avatar was nearly twice as likely to interrupt the human when speaking in English, rather than Korean, and that the human was more likely to interrupt the avatar when it was less “chatty” (i.e. spoke less). Furthermore, interruptions by both the human and the avatar were more correlated with the avatar’s ASR-based responses to human speech (i.e. its attempts to converse with the human) rather than its self-generated utterances about the game situation. Finally, the instrument data showed that English speakers generally rated the avatar more highly than the Korean speakers (in terms of perceptions of animacy, likeability etc.). On the other hand, the English speakers had higher perceptions of *social presence* with the avatar when it was more chatty, while the Korean speakers had higher perceptions when it engaged in more repetitive speech behavior. Relative to our initial study goals (see Section 1.3), these findings can be summarized as:

1) There were significant differences in human speech behavior across languages (e.g. amount of speech, speech sentiment)
2) Altering the avatar’s speech behavior produced different effects in different languages, such as turn-taking behavior (i.e. frequency of speech interruptions)
3) Human perceptions of the virtual avatar (as rated on standardized HRI instruments) were significantly affected by both #1 and #2 above

The reasons for all these differences are not entirely clear at this point, but potentially may relate to cultural differences between East Asian cultures and Western cultures in terms of the social norms that govern appropriate social interaction behavior (Masuda et al., 2008; Stankov, 2015; Bennett et al., 2014; Bennett & Weiss, 2022). We discuss the implications of this for HRI and autonomous speech systems in the next section.

4.2 Implications

The results presented in this research have a number of implications for HRI, conversational agents, and other autonomous speech systems. In short, if there are *subtly* different communication styles across languages and cultures, this will create challenges to the development of those kinds of interactive technologies. Previous research on the differences in human-human interaction are plentiful, going back to Yum’s (1988) seminal work on the subject. For instance, Confucian principles lead to stronger indirect communication and in-group/out-group signaling being embedded into East Asian languages, whereas Western languages like English have a more outcome-oriented focus with direct communication. Those kinds of differences in communication style have been shown to impact a wide array of scenarios, including work environments (Sanchez-Burks et al., 2003), responses to accidents (Liddell & Williams, 2019), and early childhood development in elementary school age children (Imada et al., 2013).

Moreover, similar to the results with the virtual avatar in this paper, we are seeing the same phenomenon of cross-cultural differences in another study using an entirely robotic platform. That study involved placing physical robotic pets into user homes in East Asia and the United States, with

the Asian participants reporting greater negative sentiment towards the interactions and a desire for the robotic pet's behaviors to be "more subdued" in nature (Bennett et al., 2022a).

Previous work, including our own, has argued for creating more culturally-aware artificial agents, which are adaptive to the kinds of *situated use* cases that occur in different cultural and linguistic settings (Lee & Sabanovic, 2014; Bruno et al., 2018; Sabanovic et al., 2014). However, more recent research has shown that cultural homophily (e.g. an agent adapted to a specific set of cultural attributes) in and of itself does not necessarily always correspond to better performance of a robot or higher ratings by human users (Lim et al., 2021). Our results here provide further support for those recent findings, suggesting that it may in fact be necessary to design social robots and interactive agents *explicitly* for different cultural and linguistic settings, with *fundamentally* different models of behavior specific to those settings.

4.3 Limitations

There are also a number of limitations to this work that need to be mentioned. First of all, we should be clear that the interpretation of the observed differences between Korean and English speakers in the results here as being related to *underlying* cultural differences is our own interpretation. It is based on existing research on the topic (see Section 4.2) and the concept of *linguistic relativity* affecting how people think and behave (Athanasopoulos & Casaponsa, 2020; Wang & Wei, 2021), as well some of our own past research on cross-cultural differences during HRI. However, it is possible that language differences are simply that, and unrelated to broader cultural differences. Definitively teasing apart those possible explanations is difficult at best, and more research is warranted on the topic.

There are also some limitations relative to our experimental design. For instance, we have no baseline condition here to compare human-human interactions during the same experimental setup. We have conducted some initial pilot tests into that with a few participants, and the results were similar to those here in this paper, but the sample size was limited. We are planning a larger full experiment in the future, but for now it remains an open question as to whether these results reflect human-human interaction patterns during gameplay more broadly, though there is some existing research that suggests that it should be reflective, at least in some situations (Banks & Bowman, 2016; Pino et al., 2021). Similarly, we did not vary the virtual avatar's appearance during the current experiments, instead using the same ethnically-ambiguous Loomie avatar for all experiments (see Figure 1). However, appearance is something that is known to impact human interactions with artificial agents, both virtual and embodied (McDonnell & Mutlu, 2021), which is something that could be explored further during speech interactions.

Finally, we note that throughout this paper we have often mentioned virtual avatars and physically-embodied robots in conjunction, though there are potentially significant differences between virtual agents and physical agents when it comes to human interaction. That is a topic that has been extensively studied in the field of HRI, including some of our own past work (Deng et al., 2019; Bennett & Sabanovic, 2014). However, it is a complicated issue, and many robots now include "digital interfaces" (e.g. screens) that incorporate both embodied interaction as well as virtual interaction on the *same* platform (i.e. "mixed reality") (Holz et al., 2009; Groechel et al., 2019; Prattico & Lamberti, 2020). Suffice it to say, it would be interesting to see if the results observed here would replicate on

physically-embodied multi-lingual robotic platforms or whether there would be any notable differences. That is an area ripe for future research.

4.4 Future Work

There are a number of potential research avenues left for future work on this subject. For instance, we are currently conducting a study with bilingual participants, to look at the effects of “language switching” during a single game session (averse to using only one language the whole time, like the experiments reported here). It is possible bilingual participants may show different effects, or that the effects may change when the language switching occurs depending on how they allocate their cognitive resources during the game in their dominant vs. non-dominant language. Along with that, we are conducting a separate study looking at the effects of “anticipatory speech”, where the virtual avatar attempts to predict future game events right before they happen and talk about them, rather than only talking about events that already occurred or currently happening. Much natural human speech involves talking about future plans, which is a core part of social cohesiveness in many cultures (De Waal & Ferrari, 2010).

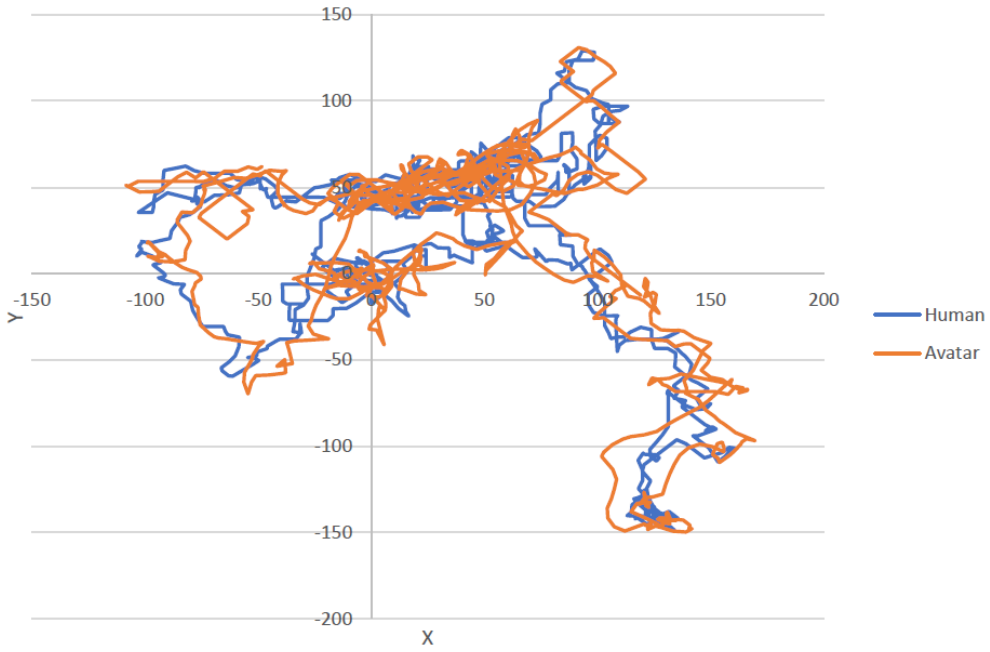

Figure 8: Distance Traveled during Gameplay (example from one participant). Represented by the human and avatar in-game position over time, where (0,0) is the fixed starting point on our customized game map

Beyond those current ongoing studies, we are also analyzing in-game actions in more detail, such as distance traveled and whether specific in-game interactions (e.g. monster encounters) affected the speech interactions. Hypothetically, one possible explanation for some of our results here is that Korean speakers were less “adventurous” than English speakers within the game, which thus lead to differences in speech. For example, Figure 8 shows how we can map the in-game movement of the human and avatar, which could then be converted into metrics to compare across languages and

conditions. The best way to go about that though is something that requires more study, in order to clearly link cooperative gameplay patterns to speech behavior.

We are also working on computer vision components to examine the interplay of verbal and non-verbal aspects (facial expressions, gestures), which was not considered in the current study. For instance, it is possible such non-verbal aspects may be capable of generating improved turn-taking capabilities by the virtual avatar, which in turn might lead to a reduced difference in the number of interruptions between languages. However, on the other hand, we do know from previous research in such cooperative game environments that direct face-to-face interactions during gameplay seem to be sparse and linked to sporadic game events (i.e. players primarily focus on the game itself) (Bennett et al., 2022b). There are also existing research challenges in regards to effectively integrating multi-modal audio and visual data streams in real time, which remain to be addressed (Tsai et al., 2019). More research is needed on these topics.

Acknowledgments

This work was supported by a grant from the National Research Foundation of Korea (NRF) (Grant number: 2021R1G1A1003801). We would also like to thank Cheda Stanojevic (Indiana University) as well as Jihong Jeong and Jaeyoung Suh (Hanyang University) for their assistance in this work.

References

1. Athanasopoulos, P., Casaponsa, A., 2020. The Whorfian brain: Neuroscientific approaches to linguistic relativity. Cognitive Neuropsychology, 37(5-6), 393-412. https://doi.org/10.1080/02643294.2020.1769050
2. Banks, J., Bowman, N.D., 2016. Avatars are (sometimes) people too: Linguistic indicators of parasocial and social ties in player–avatar relationships. New Media & Society, 18(7), 1257-1276. https://doi.org/10.1177/146144481455489
3. Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S., 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 1(1), 71-81. https://doi.org/10.1007/s12369-008-0001-3
4. Bennett, C.C., Šabanović, S., 2014. Deriving minimal features for human-like facial expressions in robotic faces. International Journal of Social Robotics, 6(3), 367-381. https://doi.org/10.1007/s12369-014-0237-z
5. Bennett, C. C., Šabanović, S., Fraune, M. R., & Shaw, K., 2014. Context congruency and robotic facial expressions: Do effects on human perceptions vary across culture? 23rd IEEE International Symposium on Robot and Human Interactive Communication (ROMAN), pp. 465-470. https://doi.org/10.1109/ROMAN.2014.6926296
6. Bennett, C.C., Stanojević, C., Kim, S., Šabanović, S., Piatt. J.A., Lee, J., Yu, J., Oh, J., 2022a. Comparison of In-home Robotic Companion Pet Use in South Korea and the United States: A

Case Study. 9th IEEE International Conference on Biomedical Robotics & Biomechatronics (BIOROB), pp.1-7. https://doi.org/10.1109/BioRob52689.2022.9925468

7. Bennett, C.C., Weiss, B., Suh, J., Yoon, E., Jeong, J., Chae, Y., 2022b. Exploring Data-Driven Components of Socially Intelligent AI through Cooperative Game Paradigms. Multimodal Technologies and Interaction, 6(2), 16.
8. Bennett, C.C., Weiss, B., 2022. Purposeful failures as a form of culturally-appropriate intelligent disobedience during human-robot social interaction. 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS): Best and Visionary Papers, pp. 84-90. https://doi.org/10.1007/978-3-031-20179-0_5
9. Biocca, F., Harms, C., Gregg, J., 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. 4th Annual International Workshop on Presence, Philadelphia, PA, pp. 1-9.
10. Bruneau, T.J., 1973. Communicative silences: Forms and functions. Journal of communication, 23(1), 17-46. https://doi.org/10.1111/j.1460-2466.1973.tb00929.x
11. Bruno, B., Menicatti, R., Recchiuto, C. T., Lagrue, E., Pandey, A. K., Sgorbissa, A., 2018. Culturally-competent human-robot verbal interaction. 15th International Conference on Ubiquitous Robots (UR), pp. 388-395. https://doi.org/10.1109/URAI.2018.8442208
12. Chahboun, S., Kvello, Ø., Page, A. G., 2021. Extending the field of extended language: A literature review on figurative language processing in neurodevelopmental disorders. Frontiers in Communication, 143. https://doi.org/10.3389/fcomm.2021.661528
13. Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F. S., Paiva, A., 2016. Just follow the suit! trust in human-robot interactions during card game playing. 25th IEEE international symposium on robot and human interactive communication (RO-MAN). pp. 507-512. https://doi.org/10.1109/ROMAN.2016.7745165
14. Deng, E., Mutlu, B., Mataric, M.J., 2019. Embodiment in socially interactive robots. Foundations and Trends® in Robotics, 7(4), 251-356. http://dx.doi.org/10.1561/2300000056
15. Dennett, D.C., 1971. Intentional systems. The Journal of Philosophy, 68(4), 87-106.
16. Deutscher, G., 2010. Through the Language Glass: Why the World Looks Different in Other Languages. Metropolitan books, New York, USA.
17. De Waal, F.B., Ferrari, P.F., 2010. Towards a bottom-up perspective on animal and human cognition. Trends in cognitive sciences, 14(5), 201-207. https://doi.org/10.1016/j.tics.2010.03.003
18. Doyle, P.R., Clark, L., Cowan, B. R., 2021. What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. CHI Conference on Human Factors in Computing Systems (CHI), pp. 1-14. https://doi.org/10.1145/3411764.3445206
19. Enfield N., 2017. How we talk. The Inner Workings of Conversation. BasicBooks, New York USA.

20. Engwall, O., Lopes, J., Åhlund, A., 2021. Robot interaction styles for conversation practice in second language learning. International Journal of Social Robotics, 13(2), 251-276. https://doi.org/10.1007/s12369-020-00635-y

21. Fischer, K., Naik, L., Langedijk, R. M., Baumann, T., Jelínek, M., Palinko, O., 2021. Initiating Human-Robot Interactions Using Incremental Speech Adaptation. Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 421-425. https://doi.org/10.1145/3434074.3447205

22. Gonzalez-Franco, M., Lanier, J., 2017. Model of illusions and virtual reality. Frontiers in psychology, 8, 1125. https://doi.org/10.3389/fpsyg.2017.01125

23. Groechel, T., Shi, Z., Pakkar, R., Matarić, M. J., 2019. Using socially expressive mixed reality arms for enhancing low-expressivity robots. 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1-8. https://doi.org/10.1109/RO-MAN46459.2019.8956458

24. Gwóźdź, M., 2019. Figurative Language Grounding in Humanoid Robots. In: Proceedings of SAI Intelligent Systems Conference. Springer, Cham, pp. 347-362. https://doi.org/10.1007/978-3-030-29513-4_25

25. Holz, T., Dragone, M., O'Hare, G. M., 2009. Where robots and virtual agents meet. International Journal of Social Robotics, 1(1), 83-93. https://doi.org/10.1007/s12369-008-0002-2

26. Honig, S., Oron-Gilad, T., 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. Frontiers in Psychology, 9, 861. https://doi.org/10.3389/fpsyg.2018.00861

27. Hutto, C., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media. 8(1), 216-225.

28. Imada, T., Carlson, S. M., Itakura, S., 2013. East–West cultural differences in context-sensitivity are evident in early childhood. Developmental Science, 16(2), 198-208. https://doi.org/10.1111/desc.12016

29. Jesso, S.T., Kennedy, W.G., Wiese, E., 2020. Behavioral cues of humanness in complex environments: How people engage with human and artificially intelligent agents in a multiplayer videogame. Frontiers in Robotics and AI, 7, 531805. https://doi.org/10.3389/frobt.2020.531805

30. Jung, M., Hinds, P., 2018. Robots in the wild: A time for more robust theories of human-robot interaction. ACM Transactions on Human-Robot Interaction (THRI), 7(1), 1-5. https://doi.org/10.1145/3208975

31. Klein, R.M., MacInnes, W.J., 1999. Inhibition of return is a foraging facilitator in visual search. Psychological science, 10(4), 346-352. https://doi.org/10.1111/1467-9280.00166

32. Kousta, S.T., Vinson, D.P., Vigliocco, G., 2008. Investigating linguistic relativity through bilingualism: the case of grammatical gender. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(4), 843. https://doi.org/10.1037/0278-7393.34.4.843

33. Koutsombogera, M., Vogel, C., 2019. Speech pause patterns in collaborative dialogs. In: Innovations in Big Data Mining and Embedded Knowledge. Springer, Cham, pp. 99-115. https://doi.org/10.1007/978-3-030-15939-9_6

34. Lee, H.R., Šabanović, S., 2014. Culturally variable preferences for robot design and use in South Korea, Turkey, and the United States. 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 17-24. https://doi.org/10.1145/2559636.2559676

35. Liddell, B.J., Williams, E.N., 2019. Cultural differences in interpersonal emotion regulation. Frontiers in Psychology, 10, 999. https://doi.org/10.3389/fpsyg.2019.00999

36. Lim, V., Rooksby, M., Cross, E. S., 2021. Social robots on a global stage: establishing a role for culture during human–robot interaction. International Journal of Social Robotics, 13(6), 1307-1333. https://doi.org/10.1007/s12369-020-00710-4

37. Lowry, P. B., Zhang, D., Zhou, L., Fu, X., 2010. Effects of culture, social presence, and group composition on trust in technology-supported decision-making groups. Information Systems Journal, 20(3), 297-315. https://doi.org/10.1111/j.1365-2575.2009.00334.x

38. Masuda, T., Ellsworth, P. C., Mesquita, B., Leu, J., Tanida, S., Van de Veerdonk, E., 2008. Placing the face in context: cultural differences in the perception of facial emotion. Journal of Personality and Social Psychology, 94(3), 365. https://doi.org/10.1037/0022-3514.94.3.365

39. McDonnell, R., Mutlu, B, 2021. Appearance. In: The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition, 105-146. https://doi.org/10.1145/3477322

40. Nafcha, O., Shamay-Tsoory, S., Gabay, S., 2020. The sociality of social inhibition of return. Cognition, 195, 104108. https://doi.org/10.1016/j.cognition.2019.104108

41. Oh, C.S., Bailenson, J.N., Welch, G.F., 2018. A systematic review of social presence: Definition, antecedents, and implications. Frontiers in Robotics and AI, 114. https://doi.org/10.3389/frobt.2018.00114

42. Park, H.M., Kim, C.H., Kim, J.H., 2020. Generating a Korean sentiment lexicon through sentiment score propagation. KIPS Transactions on Software and Data Engineering, 9(2), 53-60. https://doi.org/10.3745/KTSDE.2020.9.2.53

43. Pino, M.C., Vagnetti, R., Valenti, M., & Mazza, M., 2021. Comparing virtual vs real faces expressing emotions in children with autism: An eye-tracking study. Education and Information Technologies, 26(5), 5717-5732. https://doi.org/10.1007/s10639-021-10552-w

44. Prattico, F. G., Lamberti, F, 2020. Mixed-reality robotic games: design guidelines for effective entertainment with consumer robots. IEEE Consumer Electronics Magazine, 10(1), 6-16. https://doi.org/10.1109/MCE.2020.2988578

45. Šabanović, S., Bennett, C.C., Lee, H.R., 2014. Towards culturally robust robots: A critical social perspective on robotics and culture. Proceedings of the HRI Workshop on Culture-Aware Robotics, pp.1-6.

46. Sabanovic, S., Michalowski, M. P., Simmons, R., 2006. Robots in the wild: Observing human-robot social interaction outside the lab.]9th IEEE International Workshop on Advanced Motion Control, pp. 596-601. https://doi.org/10.1109/AMC.2006.1631758

47. Sanchez-Burks, J., Lee, F., Choi, I., Nisbett, R., Zhao, S., Koo, J., 2003. Conversing across cultures: East-West communication styles in work and nonwork contexts. Journal of personality and social psychology, 85(2), 363. https://doi.org/10.1037/0022-3514.85.2.363

48. Seok, S., Hwang, E., Choi, J., Lim, Y., 2022. Cultural differences in indirect speech act use and politeness in human-robot interaction. ACM/IEEE International Conference on Human-Robot Interaction, pp. 470-477. http://dx.doi.org/10.5555/3523760.3523823

49. Skantze, G., 2021. Turn-taking in conversational systems and human-robot interaction: a review. Computer Speech & Language, 67, 101178. https://doi.org/10.1016/j.csl.2020.101178

50. Slater, M., 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), 3549-3557. https://doi.org/10.1098/rstb.2009.0138

51. Stankov, L., 2015. Four GLOBE dimensions of perceived social norms in 33 countries. Learning and Individual Differences, 41, 30-42. https://doi.org/10.1016/j.lindif.2015.07.005

52. Suh, J., Bennett, C. C., Weiss, B., Yoon, E., Jeong, J., Chae, Y., 2021. Development of Speech Dialogue Systems for Social AI in Cooperative Game Environments. IEEE Region 10 Symposium (TENSYMP), pp. 1-4. https://doi.org/10.1109/TENSYMP52854.2021.9550859

53. Thomaz, A., Hoffman, G., Cakmak, M., 2016. Computational human-robot interaction. Foundations and Trends in Robotics, 4(2-3), 105-223. http://dx.doi.org/10.1561/2300000049

54. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences. 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558-6569. http://dx.doi.org/10.18653/v1/P19-1656

55. Wang, Y., Wei, L., 2021. Cognitive restructuring in the multilingual mind: language-specific effects on processing efficiency of caused motion events in Cantonese–English–Japanese speakers. Bilingualism: Language and Cognition, 24(4), 730-745. https://doi.org/10.1017/S1366728921000018

56. Yum, J.O., 1988. The impact of Confucianism on interpersonal relationships and communication patterns in East Asia. Communications Monographs, 55(4), 374-388. https://doi.org/10.1080/03637758809376178