Does voice anthropomorphism affect lexical alignment in speech-based human-computer dialogue?

Benjamin R. Cowan¹, Holly P. Branigan²

¹ HCI Centre, University of Birmingham, United Kingdom
² SILS, University College Dublin, Ireland

benjamin.cowan@ucd.ie, holly.branigan@ed.ac.uk

Abstract

A common observation in dialogue research is that people tend to entrain, or align, linguistically with their interlocutors. This phenomenon offers a potentially important way to shape user behavior in human-computer dialogue interactions but little is known about the mechanisms that underlie it and how they may be affected by interlocutor design. We report a Wizard of Oz study that explored how voice anthropomorphism impacts lexical alignment in speech-based human-computer dialogue. In a referential communication task, speakers showed a very strong tendency to align lexical choices with their interlocutors, whether human or computer, but this tendency was not affected by voice anthropomorphism. These results highlight the robustness of lexical alignment effects in speech based human-computer dialogues, and suggests that this effect may be impervious to at least some design cues. They also suggest that automatic priming may be an influential mechanism in explaining why we align lexically with automated dialogue partners.

Index Terms: Lexical Alignment, Entrainment, Anthropomorphism, Speech Interaction, Human-Computer Dialogue.

1. Introduction

Many studies on human-human (HHD) and human-computer dialogue (HCD) have shown that people entrain, or align, their non-linguistic (e.g. posture and gestures [1], [2]), as well their linguistic (semantic, prosodic, lexical, and syntactic [3]–[10]) behaviors in such interactions. In HCD, recent research has suggested that levels of lexical alignment are impacted by people’s judgments of a computer partner’s communicative ability when given relatively salient and explicit cues (see below). The current study extends this work by exploring whether real-world interlocutor design cues that have been shown in previous studies to potentially impact a user’s judgment of a computer as a speech-based interlocutor (here, voice anthropomorphism) affect lexical alignment in HCD in a similar way. Evidence that such cues affect levels of alignment would highlight the importance of the design of the system interface when aiming to maximize the likelihood of system-user alignment. This in turn would facilitate the creation of systems that exploit alignment for implicitly guiding users to provide speech that has desirable characteristics for the automated system. In addition to these practical implications, any findings that an interlocutor’s (perceived) identity and communicative ability affects alignment would add further support to the notion that our perceptions are important in guiding our alignment behavior. Conversely if effects of interlocutor type are not found it would be tentative evidence that interlocutor type and interlocutor design play a lesser role in eliciting alignment, adding support to more automatic priming accounts of alignment in dialogue.

1.1 Alignment in Human-Computer Dialogue

In dialogue, speakers align prosodically and acoustically with computer interlocutors [11], [10], [12] as well as aligning at the lexical [9], [13] and syntactic [8], [14] levels. Lexical alignment in particular has been shown to occur in both experimental [9] and more naturalistic contexts [15].

Recent research [9] has suggested that superficial cues to a computer partner’s abilities influence levels of lexical alignment in dialogue. Participants took part in a picture-naming and -matching task (similar to that used in this research) with an interlocutor that they believed to be either a human or a computer (in fact it was always a computer producing pre-scripted responses). In both the human and computer conditions, participants tended to name objects using the same name as their interlocutor had used (e.g., calling an object a seat [vs. a bench] when their partner had previously used seat to describe the same image). This alignment occurred in both text-based and spoken interactions, and was robust over a number of turns since the participant had heard the initial description. Crucially to this work, alignment was stronger when participants believed that they were interacting with a computer than a human, and was also impacted by users’ perceptions of the computer’s abilities. People who viewed a start-up screen with a 1987 copyright along with a fictitious review from a computer magazine stating its limited abilities (‘Basic’ computer) before the interaction showed stronger lexical alignment than those who viewed a start-up screen with a current year copyright and review stating the system’s sophistication (‘Advanced’ computer). These findings suggest not only that people may display different magnitudes of lexical alignment in HCD compared to HHD, but also that aspects that influence users’ judgments of their interlocutor’s abilities as a conversational partner could potentially affect alignment levels [9]. This supports the notion that alignment may be mediated by speakers’ beliefs about their dialogue partners [4], [9], [16]. This is analogous to the concept of audience design [17] where speakers form beliefs based on assumptions about the knowledge that they presume members of particular communities are likely to have [16], [18], as well as assessment of their interlocutors’ likely understanding based on previous language use in the interaction [9].

Previous work on partner modeling effects in lexical alignment [9] have focused on shaping partner perceptions.
through reviews and other system-external cues. This paper explores whether the design of the interlocutor itself, specifically voice anthropomorphism, affects lexical alignment in a similar way. Studies have shown that people view agents that are more anthropomorphized as intelligent, capable [19] and more lifelike [20]. Computer partners using anthropomorphized voices are seen as more advanced, flexible and competent than those using less anthropomorphemic voices [14]. Hence computer interlocutors that are seen as more anthropomorphic may be judged to be more communicatively able, which may significantly affect user’s language use in interaction, and – by extension from previous research – decrease their likelihood of using the same lexical choices.

If interlocutor design impacts lexical alignment with the computer, this would suggest that relatively superficial design cues could affect users’ alignment levels, as well as highlighting one concrete way of maximizing the likelihood of alignment. Importantly the work also has implications for theoretical understanding of the mechanisms that guide alignment in HCD. Any demonstration of increased lexical alignment with computers compared to humans, and an impact of anthropomorphism on alignment, would add further support to the notion that our perceptions of our interlocutors are an important factor guiding alignment behavior [4]. Conversely if no effect of interlocutor type or anthropomorphism is found, this would provide tentative support for an account whereby alignment in HCD is influenced less by perceptions of our interlocutor and more by automatic increased activation of representations within the language system following comprehension leading to them being more likely used in production when appropriate [21].

We hypothesize that there will be a significant alignment effect, whereby participants will be more likely to name an object in the same way as their partner. Based on previous HCD research we hypothesize that this tendency will be higher in human-computer conditions than in human partner conditions. We further hypothesize significantly higher alignment with a less anthropomorphic than a more anthropomorphic voiced computer partner.

2. Method

2.1. Participants

72 participants from the University of Birmingham community (38 Male, 34 Female, mean age 20.5 years, S.D. = 3.02 yrs) were recruited through email and a posting on the University portal. All were native English speakers. They received £7.

2.2. Conditions

Participants interacted with a female or male version of one of three interlocutors; either a Human, a Robotic Computer or an Anthropomorphic Computer (between-participants design). Human condition: participants interacted with a co-present human partner (a confederate). This was included as a control condition to allow comparison between human and human-computer dialogue conditions. Robotic Computer condition: participants interacted with a computer that projected a robotic computer voice. The partner was simulated using audio recordings of names for each object in the game using the voice options Fred (for male) and Kathy (for female) from the Vox Machina (v.1.1) text-to-speech software. Anthropomorphic Computer condition: participants played the communication game with a computer that projected an anthropomorphemic voice. The partner was simulated using audio recordings using the voice options Nick (for male) and Nina (for female) from the Festival text-to-speech system. In previous research a computer using the anthropomorphemic voice was perceived as more advanced, competent, and flexible than if it used the robotic voice, with voice gender having no effect on perceptions [14]. For the computer conditions we simulated the dialogue partner using a Wizard of Oz paradigm. An experimenter listened in to the session through Skype and played audio recordings of item descriptions on the computer in the lab remotely using Windows Remote Assistance. We also manipulated the picture names used by the partner (favored vs. disfavored Prime conditions); see below.

2.3. Materials

2.3.1. Communication Game

Participants and their partner took turns to name images of objects displayed on-screen (naming turn) and to choose an image that matched their partner’s description from an array of two images (matching turn). The procedure used is similar to that used in previous lexical alignment research [9].

2.3.2. Experimental Items

Within the game participants named 18 experimental pictures (see [9] for picture list). Each of these experimental pictures were chosen because the objects displayed could be described by the partners using a favored name (used spontaneously to name that picture more than 80% of the time) or a disfavored name (rated equally acceptable as a name for that picture, but used spontaneously less than 20% of the time; see [9] for details). An experimental item included a prime trial (where participants heard the partner naming the object in the experimental picture using either a favored or disfavored name and had to choose a matching picture out of two displayed) and a target trial (where participants were presented with the experimental picture used in the previous prime trial and were asked to name it for their partner to match- the experimental picture was highlighted by a red box; a distractor picture was also displayed alongside the experimental picture). On half the prime trials in the game the partner used a disfavored name to describe the experimental picture and the other half a favored name, randomized across the game. Alignment occurred if the participant used the same name for the object in the target trial as their partner had used to describe it in the prime trial.

2.3.3. Filler Items

The game additionally included 57 filler items to distract attention from the experimental items. For each filler item the participant saw a pair of filler pictures and selected the picture that the partner had named (matching turn) and then saw a pair of filler pictures and named the picture highlighted (naming turn). All filler objects had one acceptable name.

2.3.4. Game Structure

The game consisted of 93 pairs of matching-naming turns. Each experimental item involved two pairs of these turns. In pair 1, participants completed a prime trial (e.g. heard the disfavored prime seat from the partner and then matched an experimental picture of a bench), then in the naming turn named a filler picture. In pair 2, the participant matched a filler
picture, then produced a target utterance (target trial) that named the same experiment picture as in pair 1 (e.g. seat to name a picture of a bench). There was therefore always one filler naming turn and one filler matching turn between hearing the prime and the target trial. Three filler items intervened before the next experimental item. Further details and illustration of the procedure can be found in [9].

2.4. Procedure

The experimenter greeted the participant and then left to set up the lab, telling the participant that they were leaving to get their partner ready. The participant was then taken to the lab and asked to take a seat on one side of a table; they could see their partner (human confederate or a computer) on the other side. A screen was then placed in the middle of the table to preclude the use of non-verbal signals. The participant completed a demographic questionnaire gathering data about gender, age and whether the participant was a native English speaker. Non-native speakers were excluded from further participation.

After completing the questionnaire, the participant (and the human confederate in the human condition) was given verbal instructions. They were told that they would take turns being the matcher and describer in a picture description game. When acting as matcher, they should listen to their partner’s description and click on the matching picture, then click the ‘next’ button. When acting as describer, they should describe the image highlighted by a red box, then click the ‘next’ button to see a screen informing them of their partner’s choice. Their partner always chose the correct picture so that partners’ agreement was not the outcome of random or accidental choice. Their partner always chose the correct picture so that partners’ agreement was not the outcome of random or accidental choice. Their partner always chose the correct picture so that partners’ agreement was not the outcome of random or accidental choice.

During the experiment, the experimenter noted the participant’s descriptions in the target trials (favored or disfavored). Utterances were coded as Other if they did not use either the favored or disfavored name, and were excluded from analysis. This data is the binary outcome variable Response. The sessions were also audio-recorded to confirm coding where necessary. Participants were debriefed about the aims of the study at the end of the session.

3. Results

There were 1296 descriptions, 1005 favored and 251 disfavored. There were 40 (3.09%) utterances classed as Other, where participants used neither the favored or disfavored name for the picture. Table 1 shows the proportions and frequencies of disfavored names from the total number of favored and disfavored names by condition. To be consistent with previous alignment research [9], alignment in this table is calculated by subtracting the proportion of disfavored names used in the disfavored prime conditions from those used in the favored prime conditions.

Table 1. Proportions and frequencies of disfavored descriptions in favored and disfavored prime conditions by interlocutor

<table>
<thead>
<tr>
<th>Interloc.</th>
<th>N</th>
<th>Favored Prime</th>
<th>Disfavored Prime</th>
<th>Alignment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>24</td>
<td>.014 (3/211)</td>
<td>.35 (72/205)</td>
<td>.34</td>
</tr>
<tr>
<td>Robotic</td>
<td>24</td>
<td>.014 (3/214)</td>
<td>.43 (120/209)</td>
<td>.41</td>
</tr>
<tr>
<td>Anthrop.</td>
<td>24</td>
<td>.028 (6/214)</td>
<td>.38 (78/203)</td>
<td>.36</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>.019 (12/639)</td>
<td>.39 (239/617)</td>
<td>.37</td>
</tr>
</tbody>
</table>

We analyzed the data using logit mixed effects analysis with the lme4 package in R v.2.15.3 [22]. Response was used as the outcome variable, releveled to analyze the likelihood of participants producing a disfavored name. The Interlocutor fixed effect was releveled so that Human was used as the baseline category for comparison. The Prime fixed effect was also releveled so that the favored prime condition was used as the baseline category. The initial model included by-subject random slopes for Prime, and by-item random slopes for Prime, Interlocutor and interaction. We removed the interaction by-item random slope to facilitate convergence in the final model (see [23]). The summary of fixed and random effects for the model as well as the model syntax is shown in Table 2.

Table 2. Summary of fixed and random effects for logit mixed effects analysis

<table>
<thead>
<tr>
<th>Model: Response ~ Prime*Condition + Prime:Condition + (Prime/participant) + (Prime/item) + (Interlocutor/item)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Prime-Disfav</td>
</tr>
<tr>
<td>Interloc-Anthrop</td>
</tr>
<tr>
<td>Interloc-Robotic</td>
</tr>
<tr>
<td>Prime-Disfav*Interloc-Anthrop</td>
</tr>
<tr>
<td>Prime-Disfav*Interloc-Robotic</td>
</tr>
</tbody>
</table>

Table 2. Summary of fixed and random effects for logit mixed effects analysis

There was a significant effect of Prime on the likelihood of naming the target object using a disfavored name. That is, people were more likely to use a disfavored name for an object if they had heard their partner naming the same object using the disfavored name compared to when they had heard them.
using a favored name ($z = 2.80, p = .005$). Hence there was reliable lexical alignment. However there was no significant interaction of prime and interlocutor: Interlocutor identity (human, robotic computer or anthropomorphic computer) did not affect this lexical alignment effect.

4. Discussion

The experiment showed a significant lexical alignment effect. When interacting with a human or computer partner, participants tended to use the same name that they had previously heard their partner use to name that object. When a partner used a strongly disfavored name for that object the partner was more likely to use that same disfavored name. The results support a growing body of research that suggests that lexical alignment is not a phenomenon unique to HHD but also occurs in speech based HCD, guiding user’s word choices. However, there was no significant effect of interlocutor type. Participants aligned to a similar extent with a computer (irrespective of voice anthropomorphism) as with a human.

We therefore found no evidence that people’s judgments of their interlocutor’s abilities (i.e. interlocutor models) impacted levels of lexical alignment, in contrast to previous work [9]. A previous study using the voices established that they successfully induce people to draw different inferences (see [14]): Participants rated a computer using the robotic voice as significantly more basic, inflexible, less capable compared to a computer using a more anthropomorphic voice.

Yet these different inferences, and hence different interlocutor models, did not affect people’s tendency to lexically align. This pattern is consistent with other recent research that found no effect of perceived competence and likeability on lexical alignment in HCD [24].

Our results have important implications for dialogue systems design. Showing that people hold a robust propensity to align with their interlocutors’ choice of vocabulary during speech-based interactions offers attractive potential for implicitly guiding user behavior, making user speech more predictable for system processing (potentially leading to fewer recognition errors [25]) and potentially increasing the likelihood of successful communication [21]. Our research supports previous studies suggesting alignment is present and predictable for system processing (potentially leading to fewer recognition errors [25]) and potentially increasing the likelihood of successful communication [21]. Our research supports previous studies suggesting alignment is present and strong in speech based HCD, reinforcing the notion that alignment can be leveraged to shape users’ lexical choices.

Importantly, this work also contributes to the theoretical debate over mechanisms determining language use in HCD. The growing popularity of speech as an interface modality in a whole host of interactive products, as well as in embodied conversational and robotic agent interactions, makes it imperative to develop a theoretical understanding of the mechanisms that underpin alignment in HCD. Our results suggest that alignment is a potentially powerful determinant of language use in spoken interactions, including HCD, but is not uniformly affected by calculations about the competence of an interlocutor. A strong tendency to align with interlocutors irrespective of their identity is consistent with accounts stressing the role of automatic priming mechanisms [21], [26], [27]. These accounts propose that using a particular linguistic representation during comprehension facilitates its access in subsequent processing and thus promotes its reuse. For instance, hearing the word ‘hatchet’ (to label a picture of an axe) activates the associated lexical entry in the hearer’s mental lexicon. As this activation does not decay immediately, this lexical entry is more easily retrieved in subsequent language use and hence likely to be used to describe the same picture subsequently. Importantly this mechanism is resource-free and does not require interlocutor modeling [21], [27] unless strategically necessary (e.g. for error correction). Such automatic priming-based alignment of linguistic representations is hypothesized to lead to alignment of situation models, and hence mutual understanding, without any need in most circumstances for explicit modeling of an interlocutor’s knowledge state. Our findings are consistent with such priming-based mechanisms.

We emphasize however that such mechanisms may exist alongside mechanisms that are sensitive to beliefs about interlocutors’ knowledge states [9]. Our results suggest that such a mediated component to lexical alignment does not necessarily manifest itself in all contexts or is inevitably influenced by design cues such as voice. Indeed the findings suggest that, where previous research has emphasized the important role of accommodation on aspects such as hyperarticulation and syntax [25], [28]–[32], interlocutor modeling and accommodation may not be the only driver of user language choices in HCD, specifically in lexical choice.

It would however be premature to conclude that beliefs could never give rise to differences in alignment, or indeed linguistic choices in general. Alignment may well have both unmediated and mediated components that are dependent on contexts and dimensions of language [33]. For example, in contexts in which mutual understanding is of key importance (e.g., safety-critical situations), where breakdown in communication occurs, or where the task is more complex, beliefs about what the interlocutor will understand correctly may play a strong role in alignment. Indeed in previous research [9] the computer’s abilities were highly salient from the cues that users were given. In the present study, the cues for partner modeling - although representative of cues experienced in interaction - are subtle, and rely on participants’ drawing relevant inferences. The difference between experiments in salience of the computer’s abilities, and thus the detailed specification of the interlocutor model, could potentially account for the discrepancy in findings.

It would however be premature to conclude that beliefs could never give rise to differences in alignment, or indeed linguistic choices in general. Alignment may well have both unmediated and mediated components that are dependent on contexts and dimensions of language [33]. For example, in contexts in which mutual understanding is of key importance (e.g., safety-critical situations), where breakdown in communication occurs, or where the task is more complex, beliefs about what the interlocutor will understand correctly may play a strong role in alignment. Indeed in previous research [9] the computer’s abilities were highly salient from the cues that users were given. In the present study, the cues for partner modeling - although representative of cues experienced in interaction - are subtle, and rely on participants’ drawing relevant inferences. The difference between experiments in salience of the computer’s abilities, and thus the detailed specification of the interlocutor model, could potentially account for the discrepancy in findings.

Contexts that heighten the salience of interlocutor modeling are an important issue for future research.

5. Conclusion

The research presented found that speakers showed a strong tendency to align lexical choices with their dialogue partners, adding further support to the robustness of lexical alignment effects in computer dialogue interactions. Unlike previous research the tendency to align was not impacted by whether the partner was a human or a computer. Interlocutor design variables such as the anthropomorphism of a computer’s voice also did not significantly affect this tendency. As well as highlighting that lexical alignment seems impervious to some design cues and partner identity the findings also suggest that automatic priming may be an influential mechanism in understanding how we align lexically with automated dialogue partners.

6. Acknowledgements

This work was funded by the British Academy. We would like to thank Rajen Mistry and Asha Chauhan for their data gathering work and the reviewers for their comments.
7. References


