Lexical Alignment in Collaborative Interactions with a Conversational Agent

PAUL BLUM, University of Twente, The Netherlands

Lexical alignment is a mechanism in which participants of a conversation adapt to one another's speech by copying the other person's choice of words. This forms an important part of human interactions, as it allows for better communicative success. In light of existing evidence that alignment between humans improves collaboration and that humans align to computers, this research implements the idea of lexical alignment into a chatbot and measures its effects on collaboration with a human in terms of task performance, perceived workload and perceived fluency. An experiment was conducted in which participants were tasked to solve a collaborative map game together with a chatbot. Half of the participants were presented with a version of the chatbot that lexically aligned to them and the other half were presented a version that misaligned. The results suggest that alignment significantly improves task success and the user's perception of how much the chatbot contributes to a fluent interaction. While we did not capture a significant effect of alignment on perceived workload, we report additional insights on intelligence and companionship qualities of the chatbot.

1 INTRODUCTION

Chatbots are a type of conversational agent that interact with human users in natural language via text-based messaging interfaces [21]. Initially, they were built to entertain and mimic human conversation [21], but they have since been extended to assist users in many domains such as education, e-commerce, healthcare, finance, marketing, and business [2]. Task-oriented chatbots are aimed at assisting their users in performing a domain-specific task and have gained popularity in areas like providing customer support [27].

More than a million task-oriented chatbots have been implemented into web applications and social media platforms since 2015 [8], and the market is expected to expand aggressively until 2030 [1]. However, chatbots fail to meet expectations in terms of language skills and social behaviour causing frustration of not being understood [5, 19]. To address this issue, it is crucial to gain a deeper understanding of the language that chatbots should employ because natural language is the primary way in which the chatbots interact with the user.

Many researchers in the field of psycho-linguistics have studied the process by which two individuals involved in a conversation, known as interlocutors, can effectively establish shared comprehension and avoid breakdowns in communication. A theory that describes how humans achieve such high levels of communicative success is the Interactive alignment model [18]. According to this theory, human interlocutors naturally adapt to each other's use of language throughout a conversation as part of a natural process called linguistic alignment. They achieve a mutual understanding by copying word choices, sentence structure and style [18]. In this paper, we focus on a subset of linguistic alignment, called lexical alignment, where interlocutors copy each other's lexical items, i.e. words.

Research has been done on the role of alignment in human collaboration and studies have shown that higher alignment leads to higher task success [6, 17, 20] as well as lower perceived workload [25]. There is strong evidence that humans also align to computers and it has been found that this alignment is actually stronger than in human-human interactions [3, 22]. For example, it has been observed that computers can override significantly ingrained linguistic inclinations, prompting individuals to use a specific term for an item which they would typically refrain from using more than 85% of the time [3]. Given the evidence that collaboration is improved through alignment and that humans align to computers, it makes sense to investigate what effect alignment from a chatbot can have on the collaboration with a user. More specifically, how does lexical alignment from a chatbot during a collaborative task affect the user's task performance, perceived workload, and perceived fluency of the interaction?

To address this question, this paper presents related work from which we draw three research hypotheses. Then, we present the setup and results of a user experiment in which participants solved a collaborative map game together with a chatbot. Lastly, we seek to explain the results and understand their implications for lexical alignment in human-chatbot collaboration.

2 RELATED WORKS

2.1 Alignment Theory

Pickering and Garrod [18] argue that interlocutors understand each other when they align their model of the situation under discussion. This alignment predominantly stems from aligning at various levels of linguistic representation. For example, in a study by Garrod and Anderson [7], where pairs of participants played a cooperative maze game, the speakers converged on a common way of semantically representing the maze. More generally, Garrod and Anderson propose a principle called output/input utterance that they hypothesize to be the basis for successful communication in coordinated activities. It is the idea that formulating your utterances (outputs) according to the same principles of interpretation as your conversation partner (inputs) will lead to a mutually satisfactory description scheme with minimum collaborative effort. When both speakers conform to this principle, it "minimizes the joint pool of resources" needed to formulate and interpret the utterances. In other words, their workload is lower. While Garrod and Anderson's experiment focused on the semantic and pragmatic choices in dialogue, they also believe that it holds for lexical representations.

TScIT 39, July 7, 2023, Enschede, The Netherlands
© 2023 University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

2.2 Alignment in Human-Human Interactions

The role of alignment in human collaboration has been examined in a number of studies. Reitter and Moore [20] have analyzed the HCRC Map Task corpus [26], a set of 128 dialogues where two people solve a task-oriented cooperative exercise. In each pair, one of the speakers, known as the instruction giver, is given a map that has a path drawn on it, while the other speaker, referred to as the instruction follower, has a version of the map with no path. Without seeing each other's maps, their objective is to replicate the instruction giver's route on the instruction follower's map. The maps are not identical, and the participants are told this explicitly. Reitter and Moore [20] found that higher syntactic alignment leads to higher task success.

Further, Fusaroli et al. [6] conducted an experiment in which pairs of participants had to agree on which visual stimuli on a screen was the oddball. The results showed that lexical alignment on task-relevant vocabularies strongly correlates with collective task performance. Thomas et al. [25] have examined a corpus where in each pair, one person was assigned a series of information-seeking tasks without access to the internet while the other person had access to the internet. They found that when the pairs were more strongly aligned in terms of linguistic style, the perceived workload was lower. A study on two task-oriented corpora examined lexical alignment with respect to the most frequent words used [17]. They found that alignment is not only predictive of task success but also that the perceived naturalness and flow of the interaction was higher. This could be an indicator that the perceived fluency of the interaction is higher.

2.3 Alignment in Human-Computer Interactions

A number of studies have investigated the effect that alignment by a computer can have on interactions with humans. Among the studies that have analyzed the relationship between alignment and task success, Spillner and Wenig [23] tasked participants to find a movie that they would like to watch with the help of a chatbot. In the versions of the chatbot that aligned, they observed higher task success. Additionally, they found that perceived workload was lower and user engagement was higher when the chatbot aligned. Another study has looked at the importance of lexical alignment in troubleshooting dialogues by using a statistical user simulation model [13]. They found that lexical alignment in referring expressions is an important factor for the user's task success and decreases dialogue turns which indicates more efficient collaboration. Another study had pairs of participants solve a task based on pedestrian navigation in a simulated town [14]. In this Wizard of Oz setup, one person pretending to be a robot described their position and perspective to a user who had a full view over the environment and knew the destination. The task was for the user to navigate the 'robot' to different locations on the map through a text-based interface. The results showed that users perceived the interaction as less successful when alignment was weaker.

Among the studies that did not involve a task-based chatbot, Srivastava et al. [24] have observed the role of lexical alignment on human understanding of explanations. In this study, a chatbot was designed to explain the causes and effects of lung cancer. In a follow-up quiz, participants who interacted with the version of the chatbot that aligned, scored higher, indicating that lexical alignment is a predictor of human understanding. Chaves et al. [4] presented users with two versions of a response from a tourist assistant. Both versions were equivalent in informational content, but one version was modified to introduce linguistic features that are less likely to be appropriate to the situation. That is, the responses were adapted to the style of another corpus, which is on average more personal and oral with elements of persuasion and formality [4]. The participants were told that the tourist assistant was a chatbot. The results showed that linguistic features are a strong predictor of the user's perception of a chatbot's appropriateness, credibility and overall user experience. This suggests that a chatbot's choice of lexical items will impact how the users perceive it. Lastly, while not related to linguistic alignment, Hoffman and Breazeal [12] found that users' sense of fluency is higher when they interact with a robot that anticipates their actions.

3 HYPOTHESES

Based on the theory and evidence from related works, the hypotheses for this research are as follows.

In interactions with a chatbot that aligns lexically we expect

(1) higher task performance,
(2) lower perceived workload,
(3) higher perceived fluency.

4 METHODOLOGY

An empirical between-subject investigation was conducted wherein participants solved a task with the help of a chatbot on a designated web page. Half of the participants were presented with a version of the chatbot that employed a lexical alignment strategy and the other half was presented with a version that misaligned.

4.1 Task Design

The nature of the task was derived from the HCRC Map Task [26] setup. The human participants took the role of the instruction follower, i.e. they had to draw a specific path on a map of landmarks. The chatbot took the role of the instruction giver, i.e. it had knowledge over what that specific path was. They had to work together to reproduce the path on the participant's map. Figure 1 shows the map that all participants were presented with. The red dots served as pivot points that the participants had to use in order to draw the path. They did this by clicking on a red dot and a straight line was drawn for them from the last pivot point to that red dot. They were instructed to try to complete the task as fast as possible and that there was a time limit of ten minutes.

We conducted a small pilot study with 6 participants to gather feedback on the experiment setup and change the difficulty if needed. While most participants successfully interacted with the map and the chatbot, some participants did not understand what was asked of them. Therefore, we adjusted the task description with their feedback and rephrased the chatbot's messages to include clearer calls to action (e.g. changing from "Go around the stone slabs" to "Draw the path around the stone slabs").

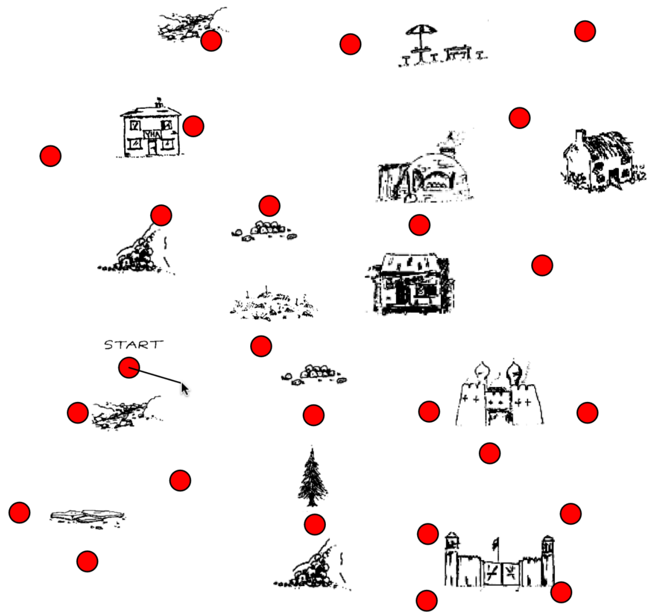

Fig. 1. Map task of the experiment

Because the task itself was found to be too easy for those who understood what to do, it was made slightly more challenging by introducing more ambiguity. This was done by grouping more similar themed landmarks close to each other on the map. For example, in the final version there is a group of features that could be classified as rocks, as well as a group of forts and houses respectively. Additionally, a time limit of ten minutes was deemed as an appropriate balance between time pressure and still being able to solve the task.

4.2 Chatbot Design

For the purpose of the experiment, a simple task-oriented chatbot has been developed. The way in which users interact with the chatbot is through different reply options provided to them. For example, in response to the chatbot asking where the starting point is, the user has two reply options:

(1) "I am just [above/north of] a(n) ______"
(2) "I am confused."

The first reply option requires the user to choose between "above" and "north of" from a drop-down menu as well as to fill in the blank using a text field. This reply option is changed throughout the dialogue to fit as a reply to the chatbot's last message. The second reply option ("I am confused") is the same throughout the entire conversation, to which the chatbot responds by rephrasing its last instruction and hinting at reading the previous messages again. The chatbot follows conditional rules based on the user's reply choice to generate its next response.

There are two versions of the chatbot: one that aligns and one that misaligns. The aligning version of the chatbot is designed to imitate the output/input principle described in Section 2.1, whereas the misaligning version is designed to violate that principle. To illustrate the differences in the two versions of the chatbot, consider the following two excerpts from dialogues that participants had with the chatbot.

Aligning condition

USER: I am just on north of a(n) valley.

CHATBOT: Okay to start, draw a path that goes to the west of the valley.

USER: Okay done.

CHATBOT: Okay. Next, do you see an obstacle south of you?

USER: Yes, I see a(n) platform.

CHATBOT: Draw the path around the platform by going west of it and then back up in the east.

Misaligning Condition

USER: I am just on north of a(n) river.

CHATBOT: Okay to start, draw a path that goes to the left of the stone creek.

USER: Okay done.

CHATBOT: Okay. Next, do you see an obstacle below you?

USER: Yes, I see a(n) puddle.

CHATBOT: Draw the path around the stone slab by going left of it and then back up on the right.

In the aligning version, the chatbot copies the referring expressions for the landmarks from the user. For example, the user refers to one of the landmarks as a 'valley' and the chatbot uses that term to say 'go west of the valley'. In the misaligning version, the chatbot ignores the referring expressions from the user and uses predetermined expressions instead. For example, the user says they are north of a 'river' but the chatbot ignores that term and says 'stone creek' instead. Additionally, in the aligning version, the chatbot matches the user's preference in cardinal directions ('north', 'east', 'south', 'west') or egocentric directions ('above', 'right', 'below', 'left') throughout the conversation while the misaligning version uses the opposite terms.

4.3 Measuring Task Performance

In order to asses task performance, we consider two measures. First, we measure task success which for this experiment is how accurately the participant draws the path. More precisely, assuming that all pivots are assigned a unique letter, the participant's sequence of pivots is compared to the correct sequence of pivots and the Levenshtein distance [15] is taken. The lower the Levenshtein distance, the higher the task success.

Second, we analyze a combination of success and time taken to complete the task because participants were asked to complete the task as fast as possible. How the participants interpreted this instruction in terms of deciding whether to prioritize speed or accuracy is unknown. For that reason, in this study we use the balanced integration score (BIS) which has been shown to be relatively insensitive to speed-accuracy trade-offs and is well suited for between-subject designs [16]. BIS can be interpreted as a measure of how much above or below average a participant in a given condition performed when compared to the entire group. It is computed by first standardizing the time taken and accuracy (task success) to bring them on the same scale, and then subtracting one from the other [16].

4.4 Measuring Perceived Workload

In order to quantify perceived workload, we use the NASA Task Load Index (TLX) [9]. The original version of the TLX uses six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration) that participants rate on scales of 100 points with increments of five points. Additionally, there is a separate section where pairwise comparisons are used to determine the relative importance of each dimension. However, many researchers have removed the section of pairwise comparisons and calculate the TLX purely as the mean of subscale ratings [10]. For this study, we discarded physical demand as it is not applicable to the task.

4.5 Measuring Perceived Fluency

Fluency as described in [11] is when humans collaborate on a shared activity, their ability to reach a high level of coordination resulting in a "well-synchronized meshing of their actions". In this study, we measure perceived fluency using a selection of the subjective fluency metrics described in [11]. They were selected based on what was applicable to the interaction in this experiment and reworded to refer to a "chatbot" instead of a "robot":

- I worked fluently together with the chatbot.
- The chatbot contributed to the fluency of the interaction.
- The chatbot was trustworthy.
- The chatbot was intelligent.
- The chatbot was committed to the task.
- The chatbot and I understood each other.
- I was confident in the chatbot's ability to help me.
- The chatbot did not understand what I was trying to accomplish. (reverse scale)
- I found what I was doing with the chatbot confusing. (reverse scale)
- The chatbot was cooperative.

The fluency score for a given participant is the sum of values that they assign to these statements on 7-point Likert scales (1='Strongly disagree', 7='Strongly agree').

4.6 Participants

A total of 58 participants took part in the experiment but 8 were excluded because they did not interact with the chatbot, making their answers irrelevant in assessing the chatbot. This leaves a total of 50 participants, of which 25 interacted with the aligning chatbot and 25 with the misaligning chatbot.

5 RESULTS

In this analysis, it is assumed that the sample groups are independent because the experiment followed a between-subject design. Because we appreciate the importance of detecting effects in the opposite direction than expected, two-sided tests are conducted. That means, the null hypothesis is that no effect is present, and the alternative hypothesis is that an effect is present (positive or negative). In the interest of quantifying the effect, we additionally perform Bayesian analyses. A significance level of 0.05 is used in all tests.

5.1 Task Success

Figure 2 shows the distribution of task success for both conditions. As expected, participants who interacted with the aligning chatbot demonstrated higher success in completing the task, as indicated by a higher mean score (M=0.792 compared to M=0.683) and a higher median score (Mdn=0.875 compared to Mdn=0.708). In order to test if this difference is significant, a Mann-Whitney U test is conducted because the assumption of normality is not met (Schapiro Wilk's test on the aligning condition returns p=0.018). The Mann-Whitney U test shows a significant difference in task success between the aligning condition and the misaligning condition (U=420, p=0.036). Furthermore, a Bayesian analysis shows that there is more than twice as much evidence for alignment affecting the user's task success relative to there being no effect (BF_{10}=2.195).

In the distribution of accuracy for the aligning condition, there appears to be a division into two groups (see Figure 2). One group scored 70% accuracy or below whereas the other group scored 83% or above. One apparent trend is that 9 out of the 11 people who made a mistake in the high performing group managed to recover from mistakes while only 1 out of the 11 participants in the low performing group managed to do so.

In Figure 4 we have graphically overlaid all paths that participants in each condition drew. A darker shade represents that more participants took a given connection between two dots and a lighter shade indicates that few people took that connection. The red dotted line is the correct path and we have labeled each dot with a unique letter. This was not visible to the participants during the experiment. The graphic visually confirms the statistical findings: the deviation from the correct path is greater in the misaligning condition than in the aligning condition.

Furthermore, we can understand at which locations on the map, many participants took wrong turns. Specifically, consider dividing the task into six logical stages, each corresponding to one instruction from the chatbot. For example, the instruction "draw a path that goes to the left of the stone creek" (as produced by the misaligning chatbot) should lead to the path that connects dots A and B. Figure 3 shows the number of participants in each condition who successfully managed to complete a given stage. We see that in all stages except stage 2, the number of successful participants is greater in the aligning condition compared to the misaligning condition. If we consider how the number of successful participants changed throughout the task, different patterns emerge for each condition. In the aligning condition, 19 participants start off with success in stage 1 immediately followed by the greatest drop in success to only 10 in stage 2. The number of successful participants then stays relatively constant with an average of 10.2 throughout stages 2 to 6. In the misaligning condition, 15 participants start off successfully but the biggest drop in success is from stage 2 (11 successful participants) to stage 3 (3 successful participants). While the success is low throughout stages 3 to 5 with an average of 3.3, we observe a great increase from stage 5 (3 successful participants) to stage 6 (9 successful participants). In fact, in the misaligning condition, 36% of the participants who made a mistake in at least one of the first 5 stages, manage to successfully complete the final stage whereas in the aligning condition, this percentage is only 24%.

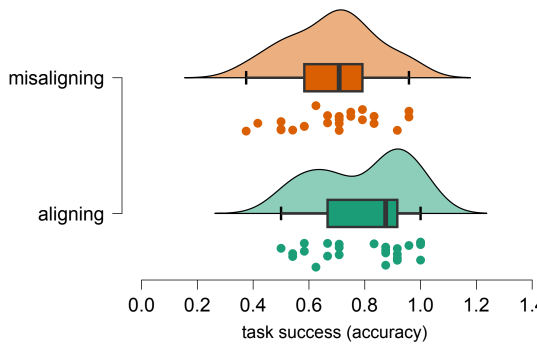

Fig. 2. Distribution of task success by alignment condition

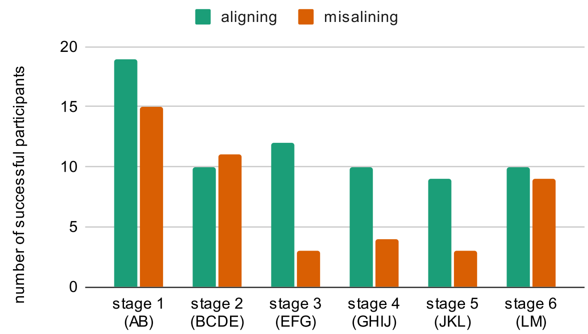

Fig. 3. Success by stage of the task

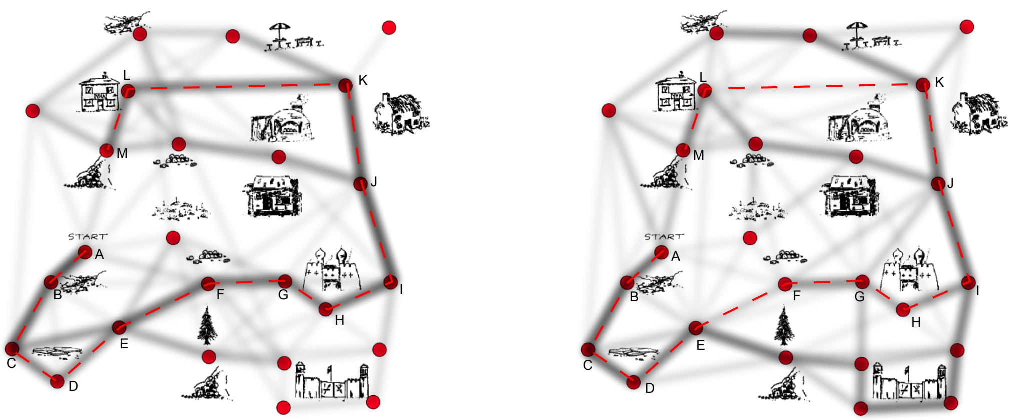

Fig. 4. Sum of paths for aligning condition (left) and misaligning condition (right)

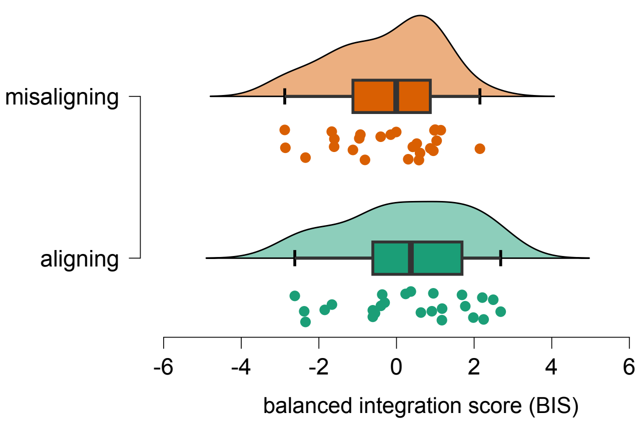

Fig. 5. Distribution of BIS by alignment condition

5.2 Balanced Integration Score

Figure 5 shows the distribution of task performance as measured by BIS. Participants who used the aligning chatbot performed higher in terms of both mean (M=0.272 compared to M=-0.272) and median (Mdn=0.370 compared to Mdn=-0.007). Note that as expected, the means add to exactly 0 due to the standardized nature of BIS [16]. Because normality and equality of variances is given, we perform a t-test which shows that the effect of alignment on BIS is non-significant (t=1.296, p=0.201). In fact, a Bayesian analysis reveals that there is 1.784 times more evidence for there being no effect of alignment on BIS.

Additionally, if we consider time taken alone, a Mann-Whitney U test reveals no significant difference (U=337, p=0.641) and a Bayesian analysis shows that there is almost three times more relative evidence for the null hypothesis (BF_{01}=2.998).

5.3 Perceived Workload

In Figure 6, we see the distribution of perceived workload for each condition. TLX is lower than in the misaligning condition (M=48.24 compared to M=54.96) and so is the median (Mdn=47 compared to Mdn=54). Since we cannot assume normality (Shapiro-Wilk's test on the misaligning condition returns p=0.034), we perform a Mann-Whitney U test. The Mann-Whitney U test returns an non-significant p-value of 0.193 so that we fail to reject the null hypothesis (U=245). In fact, there is more than 1.5 times more evidence for there being no effect according to the Bayesian analysis (BF_{01}=1.561).

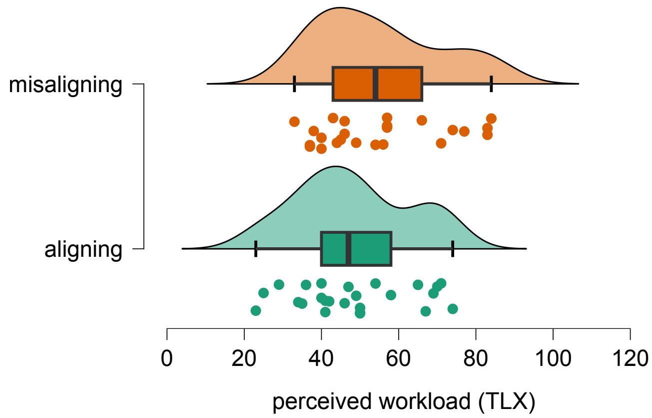

Fig. 6. Distribution of perceived workload by alignment condition

Fig. 7. Subscales of perceived workload by alignment condition

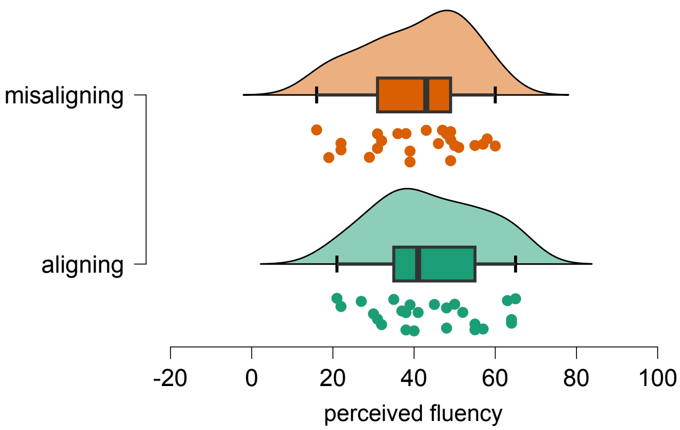

Fig. 8. Distribution of perceived fluency by alignment condition

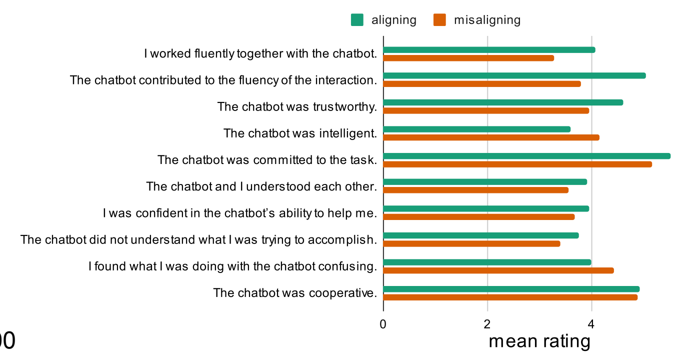

Fig. 9. Subscales of perceived fluency by alignment condition

We additionally analyze the mean subscale ratings separately as depicted in Figure 7. Participants in the misalinging condition felt that the task was more hurried or rushed and that they had to work harder to perform their level of performance. Additionally, they felt more insecure, discouraged, irritated, stressed or annoyed. When looking at how participants evaluated their success in accomplishing the task, we see that participants in the aligning condition felt like they performed better compared to the misaligning condition. Between the two conditions, there is no mean difference in how mentally demanding participants found the task to be. Therefore, all mean subscale ratings show that participants with the aligning version of the chatbot experienced lower workload except for mental demand for which the means are equal. None of the differences in means was found to be significant.

In order to understand the effect that lexical alignment has on how well users can self-evaluate their task success, we will now look at how participants in both conditions ranked the question "How successful were you in accomplishing what you were asked to do?" relative to how well they actually performed. We compute the ratio between confidence and success which will be higher than 100 if the participant thought they performed better than they actually did and lower than 100 if the participant thought they performed worse than they actually did. While not statistically significant, we find that participants in the aligning condition are more confident relative to their actual success (M=57.8, Mdn=60) compared to the misaligning condition (M=54.0, Mdn=56.8). Additionally, while in the misaligning condition there were no participants with a ratio higher than 100, we find that there were two participants in the aligning condition for who this is this case.

5.4 Perceived Fluency

On average, participants perceived the interaction with the aligning chatbot as more fluent (M=43.88) compared to the misaligning chatbot (M=40.64). Opposed to that, the median is lower for the aligning condition (Mdn=41) compared to the misaligning condition (Mdn=43). Since normality and equality of variances is given, we conduct a t-test. The result shows that there is no significant difference in overall perceived fluency between the conditions (t=0.882, p=0.382). In fact, there is more evidence for no effect being present compared to there being an effect (BF_{01} = 2.571).

Additionally, we analyze how participants rated each of the sub-questions separately as depicted in Figure 9. There is one statistically significant difference between the two alignment conditions which is in response to the statement "The chatbot contributed to the fluency of the interaction" (U=442, p=0.01). The other sub-questions show no significant differences, but the mean on eight out of the ten questions are in support of a positive effect of alignment on perceived fluency. That is, participants in the aligning condition felt more like they worked fluently together with the chatbot and that the chatbot contributed to that fluency. They also rated qualities such as trustworthiness, commitment to the task and cooperativeness higher than participants in the misaligning condition. They indicated that they understood each other with the chatbot more. They also were more confident in the chatbot's ability to help them and found what they were doing with the chatbot less confusing. The questions where we observe an effect opposite of what we hypothesized are about the intelligence of the chatbot and whether the chatbot understood what the user was trying to accomplish.

6 DISCUSSION

The results of this experiment confirm the findings of related works (see Section 2) in that lexical alignment improves task success. We were unable to capture a significant effect of lexical alignment on perceived workload and only one of the measures for perceived fluency revealed significant differences. In this section, we will point out limitations of the experiment and find possible explanations for the results. The aim is to understand the implications that this study has for lexical alignment in chatbot design.

6.1 Limitations

Some limitations of the implementation of lexical alignment in this experiment were that at times the chatbot produced ambiguous or misleading instructions. From this we will draw considerations that one should take when implementing lexical alignment into a task-oriented chatbot.

6.1.1 Ambiguous Instructions. Because the aligning chatbot in this experiment always blindly copied the referring expressions of the users, it sometimes resulted in ambiguous instructions. For example, consider the following excerpt from the aligning condition:

CHATBOT: What is north of the tree?
USER: There is a(n) stones.
CHATBOT: And what is south of the tree?
USER: There is a(n) stones.
CHATBOT: Continue drawing the path through the gap between the tree and the stones.

Here, either the gap north or south of the "tree" could be meant, so that this instruction is highly ambiguous. Hence, when implementing lexical alignment into chatbots, one should make sure the chatbot is able to clarify such ambiguity before producing instructions.

6.1.2 Misleading Instructions. We saw that participants in the aligning condition were more confident in their success relative to their actual performance than participants in the misaligning condition. This implies that lexical alignment as implemented in this study can lead to a false sense of confidence. One reason for this could be that in certain situations, the chatbot produced misleading instructions. For example, consider this excerpt from a dialogue in the aligning condition:

CHATBOT: What obstacle did you encounter?
USER: I encountered a(n) valley.
CHATBOT: From the valley, draw a path to the nearest fortress and you have reached the goal!

This user had previously made mistakes in drawing the path, and so when asked what obstacle they just encountered, they describe a landmark that is at a completely different location on the map. This mismatch in what the user and the chatbot refer to results in the chatbot producing an instruction that is highly misleading. However, the user is left to believe that they must be on the correct path because the chatbot is giving instructions that relate to the landmarks near them. In contrast, consider this excerpt from the misaligning condition:

CHATBOT: What obstacle did you encounter?
USER: I encountered a(n) beach.
CHATBOT: From the youth club, draw a path to the nearest rock fall and you have reached the goal!

If this user was to be at a location on the map with no "youth club" close to them, they would notice that they likely made a mistake. This effect is likely also the reason for why participants in the misaligning condition were better able to recover from mistakes than in the aligning condition. The aligning chatbot simply produced instructions that were not relevant to the actual user's position in the task, but the user was left to believe that they are doing well. Additionally, this could be the reason for the division into a high and low performing group in the aligning condition. Failing to have a successful start into the task lead to an overall poor performance because these participants failed to recover from their mistakes.

In conclusion, when it comes to chatbots that are designed to help the user perform a task, it is important that the chatbot can dynamically adapt to the user's current progress into the task. That is, a chatbot should be able to recognize that the user has made a mistake and adapt its instructions accordingly. Otherwise, it is possible that from one user mistake, more mistakes are doomed to happen leading to poor task success. This is even more important when implementing lexical alignment into a chatbot, because by copying the user's lexical choices in situations where the user and the chatbot are referring to different objects, the chatbot will produce misleading instructions, luring the user into a false sense of confidence.

6.2 Value of Lexical Alignment

The results of this study suggest that lexical alignment has a positive effect on task success. The reason for this can be illustrated by the following example. We saw that there was a big success discrepancy between the two conditions going from stage 2 to 3 of the task. For the misaligning condition, this corresponds to the instruction "Continue drawing the path through the gap between the old pine and the fallen cairn". If we consider Figure 4, it is clear that many participants drew the path from pivot E to south of the "old pine" instead of north of it. Hence, the description "fallen cairn" is misinterpreted, likely because it could refer to either of the landmarks north and south of the "old pine" or because the participants were unfamiliar with the term altogether. In the aligning condition, the instruction appears to be clearer. For example, after one participant termed the landmarks that are north and south of what they called "pine tree" as "small stones" and "landslide", the chatbot produces the instruction "Continue drawing the path through the gap between the pine tree and the small stones." This participant completed the stage successfully likely because in this situation, lexical alignment lead to an instruction that can be better understood.

Hence, we conclude that lexical alignment is especially valuable in situations where otherwise a given expression could be interpreted to refer to multiple objects or is unfamiliar to the user.

Additionally, there is value in lexical alignment for producing fluent interactions. We saw that participants interacting with the aligned chatbot, rated the statement "The chatbot contributed to the fluency of the interaction" significantly higher than participants in the misaligning condition. This sub-question is arguably the most important for chatbot design because it captures how well the chatbot is able to produce a fluent interaction.

6.3 Companionship Qualities of the Chatbot

Nine of the ten sub-questions for perceived fluency returned no significant differences. The reason for that could be that many of these questions focused on the chatbot's qualities in being a companion or friend. That is, how trustworthy, committed, understanding, helping and cooperative the chatbot was. In the setting of this experiment, where the dialogue focused solely on the task at hand, lexical alignment had no significant effect on these companionship or friendship qualities. More research is needed to determine whether lexical alignment does improve such qualities if the dialogue is less limited and the topic of conversation is more casual.

6.4 Intelligence of the Chatbot

One of the fluency sub-questions focused on whether the chatbot is intelligent. Participants in the misaligning condition rated the intelligence higher than participants in the aligning condition. This could be because users felt like the chatbot was more knowledgeable than them. For example, consider the following excerpt from a misaligning interaction:

USER: I am just north of a(n) unclear image, I think it's some rocks or hills.

CHATBOT: Okay to start, draw a path that goes to the left of the stone creek.

This participant rated the chatbot as highly intelligent (7 out of 7) and the reason for that could be that the chatbot was resolving their confusion. It knew what the actual term for the "unclear image" was and hence the participant might have perceived the chatbot as more knowledgeable than themselves. In contrast to that, consider the following excerpt from the aligning condition:

USER: Yes, I see a(n) something (i cannot recognize).

CHATBOT: Draw the path around the something (i cannot recognize) by going west of it and then back up in the east.

This participant rated the chatbot's intelligence the lowest possible (1 out of 7), likely because it was clear that the chatbot was just parroting them without knowledge on its own. While this is due to the implementation of lexical alignment in this experiment, further research with a more advanced implementation is needed to discover whether lexical alignment can work against how intelligent the chatbot is perceived.

6.5 Non-significant Effect on Perceived Workload

We saw that there is no significant difference between the two alignment versions in how much workload the participants perceived to have had. This result conflicts with related theory and studies (see Section 2) in that lexical alignment should decrease perceived workload. More research is needed to reconcile these differences but one possible explanation is that factors outside of just lexical alignment might have significantly contributed to how demanding participants found the entire experiment to be. For example, the participants' familiarity with chatbots or using computers in general could have overwritten an effect on workload caused by lexical alignment alone. Furthermore, the non-significant difference in temporal demand could be explained by that participants in both conditions committed similar amounts of time to the task and hence felt like the task was similarly rushed or hurried. Another explanation could be that participants were not even aware of a time limit as they might have skipped reading the task description and not paid attention to the countdown in the corner of the screen. This is supported by the fact that there was no significant correlation between how long the participants took and how temporally demanding they found the task to be.

6.6 Non-significant Effect on BIS

In contrast to task accuracy, we were unable to detect a significant difference in BIS between the conditions. This could be due to the fact that there was no significant difference in time taken as discussed above. Since BIS weighs time as 50% of the score, the significant difference in accuracy was likely overshadowed. Therefore, when trying to make human-chatbot interactions more speedy, lexical alignment does not necessarily contribute to that goal.

7 CONCLUSION

In this paper, we presented an experiment to investigate the effect of lexical alignment from a chatbot on task success, perceived workload and perceived fluency. We found that lexical alignment significantly improves the user's success in solving a collaborative task with the chatbot. While we did not capture a significant effect on perceived workload or companionship qualities in the setting of this experiment, we found that a chatbot that lexically aligns is perceived as contributing to a more fluent interaction. In conclusion, lexical alignment should be taken into consideration in designing task-oriented chatbots, because it can significantly improve human-chatbot collaboration.

REFERENCES

[1] [n. d.]. Chatbot Market Size, Share, Trends & Growth Report, 2030. https://www.grandviewresearch.com/industry-analysis/chatbot-market

[2] Ahlam Alnefaie, Sonika Singh, Baki Kocaballi, and Mukesh Prasad. 2021. An Overview of Conversational Agent: Applications, Challenges and Future Directions:. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 388–396. https://doi.org/10.5220/0010708600003058

[3] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (Sept. 2010), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

[4] Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2022. Chatbots Language Design: The Influence of Language Variation on User Experience with Tourist Assistant Chatbots. *ACM Transactions on Computer-Human Interaction* 29, 2 (Jan. 2022), 13:1–13:38. https://doi.org/10.1145/3487193

[5] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on human-chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (May 2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438 arXiv:1904.02743 [cs].

[6] Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science* 23, 8 (Aug. 2012), 931–939. https://doi.org/10.1177/0956797612436816

[7] Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27, 2 (Nov. 1987), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7

[8] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. https://doi.org/10.1145/3290605.3300439

[9] Sandra G. Hart. 1986. NASA Task Load Index (TLX). https://ntrs.nasa.gov/citations/20000021488 NTRS Author Affiliations: NASA Ames Research Center NTRS Document ID: 20000021488 NTRS Research Center: Ames Research Center (ARC).

[10] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. https://doi.org/10.1177/154193120605000909 Publisher: SAGE Publications Inc.

[11] Guy Hoffman. 2019. Evaluating Fluency in Human–Robot Collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (June 2019), 209–218. https://doi.org/10.1109/THMS.2019.2904558

[12] Guy Hoffman and Cynthia Breazeal. 2007. Cost-Based Anticipatory Action Selection for Human–Robot Fluency. *Robotics, IEEE Transactions on* 23 (Nov. 2007), 952–961. https://doi.org/10.1109/TRO.2007.907483

[13] Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions in spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG '09)*. Association for Computational Linguistics, USA, 74–81.

[14] Theodora Koulouri, Stanislao Lauria, and Robert D. Macredie. 2016. Do (and Say) as I Say: Linguistic Adaptation in Human–Computer Dialogs. *Human–Computer Interaction* 31, 1 (Jan. 2016), 59–95. https://doi.org/10.1080/07370024.2014.934180

[15] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (Feb. 1966), 707. https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L ADS Bibcode: 1966SPhD...10..707L.

[16] Heinrich René Liesefeld and Markus Janczyk. 2019. Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods* 51, 1 (Feb. 2019), 40–60. https://doi.org/10.3758/s13428-018-1076-x

[17] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. *High Frequency Word Entrainment in Spoken Dialogue*. https://doi.org/10.3115/1557690.1557737 Journal Abbreviation: Proceedings of the ACL/HLT 2008 Pages: 172 Publication Title: Proceedings of the ACL/HLT 2008.

[18] Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, 2 (April 2004), 169–190. https://doi.org/10.1017/S0140525X04000056 Publisher: Cambridge University Press.

[19] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151 (July 2021), 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

[20] David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language* 76 (Oct. 2014), 29–46. https://doi.org/10.1016/j.jml.2014.05.008

[21] Bayan Shawar and Eric Atwell. 2007. Chatbots: Are they Really Useful? *LDV Forum* 22 (July 2007), 29–49. https://doi.org/10.21248/jlcl.22.2007.88

[22] Huiyang Shen and Min Wang. 2023. Effects of social skills on lexical alignment in human-human interaction and human-computer interaction. *Computers in Human Behavior* 143 (June 2023), 107718. https://doi.org/10.1016/j.chb.2023.107718

[23] Laura Spillner and Nina Wenig. 2021. Talk to Me on My Level – Linguistic Alignment for Chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM, Toulouse & Virtual France, 1–12. https://doi.org/10.1145/3447526.3472050

[24] Sumit Srivastava, Mariët Theune, and Alejandro Catala. 2023. The Role of Lexical Alignment in Human Understanding of Explanations by Conversational Agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 423–435. https://doi.org/10.1145/3581641.3584086

[25] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and Alignment in Information-Seeking Conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 42–51. https://doi.org/10.1145/3176349.3176388

[26] Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The HCRC Map Task corpus: natural dialogue for speech recognition. In *Proceedings of the workshop on Human Language Technology (HLT '93)*. Association for Computational Linguistics, USA, 25–30. https://doi.org/10.3115/1075671.1075677

[27] Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to Guide Task-oriented Chatbot Users, and When: A Mixed-methods Study of Combinations of Chatbot Guidance Types and Timings. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3501941