INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Rapid Lexical Alignment to a Conversational Agent

Rachel Ostrand[1], *Victor S. Ferreira*[2], *David Piorkowski*[1]

[1]IBM Research, Yorktown Heights, NY, USA
[2]University of California - San Diego, San Diego, CA, USA
rachel.ostrand@ibm.com, vferreira@ucsd.edu, david.piorkowski@ibm.com

Abstract

Conversational partners modify their language to be more similar to each other during interactions. This phenomenon, known as alignment, has been shown in human-human interactions, but there is little work on lexical alignment in human-computer interactions. We investigate whether people lexically align to a conversational agent, and whether the degree of alignment depends on feedback from the agent. This study compared three feedback conditions for how the agent responded to users' word choice: (1) the agent only understood the specific words that it produced itself; (2) the agent understood the words that it produced as well as more appropriate synonyms; (3) the agent's understanding of words that it did not produce was random. Participants significantly aligned to the agent in all conditions, and aligned more when they learned that the agent's comprehension was contingent on their alignment. Thus, inducing lexical alignment may be an effective way to increase dialogue success.

Index Terms: alignment, lexical entrainment, human-computer interaction, conversational agents

1. Introduction

An important feature of human dialogue is that people modify their own language production to converge on properties of their interlocutor's production. This linguistic modification is variously referred to as *alignment*, accommodation, adaptation, convergence, or entrainment. It has been observed at all linguistic levels, including acoustics [1], phonetics [2], temporality [3], lexical choice [4], syntax [5], and discourse [6]. Some theories posit that alignment occurs for social reasons, to signal affiliation with an interlocutor [7], [8], or for communicative reasons, as repeating a conversational partner's linguistic choices is a way to ensure that one's partner will understand [9]. Alignment of word choice (often termed *lexical entrainment*) is thought to occur as a result of conversational partners forming a *conceptual pact*, in which they implicitly build a shared agreement of how to conceptualize or refer to an object during the interaction [10]. As a result, whether or not a speaker engages in lexical entrainment tends to be strongly dependent on whether their listener is likely to understand and benefit [11], [12]. Alignment occurs automatically and without conscious design, and helps language users navigate the substantial, multidimensional linguistic variability that they are confronted with in daily life.

The current study investigates lexical entrainment between a human and a conversational agent, to understand the degree to which a human's language production can be affected by language from an agent. We address three research questions (RQ). RQ1: Can users' word choice to a conversational agent be modulated by exposing them to the agent's lexical preferences in a task-oriented conversational setting? RQ2: Does real-time feedback from a conversational agent about its understanding affect users' lexical choices? RQ3: How quickly do users align to the conversational agent?

2. Related work

Although most prior work on alignment has investigated human-human interactions, there is evidence that people engage in lexical alignment with non-human partners as well, including conversational agents [13], [14]. In fact, alignment is stronger when interacting with an automated partner as compared to another human, and stronger still when interacting with a computer that is allegedly "basic" versus "advanced" [15]. In general, people align more to partners that they believe are less linguistically competent, to make themselves more easily understood by their interlocutor [16]–[18]. This behavior, a form of *audience design*, occurs because the best predictor of what a partner is able to understand is the linguistic forms they have previously produced themselves.

A critical factor in assessing performance of a conversational agent is intent prediction – the agent's ability to map the user's input to the correct predefined intent; namely, to understand what the user wants. A failure occurs when the system selects the wrong intent, or fails to map the user's input onto an existing meaning [19], [20]. Some previous studies have investigated alignment as a tool to nudge users' input towards linguistic properties that are easier for the system to process [21], [22]. Of particular relevance to the current work, [14] studied a spoken dialogue system which changed its lexical production over time to produce more words that users had aligned to (i.e., the system aligned to the users' alignment to the system), causing its automatic speech recognition to improve.

However, prior research has not investigated situations where the conversational agent gave real-time feedback about its comprehension ability. Thus there remains an open question about whether people differentially align as a function of an automated partner's demonstrated level of understanding.

Prior work has shown that *other* forms of feedback from an automated system or agent can be used to affect users' behavior. Providing anthropomorphic cues [23] and giving a chatbot a human name [24] are shown to increase user comfort and increase the likelihood of a user disclosing personal information. Research in explainable AI has investigated ways that explaining a conversational agent's decision can influence users' decision-making behaviors [25]–[27], including dark patterns to manipulate users towards a certain action [28].

The present study builds on this prior work by investigating whether lexical entrainment can be deployed as a method to modify users' behavior when interacting with a conversational agent. If users can be nudged away from their default lexical preferences, and instead implicitly induced to produce the

10.21437/Interspeech.2023-2332

agent's preferred words merely via exposure from the agent's own lexical production, this could be a fruitful method for improving dialogue success rate for conversational systems.

3. Method

3.1. Participants

The participants were 120 students at UC San Diego, who completed the experiment for course credit (88 female, 30 male, 2 nonbinary; age mean: 21, range: 18-52 years). All reported learning English before age 7. An additional 9 participants were excluded and replaced before data analysis, for providing inappropriate responses (1) or taking more than 25 minutes (20 minutes for the control experiment) (8) to complete the experiment. All participants were treated in accordance with the guidelines for ethical treatment of human subjects and provided written informed consent, as approved by the UC San Diego Institutional Review Board.

3.2. Materials

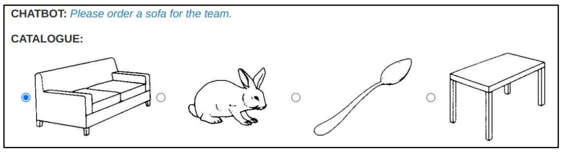
Stimuli consisted of 48 black-and-white line drawings of everyday objects, drawn from the International Picture Naming Project database [29] and supplemented with clipart. Of these, 23 were critical stimuli which each had two acceptable names in American English, one dominant (e.g., *couch*) and the other secondary (e.g., *sofa*). The other 25 pictures were filler stimuli.

Name dominance for the critical stimuli was determined in a separate pilot experiment, with unique participants from the same population (N = 80). Pilot participants were shown each picture and performed a (a) production task, to elicit free-response picture naming, and (b) acceptability judgement task, to determine whether the given word was an appropriate name for the picture. For a picture to be used as a critical stimulus in the main experiment, all of the following criteria had to be met: (1) the dominant and secondary names were *each* produced by at least 10% of participants; (2) either the dominant or secondary name (as opposed to a third name) was produced by at least 75% of participants; (3) each name was *judged acceptable* by at least 90% of participants; (4) the dominant and secondary names were lexically distinct (e.g., excluding *rocket* and *rocket ship*). Filler stimuli had at least 96% name agreement in the production task (e.g., *apple*, *hammer*).

3.3. Procedure

Participants interacted with a simulated conversational agent implemented in the Qualtrics survey software. Their task was to work with the agent to order supplies for a company. The experiment was conducted online, and participants progressed through the experiment at their own pace.

The conversational agent and participant alternated roles as *orderer* and *matcher*. The orderer told the matcher which item to purchase, and the matcher selected a picture of that item from a catalogue, with the goal of matching the orderer's intention.

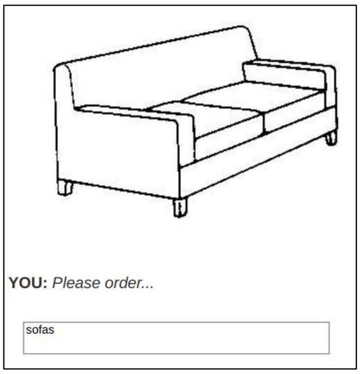


CHATBOT: *Please order a sofa for the team.*

CATALOGUE:

● ○ ○ ○

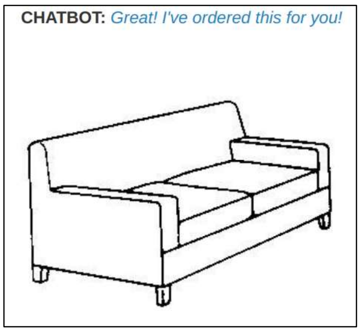
Figure 1: *One trial in the Exposure Phase. The participant has correctly selected the item (sofa) that the conversational agent ordered.*



YOU: *Please order...*

sofas

Figure 2: *One trial in the Test Phase. The participant has typed their response into the textbox: "sofas."*



CHATBOT: *Great! I've ordered this for you!*

Figure 3: *Response from the conversational agent after the participant used a word that the agent understood.*

The experiment consisted of six rounds, each with an Exposure Phase followed by a Test Phase. In the Exposure Phase, the conversational agent told the participant the names of items to order from the catalogue, and the participant selected the correct picture from a set of four (Figure 1). There were eight trials in each Exposure Phase, presented in a randomized order to each participant. After eight Exposure trials, the roles switched and the participant became the orderer. They were shown the same eight items as in the Exposure Phase (one at a time), and had to type the name of the item, to tell the agent which item to purchase (Figure 2). The agent then responded to the participant either by showing the picture of the item it had selected (Figure 3), or displaying an error message if it could not identify which item the participant had ordered.

One set of eight Exposure trials in a row followed by eight Test trials in a row comprised one round. There were six rounds in the experiment; each round had a different set of eight pictures (four criticals and four fillers, except for Round 6 which had three criticals and five fillers). Thus, the agent and participant each named the same 48 pictures to the other across the experiment. (The experiment began with a practice round of four items, which are not included in the analyses below.)

The dependent variable was whether the participant used the same label as the conversational agent for each critical item. The agent always produced the secondary name (*sofa*) for each item. This was to give more opportunity for participants to align to the agent's lexical use, as most participants' default lexical choice should be the dominant name (*couch*).

When the participant did align to the agent – that is, when the participant produced an item's secondary name (*sofa*) – the agent successfully selected that item (Figure 3). There were three between-participant experimental conditions which varied how the agent responded on trials for which the participant did *not* align, and instead produced the item's dominant name (*couch*).

In the Correct100 condition, the agent understood, and chose the correct item, when the participant produced the secondary name (aligned to the agent), or the dominant name (did not align to the agent). In the Correct0 condition, the agent only understood when the participant produced the secondary name (aligned), but not when the participant did not align. In the Correct50 condition, the agent understood when the participant produced the secondary name (aligned), and randomly with 50% probability either understood or did not when the participant produced the dominant name (did not align). In all conditions, if the participant produced a name that was neither the dominant nor secondary name, regardless of whether it was acceptable (*loveseat*) or incorrect (*tree*), the agent did not understand and produced the error message.

In addition to the main experiment with the three experimental conditions, a Control experiment was also conducted. The Control experiment's procedure was identical to that of the main experiment, except that there was no Exposure Phase in each round. Thus, participants were simply presented with the pictures one at a time, and asked to tell the agent the name of that item to purchase (as in Figure 2). Thus, the Control experiment elicited baseline rates of how frequently participants referred to each critical item using the dominant versus secondary name, and served as an additional norming experiment within the current experimental procedure.

Participants' responses were automatically processed by the agent as they submitted them, in order for the agent to respond appropriately. Responses were counted as the dominant or secondary name if they contained that word (i.e., the response did not need to be an exact string match). For example, if the participant told the agent to order "sofas" or "a sofa for the team," this was processed by the agent as *sofa*, the secondary name, and thus a success.

There were 30 unique participants in each of the three experimental conditions and the Control experiment, for a total of 120 participants.

To control for possible item order effects, the order of the eight items in each Exposure Phase was randomized for each participant, and separately, the order of the eight items in each Test Phase was randomized. In addition, the order of rounds was counterbalanced between participants, such that half of the participants in each condition saw a certain item in the first round and half saw that item in the sixth round, and so on.

After the six experimental rounds, participants completed demographics, language history, and debriefing questionnaires.

4. Analysis

Participants' responses were analyzed using generalized logit mixed-effects models (GLMM) in R (version 3.6.2) [30] using the *lme4* package (version 1.1.21) [31]. De-identified data and scripts for running analyses are available at https://osf.io/jr5k8/.

The omnibus model included two categorical independent variables: Feedback Condition (between-participants) and Tercile (within-participants). Feedback Condition was a 4-level factor, and was treatment-coded with Control as the reference level and Correct0, Correct50, and Correct100 as treatment levels. Tercile (diving the experiment into three equal time spans) was a 3-level factor, and was sum-coded in the model as Tercile1 = (-0.5, 0.0); Tercile2 = (0.0, -0.5); Tercile3 = (+0.5, +0.5). The dependent variable was whether the participant's word matched the agent's word (i.e., did the participant produce the secondary name for the item), and was categorical. Analysis was conducted on critical items, for which participants had multiple lexical options for naming the picture.

The model structure included the maximal random effects structure with the *bobyqa* optimizer to aid convergence. The initial model failed to converge. First, correlations between random effects were removed, and then all random factors which accounted for less than 1% of the model's variance were removed; this reduced model did converge. Test statistics and statistical significance for each effect were determined using the *lmerTest* (version 3.1.2) [32] and *emmeans* (version 1.5.0) [33] packages, employing Satterthwaite's method for approximating degrees of freedom. The final converged model structure is shown in the following equation:

*match_Name2 ~ FeedbackCondition * Tercile*
+ (1 + Tercile.num1 + Tercile.num2 || Participant)
+ (1 + FeedbackCondition.num3 + Tercile.num1 + Tercile.num2 || Item)

5. Results

In the omnibus model, there was a significant main effect of Feedback Condition (F = 62.46, p < .0001), no main effect of Tercile (F < 1), and no Feedback Condition x Tercile interaction (F < 1). Note that although the statistical models were conducted in log-odds space using GLMMs, the figure and conditions means reported in the text show untransformed percentage data, as this scale makes the interpretation of effect sizes easier. See Figure 4 for results from the experiment.

RQ1 explored whether people's word choice can be modified to match those produced by a conversational agent simply by exposing them to the agent's word choice. Pairwise contrasts were conducted within the omnibus model, comparing each of the three experimental conditions against the Control condition, with p-values adjusted for multiple comparisons using the Tukey correction. Participants showed substantial alignment to the agent's words, producing significantly more secondary names in the Correct0 condition (76.4%; z = 12.21, p < .0001), the Correct50 condition (72.2%; z = 11.25, p < .0001), and the Correct100 condition (66.5%; z = 9.95, p < .0001) compared to the baseline Control experiment (21.7%). This demonstrates that *participants aligned their language production to the agent's word choice to a substantial degree*, independent of whether the agent could understand the dominant, non-aligned item names.

RQ2 asked whether feedback from the agent about its comprehension affected how much participants aligned to the agent's word choice. A model was constructed comparing degree of alignment in the three experimental conditions pairwise against each other (adjusting p-values for multiple comparisons using the Tukey correction). *Participants aligned significantly more in the Correct0 condition (76.4%), in which non-alignment caused communicative failure, compared to the Correct100 condition (66.5%), in which non-alignment nevertheless resulted in communicative success* (z = 2.43, p < .04). Neither of the other pairwise comparisons (Correct0 vs. Correct50 or Correct50 vs. Correct100) was significant.

RQ3 investigated whether participants adapted their word choice based on the agent's feedback over the course of the experiment. That is, did participants in the Correct0 and Correct50 conditions (which at least sometimes needed alignment for the agent to understand) need ongoing exposure trials to align, or did they adapt right away, within the first few trials of the experiment? Surprisingly, *participants adapted their word choice to match the agent's immediately*, and did not require substantial feedback to do so (no main effect of Tercile; F < 1). In addition, participants did not show any differential learning and adaptation effects between the conditions (no Feedback Condition x Tercile interaction; F < 1; additionally, no pairwise contrasts between the experimental conditions were significant at any Tercile).

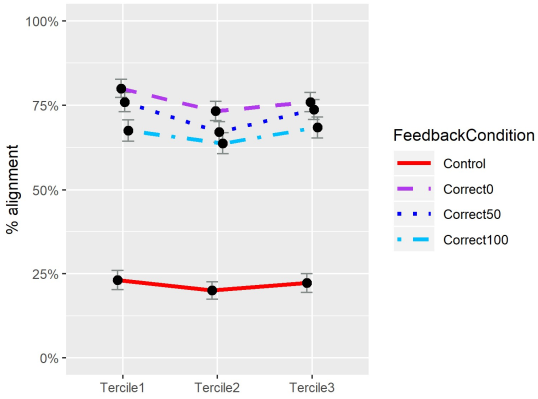


Figure 4: *Degree of alignment to the agent, as a function of Feedback Condition and Tercile. Error bars show standard error of the mean.*

6. Discussion

This experiment investigated whether and how people modify their word choice based on feedback from a conversational agent. Participants strongly aligned their lexical production to the agent. The fact that the alignment effect was so substantial – participants produced the agent's picture name more than three times as often in the experimental conditions (when they had been exposed to the agent's preferred word) as in the Control condition (when they did not know the agent's preferred word) – suggests users may *enter* into dialogue with conversational agents with the assumption that alignment is necessary for communicative success. Users aligned their word choice to the agent seemingly by default, even without feedback that it did not understand non-aligned words. This is consistent with prior work showing greater alignment to less linguistically competent partners [15]–[18], and thus, the mere knowledge that one's interlocutor is a non-human conversational agent may be enough to influence users' lexical production.

Interacting with an agent which only understood the words that it had previously produced induced a small but significant additional degree of alignment from participants. There are two important take-aways from this behavior. First, the amount that participants increased their alignment based on specific feedback from the agent was substantially smaller than the amount that participants increased their alignment after merely being exposed to the agent's preferred words. This suggests that preexisting assumptions about a conversational agent's (low) level of understanding play a much larger role in determining a user's lexical choice, than does observation of this *particular* agent's level of understanding. It is possible that users' prior experience with conversational agents led them to expect low comprehension ability, and thus producing words that this agent demonstrably can understand was a better strategy than potentially suffering the consequences of non-comprehension. Second, adaptation to the agent's response behavior happened quickly, as there was no difference in alignment as a function of Feedback Condition over time. As soon as participants received just a few examples of the agent either understanding or not understanding the non-aligned, dominant picture names, they adapted to that comprehension behavior.

It is important to note that the observed alignment effects are unlikely to be attributable merely to repetition priming (in which exposure to a word decreases the activation necessary for subsequently accessing it, and thus increases the likelihood of future production). Due to the randomization of item order in both the Exposure and Test phases, a particular item was named by the participant in the Test Phase with on average seven items intervening after seeing it named by the agent in the Exposure Phase, reducing any immediate effects of priming. In addition, if the observed alignment effects were caused exclusively by repetition priming, then the degree of alignment should be equivalent across the three experimental conditions. Thus, the spread between these conditions makes a pure repetition priming-driven mechanism unlikely.

In future work, it will be important to explore whether alignment behavior differs between participants who have different characteristics. People who have more experience working with – and even training – conversational agents may come into the interaction with stronger expectations of the range of agents' comprehension abilities, and thus may be more receptive to feedback about this agent's level of comprehension. In addition, it will be important to investigate how non-native or low language proficiency users align to an agent, as they likely have a less diverse vocabulary themselves.

7. Conclusions

The current experiment investigated lexical alignment between a human and conversational agent. The results showed that people massively modulated their lexical production in order to match the words that the agent had previously produced. The alignment effect was heightened when interacting with an agent which *only* understood the particular word that it had used to refer to a particular item. However, this alignment occurred practically immediately, and did not require much exposure for users to adapt.

When conversational agents fail to recognize the user's intent, it is often due to lexical failure – because the agent did not correctly map the user's word choice to the correct intent, or any intent at all. The results from the current study suggest a practical way to ameliorate this problem, and improve the recognition success rate for automated conversational agents. If a system can expose to users the words that it uses itself, and the mappings from particular words to referents or intents, then users are likely to produce those same words back to the agent, and thus have more successful conversational interactions.

8. References

[1] M. Babel and D. Bulatov, “The Role of Fundamental Frequency in Phonetic Accommodation,” *Language and Speech*, vol. 55, no. 2, pp. 231–248, Jun. 2012, doi: 10.1177/0023830911417695.

[2] J. S. Pardo, I. C. Jay, and R. M. Krauss, “Conversational role influences speech imitation,” *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2254–2264, Nov. 2010, doi: 10.3758/BF03196699.

[3] F. Bonin *et al.*, “Investigating fine temporal dynamics of prosodic and lexical accommodation,” in *Proc. INTERSPEECH 2013*, Lyon, France, 2013, pp. 539–543. doi: 10.21437/Interspeech.2013-151.

[4] H. H. Clark and D. Wilkes-Gibbs, “Referring as a collaborative process,” *Cognition*, vol. 22, no. 1, pp. 1–39, Feb. 1986, doi: 10.1016/0010-0277(86)90010-7.

[5] H. P. Branigan, M. J. Pickering, and A. A. Cleland, “Syntactic co-ordination in dialogue,” *Cognition*, vol. 75, no. 2, pp. B13–B25, May 2000, doi: 10.1016/s0010-0277(99)00081-5.

[6] S. Garrod and A. Anderson, “Saying what you mean in dialogue: A study in conceptual and semantic co-ordination,” *Cognition*, vol. 27, no. 2, pp. 181–218, 1987, doi: 10.1016/0010-0277(87)90018-7.

[7] H. Giles, N. Coupland, and J. Coupland, “Accommodation theory: Communication, context, and consequence,” in *Contexts of accommodation: Developments in applied sociolinguistics*, Cambridge: Cambridge University Press, 1991, pp. 1–68. doi: 10.1017/CBO9780511663673.001.

[8] M. Babel, “Dialect divergence and convergence in New Zealand English,” *Language in Society*, vol. 39, pp. 437–456, Sep. 2010, doi: 10.1017/s0047404510000400.

[9] M. J. Pickering and S. Garrod, “Toward a mechanistic psychology of dialogue,” *Behavioral and Brain Sciences*, vol. 27, no. 02, pp. 169–190, Apr. 2004, doi: 10.1017/s0140525x04000056.

[10] S. E. Brennan and H. H. Clark, “Conceptual pacts and lexical choice in conversation.,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, no. 6, pp. 1482–1493, Nov. 1996, doi: 10.1037/0278-7393.22.6.1482.

[11] S. O. Yoon and S. Brown-Schmidt, “Adjusting conceptual pacts in three-party conversation.,” *Journal of Experimental Psychology. Learning, Memory, and Cognition*, vol. 40, no. 4, pp. 919–937, Jul. 2014, doi: https://doi.org/10.1037/a0036161.

[12] S. O. Yoon and S. Brown-Schmidt, “Aim Low: Mechanisms of Audience Design in Multiparty Conversation,” *Discourse Processes*, vol. 55, no. 7, pp. 566–592, Mar. 2017, doi: 10.1080/0163853x.2017.1286225.

[13] G. Parent and M. Eskenazi, “Lexical Entrainment of Real Users in the Let’s Go Spoken Dialog System,” in *Proc. INTERSPEECH 2010*, Makuhari, Chiba, Japan, Sep. 2010, pp. 3018–3021. doi: 10.21437/Interspeech.2010-49.

[14] J. Lopes, M. Eskenazi, and I. Trancoso, “Automated two-way entrainment to improve spoken dialog system performance,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 8372–8376. doi: 10.1109/icassp.2013.6639298.

[15] H. P. Branigan, M. J. Pickering, J. Pearson, J. F. McLean, and A. Brown, “The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers,” *Cognition*, vol. 121, no. 1, pp. 41–57, Oct. 2011, doi: 10.1016/j.cognition.2011.05.011.

[16] Z. G. Cai, Z. Sun, and N. Zhao, “Interlocutor modelling in lexical alignment: The role of linguistic competence,” *Journal of Memory and Language*, vol. 121, p. 104278, Dec. 2021, doi: 10.1016/j.jml.2021.104278.

[17] I. Ivanova, H. Branigan, J. McLean, A. Costa, and M. Pickering, “Lexical Alignment to Non-native Speakers,” *Dialogue & Discourse*, vol. 12, no. 2, pp. 145–173, Oct. 2021, doi: 10.5210/dad.2021.205.

[18] E. Suffill, T. Kutasi, M. J. Pickering, and H. P. Branigan, “Lexical alignment is affected by addressee but not speaker nativeness,” *Bilingualism: Language and Cognition*, vol. 24, no. 4, pp. 746–757, 2021, doi: 10.1017/S1366728921000092.

[19] K. Kvale, O. A. Sell, S. Hodnebrog, and A. Følstad, “Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues,” in *Chatbot Research and Design: CONVERSATIONS 2019*, Cham, 2020, pp. 187–200. doi: 10.1007/978-3-030-39540-7_13.

[20] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, “Benchmarking Natural Language Understanding Services for Building Conversational Agents,” in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, vol. 714, E. Marchi, S. M. Siniscalchi, S. Cumani, V. M. Salerno, and H. Li, Eds. Springer Singapore, 2021, pp. 165–183.

[21] L. Bell, J. Gustafson, and M. Heldner, “Prosodic adaptation in human–computer interaction,” in *Proceedings of The 15th International Congress of Phonetic Sciences*, Barcelona, Spain, Aug. 2003, pp. 2453–2456.

[22] A. Fandrianto and M. Eskenazi, “Prosodic entrainment in an information-driven dialog system,” in *Proc. INTERSPEECH 2012*, Portland, OR, USA, Sep. 2012, pp. 342–345. doi: 10.21437/Interspeech.2012-85.

[23] C. Ischen, T. Araujo, H. Voorveld, G. van Noort, and E. Smit, “Privacy Concerns in Chatbot Interactions,” in *Chatbot Research and Design: CONVERSATIONS 2019*, Cham, 2020, vol. 11970, pp. 34–48. doi: 10.1007/978-3-030-39540-7_3.

[24] M. Ng, K. P. L. Coopamootoo, E. Toreini, M. Aitken, K. Elliot, and A. van Moorsel, “Simulating the Effects of Social Presence on Trust, Privacy Concerns & Usage Intentions in Automated Bots for Finance,” in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, Genoa, Italy, Sep. 2020, pp. 190–199. doi: 10.1109/EuroSPW51379.2020.00034.

[25] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 295–305. doi: 10.1145/3351095.3372852.

[26] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, “Explaining models: an empirical study of how explanations impact fairness judgment,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York, NY, USA, Mar. 2019, pp. 275–285. doi: 10.1145/3301275.3302310.

[27] M. Nauta *et al.*, “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI,” *ACM Comput. Surv.*, p. 3583558, Feb. 2023, doi: 10.1145/3583558.

[28] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek, “Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems,” in *Joint Proceedings of the ACM IUI 2019 Workshops*, Los Angeles, USA, Mar. 2019.

[29] A. Szekely *et al.*, “A new on-line resource for psycholinguistic studies,” *Journal of Memory and Language*, vol. 51, no. 2, pp. 247–250, Aug. 2004, doi: 10.1016/j.jml.2004.03.002.

[30] R Core Team, “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[31] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models Using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, Oct. 2015, doi: 10.18637/jss.v067.i01.

[32] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest Package: Tests in Linear Mixed Effects Models,” *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, Dec. 2017, doi: 10.18637/jss.v082.i13.

[33] Russell Lenth, “emmeans: Estimated Marginal Means, aka Least-Squares Means.” 2020. [Online]. Available: https://CRAN.R-project.org/package=emmeans