Proceedings of the Human Factors and Ergonomics Society Annual Meeting
2025, Vol. 69(1) 395–397

DOI: 10.1177/10711813251357897
journals.sagepub.com/home/pro
Sage

Less Trust, More Mirroring Words: Lexical Alignment as a Response to Uncertainty in Human-AI Teams

Mengyao Li[1] and Emanuel Rojas[1]

Abstract

This study investigates how trust contagion in human-AI teams is reflected through linguistic alignment. In a resource allocation game, 42 participants collaborated with an AI and a confederate teammate trained to express high, neutral, or low trust in the AI. We analyzed lexical and structural alignment in participants' responses. Results showed significantly higher lexical alignment when confederates expressed low trust, suggesting participants mirrored more words in response to uncertainty. Structural alignment did not vary across conditions. These findings suggest lexical alignment serves as a social adaptation to low-trust environments, potentially as a compensatory or affiliative response. Real-time tracking of lexical alignment could inform adaptive AI interfaces to detect and mitigate negative trust contagion. Future work should investigate non-verbal alignment and longer interactions to capture broader trust dynamics.

Keywords

trust in automation, alignment, mirroring, text analysis

Introduction

Understanding trust contagion—how trust spreads among human operators toward an AI teammate—is critical for enhancing cooperation in human-AI teams. For example, users who initially distrust an AI teammate may increase their trust after observing their human teammate who interacts positively with the AI teammate. This suggests that trust in automation is shaped not only by direct experiences with and characteristics of the AI teammate, but also by social influences from human teammates (Guo et al., 2023). While previous research has examined trust contagion (Rojas & Li, 2024), the role of conversational alignment, specifically *lexical and structural alignment*, in facilitating this trust contagion process has received less attention. This study investigates how lexical and structural alignment contribute to this trust contagion process. By manipulating a confederate's expressed trust level (high, low, neutral) toward an AI teammate, we analyzed how this influenced the linguistic alignment patterns of human teammates in collaborative interactions. Interestingly, results showed that low trust conditions exhibited higher lexical alignment compared to neutral and high trust conditions, suggesting that lower trust levels may trigger greater linguistic effort to align as a compensatory mechanism for uncertainty. However, structural alignment scores did not show significant main effects. Understanding these alignment mechanisms offers practical insights into how trust spreads within teams and can inform the design for conversational agents.

Background

Effective communication in such teams often exhibits linguistic alignment, wherein conversational partners unconsciously adapt their language to one another (Pickering & Garrod, 2004). This alignment occurs at multiple levels, including lexical (word choice) and structural (syntactic patterns) alignment. *Lexical alignment* refers to the degree to which the word choice in a conversational turn reflects that of the preceding turn (Srivastava et al., 2024). For example, if a confederate says, "The AI seems unreliable," and the participant responds, "Yes, it's definitely unreliable," the repeated use of "unreliable" demonstrates strong lexical alignment. In contrast, a response like "I don't think it's consistent" lacks shared words, resulting in weaker lexical alignment. *Structural alignment*, on the other hand, evaluates the syntactic similarity between consecutive conversational

[1]School of Psychology, Georgia Institute of Technology, Atlanta, USA

Corresponding Author:
Mengyao Li, School of Psychology, Georgia Institute of Technology, 654 Cherry Street NW, Atlanta, GA 30332, USA.
Email: mengyao.li@gatech.edu

turns, measuring how closely the structure of one utterance mirrors that of the preceding, priming utterance. For instance, an utterance like "I am skeptical" would be represented as "PRON VERB ADJ," while "I am skeptical too" extends the pattern to "PRON VERB ADJ ADV." Prior research suggests that linguistic alignment enhances communication efficiency, increases perceptions of integrity and trustworthiness, and reduces task workload (Linnemann & Jucks, 2018; Spillner & Wenig, 2021).

In human-AI teams, lexical and structural alignment may serve as cognitive mechanisms for trust contagion. When trust-positive lexical choices (e.g., "accurate," "reliable") and syntactic structures are mirrored, trust contagion is amplified, leading to greater confidence in joint decision-making. Conversely, alignment on distrust-related language (e.g., "risky," "unpredictable") fosters negative trust contagion, increasing skepticism toward AI or human teammates. This paper investigates how lexical and structural alignment patterns are influenced by manipulated levels of trust expressed by a confederate in human-AI team interactions. Specifically, we examined whether low, neutral, or high trust conditions elicit differences in alignment behaviors.

Approach

A 2 (AI reliability: high vs. low, within-subjects) × 3 (confederate trusting: high, low, neutral, between-subjects) mixed-subject design was conducted. Each team consisted of a participant, a confederate, and an AI teammate, collaborating on a 10-round resource allocation game requiring joint decision-making. The AI performed with 100% accuracy in the high-reliability condition and 60% in the low-reliability condition, with all participants experiencing the high-low fixed order to first build trust and then erode trust. Trust contagion was manipulated by training the confederate to display varying levels of trust toward the AI: fact-based and neutral comments (neutral condition), positive attitudes (high-trusting condition), or skeptical remarks (low-trusting condition). Participants provided trust ratings for both the AI teammate and their human teammate on a Likert scale ranging from 1 to 7 at around 1, 5, and 10 (Malle & Ullman, 2021). A total of 42 subjects were recruited.

Game session interactions between the confederate and the participant were transcribed, filtered, and cleaned by converting text to lowercase, removing punctuation, whitespace, and common stop words, and excluding trivial responses such as single words "okay" or "yeah." The cleaned text was tokenized to construct conversational pairs and filtered to only include the direction that the participant aligns with the confederate's utterances. We calculated the lexical alignment score for each message as the ratio of tokens in that message that also appear in the preceding message, referred to as the "priming message." This approach specifically analyzed turn-by-turn alignment when speaker roles alternated, focusing on responses rather than initiating statements. The lexical alignment score ranges from 0 to 1, with higher scores indicating greater alignment in word choice. Such differences illustrate how lexical alignment can vary across conversational contexts. This alignment measure captures the extent to which interlocutors reuse vocabulary, reflecting their cognitive and social coordination.

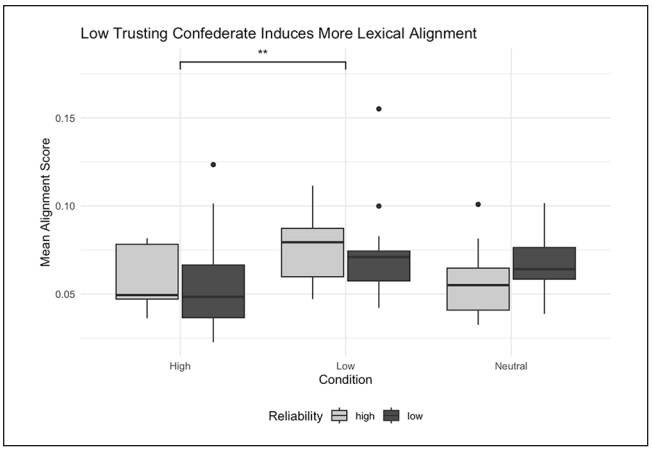

Figure 1. Lexical alignment by trusting condition and AI reliability.
**Low Trusting Confederate Induces More Lexical Alignment.

For structural alignment, each utterance was annotated using the UDPipe Universal Dependencies model to extract Part-of-Speech (POS) patterns (e.g., "PRON VERB ADJ" for "I am happy") and dependency relations (e.g., subject, object, modifier; Straka & Straková, 2017). Structural similarity scores were computed using a string similarity function applied to the POS and dependency patterns of paired utterances. The final structural alignment score for each turn was calculated as the average similarity across POS and dependency structures.

Outcome

Linear mixed-effects models were fitted to predict both lexical alignment and structural alignment scores, with confederate trusting condition and reliability as fixed effects and random intercepts and slopes for Reliability within Participant ID to account for individual variation. As shown in Figure 1, the Reliability Condition (Low Trust) significantly increased lexical alignment ($\beta = .02$, 95% CI [.0009, .03], $t[1453] = 2.08$, $p = .038$), with a standardized effect size of 0.18 (95% CI [.01, .35]). Notably, the result suggests that participants mirrored more words when their teammate expressed low trust in the AI teammate. Reliability (Low) had no significant effect, $p > .05$, suggesting that AI performance reliability did not directly influence lexical alignment. For structural alignment scores, no significant main effects were found, $p > 0.05$.

To test whether lexical alignment scores predicted trust ratings, a multiple linear regression analysis was conducted to examine the predictors of trust in AI. The model was

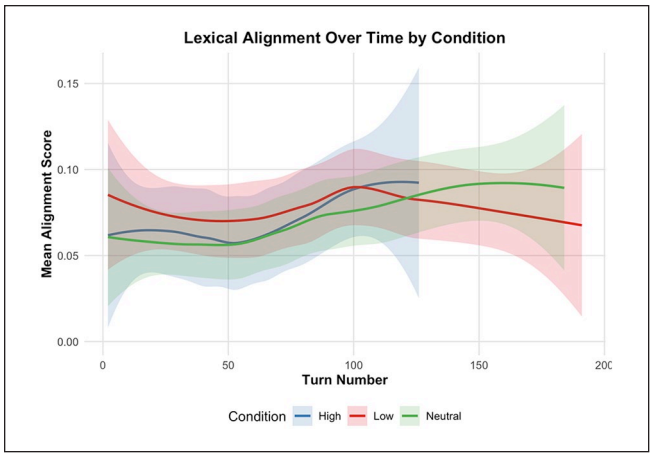

Figure 2. Lexical alignment over time grouped by trusting condition. High trusting condition had much shorter conversation and a lower initial alignment score.

statistically significant, $R^2 = .35$, adjusted $R^2 = .31$, $F(5, 74) = 8.11$, $p < .001$. The number of conversational turns negatively predicted trust, $\beta = -.005$, $p < .001$, as shown in Figure 2, suggesting that longer discussions may reinforce skepticism rather than build trust. Lexical alignment did not significantly predict trust in AI, $\beta = .172$, $p = .563$, suggests that alignment does not directly predict trust shifts and may function as a compensatory behavior rather than a trust-building mechanism. The confederate trust condition significantly influenced trust in AI, with the low-trust condition associated with significantly lower AI trust, $\beta = -.844$, $p < .001$. AI reliability also significantly affected trust in AI, $\beta = -.313$, $p = .003$, with low reliability reducing trust. Significant interaction effects were found such that the low-trust condition combined with low AI reliability resulted in the lowest AI trust, $\beta = -.807$, $p < .001$.

Discussion and Conclusion

Our findings indicate that low-trust teammates drive greater lexical alignment and induce more conversational turns, likely as a compensatory mechanism for increased uncertainty. This pattern may reflect a need to establish common ground when trust in the AI is questioned. Another explanation is that participants subconsciously mirrored their teammate's language through social mimicry—a low-effort affiliative response to maintain social rapport. Distinguishing between these mechanisms will require future studies with refined manipulations, such as cognitive load measures or process-tracing techniques, to disentangle intentional alignment from automatic adaptation.

Trust contagion operates primarily at the lexical level, with no significant structural alignment effects. One possible explanation is that lexical alignment requires less cognitive effort and occurs more rapidly, whereas structural adaptation may emerge over longer interactions. Additionally, the collaborative task may have elicited limited syntactic variability, further reducing the detectability of structural alignment. Future research should explore longer interaction periods, alternative syntactic similarity metrics, and individual differences in linguistic flexibility to better understand the role of structural alignment in human-AI trust dynamics.

These findings can inform the design of adaptive AI systems that monitor real-time linguistic alignment to infer team trust. Such systems could mitigate negative trust contagion by fostering constructive language in low-trust conditions. Expanding to non-verbal alignment (e.g., gaze, gestures, tone) would provide a more comprehensive model of trust contagion in human-human-AI interactions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Guo, Y., Yang, X. J., & Shi, C. (2023). *TIP: A trust inference and propagation model in multi-human multi-robot teams* [Conference session]. Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, pp. 639–643.

Linnemann, G. A., & Jucks, R. (2018). "Can I trust the spoken dialogue system because it uses the same words as I do?"—Influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction. *Interacting with Computers*, *30*(3), 173–186.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3–25). Elsevier.

Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, *27*(2), 212–225.

Rojas, E., & Li, M. (2024). Trust is contagious: Social influences in human-human-AI team. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 317–322.

Spillner, L., & Wenig, N. (2021). *Talk to me on my level–linguistic alignment for chatbots* [Conference session]. Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction, Athens, Greece, pp. 1–12.

Srivastava, S., Wentzel, S. D., Catala, A., & Theune, M. (2024). Measuring and implementing lexical alignment: A systematic literature review. *Computer Speech & Language*, *90*, Article 101731.

Straka, M., & Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 88–99).