Cognitive and Semantic Alignment: Evaluating the Effectiveness of Proactive AI Scaffolding in Non-Native Speaker Turn-Taking

Psyche Wanqing He
Department of Information Science, Cornell University
wh385@cornell.edu

Abstract

Turn-taking represents one of the most cognitively demanding processes in human communication, requiring speakers to simultaneously comprehend incoming speech, predict upcoming content, and formulate responses under tight temporal constraints. For non-native speakers (NNS), this pressure is amplified significantly by the need to navigate linguistic gaps and manage cultural nuances in real-time. Recent advancements in AI-mediated communication (AI-MC) have introduced proactive systems designed to offload these burdens, yet evaluating their success remains a significant methodological challenge. This proposal outlines a research project that moves beyond subjective satisfaction to evaluate interaction quality through a dual-lens framework of content integration and cognitive fluency. Utilizing a dataset from a Wizard-of-Oz study of the XPLAIN system, this research posits that effective AI scaffolding should manifest in two specific behavioral signals: Semantic Integration, where users integrate AI suggestions into their own context rather than copying them, and Cognitive Fluency, where the offloading of predictive tasks results in a measurable reduction in speech disfluencies. By employing high-dimensional text analysis methods alongside acoustic disfluency analysis, this research aims to uncover the "Integration Zone"—an optimal state of human-AI collaboration where algorithmic assistance reduces cognitive load without effectively replacing the human speaker.

ACM Reference Format:
Psyche Wanqing He. 2026. Cognitive and Semantic Alignment: Evaluating the Effectiveness of Proactive AI Scaffolding in Non-Native Speaker Turn-Taking. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Turn-taking represents one of the most cognitively demanding processes in human communication. To maintain a fluid conversation, speakers must simultaneously comprehend incoming speech, predict upcoming content, and formulate their own responses under tight temporal constraints [11]. For non-native speakers (NNS), this pressure is amplified significantly by the need to navigate linguistic gaps, retrieve vocabulary, and manage cultural nuances in real-time. The cognitive load required to manage these parallel processes often results in communication breakdowns, manifested as disfluencies, hesitations, or withdrawal from the conversation.

Recent advancements in AI-mediated communication (AI-MC) have introduced proactive systems designed to offload these burdens. Systems such as XPLAIN [5] aim to "scaffold" interaction by providing real-time lexical clarifications, idea suggestions, and topic summaries. These tools theoretically function as a cognitive prosthetic, bridging the gap between an NNS's communicative intent and their linguistic performance. However, evaluating the success of such tools remains a significant methodological challenge in the field of Human-Computer Interaction (HCI). Traditional evaluation metrics often rely heavily on post-hoc self-reported satisfaction surveys or user interviews. While subjective feedback is valuable, it fails to capture the granular, moment-to-moment dynamics of human-AI collaboration. A user may report high satisfaction because the AI completed the task for them, even if their own agency and active participation were diminished in the process.

This proposal outlines a research project that moves beyond subjective satisfaction to evaluate interaction quality through a dual-lens framework of *content integration* and *cognitive fluency*. Utilizing the dataset from the previous Wizard-of-Oz study of XPLAIN [5], this research posits that effective AI scaffolding should manifest in two specific behavioral signals: *Semantic Integration*, where users do not merely read out AI suggestions but integrate them into their own speech context, and *Cognitive Fluency*, where the offloading of predictive tasks results in a measurable reduction in speech disfluencies [14]. By employing high-dimensional text analysis methods, specifically semantic similarity using transformer-based embeddings, alongside acoustic disfluency analysis, this research aims to uncover the "Critical Integration Level." This theoretical construct represents an optimal state of human-AI collaboration where algorithmic assistance reduces cognitive load without effectively replacing human's critical thinking.

2 Theoretical Background

2.1 Semantic Alignment as a Proxy for Social Connection

The Interactive Alignment Model, a foundational theory in the psychology of dialogue, posits that successful conversation is achieved through the automatic, resource-free priming of linguistic representations between interlocutors [12]. From phonetics to syntax to semantics, speakers unconsciously converge at multiple levels, creating a shared mental model that facilitates mutual understanding. Extending this theory to the sociolinguistic domain, computational research analyzing longitudinal text messages has shown that as relationships deepen, the semantic similarity of partners' language, often measured via Cosine Similarity, significantly increases [2]. Conversational alignment suggests that "sounding alike" is a fundamental mechanism of efficient communication with social connotations.

In the context of AI-Mediated Communication, these principles provide a powerful framework for evaluation. The system is no longer just a reactive tool, but acts partially or fully as an interlocutor. If an NNS user aligns semantically with the AI assistant,

Conference'17, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

it suggests that the system is successfully providing content that matches the user's communicative intent, mirroring the natural convergence seen in human-human dialogue. Therefore, measuring the cosine similarity between an AI's suggestion and the user's subsequent utterance provides a quantifiable proxy for the "uptake"—a crucial first step in confirming that the AI is actively and relevantly shaping the conversation. Recent benchmarks for AI-generated text have adopted this very approach, using semantic similarity to human baselines to determine if AI "smart replies" are socially appropriate and contextually relevant [9].

2.2 The "Mimicry Trap" and the Risk of Lost Agency

However, high semantic similarity could be controversial in its interpretation. While it can signal alignment, it can also indicate a potential over-reliance on AI and loss of user agency, diminishing user's own linguistic development. Literature on human-AI interaction warns that users often rely on heuristics to judge AI outputs; given that AI-generated text often looks polished and authoritative, users can be "manipulated" into accepting suggestions uncritically [10]. Hohenstein and Jung [7] describe this as a "transfer of agency", where the temporal pressure of conversation leads users to defer to the AI's "smart" suggestions, effectively letting the algorithm speak for them. This creates a "Mimicry Trap": a user might exhibit a high semantic alignment score not because they thoughtfully agreed with the AI, but because they mindlessly copied its text to survive the turn-taking pressure.

This distinction is critical for evaluating scaffolding tools, which are meant to empower, not replace critical thinking. True scaffolding helps users construct their own contributions. Heer [6] argues for a design philosophy of "agency plus automation," where systems are designed to augment human intellect rather than replace it. One method for achieving this is through "cognitive forcing functions", which are design elements that require active user engagement. Buçinca et al. [3] show that systems requiring users to interact with and modify suggestions reduce overreliance on AI and increase psychological ownership. This directly informs our hypothesis: the optimal level of scaffolding is not simply reflected as high similarity, but moderate or high similarity combined with moderate textual modification, indicating thoughtful engagement.

2.3 Cognitive Load: The Bottleneck of Fluency

The second theoretical aspect of this proposal addresses *why* NNS speakers struggle and *how* AI can help. Psycholinguistic studies on second language (L2) distinguish between *Cognitive Fluency* (i.e., the efficiency of underlying mental processes like lexical retrieval and speech planning) and *Utterance Fluency* (the acoustic smoothness of speech production) [13]. The relationship between the two is causal: when there is a cognitive "bottleneck" in retrieving a term, the production system fails to formulate smooth speech, resulting in disfluencies. Different disfluency types map to different cognitive struggles: shorter filled pauses like "um" and "ah" are strongly linked to lexical search difficulty, whereas longer silent pauses (often accompanied by filler words) are more likely to be associated with higher-level macro-planning such as ideation and conceptualization [1].

Proactive AI tools like XPLAIN are theoretically positioned to clear these bottlenecks. By providing "Lexical Clarifications" (a contextual explanation more than a definition) or "Idea Suggestions" (a contextually grounded sentence starter), the AI offloads the extraneous cognitive load associated with searching for words and formulating basic ideas in L2. This frees up the user's limited cognitive resources to be spent on *germane load*, enabling NNS to focus on higher-order communicative goals, refine their argument, and engage with the conversational partner. Direct empirical evidence for this mechanism in educational technology comes from studies on "Reading Assistant Software", where providing visual textual support significantly reduces filled pauses and repairs while increasing the length of fluent speech runs [8]. Similarly, AI writing assistants reduced intrinsic cognitive load associated with linguistic production, allowing learners to focus on higher-level writing tasks [4]. This literature underscores the potential for AI-powered interventions to enhance L2 fluency by mitigating cognitive load, suggesting that effective AI scaffolds should be reflected as physical, measurable changes in the acoustic signals.

3 Methodology

This research will utilize the rich multi-modal dataset collected during the evaluation of the XPLAIN system [5]. The dataset consists of dyadic conversations involving 27 NNS of English, individually interacting with a confederate in a simulated online meeting environment. The data includes time-stamped transcripts of user speech, time-stamped AI interventions (i.e., clarifications, suggestions, summaries), high-quality video recordings, and post-interaction surveys and interviews. Two interrelated analytical studies are proposed.

3.1 Study 1: Semantic Integration and Modification Analysis

The first study focuses on measuring the uptake and integration of AI suggestions. The core methodological challenge is to quantify the extent to which a user adopts an AI's suggestion without relying on simple keyword matching. To address this, I will employ high-dimensional vector representations of text. I will use **Sentence-BERT (SBERT)** to generate dense vector embeddings for both the AI-generated suggestions and the user's subsequent spoken turns. SBERT is chosen over older methods like TF-IDF or classic Word2Vec because it is fine-tuned to create sentence-level representations that excel at capturing semantic similarity for short conversational turns, effectively handling the context and polysemy that bag-of-words models miss.

For each interaction where an idea or sentence suggestion was displayed, we will compute two primary metrics:

(1) **Semantic Similarity (Cosine Similarity):** This continuous metric quantifies the conceptual alignment between the AI's prompt and the user's utterance. A high score indicates that the user effectively incorporated the informational content of the suggestion.
(2) **Surface Modification (Levenshtein Edit Distance):** This metric calculates the character-level edit distance between the suggested text and the final spoken text, normalized by the length of the string. This measures the degree of literal modification.

The analytical strategy involves testing for a non-linear relationship between these metrics and user satisfaction. I hypothesize that the interactions of the highest quality on the "Critical Integration Level" will be characterized by high semantic similarity (indicating the help was relevant) but moderate edit distance (indicating the user maintained agency by adapting the text). This contrasts with "passive copying" (high similarity, near-zero edits) and "rejection" (low similarity, near-all edits). Utilizing these computational text analysis techniques helps to construct a robust proxy for user agency to complement standard performance metrics.

3.2 Study 2: Disfluency and Cognitive Load Reduction

The second study aims to quantify the cognitive benefits of the system by analyzing acoustic markers of load. If the XPLAIN system effectively scaffolds the conversation, it is expected to observe a causal reduction in disfluency rates. The analysis will proceed by segmenting the audio data into "AI-Scaffolded Turns" (utterances occurring within a 5-second window following an interaction) and "Baseline Turns" (utterances generated without immediate AI support or rejected support). We will annotate these segments for filled pauses, unfilled silent pauses, and different types of repairs, following established coding schemes and ensuring reliability through inter-rater agreement checks (Cohen's Kappa).

To enable valid comparisons across speakers, raw disfluency counts will be normalized by speech rate (syllables per second). The statistical analysis will employ a linear mixed-effect (LME) model. This approach is essential for this type of panel data as it allows for unobserved heterogeneity among participants whose baseline disfluency could differ substantially. The model will specify the *condition* (Scaffolded vs. Baseline) as a fixed effect while including a random intercept for *participants*, isolating the effect of the AI intervention while accounting for variance in individual speaking styles. This within-subjects design inherently controls for all time-invariant confounders (e.g., a person's baseline L2 proficiency, speech fluency, or personality).

Furthermore, to strengthen the causal claim, I will conduct a pre-treatment check for selection bias. Users may be more likely to engage with the AI when they are *already* experiencing high cognitive load. To test this, I can model the likelihood of a user clicking on a suggestion as a function of their disfluency rate in the preceding turn. Finding no significant relationship would bolster the argument that the scaffolding is an exogenous treatment, strengthening the validity of the LME model results.

4 Discussion and Expected Contributions

This research proposal aims to synthesize computational methods with psycholinguistic theory to provide a more nuanced evaluation of Human-AI collaboration. I anticipate three potential patterns of results, each offering distinct insights into the dynamics of AI-MC.

First, a finding of **High Similarity with Low Satisfaction** would suggest a "Loss of Agency" effect. In this scenario, users are successfully utilizing the tool to survive the conversation but feel alienated by the process, acting as passive spokespeople for the algorithm. This would confirm the concerns raised by Hohenstein and Jung [7] regarding the homogenizing effect of smart replies.

Second, a finding of **High Satisfaction with No Disfluency Reduction** would indicate a "Cognitive Trade-off." The cognitive effort saved on lexical retrieval is immediately reinvested in the new task of reading and evaluating the suggested output. While the user feels supported (high satisfaction), the actual acoustic fluency does not improve because the modality of the cognitive load has shifted from production to AI evaluation, an extra cognitive stage specific to AI-MC.

The third point is the identification of the "**Critical Integration Level**." I hypothesize that the most successful interactions will be characterized by a measurable reduction in search-based disfluencies (e.g., fewer "um"s) combined with moderate text modification. This result would empirically validate the concept of "scaffolding" in AI-MC, demonstrating that the system successfully provides the conversational structure with resources of content and vocabulary upon which the user builds their own authentic communicative acts.

By moving beyond simple performance metrics, self-reported surveys, and subjective user interviews, this proposal advocates for a behavioral approach to evaluating AI systems. The combination of semantic embeddings to measure *what* is said, and disfluency analysis to measure *how* it is said, offers a comprehensive toolkit for understanding the complex dynamics of trust, agency, and cognition in the era of AI-mediated communication.

References

[1] Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech* 44, 2 (2001), 123–147.

[2] Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: a new approach to personality assessment. *Current Opinion in Behavioral Sciences* 18 (2017), 63–68.

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[4] John Gayed, May Kristine Carlon, Angelu Oriola, and Jeffrey Cross. 2022. Exploring an AI-based writing Assistant's impact on English language learners. *Computers and Education Artificial Intelligence* 3 (02 2022), 100055. doi:10.1016/j.caeai.2022.100055

[5] Wanqing Psyche He and Susan R Fussell. 2025. Proactivity in Scaffolding Comprehension and Production in Real-Time Turn-Taking: A Case Study of Bridging Communication Gaps for Non-Native Speakers. In *Companion of the 2025 Computer-Supported Cooperative Work and Social Computing*. ACM.

[6] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[7] Jess Hohenstein and Malte Jung. 2020. AI-supported messaging: an investigation of human-human text conversation with AI support. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[8] Nakhon Kitjaroonchai and Suzanna Maywald. 2024. The Effects of Reading Assistant Software on the Speech Fluency and Accuracy of EFL University Students. *JET (Journal of English Teaching)* 10 (06 2024), 183–197. doi:10.33541/jet.v10i2.5763

[9] Firstname Lastname and Firstname Lastname. 2024. EnronSR: A Benchmark for Evaluating AI-Generated Email Replies. In *Proceedings of the AAAI Conference on Web and Social Media*.

[10] M Leib et al. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* (2023).

[11] Stephen C Levinson. 2016. Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences* 20, 1 (2016), 6–14.

[12] Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences* 27, 2 (2004), 169–190.

[13] Norman Segalowitz. 2010. *Cognitive bases of second language fluency*. Routledge.

[14] Shungo Suzuki and Judit Kormos. 2023. The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition* 45, 1 (2023), 38–64. doi:10.1017/S0272263121000899