That's (not) Correct. Unveiling Cohesive Devices in Conversational AI

Anonymous ACL submission

Abstract

Cohesion is a fundamental element of conversations, reflecting the ability of interlocutors to establish meaningful connections across conversational turns. We investigate the linguistic alignment of cohesive devices in the form of anaphoric references in dyadic dialogues between users and conversational agents. Beyond their role as cohesive device, anaphoric references serve as a meta-commentary. By examining the communicative goals that are conveyed through the lens of this meta-commentary, we uncover notable differences across pragmatic dimensions, including linguistic choices, modulation strategies, and polarity trends. Our investigation reveals a fundamental distinction in the use of cohesive devices: users primarily employ anaphoric references for fact-checking, contrasting with conversational agents focusing on descriptive clarity, often utilizing discursive markers, emotive elements, and moral framing.

1 Introduction

Conversational AI, fueled by advancements in language models (Brown et al., 2020; Chowdhery et al., 2023), has revolutionized human-computer interaction, achieving remarkable proficiency in mastering natural language and performing diverse tasks. Seamlessly integrated into everyday routines, conversational agents [1] have evolved into indispensable tools, supporting users in managing personal, professional, and educational activities.

Cohesion (Halliday and Hasan, 1976) is crucial for conversational interactions, enabling users and agents to navigate and contribute meaningfully to a conversation. However, subtle misalignments in cohesive devices can disrupt conversational flow and lead to frustration (Chaves et al., 2022). While prior studies have explored various aspects of conversational interaction such as linguistic alignment (Koulouri et al., 2016; Duplessis et al., 2021; Spillner and Wenig, 2021; Wang et al., 2023; Sandler et al., 2024) and stylistic variation (Thomas et al., 2018; Elsholz et al., 2019), showing that conversational agents mirror the language of users and adapt their tone to match specific contexts or situations, there is a lack of focused studies on linguistic constructs that underpin cohesion.

To examine linguistic cohesion at scale, we narrow our focus on *anaphoric references* (Garnham, 2016), which serve as cohesive devices that connect discourse to prior conversational turns, promoting coherence and avoiding redundancy. Beyond their cohesive function, we argue that interlocutors frequently utilize anaphoric references to convey meta-commentary on prior conversational turns. This meta-commentary encompasses a spectrum of communicative purposes, ranging from signaling agreement to requesting elaboration.

Contribution. We contribute a detailed profile of anaphoric references in user-agent conversations, acting as a pragmatic lens onto cohesion due to their integral role in linking conversational turns. Specifically, we employ lexico-grammatical pattern matching (Schmid et al., 2021) to extract anaphoric references from `LMSYS-Chat-1M` (Zheng et al., 2023), a large-scale corpus of dyadic dialogues between users and agents in the wild. This scale and authenticity facilitate an linguistic investigation that is representative and grounded in everyday conversational contexts. Enriched by communicative goals embedded within anaphoric references, this offers insights into how users and agents construct and sustain cohesion in multi-turn conversations, serving as the basis for identifying misalignment in their use of cohesive devices.

2 Background

Conversations are a fundamental mode of human communication, characterized by the dynamic ex-

[1] For brevity, we use 'agent' and 'chatbot' interchangeably throughout this study to refer to conversational agents.

change of meaning between interlocutors. Linguistically, conversations are structured as a series of *turns* aimed at establishing mutual understanding (Clark and Schaefer, 1989), guided by principles of cooperation and relevance (Grice, 1975). These principles emphasize that conversations are cooperative efforts where interlocutors recognize a shared purpose or common direction in the dialogue.

The structure of conversations is organized by *turn-taking* (Sacks et al., 1974; Wilson et al., 1984), which allows interlocutors to coordinate their contributions in a structured and orderly manner, ensuring that conversational flow is maintained.

Cohesion plays a central role in conversational turn-taking (Schegloff, 2007), contributing to contextual meaning and logical consistency. Halliday and Hasan (1976) argue that cohesion is often accomplished through cohesive devices such as references substitution, ellipsis, and conjunction.

- **Reference** involves elements whose interpretation is contingent upon their connection to another element. By obviating the need to repeat full details, references through articles, pronouns, demonstratives, and comparatives play a crucial role in reducing cognitive load.

- **Substitution** refers to the act of replacing an element by another element to avoid repetition. Unlike references that connect meanings on the semantic level, substitution typically replaces words on the lexical level.

- **Ellipsis** defines the process of omitting elements, effectively replacing them with nothing. Since the element can be recovered from prior discourse, this process is anaphoric.

- **Conjunctions** serve to organize textual meaning through logical connectors. By marking relations between elements, conjunctions operate at clause level rather than lexical level.

3 Related Work

Nass and Moon (2000) established that users readily anthropomorphize computer systems, applying conversational norms as if interacting with social actors. Hence, this phenomenon was termed *computers are social actors* and laid the groundwork for research into human-computer interaction.

For investigations into linguistic characteristics of human-computer interaction, *discourse analysis* (Paltridge, 2021) became a key method for examining how agents replicate or diverge from natural communication. A central focus of this branch has been on linguistic alignment (Giles et al., 1991), describing the tendency for interlocutors to accommodate each other's linguistic choices and stylistic variations over the course of their conversation, leading to mutual convergence across various levels of linguistic representation. This convergence has been observed in human-human communication across lexical choices (Brennan and Clark, 1996; Metzing and Brennan, 2003), syntactic structures (Branigan et al., 2000; Hartsuiker et al., 2004), and semantic schemes (Garrod and Anderson, 1987).

Numerous studies explored whether this linguistic behavior extends to human-computer interaction (Branigan et al., 2010; Bergmann et al., 2015). Early work by Hill et al. (2015) observed that users adapt their language when interacting with early-stage agents, simplifying their vocabular and employing more directive expressions, seemingly in response to perceived limitations in linguistic comprehension (Pearson et al., 2006). While this linguistic behavior suggests a unidirectional adaptation from users to agents, recent studies highlight a more reciprocal adaptation. Koulouri et al. (2016) found that agents gradually reduce their vocabulary over the course of a conversation, while Duplessis et al. (2021) emphasized alignment through reciprocal repetition of words (Tannen, 1989). Wang et al. (2023) further noticed that modern agents not only mimic user language by reusing their expressions but also actively influence users to repeat certain lexical choices, resulting in shared expressions becoming more frequent within a conversation.

While most studies on human-computer interaction emphasize alignment (Duplessis et al., 2021; Wang et al., 2023) and perception (Mariacher et al., 2021), less attention has been paid to the specific linguistic devices employed by users and agents, especially for constructing and sustaining cohesion.

4 Methodology

To investigate cohesive devices in conversational discourse at scale, we employ lexico-grammatical pattern matching (Schmid et al., 2021) to extract anaphoric references. The choice of anaphoric references as a proxy for cohesion is motivated by their role as a manageable yet meaningful subset of cohesive devices, which enable interlocutors to succinctly and pragmatically express some sort of

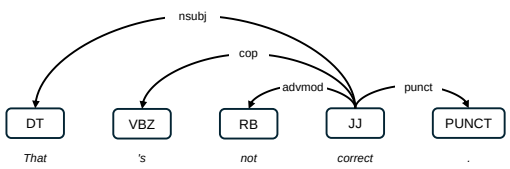

Figure 1: Lexico-grammatical template for automatic extraction of anaphoric references from user-agent conversations.

Figure 2: Distribution of communicative goals based on the interpretation of adjectives, differentiated for users and agents.

comment while referring it to an entire segment of preceding discourse (Halliday and Hasan, 1976).

Common forms of anaphoric references are manifested by *pronouns* or *demonstratives*, linking an expression to a preceding element in discourse. To analyze a pragmatically rich subset of anaphoric references, we targeted phrases in which a demonstrative is followed by the (contracted) form of the copula and an adjective. Figure 1 shows the lexico-grammatical pattern, where each component contributes distinct functions: (1) the demonstrative establishes the anaphoric reference, (2) the adjective carries the communicative purpose, encapsulating a form of feedback, while (3) an optional adverbial modifier (e.g., *not*) adds nuances to this feedback.

Note that we deliberately designed this pattern to balance computational tractability and linguistic interpretability, as its clear formulation ensures consistent and accurate extraction while affording rich insights into pragmatic functions.

We apply our pattern to `LMSYS-Chat-1M` (Zheng et al., 2023), a large-scale corpus of one million dyadic dialogues between users and agents, chosen for its its scale and representativeness. Since `LMSYS-Chat-1M` contains multilingual conversations, we restricted the extraction to conversations held in English. Following this language restriction, we leverage the `spaCy` library to systematically extract instances of anaphoric references that conform to our lexico-grammatical pattern. This extraction process yielded an initial set of 2,146 samples, which were manually reviewed to ensure relevance and appropriateness. After filtering out misleading and irrelevant cases using a blacklist of 44 words, a set of 2002 samples were retained for discourse analysis. For each occurrence, we meticulously recorded the full attestation (e.g., *that's not correct*), the slot-filling adjective (e.g., *correct*), and optionally the adverbial modifier (e.g. *not*).

Once we extracted the anaphoric references, we categorized adjectives and modifiers based on their communicative purpose along with their polarity.

- For the obligatory adjectives, we employ a one-to-one correspondence according to *six* dominant usage types (Schmid et al., 2021): **discursive** (e.g., *okay*) for signaling uptake, **factual** (e.g., *correct*) for judging accuracy, **descriptive** (e.g., *long*) for providing descriptions, **evaluative** (e.g., *amazing*) for expressing opinions, **emotive** (e.g., *frustrating*) for reflecting emotions, and **ethical** (e.g., *mean*) for indicating moral values.

- For the optional modifiers, we draw upon established typologies of adverbial modification (Huddleston and Huddleston, 1988), accounting for *four* dimensions of modulation: **negation** for reversing the meaning (e.g., *not*), **degree** modifiers for adjusting the intensity (e.g., *very* or *kinda*), **scope** modifiers for delimiting focus and continuity (e.g., *just* or *still*), and **tone** modifiers for expressing stance through attitude markers (e.g., *surprisingly*).

5 Findings

Through the lens of anaphoric references, we aim to investigate the alignment and divergence of cohesive devices employed by users and agents in dyadic dialogues. Our investigation reveals remarkable linguistic variations in terms of communicative goals, richness, modulation, and polarity.

5.1 Communicative Goals

We commence with a comparison of communicative goals. Figure 2 illustrates linguistic profiles separate for users and agents, uncovering fundamental deviations in their communicative goals.

We notice that users appear primarily driven by a desire for factual verification ($n = 518$) in tan-

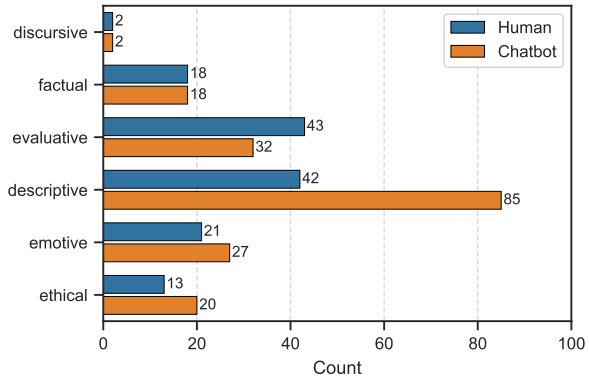

Figure 3: Count of unique adjectives per communicative goal, illustrating lexical richness and diversity for users and agents.

dem with evaluative opinions ($n = 300$), making up 61% of the usage of anaphoric references. This indicates that users actively engage in commenting agent's responses through objective and affective lenses. Conversely, agents demonstrate a more diverse communicative approach when reacting to user responses. While delivering factual statements remain a key component ($n = 336$), agents prioritize descriptive information ($n = 212$) over evaluative statements ($n = 140$). This tendency underscores the agents' role as information providers (Thomas et al., 2018), as they focus on reaffirming their stance by offering supporting clarifications or elaborations. Moreover, agents employ a higher proportion of discursive markers to acknowledge previous turns and maintain cohesion. By conveying emotive reactions and ethical comments, agents further adhere to a communicative style that is inclined to incorporate moral framing.

5.2 Lexical Richness and Diversity

To provide insights into the lexical richness of users and agents, Figure 3 presents the count of unique adjectives employed within each communicative goal. Although users ($n = 139$) and agents ($n = 159$) demonstrate comparable lexical diversity, they allocate their linguistic resources differently in relation to the communicative goals.

We observe nearly identical lexical diversity for discursive and factual language. For discursive language, this uniformity suggests a distinctive conventionality, reflecting a shared reliance on standardized expressions to signal acknowledgment. For factual language, the parity indicates that both groups draw from a similar range of lexical items, suggesting comparable strategies in addressing issues of correctness. However, a remarkable divergence emerges for other communicative goals, revealing a striking difference in lexical diversity. Comparing the vocabulary, users tend to exhibit a higher degree of diversity in their evaluative language compared to agents, suggesting that users tend to express their subjective opinions with a richer and more varied vocabulary, while agents appear to limit their range of evaluative terms. In contrast, agents display substantially more diversity in their descriptive language compared to users, reflecting a pronounced emphasis on utilizing a wider range of adjectives to provide more detailed and varied descriptions. For emotive and ethical language, agents marginally surpass users' lexical richness, albeit within a generally moderate level of diversity. This suggests that agents employ a broader range of expressions to convey affect and moral nuances, although expressions in this category remain relatively constrained overall.

Table 1 corroborates our analysis of richness and diversity by listing the most frequently employed adjectives according to their communicative goal, providing insights into preferences and the potential dominance of specific terms. A closer examination of the distribution of specific terms reveals varying degrees of dominance. Certain communicative goals display a higher degree of lexical concentration, with a single term accounting for a large proportion of occurrences, while other communicative goals lack a dominant term, indicating a greater degree of lexical variation. Accounting for 81%, discursive language is heavily dominated by *okay*. This indicates a shared preference for this simple acknowledgment marker and standard, low-variability responses in establishing and managing conversational flow. The overwhelming frequency of *correct* in factual language also points to a strong consensus in confirming information.

5.3 Modulation Strategies

We continue with the examination of modulation strategies through adverbial modifiers. From a total of 2,002 expressions, users employed 479 modifiers, whereas agents utilized 348 modifiers. While both users and agents utilize a range of modifying elements, this initial observation of modifier usage suggests clear areas of alignment and discrepancy.

Figure 4 shows adverbial elements that serve to modify the *meaning*, *degree*, *scope*, and *tone*.

Regarding meaning modulation, we observe a stark contrast in the prevalence of negation, which is significantly more frequently employed by users

Table 1: Frequently occurring adjectives associated with each communicative goal, highlighting preferences and variations.

Discursive	n	**Factual**	n	**Evaluative**	n	**Descriptive**	n	**Emotive**	n	**Ethical**	n
okay	35	correct	307	awesome	74	ridiculous	36	amazing	63	unacceptable	32
fine	7	incorrect	146	great	60	important	28	frustrating	25	humiliating	13
alright	1	wrong	144	good	37	insane	28	disgusting	16	inappropriate	13
-	-	true	123	bad	32	normal	26	embarrassing	15	appropriate	11
-	-	right	51	incredible	23	crazy	16	exciting	14	toxic	6

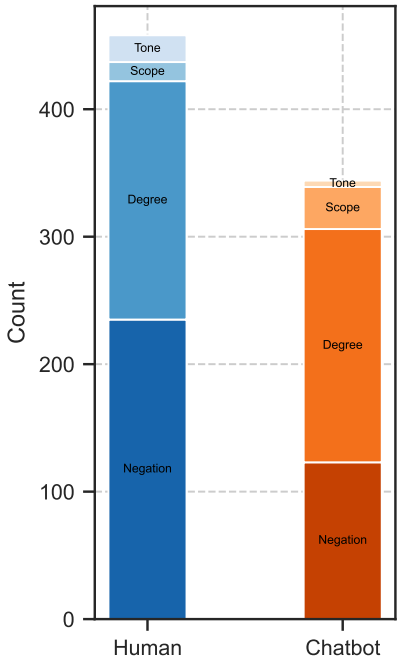

Figure 4: Distribution of modulation strategies derived from adverbial modifiers, shown separately for users and agent.

($n = 235$) than agents ($n = 123$). This more distinctive use of negation underscores the proactive role of users in initiating repair sequences and their willingness to actively engage in discourse correction by refuting or rectifying prior responses.

Degree modulation, as realized through intensifiers and quantifiers, serves to amplify or attenuate meaning. We find that users ($n = 187$) and agents ($n = 183$) employ these modifiers at comparable rates, indicating a shared strategy in adjusting their utterances while maintaining a cooperative tone.

Examining scope modulation, we observe that users favor continuity modifiers ($n = 12$), while agents exhibit a clear preference for focus modifiers ($n = 27$). This contrast implies that users reinforce narrative coherence by acknowledging the ongoing discourse, whereas agents emphasize references to pinpoint the position within a discourse.

Concerning tone modulation, we note that users ($n = 21$) employ attitude markers considerably more frequently than agents ($n = 4$). This pronounced asymmetry in the utilization of attitude modifiers suggests that users readily integrate affective signals into their communication, enriching the emotional texture of the dialogue. Conversely, agents appear to adopt an objective stance regarding tone, reflective of design choices aimed at maintaining objectivity and reducing ambiguity.

The differential pattern in modifier usage reflects distinct communicative goals. While users show a tendency to provide contextual nuance, agents prioritize clarity and establishing logical connections.

5.4 Polarity Trends

Figure 5 reveals a shared polarity trend between users and agents, with few distinctions across specific communicative goals. Discursive expressions generally carry positive polarity, which can be attributed to their function as uptake markers signaling acknowledgment. However, the polarity for factual expressions diverge. We derive that users display facts in a negative tone, reflecting their tendency to correct errors and challenge misinformation, whereas agents maintain a predominantly neutral tone. In comparison, both groups demonstrate varied polarity for evaluative, descriptive, and emotive expressions. This polarity distribution suggests that sentiment in these categories is not fixed but rather adjusted based on conversational context. Ethical expressions, while generally neutral, show occasional strong polarity shifts, indicating moments of pronounced moral judgment.

Despite an overall similarity in polarity between users and agents, our findings suggest certain differences in epistemic approach. While users employ factual statements mainly for criticism rather than endorsement, agents prioritize affirmation and engagement through expressions with neutral stance.

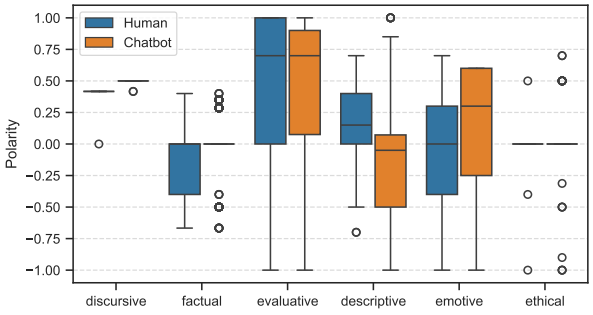

Figure 5: Distribution of polarity trends across communicative goals, calculated separately for users and agents.

6 Conclusion

Cohesion, central to effective user-agent communication, reveals notable pragmatic divergence when viewed through the lens of anaphoric reference. Examining communicative goals, reflected in metacommentary, unveils a fundamental functional contrast: while both users and agents utilize factual verification through anaphoric reference, users further convey subjective opinions, whereas agents prioritize clarity by providing objective descriptions. This asymmetry in the deployment of cohesive devices underscores a critical linguistic misalignment in user-agent conversational systems.

Limitations. We recognize that this study is subject to design choices limiting its generalizability.

One key limitation is our reliance on a specific lexico-grammatical pattern for extracting cohesive devices. While this methodological approach ensures computational efficiency and consistency in identifying anaphoric references, it inevitably excludes other forms of cohesion that do not adhere to this specific pattern. Expanding the scope by incorporating other grammatical constructions, such as interrogatives or imperatives, could provide a more comprehensive view on conversational coherence.

Another notable constraint is the lack of demographic differentiation. Our analysis does not account for potential variations in linguistic choices influenced by factors such as language proficiency, communication style, educational background or cultural norms. Without access to metadata on user demographics, our findings remain generalizable only within a undifferentiated population.

References

Kirsten Bergmann, Holly P Branigan, and Stefan Kopp. 2015. Exploring the alignment space–lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, 2:7.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of pragmatics*, 42(9):2355–2368.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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):1–38.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frédéric Landragin. 2021. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55:353–388.

Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 301–305.

Alan Garnham. 2016. Reference and anaphora. In *Language processing*, pages 335–362. Psychology Press.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:43–58.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge.

Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological science*, 15(6):409–414.

Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250.

Rodney Huddleston and Rodney D Huddleston. 1988. *English grammar: An outline*. Cambridge University Press.

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):59–95.

Natascha Mariacher, Stephan Schlögl, and Alexander Monz. 2021. Investigating perceptions of social intelligence in simulated human-chatbot interactions. *Progresses in Artificial Intelligence and Neural Systems*, pages 513–529.

Charles Metzing and Susan E Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Brian Paltridge. 2021. *Discourse analysis*. Springer.

Jamie Pearson, Jiang Hu, Holly P Branigan, Martin J Pickering, and Clifford I Nass. 2006. Adaptive language behavior in hci: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1177–1180.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simple systematic for the organisation of turn taking in conversation. *Language*, 50:696–735.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations. *arXiv preprint arXiv:2401.16587*.

Emanuel A Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press.

Hans-Jörg Schmid, Quirin Würschinger, Sebastian Fischer, and Helmut Küchenhoff. 2021. That's cool. computational sociolinguistic methods for investigating individual lexico-grammatical variation. *Frontiers in Artificial Intelligence*, 3:547531.

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*, pages 1–12.

Deborah Tannen. 1989. *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse*. Cambridge University Press.

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*, pages 42–51.

Boxuan Wang, Mariët Theune, and Sumit Srivastava. 2023. Examining lexical alignment in human-agent conversations with gpt-3.5 and gpt-4 models. In *International Workshop on Chatbot Research and Design*, pages 94–114. Springer.

Thomas P Wilson, John M Wiemann, and Don H Zimmerman. 1984. Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3(3):159–183.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Twelfth International Conference on Learning Representations*.