

<https://doi.org/10.25312/j.10781>

Karolina Jankowska  <https://orcid.org/0000-0002-5023-8952>
Polish Phonetics Association
e-mail: karolina.jankowska@wsjo.pl

Communicative convergence in human–LLM written dialogues: A review of linguistic adaptation in interactions with large language models

Konwergencja komunikacyjna w dialogach pisanych człowiek–LLM – przegląd adaptacji językowej w interakcjach z dużymi modelami językowymi

Abstract

This review article provides a comparative analysis of communicative convergence in written dialogues between humans and Large Language Models (LLMs). It draws on theoretical work on communication accommodation and interactive alignment as well as empirical research on linguistic adaptation in human–computer and human–LLM interaction. The reviewed studies address lexical, syntactic, stylistic, pragmatic, and ideological alignment, together with the effects of prompt wording and dialogue history on model responses. The comparison shows that similar manifestations of convergence arise from different mechanisms: in humans, they may reflect affiliation, priming, and interactional coordination, whereas in LLMs they result from probabilistic text generation and sensitivity to context. Convergence is bidirectional but asymmetrical: users adapt their language to the perceived characteristics and constraints of the system, while models reproduce or amplify cues present in user input. Apparent alignment should not be treated as evidence of communicative intention or shared understanding. Instead, human–LLM convergence is best understood as an interactional effect jointly produced by human social-cognitive mechanisms and the model’s contextual responsiveness.

The review also identifies a shortage of longitudinal research and studies involving morphologically rich languages, including Polish.

Keywords: artificial intelligence, communicative convergence, human–LLM interaction, language adaptation, language convergence, Large Language Models (LLMs)

Streszczenie

Artykuł ma charakter przeglądowo-analityczny i dotyczy konwergencji komunikacyjnej w pisanych dialogach między ludźmi a dużymi modelami językowymi (LLM). Podstawę analizy stanowią prace teoretyczne nad akomodacją komunikacyjną i interakcyjnym dopasowaniem oraz badania empiryczne dotyczące adaptacji językowej w interakcjach człowiek–komputer i człowiek–LLM. Analizowane badania obejmują dopasowanie leksykalne, składniowe, stylistyczne, pragmatyczne i ideologiczne, a także wpływ sposobu sformułowania promptu oraz historii dialogu na odpowiedzi modeli. Porównanie pokazuje, że podobne przejawy konwergencji wynikają z odmiennych mechanizmów: u ludzi wiążą się m.in. z afiliacją, *primingiem* i koordynacją interakcji, natomiast w LLM są efektem probabilistycznego generowania tekstu i wrażliwości na kontekst. Konwergencja ma charakter dwukierunkowy, lecz asymetryczny: użytkownicy dostosowują język do postrzeganych cech i ograniczeń systemu, a modele odtwarzają lub wzmacniają sygnały zawarte w wypowiedziach użytkownika. Pozornego dopasowania nie należy utożsamiać z intencją komunikacyjną ani wspólnym rozumieniem. Zjawisko to najlepiej ujmować jako efekt interakcyjny współtworzony przez społeczno-poznawcze mechanizmy człowieka oraz kontekstową responsywność modelu. Przegląd wskazuje również na niedostatek badań podłużnych i dotyczących języków bogatych morfologicznie, w tym polszczyzny.

Słowa kluczowe: adaptacja językowa, duże modele językowe (LLM), interakcja człowiek–LLM, konwergencja komunikacyjna, konwergencja językowa, sztuczna inteligencja

Introduction

The rapid proliferation of Large Language Models (LLMs) has transformed communication between humans and machines. Systems such as ChatGPT, Claude, Gemini, and Llama are increasingly embedded in education, professional writing, programming, healthcare communication and everyday information retrieval. These interactions are overwhelmingly text-based and dialogic in structure. As such, they provide a unique context for examining whether established linguistic theories of communicative convergence extend beyond human interlocutors. Communicative convergence refers to the process by which interlocutors adapt aspects of their linguistic behaviour to one another during interaction. Such adaptation may involve lexical choice, syntactic structure, speech rate, discourse markers, politeness strategies, or broader stylistic parameters (Giles, Coupland, Coupland, 1991). In human communication, convergence is often motivated by social alignment, cooperation, identity negotiation, or cognitive priming mechanisms (Pickering, Garrod, 2004).

The emergence of LLM-mediated dialogue raises fundamental theoretical questions concerning whether statistical language models can exhibit convergence, whether humans accommodate to AI systems, whether convergence effects are symmetrical or asymmetrical, how these phenomena vary across languages and what implications they hold for morphologically rich languages such as Polish. This review addresses these issues through an interdisciplinary synthesis of linguistic, computational, and human–LLM interaction research.

Theoretical foundations of communicative convergence

Communication Accommodation Theory (CAT)

Communication Accommodation Theory (CAT) is a socio-psychological framework that explains how and why people modify their communicative behaviour during interactions with others. The theory argues that individuals adapt the way they speak or write depending on their interlocutors, with these adjustments being influenced by social identity, relational objectives, and situational context (Giles, Coupland, Coupland, 1991; Giles, Ogay, 2007).

Accommodation can occur both consciously and unconsciously. Speakers may intentionally adjust their style to gain approval or improve clarity, but many changes also happen automatically through cognitive processes and social alignment mechanisms. Within CAT, a key distinction is made between convergence, where individuals become more similar in their communication style, and divergence, where differences are emphasised (Giles, Coupland, Coupland, 1991; Giles, Ogay, 2007). Convergence is typically associated with a desire for social acceptance, affiliation, or mutual understanding, while divergence may function to express social distance or reinforce distinct group identity.

Communication accommodation is especially common in contexts involving status disparities, intergroup interaction, intercultural communication, or institutional relationships. Convergent behaviour can be seen in features such as accent, vocabulary choice, syntax, and discourse structure. In written communication, it may appear through changes in register, word selection, sentence complexity, tone, use of emojis or punctuation, and even formatting conventions. Writers often mirror lexical choices, syntactic patterns, or levels of formality, which is particularly evident in online communication and increasingly relevant in human–LLM interaction. Empirical findings show evidence of lexical alignment even in very simple interactive settings (Brennan, Clark, 1996).

Interactive Alignment Model

The Interactive Alignment Model proposes that effective conversation results from an automatic process of alignment between participants across several levels of representation, including lexical, syntactic, semantic, and situational aspects (Pickering, Garrod, 2004). The model suggests that this alignment is primarily driven by automatic priming mechanisms, in which exposure to a specific linguistic form increases the probability of reusing it later, which in turn reduces cognitive load and supports smoother mutual understanding (Pickering, Garrod 2004; Garrod, Pickering, 2009).

Initial experimental research strongly supports this mechanism, especially through evidence of syntactic priming between interlocutors (Branigan, Pickering, Cleland, 2000), demonstrating that speakers unconsciously replicate each other's sentence structures. Later studies have broadened these findings to include lexical and phonological alignment (Brennan, Clark, 1996; Horton, 2017), showing that alignment operates across multiple layers of language in both spoken and written communication. Taken together, this body of research indicates that alignment is not purely a deliberate communicative strategy but also a partially automatic cognitive process emerging from interactive dynamics and shared mental representations.

Furthermore, alignment appears to affect not only language production but also comprehension, enabling interlocutors who are aligned to predict and interpret each other's utterances more efficiently (Healey, Purver, Howes, 2014). When applied to artificial intelligence, the model raises significant questions about whether transformer-based language models reproduce human-like alignment patterns through probabilistic continuation of contextual cues and to what extent such behaviour resembles the adaptive, interactive mechanisms found in human dialogue (Jurafsky, Martin 2009; Vaswani et al., 2017). Exploring these similarities and differences is essential for assessing how AI systems may achieve alignment that facilitates effective communication and shared understanding.

Convergence in written dialogue and lexical entrainment in human–LLM interaction

Linguistic convergence in written language refers to the tendency of interlocutors to adapt their communicative behaviour to one another across multiple levels of textual expression. In written interaction, convergence may involve adjustments in lexical choice, syntactic structures, stylistic features, discourse organization, level of formality, punctuation, and formatting conventions. One well-documented phenomenon within this domain is lexical entrainment, understood as the process by which interlocutors gradually adopt shared referential expressions over the course of interaction (Brennan, Clark, 1996). Through repeated usage, specific terms or descriptions become jointly established between participants, which helps to minimise ambiguity and lower cognitive effort, while also strengthening a shared mutual understanding.

Importantly, convergence is not restricted to spoken interaction. Research in computer-mediated communication has shown that stylistic convergence also occurs in asynchronous, text-based environments (Danescu-Niculescu-Mizil, Gamon, Dumais, 2011). These results suggest that alignment processes remain strongly active in written discourse, even in the absence of prosodic features and nonverbal cues that are typically present in face-to-face communication. In digitally mediated settings, convergence can appear through the repetition of terminology, reuse of salient phrases, similarity in sentence complexity, or the adoption of comparable pragmatic strategies such as hedging or directness.

These patterns are especially relevant in human–LLM communication, which is predominantly text-based. In computational dialogue systems, phenomena similar to lexical entrainment may arise through conditioning on previous conversational context, as language models generate responses probabilistically based on earlier expressions and surrounding discourse patterns. As a result, interaction with language models in written form offers a useful environment for studying convergence as both a structural and distributional property of text.

Large Language Models: Architecture, training, and communicative capacities

Understanding communicative convergence in human–LLM dialogue requires, at the very least, a general description of how Large Language Models work. Although LLMs are often explained using technical jargon, from a linguistic perspective they can be seen as probabilistic systems trained on massive datasets to capture patterns of how linguistic elements co-occur. Their apparent ability to engage in conversation does not arise from any underlying communicative intention, but instead emerges from statistical regularities learned during training and reflected in their generated outputs.

Transformer architecture and context modelling

Contemporary Large Language Models are based on the transformer architecture introduced by Vaswani et al. (2017), which marked a major shift in Natural Language Processing. Earlier neural models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) processed text sequentially, meaning each word depended on the previous ones (Hochreiter, Schmidhuber, 1997). This approach worked well for many tasks, but it made it harder to capture long-range dependencies and limited the ability to process long texts efficiently. LSTMs were designed to reduce these problems, but they still rely on sequential processing, which constrains scalability compared to transformers.

Transformers depart from this sequential paradigm by introducing a mechanism known as ‘self-attention’. Self-attention allows the model to compute weighted relationships between all tokens within an input sequence simultaneously, rather than in strictly linear order. In practical terms, when generating a response in a dialogic exchange, the model does not rely solely on the immediately preceding word or sentence. Instead, it evaluates the entire accessible context and dynamically assigns relevance weights to prior elements. This capacity enables the system to model long-distance syntactic dependencies, thematic continuity, and discourse-level coherence with far greater efficiency than earlier architectures (Vaswani et al., 2017).

Subsequent refinements, such as bidirectional contextual encoding in BERT (Devlin et al., 2018) and large-scale autoregressive scaling in GPT-style models (Brown et al., 2020), demonstrated that the transformer framework is highly adaptable across linguistic tasks. Importantly, scaling laws in neural language modelling revealed that increasing model parameters and training data systematically improves contextual prediction capacity, thereby enhancing coherence and stylistic stability in generated discourse (Kaplan et al., 2020).

From a linguistic perspective, the implications of transformer-based context modelling are substantial. The ability to maintain distributed representations of prior discourse allows LLMs to preserve topic continuity, track referential expressions across turns, and replicate stylistic or terminological patterns introduced by the user. In extended dialogues, models

can maintain lexical consistency and discourse framing across multiple conversational turns, a feature previously difficult to achieve computationally.

At the same time, it is important to note that the *context* modelled by transformers is statistical rather than conceptual. The architecture encodes patterns of token co-occurrence within high-dimensional vector spaces, enabling semantic similarity to emerge from distributional properties of language (Mikolov et al., 2013). Thus, when a model appears to preserve thematic or stylistic coherence, it does so through probabilistic weighting mechanisms rather than through mental representation or discourse planning in the human sense.

In dialogic settings, this architecture makes large-scale contextual sensitivity computationally feasible. What may appear to users as stylistic awareness, responsiveness, or dialogic alignment is grounded in attention-based weighting of prior linguistic material. The transformer does not recognise alignment as a communicative strategy. Rather, it amplifies patterns present in the input sequence that are statistically predictive of coherent continuation. The outcome, such as consistent terminology, register stability, referential tracking, closely resembles interactional alignment phenomena described in human dialogue research (Pickering, Garrod, 2004).

In this sense, the transformer architecture provides the structural precondition for observable convergence in human–LLM written dialogue. Without global context modelling and self-attentive weighting, lexical and stylistic mirroring across turns would be substantially weaker. The architecture therefore constitutes not merely a technical substrate but a foundational element shaping the communicative affordances of contemporary Large Language Models.

Training paradigms and alignment procedures

The communicative behaviour of Large Language Models is shaped not only by architectural design, but also by the procedures through which these systems are trained and subsequently aligned with human expectations. While the transformer architecture provides the structural capacity for contextual modelling, it is the training paradigm that determines how this capacity is realised in actual interaction.

Most contemporary LLMs undergo an initial phase of large-scale self-supervised training. During this stage, models are exposed to massive corpora of written text and trained to predict the most probable continuation of a sequence given its preceding context (Brown et al., 2020). This objective does not involve explicit annotation of semantic meaning, pragmatic function, or communicative intention. Rather, the system internalises statistical regularities across linguistic data, capturing distributional patterns of lexical co-occurrence, syntactic structure, and discourse sequencing. From a linguistic point of view, this process operationalises a radical version of the *distributional hypothesis* (Harris, 1954) where meaning emerges from patterns of contextual usage. However, it is important to emphasise that such distributional learning yields probabilistic representations, not conceptual understanding. The model encodes correlations among tokens within high dimensional vector spaces but does not construct mental models of interlocutors, intentions, or shared communicative goals.

Following pretraining, many state-of-the-art systems undergo additional alignment procedures designed to refine their interactional behaviour. One widely adopted method is instruction tuning, in which models are fine-tuned on curated datasets of prompts paired with high quality responses (Chung et al., 2022). This process enhances the model’s ability to respond coherently to task-oriented queries and conversational instructions. A further stage often involves Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022). In RLHF, human evaluators rank model outputs according to criteria such as helpfulness, clarity, harmlessness and relevance. These preferences are then incorporated into the model’s optimisation process, effectively reshaping its probability distributions to favour outputs that align with human normative judgments. Alignment, in this technical sense, refers not to social synchronisation, but to preference optimisation within a reinforcement learning framework.

This distinction is critical for understanding communicative convergence in human–LLM dialogue. Alignment procedures encourage models to produce responses that are contextually appropriate and perceived as responsive by users. Statistically, this increases the likelihood of lexical reuse, structural mirroring, and stylistic continuity-features commonly interpreted as signs of conversational alignment. However, these effects arise not from intentional accommodation but from the optimisation of response likelihood under human-evaluated reward signals. Moreover, alignment training can amplify context sensitivity. Because human evaluators tend to reward outputs that explicitly address the user’s phrasing and conceptual framing, models learn that reusing salient lexical items or maintaining consistent terminology is advantageous. In this way, convergence becomes instrumentally reinforced during training. The system does not *decide* to align, rather, alignment patterns are probabilistically favoured outcomes within the learned response distribution. Recent discussions in AI alignment research emphasise that aligning models with human values remains an ongoing technical challenge grounded in reward modelling and preference optimisation rather than in the development of intentional agency (Amodei et al., 2016; Bommasani et al., 2021; Ouyang et al., 2022). From a linguistic perspective, this underscores the fact that LLM convergence operates within the constraints of reward optimisation and large-scale statistical conditioning rather than social cognition (Bender et al., 2021). The model’s responsiveness reflects statistical adaptation to evaluative feedback and contextual input, not interpersonal negotiation, theory of mind, or shared intentionality (Pickering, Garrod, 2004).

Consequently, when lexical or stylistic convergence is observed in human–LLM interaction, it should be interpreted as an emergent property of large-scale probabilistic conditioning combined with preference-based fine-tuning. The training paradigm predisposes the model toward contextual coherence and terminological consistency, thereby producing outputs that resemble accommodation phenomena documented in human dialogue research. Yet the underlying mechanisms differ fundamentally, convergence in LLMs is computationally induced rather than socially motivated. Understanding this distinction is essential for theoretical clarity. Without attention to training paradigms and alignment procedures, empirical observations of convergence risk being misattributed to cognitive or communicative capacities that the system does not, in fact, possess.

Emergent dialogue behaviour and apparent adaptation

Scaling studies have demonstrated that increasing model size and training data volume leads to the emergence of capabilities not explicitly programmed into the system (Brown et al., 2020; Chung et al., 2022). These scaling laws suggest that qualitative shifts in performance may arise from quantitative increases in parameters and data exposure. Among the capabilities associated with such scaling are few-shot learning, contextual coherence across extended passages, stylistic modulation, and improved discourse tracking (Bommasani et al., 2021; Chowdhery et al., 2022). In dialogue settings, these emergent properties can give rise to behaviour that resembles conversational alignment – models maintain terminological consistency, adjust register, and reproduce discourse structure introduced by the user, often across multiple turns.

However, it is crucial to distinguish between functional resemblance and cognitive equivalence. In human interaction, convergence is linked to social cognition, cooperative intention, and the negotiation of interpersonal alignment (Giles, Coupland, Coupland, 1991; Pickering, Garrod, 2004). Experimental research on dialogue demonstrates that alignment emerges through interactive priming mechanisms embedded in shared communicative contexts. In contrast, LLMs operate without mental states, intentions, or awareness of interlocutors. Their adaptation is the byproduct of conditional probability optimisation within a high-dimensional representational space (Bender et al., 2021). This distinction has theoretical consequences. If convergence in human dialogue is partly motivated by social affiliation, cooperative goals, or communicative efficiency, convergence in LLM output must be conceptualised differently. It is not accommodation in the sociolinguistic sense but rather context-sensitive statistical continuation shaped by training objectives and optimisation constraints (Amodei et al., 2016). Yet from the user’s perspective, the observable effect–lexical mirroring, stylistic consistency, discourse coherence–can be strikingly similar to human alignment. Studies of human perception of AI-generated text suggest that users readily attribute intentionality and communicative awareness to systems exhibiting coherent stylistic adaptation (Nass, Lee, 2000).

The resulting ambiguity lies at the heart of research on communicative convergence in human–LLM interaction as the system produces outputs that behave as if they were aligned, while lacking the social-cognitive mechanisms that underlie human accommodation. Understanding this tension between emergent functional alignment and the absence of socio-cognitive grounding is essential for interpreting empirical findings on lexical, syntactic, and stylistic convergence in subsequent sections.

Recent studies on convergence in human–computer interaction

The tendency of users to adapt their linguistic behaviour to technology is strongly grounded in the Computers Are Social Actors (CASA) paradigm. Nass and Moon (2000) demonstrate that humans automatically apply social rules to computers, even when fully aware

that they are non-human systems. In their experiments, users applied politeness norms, attributed social competence, and engaged in interaction patterns typical of human–human communication (Nass, Moon, 2000). This automatic activation of social cognition provides a foundational explanation for why convergence processes emerge so readily in human–LLM interaction.

Building on this framework, research on linguistic convergence and alignment shows that interlocutors adapt to each other’s lexical, syntactic, and pragmatic patterns. In the context of AI systems, this phenomenon becomes multidimensional and increasingly pronounced with modern Large Language Models. Park et al. (2023) demonstrate that LLMs exhibit strong social alignment to user style, including syntactic mirroring and tonal adaptation, which can lead to sycophantic behaviour, where models reinforce user beliefs even when they are factually incorrect (Park et al., 2023). Similarly, Santurkar and other authors (2023) show that LLM outputs are highly sensitive to prompt framing and tend to reflect the values, opinions, and ideological stance embedded in user input, indicating that alignment extends beyond linguistic form into the semantic and normative domain (Santurkar et al., 2023).

From the user side, van der Goot and other authors (2021) demonstrate that humans also exhibit strong adaptation to perceived agent personality. When chatbots are designed as polite, formal, or highly structured, users systematically adjust their own writing style, becoming more precise, more formal and more polite in their prompts (van der Goot et al., 2021). This indicates that alignment is bidirectional, with both humans and systems adapting dynamically depending on perceived conversational roles.

Further evidence comes from *Same Meaning, Different Scores: Lexical and Syntactic Sensitivity in LLM Evaluation* (Kostić et al., 2026) which shows that Large Language Models can react in a surprisingly strong way to relatively small changes in wording or sentence structure, even when the overall meaning of the text remains the same. In other words, if a sentence is rewritten using synonyms or slightly reorganised syntactically, the model may produce noticeably different outputs or assign different scores, despite the semantic content being equivalent. This suggests that LLMs are not fully stable in how they process meaning and can be influenced by surface-level linguistic form. The study therefore indicates that these systems rely not only on deep semantic understanding, but also heavily on patterns present in the exact wording and structure of input text (Kostić et al., 2026).

Beyond structural and lexical adaptation, alignment also operates at the pragmatic and cognitive level. Recent work on human–LLM collaboration shows that iterative prompt – response interaction leads to shared understanding of task goals and gradual convergence of user and system mental models. Over time, users refine their expectations of system capabilities, while models increasingly adapt to inferred intent, resulting in tighter coordination of goals and interpretation between humans and AI systems (Ji et al., 2025; Shen et al., 2025).

In voice-based interaction, convergence extends into the acoustic-prosodic domain. Users adapt speech rate, pitch, and intensity to synthetic voices. Lubold, Walker, and Pon-Barry (2016) show that pitch-matching in robotic tutoring increases trust and improves

learning outcomes (Lubold, Walker, Pon-Barry, 2021). Cowan and other authors (2015) further demonstrate system-led priming effects, where system constraints lead users to shorten utterances and simplify syntax, effectively aligning communication toward machine-imposed structures (Cowan et al., 2015).

With the emergence of Large Language Models, convergence has shifted from local stylistic adaptation to broader structural and ideological homogenization. McGuire and other authors (2024) show that sustained interaction with models such as GPT-4 leads to long-term convergence toward the model’s average stylistic output, suggesting that linguistic adaptation may extend beyond individual dialogues into persistent changes in user writing behaviours (French, D’Mello, von der Wense, 2024). Complementary studies on neural dialogue systems (Hoegen et al., 2019) and OpenReview findings (Srivastava et al., 2023) confirm that alignment operates across multiple linguistic levels, including discourse structure, argument organisation, and information density.

Further evidence for the mechanisms underlying linguistic convergence comes from research on automated analysis of L2 learner production, which highlights how systematically users adapt their linguistic output under feedback-driven conditions. The study demonstrates that language learners consistently reproduce and stabilise recurring error patterns that emerge from interaction with language processing systems and structured feedback environments, suggesting that linguistic behaviour is highly sensitive to external correction signals and input regularities. Importantly, automated analysis reveals that learner language is not random but shows predictable convergence toward simplified and system-influenced forms, particularly in lexical choice and morphosyntactic structure. These findings indicate that exposure to technologically mediated linguistic input can shape production patterns in a way that mirrors alignment effects observed in human–LLM interaction, reinforcing the idea that adaptation is driven by both cognitive constraints and interactional feedback loops (Li, 2010; Mizumoto, 2025).

Summary

This review has synthesised interdisciplinary research on communicative convergence in written human–LLM interactions, integrating insights from sociolinguistics, psycholinguistics, computational linguistics, and human–LLM interaction. It has examined whether alignment phenomena known from human–human dialogue, such as lexical, syntactic, stylistic, and pragmatic convergence, also occur in interactions involving Large Language Models.

Theoretical frameworks, including Communication Accommodation Theory and the Interactive Alignment Model, suggest that convergence in human communication is driven by social, cognitive, and interactive mechanisms. In contrast, LLMs generate text through probabilistic prediction based on large-scale training data and alignment procedures such as instruction tuning and reinforcement learning from human feedback. While these systems can produce outputs that resemble human-like adaptation, the underlying mechanisms are fundamentally different, they are statistical rather than social or intentional.

Empirical research reviewed in this paper indicates that convergence in human–LLM interaction is observable at multiple linguistic levels, particularly lexical and stylistic alignment. However, this convergence is asymmetrical. Humans tend to adapt to perceived conversational agents, while LLMs adjust outputs based on contextual conditioning and optimisation objectives rather than communicative intent. Evidence from human–LLM interaction studies also shows that alignment extends beyond form into pragmatic and ideological dimensions, although such effects are highly sensitive to prompt structure and interaction design.

A notable gap in the literature concerns cross-linguistic variation, especially in morphologically rich languages such as Polish and other Slavic languages, where structural complexity may significantly influence alignment dynamics.

Conclusion

The analysis suggests that communicative convergence in human–LLM interaction is largely a surface-level phenomenon. Although large language models can reproduce patterns of lexical, syntactic, and stylistic alignment similar to those found in human dialogue, the underlying mechanisms are fundamentally different.

In human communication, convergence arises from shared intentionality, social affiliation, and interactive cognitive processes. In contrast, LLM-generated convergence results from probabilistic modelling, large-scale data exposure, and alignment optimization. This means that similar observable behaviours emerge from distinct foundations. Recognizing this distinction is important to avoid anthropomorphizing machine communication and to better interpret linguistic adaptation in human–LLM systems.

Future research should move beyond short-term, English-centred studies and focus on more diverse, long-term contexts. Morphologically rich languages like Polish provide valuable opportunities to examine convergence not only at the lexical level, but also in morphology, syntax, and pragmatics. It is also important to investigate how users and models adapt to each other over time, including changes in writing style and language use.

Longitudinal and multilingual studies, especially those involving Polish–English mixing, could reveal deeper patterns of convergence. Additionally, more advanced methodological tools are needed to capture these effects beyond simple lexical similarity, helping to build a more comprehensive theory of interaction in the era of Large Language Models.

References

- Amodei D., Olah C., Steinhardt J., Christiano P., Schulman J., Mané D. (2016), *Concrete problems in AI safety*, “arXiv”, <https://doi.org/10.48550/arXiv.1606.06565>
- Bender E.M., Gebru T., McMillan-Major A., Shmitchell S. (2021), *On the dangers of stochastic parrots: Can language models be too big?*, [in:] *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, pp. 610–623, <https://doi.org/10.1145/3442188.3445922>

- Bommasani R. et al. (2021), *On the opportunities and risks of foundation models*, “arXiv”, <https://doi.org/10.48550/arXiv.2108.07258>
- Branigan H.P., Pickering M.J., Cleland A.A. (2000), *Syntactic co-ordination in dialogue*, “Cognition”, vol. 75(2), pp. B13–B25, [https://doi.org/10.1016/S0010-0277\(99\)00081-5](https://doi.org/10.1016/S0010-0277(99)00081-5)
- Brennan S.E., Clark H.H. (1996), *Conceptual pacts and lexical choice in conversation*, “Journal of Experimental Psychology: Learning, Memory, and Cognition”, vol. 22(6), pp. 1482–1493, <https://doi.org/10.1037/0278-7393.22.6.1482>
- Brown T.B. et al. (2020), *Language models are few-shot learners*, “arXiv”, <https://doi.org/10.48550/arXiv.2005.14165>
- Chowdhery A. et al. (2022), *PaLM: Scaling language modeling with Pathways*, “arXiv”, <https://doi.org/10.48550/arXiv.2204.02311>
- Christiano P., Leike J., Brown T.B., Martic M., Legg S., Amodei D. (2017), *Deep reinforcement learning from human preferences*, “arXiv”, <https://doi.org/10.48550/arXiv.1706.03741>
- Chung H.W. et al. (2022), *Scaling instruction-finetuned language models*, “arXiv”, <https://doi.org/10.48550/arXiv.2210.11416>
- Cowan B.R., Branigan H.P., Obregón M., Bugis E., Beale R. (2015), *Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human–computer dialogue*, “International Journal of Human-Computer Studies”, vol. 83, pp. 27–42, <https://doi.org/10.1016/j.ijhcs.2015.05.008>
- Danescu-Niculescu-Mizil C., Gamon M., Dumais S. (2011), *Mark my words!: Linguistic style accommodation in social media*, [in:] S. Srinivasan, K. Ramamritham, A. Kumar, M.P. Ravindra, E. Bertino, R. Kumar (eds.), *Proceedings of the 20th International Conference on World Wide Web*, New York: Association for Computing Machinery, pp. 745–754, <https://doi.org/10.1145/1963405.1963509>
- Devlin J., Chang M.-W., Lee K., Toutanova K. (2018), *BERT: Pre-training of deep bidirectional transformers for language understanding*, “arXiv”, <https://doi.org/10.48550/arXiv.1810.04805>
- French D., D’Mello S., von der Wense K. (2024), *Aligning to adults is easy, aligning to children is hard: A study of linguistic alignment in dialogue systems*, [in:] N. Soni, L. Flek, A. Sharma, D. Yang, S. Hooker, H.A. Schwartz (eds.), *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, Kerville: Association for Computational Linguistic, pp. 81–87, <https://doi.org/10.18653/v1/2024.hucllm-1.7>
- Garrod S., Pickering M.J. (2009), *Joint action, interactive alignment, and dialog*, “Topics in Cognitive Science”, vol. 1(2), pp. 292–304, <https://doi.org/10.1111/j.1756-8765.2009.01020.x>
- Giles H., Coupland N., Coupland J. (1991), *Accommodation theory: Communication, context, and consequence*, [in:] H. Giles, J. Coupland, N. Coupland (eds.), *Contexts of accommodation*, Cambridge: Cambridge University Press, pp. 1–68, <https://doi.org/10.1017/CBO9780511663673.001>
- Giles H., Ogay T. (2007), *Communication accommodation theory*, [in:] B.B. Whaley, W. Samter (eds.), *Explaining communication: Contemporary theories and exemplars*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 293–310.
- Goot van der R., Sharaf I., Imankulova A., Üstün A., Stepanović M., Ramponi A., Khairunnisa S.O., Komachi M., Plank B. (2021), *From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding*, [in:] K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell,

- T. Chakraborty, Y. Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg: Association for Computational Linguistics, pp. 2479–2497, <https://doi.org/10.18653/v1/2021.naacl-main.197>
- Harris Z.S. (1954), *Distributional structure*, “Word”, vol. 10(2–3), pp. 146–162, <https://doi.org/10.1080/00437956.1954.11659520>
- Healey P.G.T., Purver M., Howes C. (2014), *Divergence in dialogue*, “PLoS ONE”, vol. 9(6), article e98598, <https://doi.org/10.1371/journal.pone.0098598>
- Hochreiter S., Schmidhuber J. (1997), *Long short-term memory*, “Neural Computation”, vol. 9(8), pp. 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hoegen R., Aneja D., McDuff D., Czerwinski M. (2019), *An end-to-end conversational style matching agent*, [in:] Gebhard P., Schneeberger T., Mehlmann G., Baur T., André E. (eds.), *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, New York: ACM, pp. 111–118, <https://doi.org/10.1145/3308532.3329473>
- Horton W.S. (2017), *Theories and approaches to the study of conversation and interactive discourse*, [in:] M.F. Schober, D.N. Rapp, M.A. Britt (eds.), *The Routledge handbook of discourse processes*, London: Routledge, pp. 22–68, <https://doi.org/10.4324/9781315687384-3>
- Ji J. et al. (2025), *AI alignment: A comprehensive survey*, “arXiv”, <https://doi.org/10.48550/arXiv.2310.19852>
- Jurafsky D., Martin J.H. (2009), *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2nd ed., Upper Saddle River, NJ: Pearson Prentice Hall.
- Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D. (2020), *Scaling laws for neural language models*, “arXiv”, <https://doi.org/10.48550/arXiv.2001.08361>
- Kostić B., Fallon C., Risch J., Löser A. (2026), *Same meaning, different scores: Lexical and syntactic sensitivity in LLM evaluation*, “arXiv”, <https://doi.org/10.48550/arXiv.2602.17316>
- Li S. (2010), *The effectiveness of corrective feedback in SLA: A meta-analysis*, “Language Learning”, vol. 60(2), pp. 309–365, <https://doi.org/10.1111/j.1467-9922.2010.00561.x>
- Lubold N., Walker E., Pon-Barry H. (2021), *Effects of adapting to user pitch on rapport perception, behavior, and state with a social robotic learning companion*, “User Modeling and User-Adapted Interaction”, vol. 31(1), pp. 35–73, <https://doi.org/10.1007/s11257-020-09267-3>
- Mikolov T., Chen K., Corrado G., Dean J. (2013), *Efficient estimation of word representations in vector space*, “arXiv”, <https://doi.org/10.48550/arXiv.1301.3781>
- Mizumoto A. (2025), *Automated analysis of common errors in L2 learner production: Prototype web application development*, “Studies in Second Language Acquisition”, vol. 47(3), pp. 867–884, <https://doi.org/10.1017/S0272263125100934>
- Nass C., Lee K.M. (2000), *Does computer-generated speech manifest personality? An experimental test of similarity-attraction*, [in:] T. Turner, G. Szwillus (eds.), *Proceedings of the SIG-CHI Conference on Human Factors in Computing Systems (CHI '00)*, New York: Association for Computing Machinery, pp. 329–336, <https://doi.org/10.1145/332040.332452>
- Nass C., Moon Y. (2000), *Machines and mindlessness: Social responses to computers*, “Journal of Social Issues”, vol. 56(1), pp. 81–103, <https://doi.org/10.1111/0022-4537.00153>

Ouyang L. et al. (2022), *Training language models to follow instructions with human feedback*, “arXiv”, <https://doi.org/10.48550/arXiv.2203.02155>

Park J.S., O’Brien J., Cai C.J., Morris M.R., Liang P., Bernstein M.S. (2023), *Generative agents: Interactive simulacra of human behavior*, [in:] S. Follmer, J. Han, J. Steimle, N. Henry Riche (eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)*, New York: Association for Computing Machinery, pp. 1–22, <https://doi.org/10.1145/3586183.3606763>

Pickering M.J., Garrod S. (2004), *Toward a mechanistic psychology of dialogue*, “Behavioral and Brain Sciences”, vol. 27(2), pp. 169–190, <https://doi.org/10.1017/S0140525X04000056>

Santurkar S., Durmus E., Ladhak F., Lee C., Liang P., Hashimoto T. (2023), *Whose opinions do language models reflect?*, “arXiv”, <https://doi.org/10.48550/arXiv.2303.17548>

Shen H. et al. (2025), *Position: Towards bidirectional human-AI alignment*, “arXiv”, <https://doi.org/10.48550/arXiv.2406.09264>

Srivastava A. et al. (2023), *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*, “arXiv”, <https://doi.org/10.48550/arXiv.2206.04615>

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. (2017), *Attention is all you need*, [in:] I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, New York: Curran Associates, Inc., pp. 5998–6008.