Syntactic Alignment in Conversations with Large Language Models: Do LLMs Adapt their Syntax Over the Long Term Similar to Humans?

Anonymous ACL submission

Abstract

This paper explores the effects of long-term syntactic alignment in Large Language Models (LLMs). Using OpenAI's GPT-4o, artificial conversations were generated, addressing a lack in existing research of long natural conversations with LLMs. A statistical analysis on syntactic structures present in these conversations reveals that syntactic alignment occurs in LLMs over extended periods. A second analysis further explores how the process of alignment evolves throughout a conversation, showing that LLMs progressively adjust their syntax, with the largest changes occurring early on. The results indicate that LLMs are not only influenced by the linear order in which tokens of their inputs appear, but also that its influence becomes continuously larger with increasing context lengths.

1 Introduction

Alignment in human language and communication is a widely studied process, in which people adapt to their communication partner by coordinating their behavior and language. These adaptation processes not only appears on a surface level, such as gestures, postures or the speech rate (Holler and Wilkin, 2011, Shockley et al., 2009, Jungers and Hupp, 2009), but also on more underlying levels, e.g. the semantics or syntax (Bock, 1986, Garrod and Anderson, 1987). Under these latter two aspects, artificial language generation has become almost indistinguishable from human language in recent years; Large language models (LLMs) are trained to produce texts that seem as coherent as human language. The extend to which they resemble human behaviour, however, has only recently gained focus (Cai et al., 2024). As LLMs become increasingly popular, it is pivotal to understand the extend of their human-like behaviour for better understanding of their societal impact and their potential psycholinguistic implications.

Although LLMs are never explicitly guided to exhibit such behaviour, do large language models nonetheless exhibit syntactic alignment in their text production, similar to us?

1.1 Priming and Alignment

Research on human adaptation processes in language and communication has covered many different of its aspects. This paper puts its focus on syntactic adaptation. Reitter, 2008 showed that such adaptation correlates with success in goal-oriented conversational tasks when this process appears over longer periods. On a theoretical side, the difference between long-term and short-term effects are explained by two opposing (although not exclusive) camps: One explains alignment as a result of conscious, cooperative decisions made during communication (Brennan and Clark, 1996), the other by an automatic, mechanistic process occurring across various linguistic levels (Rasenberg et al., 2020). This latter perspective has its theoretical foundation in Pickering and Garrod, 2004's interactive alignment model (IAM). Under this view, alignment is used to refer to a process in which situational cognitive models of speakers approach each other, such that they develop shared representations on different levels. The process is driven by a priming mechanism, automatic repetitions that occur in a short term, in which encountering an utterance will activate a representation increasing the likelihood of reproducing an utterance that uses the same representation.

As such theoretical views are not applicable to language models, the terms will be used in a more general sense. Alignment will refer to more robust adaptation over a longer period, whereas priming will refer to short-term repetitions. Based on psycholinguistic experiments, the terms *prime* and *target* are taken to refer to the first appearance of a linguistic structure and its subsequent repetition, detached from any theoretical implications.

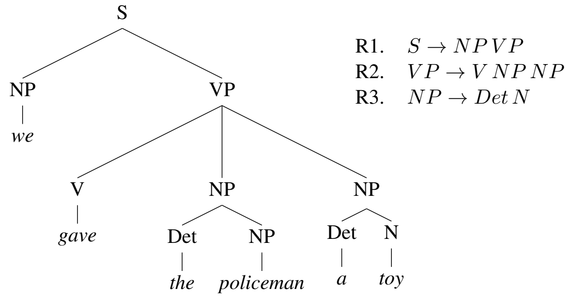
S
NP
VP
we
V
NP
NP
gave
Det
NP
Det
N
the
policeman
a
toy
R1. $S \rightarrow NP\,VP$
R2. $VP \rightarrow V\,NP\,NP$
R3. $NP \rightarrow Det\,N$

Figure 1: Example Phrase Structure Tree and Rules

2 Methodology

2.1 Structural Annotations

To analyze syntactic structures, corpora must be annotated accordingly. Syntactic annotations result in phrase structure trees (Figure 1), which can be interpreted in terms of their phrase structure rules, e.g. the subphrases that a phrase encompasses. Lexical and unary rules (phrases with a single child) are omitted here, as they don't contribute to the syntactic structure of sentences.

2.2 Analysing Syntactic Alignment

We analyse effects across whole conversations to capture adaptation processes that occur over the long term, following Reitter, 2008. His analysis introduced a method to measure alignment in corpora of natural conversations, eliminating the need for controlled experimental setups: Syntactic alignment takes place if occurences of syntactic structures become more likely, given that they have been used before inside a conversation. Conversely, there must be syntactic alignment if the presence of a syntactic structure is predictive for that structure to have occured beforehand. We use this paradigm, following Reitter's (2008) analysis:

First, to establish which phrase structures are deemed primes and targets, conversations are split in half, removing a small portion in the middle to eliminate short-term priming effects. Phrase structure rules from the remaining halves form two sets of rules for each conversation. We will refer to the set of the first half as PRIME and to the second as TARGET. For every rule in TARGET, we check whether that rule has been uttered by the other speaker in PRIME and encode the presence in a binary variable *Prime*. To establish whether such occurences are predictive for repetitions inside a conversation, we choose a control by sampling from another randomly chosen conversation, checking whether the rule is in its PRIME. Datasamples capture this difference in another binary variable *SameConversation*. This process leads to two samples for each rule in TARGET of every conversation. Fitting a mixed effects logistic regression, we check whether there is an effect of *SameConversation* on *Prime*. A positive effect is only expected if there is syntactic alignment.

The analysis further includes the logarithmic frequency of rules across all conversations (*ln(Frequency)*) and the logarithmic size of PRIME (*ln(Size)*), the number of words uttered by the speaker sampled from. Rules that appear only once are excluded.

The analysis improves on Reitter's (2008) approach in two crucial aspects: First, the sampling process limits itself on the other speaker of every conversation, eliminating effects that solely stem from speaker idiosyncrasies. Second, the analysis directly includes the size of PRIME that was sampled from, which has definite impact on the likelihood of a rule to be present.

3 Verifying on Human Conversations

To verify the adapted analysis and to reproduce Reitter's (2008) findings, we first apply it to a dataset of human-human conversations. We use the Switchboard Corpus (Marcus et al., 1994) which consists of 650 annotated telephone conversations. The analysis follows the method described in section 2.

The mixed effects logistic regression was applied using the generalized linear mixed models (GLMM) of python's pymer4 package, including a random intercept for speakers and a random slope for *ln(Frequency)*. Outliers, rules with a frequency > 12000, were excluded. Results are shown in Table 1.

3.1 Results: Switchboard

As is expected, *ln(Frequency)* ($\beta = 1.174, p < 0.001$) and *ln(Size)* ($\beta = 1.402, p < 0.001$) largely increase the odds of *Prime*. *SameConversation* ($beta = 0.228, p < 0.001$) has a significant positive effect, supporting the findings of Reitter, 2008, although with slightly lower values ($\beta_{Reitter} = 1.064$, $OR_{Reitter} = 2.90$). The interaction between *ln(Frequency)* and *SameConversation* ($\beta = -0.101, p < 0.001$) indicates that alignment is stronger for less frequent syntactic structures. The interaction of *ln(Frequency)* and *ln(Size)* ($\beta = 0.068, p < 0.004$) is neglegible.

	β	SE	OR	z	$P > \|z\|$
Intercept	-2.927	0.018	0.054	-158.847	0.000
ln(Frequency)	1.174	0.008	3.184	143.182	0.000
SameConversation	0.228	0.023	1.201	9.950	0.000
ln(Size)	1.402	0.033	3.804	41.921	0.000
ln(Frequency):SameConversation	-0.101	0.010	0.885	-9.757	0.000
ln(Frequency):ln(Size)	0.068	0.015	1.041	4.699	0.004

Table 1: The regression model for the Switchboard Corpus. Fixed effects, except *SameConversation*, are centered. The model with the lowest AIC was taken ($\Delta AIC > 6$ compared to the second-best model).

3.2 Preliminary Discussion

The analysis captures alignment effects in human conversations, verifying the results found in Reitter, 2008 on a different dataset and with a refined sampling method. The lower effect of *SameConversation* can be largely explained by this adaptation: Only sampling from the other speaker reduces the chances of finding a prime by more than half.

Adding on previous findings, the results underpin that alignment happens in-between speakers of conversations that are not goal-oriented. There are limitations, specifically concerning different topicality across conversations, which could be hypothesized to impact a speaker's used syntax. A discussion of this can be found in **??**, as the focus here is on LLM-generated conversations, where topicality can be mostly controlled.

4 Experiment 1: Long-Term Syntactic Alignment in LLMs

4.1 Dataset Generation

Identical to the analysis on humans, the same analysis was run on conversations generated using OpenAI's GPT-4o. Existing datasets proved to be unsuitable for two main reasons: Datasets containing conversations between LLMs contain interactions that are too short to analyze long-term effects. Datasets of human-LLM interactions are much too diverse and far from natural conversations to be suitable.

Conversations were generated by pairing different agents. Agents share a system prompt that instructs them on the conversational task. Each agent is further given a unique language specification to vary their individual syntax Reference Prompt templates. Reference the differences in their syntax (jsd matrix). Their responses are iteratively used to prompt the other agent, gradually building the context of a conversation. The process is run until a certain number of words is reached.

Overall 17 agents were created using ChatGPT. Those agents were cross-matched to create a total of 136 conversations with unique pairings, out of which 12 were excluded as they ended in repeating patterns. All conversations have the same topic: "What makes a day a good day?". The remaing 124 conversations were used for the analysis.

4.2 Method

The analysis is described in section 2. The sampling process was adapted to exclude sampling from identical agents of other conversations. Outliers, rules with a frequency > 2000, were excluded. The results are shown in Table 2.

4.3 Results

SameConversation has a significant positive effect on *Prime* ($\beta = 0.198$, $p < 0.001$), increasing the odds by around 1.2. As is expected, *ln(Frequency)* and *ln(Size)* also show strong positive effects. Similar to human conversations, *ln(Frequency):SameConversation* has a negative effect on *Prime*, indicating that alignment is stronger on less frequent rules. The interaction between *ln(Frequency)* and *ln(Size)* deviates from the results found for human conversations with a significant positive effect ($\beta = 0.266$, $p < 0.001$).

4.4 Discussion

The results show that LLMs exhibit syntactic alignment over long periods, indicating that they adapt to the syntax present in their input. Clearly, the mechanism in LLMs for text production is different from humans, making the results highlight a shared feature, which appears rather coincidental than by any shared mechanism.

The data was generated on the instruction of having a conversation, which potentially induces this behaviour, as conversational data of humans follows this pattern. Further analysis is needed to explore, whether this behaviour reflects a bias in

	β	SE	OR	z	$P > \lvert z \rvert$
Intercept	-2.031	0.048	0.131	-42.488	0.000
ln(Frequency)	1.275	0.028	3.580	45.582	0.000
SameConversation	0.198	0.056	1.219	3.538	0.000
ln(Size)	1.175	0.107	3.240	10.972	0.000
ln(Frequency):SameConversation	-0.146	0.035	0.864	-4.204	0.000
ln(Frequency):ln(Size)	0.266	0.062	1.305	4.297	0.000

Table 2: The regression model for the GPT-4o-generated conversations. Fixed effects, except *SameConversation*, are centered. The model with the lowest AIC was taken ($\Delta AIC > 4$ compared to the second-best model).

their training data, which is left for future directions.

The focus of the second experiment will be on how LLMs adapt their syntax over the course of a conversation.

5 Experiment 2: Progression of Syntactic Alignment in LLMs

To explore the process of alignment throughout a conversation, we analyze the discrete distributions of used syntactic structures. We use the Jensen-Shannon Divergence (JSD) as distance measurement between two different agents, modelling how the syntactic similarity of their languages evolves.

5.1 Method

Conversations between two different agents are analyzed across different sections. For each section, the respective distribution of rules for both agents are extracted. Then, for each section, their Jensen-Shannon Divergence is calculated, resulting in a series of similarity measurements over the course of the conversation.

To estimate accurate rule distributions across sections, there needs to be sufficient data available. We ran a pre-analysis of variance in JSD values with variable amounts of data to determine a suitable section size (see Appendix). Following these results, a single conversation between two different agents was run for 520 trials, each containing an identical prompting. Out of these 520 conversations, 14 were excluded due to repeating patterns. The remaining 506 conversations were used for the analysis.

5.2 Results

Figure 2 shows that the probability distributions of agents become progressively closer across splits. Most alignment happens primarily between the first and second splits. Additionally, the divergence score for the first split is already much lower than the expected 0.37 between Agents 7 and 8 **??**. This suggests that agents align their language predominantly on just a few initial exchanges. Overall, these findings align with the results from section 4.

5.3 Discussion

The process of alignment in LLMs follows a continuous pattern - syntactic structures gradually converge towards the distribution given in their input context. The effects are largest at the beginning and fall off quickly. This behaviour is expected for continuous convergence of two approximating values: The further they are apart, the larger the change. It should be noted however, that the JSD naturally follows a non-linear decay (Lin, 1991). Further analysis is required to determine the exact decay rate for the alignment process.

Nonetheless, we see clear evidence that LLMs not only adapt their syntax to the context that they are given, but also that this process correlates with the amount of syntactic structures that are available in their input.

6 Conclusion

We have shown that LLMs adapt to the syntax of their input in a scenario of aritifical conversations. Unlike humans, who exhibit variability far from stochastic behavior (Kilgarriff, 2005), the findings support that LLMs operate in a much more deterministic way. While short-term and long-term alignment effects in humans are thought to arise from distinct mechanisms (Reitter and Moore, 2014, see discussion in Rasenberg et al., 2020), the results underpin the intuition that short-term effects, like those reported in Cai et al., 2024, are driven by the same principles as long-term alignment in LLMs; LLMs are not only influenced by the semantics of their input vectors, but also by the order in which these token embeddings appear.

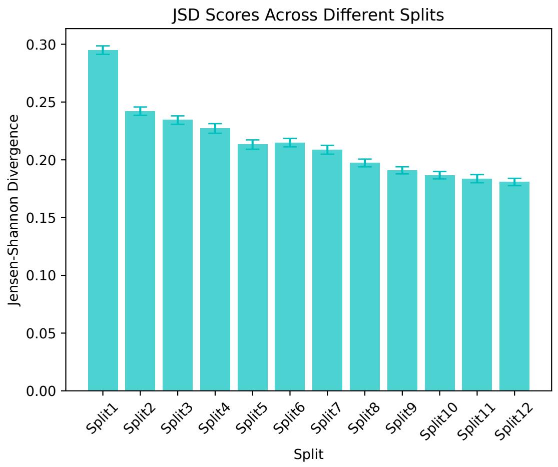

Figure 2: Jensen-Shannon Divergence scores for each split of the conversation. Values are averaged over 100 bootstraps of 506 conversations. Errorbars show their standard deviation.

References

J.Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

S. E. Brennan and H. H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Zhenguang G. Cai, Xufeng Duan, David A. Haslett, Shuqi Wang, and Martin J. Pickering. 2024. Do large language models resemble humans in language use? *Preprint*, arXiv:2303.08014.

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.

Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35(2):133–153.

Melissa K Jungers and Julie M Hupp. 2009. Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, 24(4):611–624.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, 44(11):e12911.

David Reitter. 2008. Context effects in language production: Models of syntactic priming in dialogue corpora.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Kevin Shockley, Daniel C. Richardson, and Rick Dale. 2009. Conversation and coordinative structures. *Topics in Cognitive Science*, 1(2):305–319.

A Example Appendix

This is an appendix.