

Syntactic Complexity Convergence in Dialogue: Analysis of the Phenomenon and Application to LLM Detection

Yu Wang^{1,2} Yanran Chen³ Yifan Wang⁴

Steffen Eger³ Hendrik Buschmeier¹

¹Digital Linguistics Lab, Department of Linguistics, Bielefeld University

²Institute for Language Sciences, Utrecht University

³NLLG Lab, University of Technology Nuremberg

⁴Department of Language Science and Technology, Saarland University

Abstract

We revisit the under-explored phenomenon of syntactic complexity convergence in dialogue. Extending prior topic-level analyses to full conversations, we examine five human-human datasets and a set of human-LLM dialogues. We integrate and evaluate a multidimensional set of syntactic complexity metrics to quantify structural variation in dialogue. Our analyses reveal robust syntactic complexity convergence across human-human datasets, especially in task-oriented interactions, and statistical tests confirm that these effects are not due to chance. In contrast, human-LLM dialogues show no significant convergence under individual metrics; however, linear-CKA uncovers a distinct cross-metric convergence pattern. Motivated by these differences, we use syntactic complexity as a feature for detecting LLM-involved dialogue, achieving 96% accuracy. Our classification results demonstrate its potential without relying on likelihood-based signals such as surprisal, which require a language model as the estimator.

specific mechanism of syntactic complexity convergence, where interlocutors mirror the structural density of each other’s language, remains less understood in dialogue. Previous research on complexity matching has primarily used small datasets and mainly focused on short linguistic units, e.g., the so called T-unit, the shortest grammatically allowable sentence where writing can be split (Hunt, 1970), or topic episodes (Xu and Reitter, 2016). Thus, it remains unclear whether these patterns hold across entire conversations or across different languages. The rapid integration of Large Language Models (LLMs) into everyday communication also raises a new question: Do human-LLM conversations show complexity matching patterns similar to those in human-human dyads? In this paper, we revisit syntactic complexity convergence by analysing a diverse corpus of natural and LLM-involved dialogues. We address gaps in the literature by examining the phenomenon across conversations rather than short linguistic units. Our contributions are:

1 Introduction

In dialogue, mutual understanding is achieved through adaptive processes that reduce the communicative effort of participants and facilitate successful interaction (Pickering and Garrod, 2004, 2006; Reitter and Moore, 2007, 2014). Extensive research has shown that humans adjust their language use to align with their conversation partners (e.g., Reitter and Moore, 2014; Buschmeier and Kopp, 2018; Giulianelli and Fernández, 2021). This co-adaptivity in interlocutors’ verbal behaviour manifests in linguistic representations across multiple levels of processing (Pickering and Garrod, 2004, p. 176). As interlocutors develop mutual understanding, their linguistic patterns tend to align. Such convergence has been measured at a more abstract level through shifts in syntactic complexity (Xu and Reitter, 2016) and semantic similarity (Xu, 2021). While lexical and semantic alignment are well-documented, the

- **A multidimensional metric and large-scale multilingual analysis of syntactic complexity:** We integrate a comprehensive metric that captures syntactic complexity from multiple structural perspectives. We verify its efficacy by conducting inter-correlation analysis on the metrics and using an open parallel corpus containing sentence pairs controlled for syntactic complexity variation, and apply it to five human-human dialogue datasets across three languages (English, German, Japanese) to examine syntactic complexity convergence on a large scale.
- **A comparison of syntactic complexity in human-human and human-LLM dialogue:** We analyse the difference between human-human and human-LLM dialogue from the perspective of syntactic complexity. The analysis provides the basis for an

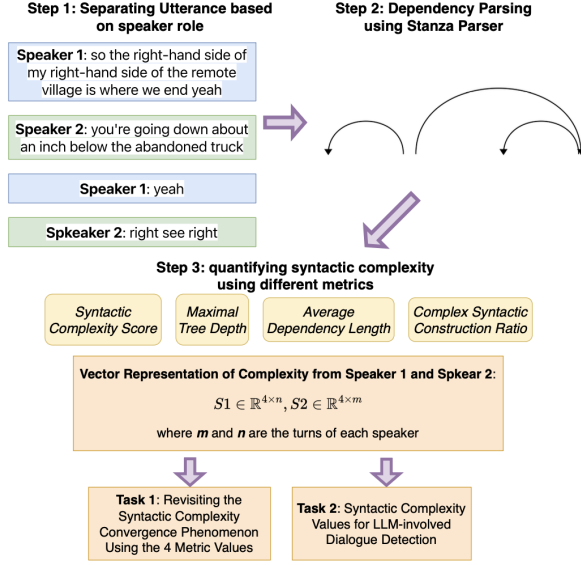


Figure 1: Workflow for the syntactic complexity metrics (example dialogue taken from the Map Task dataset).

LLM-involved dialogue detection task.¹

2 Background

2.1 Complexity Matching as a Type of Linguistic Alignment

The concept of ‘complexity matching’ was first studied by Abney et al. (2014), who analysed speech signals from conversational speech in terms of time series of acoustic onset events. As a type of abstract behavioural matching, it was later further tested by Xu and Reitter (2016) and Xu (2021) who found evidence for complexity convergence on the syntactic and the semantic level. Matching of complexity in dialogue is theoretically grounded in the *Interactive Alignment Model* (Pickering and Garrod, 2004), which posits that linguistic alignment facilitates mutual understanding.

While previous research has primarily focused on the development of linguistic alignment in human interaction, recent studies have begun to explore these dynamics within Large Language Models. For example, Frisch and Giulianelli (2024) show that certain types of persona-conditioned LLMs can be adapted to exhibit linguistic alignment toward their conversational partner. Further, clear

¹We consider this to be our novel point: to the best of our knowledge, the most relevant prior work, Bhatt and Rios (2021), is a chatbot-detection study focused on dialogues generated by pre-LLM language models. More recent LLM-era studies such as DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2024) address LLM-generated text detection in general, but not specifically dialogue.

evidence is provided by Molnar et al. (2023), whose empirical experiments uncover that LLMs exhibit language use behaviour (lexical repetition in the interaction) that is similar to humans. Moreover, Kandra et al. (2025) demonstrate that two LLM agents acting as conversation partners, exhibit mutual syntactic adaptation. Extending this to different setting, Blevins et al. (2026) investigate convergence through dialogue completion tasks, finding that syntactic accommodation (measured mainly through utterance length) is significantly more pronounced in human–LLM interactions than in human baselines. Building on this previous work, we investigate syntactic complexity convergence as a high-level behavioural phenomenon and focus on the change in syntactic complexity between dialogue participants across interactional contexts. By analysing dialogues ranging from human–human to human–LLM dyads, our study aims to characterize how syntactic complexity, as a potential linguistic feature, is negotiated and aligned in both natural dialogue as well as when one interlocutor is an LLM-based conversational agent. If obvious differences exist in syntactic complexity between human–human and human–LLM dialogue, we will discuss the feasibility of using these differences to detect LLM-involved dialogue.

2.2 Syntactic Complexity Quantification

Syntactic complexity is a formal linguistic property associated with increased cognitive load during parsing (Szmrecsányi, 2004; Lin, 1996). Researchers typically evaluate syntactic complexity through three lenses: **word count** (length as a proxy for structural volume), **node count** (phrasal density), and the **Index of Syntactic Complexity**, which tracks clausal subordination and morphological embeddedness (Beaman, 1984; Givón, 1991). To quantify syntactic complexity, Xu and Reitter (2016) utilize three structural measures: (i) **Sentence Length**, a proxy for structural size; (ii) **Tree Depth**, the maximum vertical path in the parse tree, where deeper trees indicate greater syntactic complexity; and (iii) **Branching Factor**, the average number of children per non-leaf node, reflecting constituent density within a sentence. The quantification methods proposed in Xu and Reitter (2016) were later adapted and used in follow-up work. Chen et al. (2026), for example, look at syntactic features, such as tree depth, to analyse the syntactic change in German, and Bagdasarov and

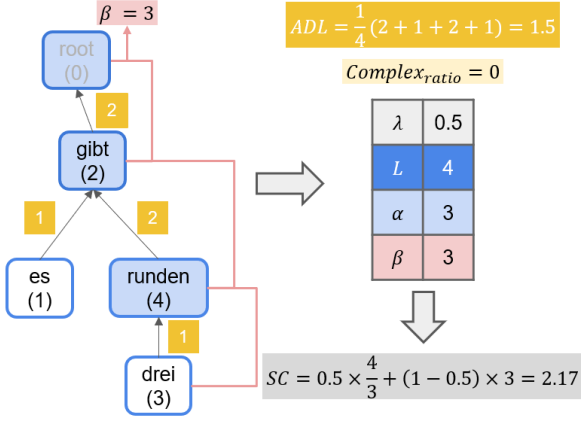


Figure 2: Illustration of calculating/observing β , ADL , SC , $Complex_{ratio}$ for the German example sentence “Es gibt drei Runden” (*There are three rounds*). Values in parentheses denote token position indices. Edge labels indicate the dependency distance for each dependency.

Alves (2025) use complexity quantification methods to build a syntax-aware classifier for detecting LLM-generated text.

3 Syntactic Complexity Metrics

In this section, we introduce four metrics for quantifying syntactic complexity of utterances from different perspectives, based on prior work. For each metric, larger values indicate greater syntactic complexity. As an example, we illustrate how four metrics are computed using the German sentence “Es gibt drei Runden.” (*there are three rounds*; see Figure 2). All metrics are derived from dependency parsing (Nivre, 2010), which identifies grammatical relations between tokens in a sentence by assigning each dependent token a head and a relation label. The resulting structure forms a directed tree (Figure 2, left), where each dependent points to its head and the root token is attached to a pseudo-root. In this work, we use Stanza (Qi et al., 2018) from the StanfordNLP toolkit to segment utterances into sentences and obtain their dependency parses.

Human dialogue data often contain syntactically incomplete utterances, disfluencies (e.g., self-repairs or repetitions), and discourse markers such as fillers (“uh”, “um”) or backchannels (“uh-huh”). We choose Stanza because it can parse incomplete or grammatically irregular sentences and treats such discourse markers as, e.g. modifier tags. Based on the resulting dependency structures, we compute syntactic complexity using the following measures:

Maximal Tree Depth (β) Tree depth is widely used to measure syntactic complexity (e.g., Xu and

Reitter, 2016; Chen et al., 2026). The depth of a node is defined as its distance from the root, measured by the number of edges on the path between them. The maximal tree depth is the largest node depth in the tree. In Figure 2, the maximal tree depth ($\beta = 3$) occurs along the path (root) $\rightarrow$ *gibt* $\rightarrow$ *runden* $\rightarrow$ *drei*. The algorithm for maximal tree depth calculation is provided in Appendix H.

Average Dependency Length (ADL) Dependency distance is the linear distance between a head and its dependent, calculated as the absolute difference between their position indices. The average dependency length is commonly used as an indicator of syntactic complexity (Liu, 2008; Futrell et al., 2015). For example, the distance between (*gibt*, *es*) in Figure 2 is 1. We compute this distance for each dependency pair and average over all dependencies to obtain a sentence-level complexity score:

$$ADL = \frac{1}{|D|} \sum_{(h,d) \in D} |h - d|,$$

where h and d denote the position indices of the head and dependent tokens, and D is the set of dependencies in the sentence. In the example in Figure 2, $ADL = 1.5$.

Quantified Syntactic Complexity Value (SC) We compute the syntactic complexity SC of a sentence following Wang and Buschmeier (2024):

$$SC = \begin{cases} \lambda \cdot \frac{L}{\alpha} + (1 - \lambda) \cdot \beta & \text{if } \alpha > 0, \\ (1 - \lambda) \cdot \beta & \text{otherwise.} \end{cases}$$

Here, L denotes sentence length, α the number of heads, β the maximal tree depth, and λ a tuning factor (set to 0.5). The formula combines the effects of sentence length and syntactic structure. Longer sentences tend to have higher syntactic complexity, often resulting in deeper parse trees. The term $\lambda \cdot \frac{L}{\alpha}$ captures how many constituents are dominated by a head, similar to the branching factor proposed by Xu and Reitter (2016). For the example sentence in Figure 2, $L = 4$, $\alpha = 3$ (including the root), and $\beta = 3$, resulting in $SC = 2.17$.

Complex Dependency Structure Ratio ($Complex_{ratio}$) In addition to sentence length, syntactic complexity can also be estimated by the proportion of complex structures (Lu, 2010). We compute $Complex_{ratio}$ as the proportion of *long* dependency edges, defined as edges with distance > 3 . Such long dependencies often

reflect embedded or inflected constructions. In the example sentence, no dependency edge has a distance greater than 3, so no edge is counted as long and $Complex_{ratio} = 0$, reflecting the simplicity of the utterance.

Figure 4 in Appendix B illustrates this with two dependency trees: one for the simple example sentence (Figure 2) and another for a more complex sentence. The simple sentence has $\beta = 3$, $ADL = 1.5$, $SC = 2.17$, and $Complex_{ratio} = 0$. In contrast, the complex sentence in Figure 4b has $\beta = 4$, $ADL = 2.428$, $SC = 3$, and $Complex_{ratio} = 0.48$.²

3.1 Integrated Complexity Metrics

When integrating the above four metrics to quantify syntactic complexity, we notice that previous work proposing these individual metrics lack systematic evaluation of their validity in combination. Therefore, we validate our integrated metric framework through three complementary analyses: (i) an inter-correlation analysis confirming that the metrics capture distinct dimensions of syntactic complexity; (ii) a corpus-level evaluation on **DEplain**, a parallel dataset of German sentences and their simplified equivalences (Stodden et al., 2023), confirming directionally correct behaviour; (iii) a qualitative sentence-pair analysis confirming sensitivity to explicit syntactic variation.

For the inter-correlation analysis, we examine whether any single metric sufficiently represents the others. The results reveal consistently weak correlations between several metric pairs, most notably between Tree Depth and Complex Dependency Ratio across datasets, which indicates that each metric contributes unique information and thus all four metrics are necessary to capture syntactic complexity from different structural perspectives (details in Appendix F and Figure 6).

For the corpus-level evaluation, we apply our metrics to the **DEplain** corpus. We acknowledge that simplified sentences in **DEplain** may undergo lexical or semantic simplification rather than purely structural reduction, which means the corpus provides a conservative test of our metrics’ sensitivity to syntactic complexity specifically. Nevertheless, as reported in Table 5, all metrics except $Complex_{ratio}$ consistently assign lower complexity scores to simplified sentences in both training and

test sets, demonstrating directionally correct behaviour at the corpus level.

Moreover, to ensure our metrics are specifically sensitive to syntactic rather than lexical simplification, we additionally sampled a subset of sentence pairs with explicit structural complexity differences for qualitative analysis. As shown in Figure 7 (Appendix F), all four metrics successfully capture the complexity reduction in these cases, further supporting empirical validity of the four metrics for measuring syntactic complexity in dialogue data.

4 Data

We run our experiments on multiple human–human dialogue datasets. Some of them are task-oriented dialogues, while others are spontaneous conversations. We also incorporate human–LLM dialogues to examine whether LLMs exhibit distinct patterns of syntactic complexity in human–LLM dialogues compared to their human counterparts in human–human dialogues during the development of the conversation. A criterion for datasets selection was that dialogues are transcribed such that turns, (anonymous) speakers IDs, and smaller aspects of language use (e.g., stuttered speech, pauses, backchannels) are available. In total, we use five human dialogue datasets, which cover the languages German, Japanese and English: *MUNDEX* (ger; Türk et al., 2023), *Hazumi* (jpn; Komatani and Okada, 2021), *BTSJ* (jpn; Usami, 2023), *CANDOR* (eng; Reece et al., 2023), and *Map Task* (eng; Anderson et al., 1991). For the human–LLM dialogues, we use the *LMSYS-Chat-1M* dataset (Zheng et al., 2024), from which we select the English and German parts (given their comparatively large proportion among the other language resources). Detailed descriptions of the datasets, pre-processing, and statistics of the used dialogues can be found in Appendix A and B.

The dialogue datasets in this study can be categorised into the following three types: (i) task-oriented dialogue or dialogue with fixed topics (*MUNDEX*, *MAP TASK*, *Hazumi*); (ii) everyday conversation dialogue (*BTSJ*, *CANDOR*); (iii) LLM-involved dialogues (*LMSYS-Chat-1M*). We define the dialogical roles as ‘initiator’ and ‘follower’ as follows: for the human–human dialogues, the person who initiates the conversation is the initiator and the others are followers; for the human–LLM dialogues, human users are considered initiators and LLM agents followers.

²If an utterance contains multiple sentences, we average their values to represent its syntactic complexity (see Appendix B).

5 Results and Discussion

5.1 Trend of the Syntactic Complexity in Dialogues Based on Mean Value

One of the challenges for tracking the general trend of syntactic complexity convergence in dialogue is that dialogues differ in length, which makes their comparison inconsistent and thus suboptimal. A common practice is to divide dialogue into different time steps or relative positions within dialogue (see, e.g., Xu, 2021; Blevins et al., 2026). We use a similar solution introduced by Fuscone et al. (2020), where dialogues are split up into multiple phases with each phase containing the same number of turns, disregarding the variation in length of dialogues when verifying convergence. As we consider convergence as a phenomenon in dialogue which can only be observed if the dialogue contains enough turn-shifts, we set the number of phases to 8 in order to provide enough granularity for checking the general trend.³

Figures 3 and 5 in Appendix C together plot the mean values of our metrics across 8 dialogue phases in the datasets. For each phase in each qualified dialogue, we have four time series of quantified syntactic complexity from every conversation participants, and we calculate the average value on each time series to represent the syntactic complexity for that phase for each dialogue role (either initiator or follower). Among the five human dialogue datasets, *MUNDEX* and *Hazumi* exhibit the most pronounced convergence (Figure 3): in the *MUNDEX* dataset, the syntactic complexity of initiators gradually decreases and converges with that of followers. This tendency is clearly evident in the first three metrics. In contrast to that, in the *Hazumi* dataset, the syntactic complexity from initiators increases and adapts to the equivalences from the followers. This tendency is clearly evident in the first three metrics as well. While *BTSJ* and *Candor* (Figures 5a and Figure 5b in Appendix A) also display a discernible tendency toward convergence, the *Map Task* dataset as human–human dialogue dataset shows no observable trend of syntactic complexity convergence, and similarly, both human–LLM dialogue datasets have no observable trend of syntactic complexity convergence (Figure 5c to 5e in Appendix A).

With regard to the reason why *MUNDEX* and

Hazumi exhibit more pronounced convergence, we have taken a closer look into the dialogues and came up with the following hypothesis: in both dialogue datasets, one of the speakers may have better topic-related knowledge compared to the others. In case of the *MUNDEX* corpus, where explainers, as dialogue initiators, explain a board game to the explainees, explainers tend to adjust their explaining strategies, i.e., simplifying their utterances, to make the opponents pick up the game rules more easily. In response, explainees start to initiate more complex questions (as shown in the plot, a peak in phase 6). For the *Hazumi* corpus, the dialogue initiators, who are familiar with the topics, initiate the talk on one topic with the followers and then expand the conversation based on the followers’ utterances. In response, the dialogue initiators’ utterances become more syntactically complex.

Takeaway: Through visualisation, we notice that syntactic complexity convergence appears to be more obvious in human–human dialogue than human–LLM dialogue.

5.2 Convergence Patterns and Ratio

As discussed in Section 5.1, when we take the average value of quantified syntactic complexity on each pre-define phase, syntactic complexity convergence can be observed in different natural dialogue datasets. However, more evidence is still needed in order to ensure the validity of our observation. Our next goal is to check, to what degree the phenomenon of syntactic complexity convergence can be confirmed. Specifically, we want to analyse (i) the patterns of convergence, and (ii) the concrete ratio of convergence in each dataset.

To verify the ratio of convergence as well as the convergence patterns, we need to calculate the complexity difference between the initiator and follower along the 8 different phases. First, in order to verify that convergence occurs, in each selected dialogue, we check if the difference between the two speakers decreases between the first and the last phase: $|P_{init}^1 - P_{fol}^1| > |P_{init}^8 - P_{fol}^8|$. The second constraint requires the sequence of differences to show a consistent downward trend, with at least two decreases across phases. We consider two convergence patterns. In Pattern 1, the syntactic complexity of both speakers decreases and converges to a lower value, which may occur when a conversation ends and speakers produce simpler utterances. In Pattern 2,

³Dividing each dialogue into eight phases also means that a dialogue can only qualify for our evaluation if both speakers have more than eight turns (see Appendix B for further explanation on setting the number of phases to 8).

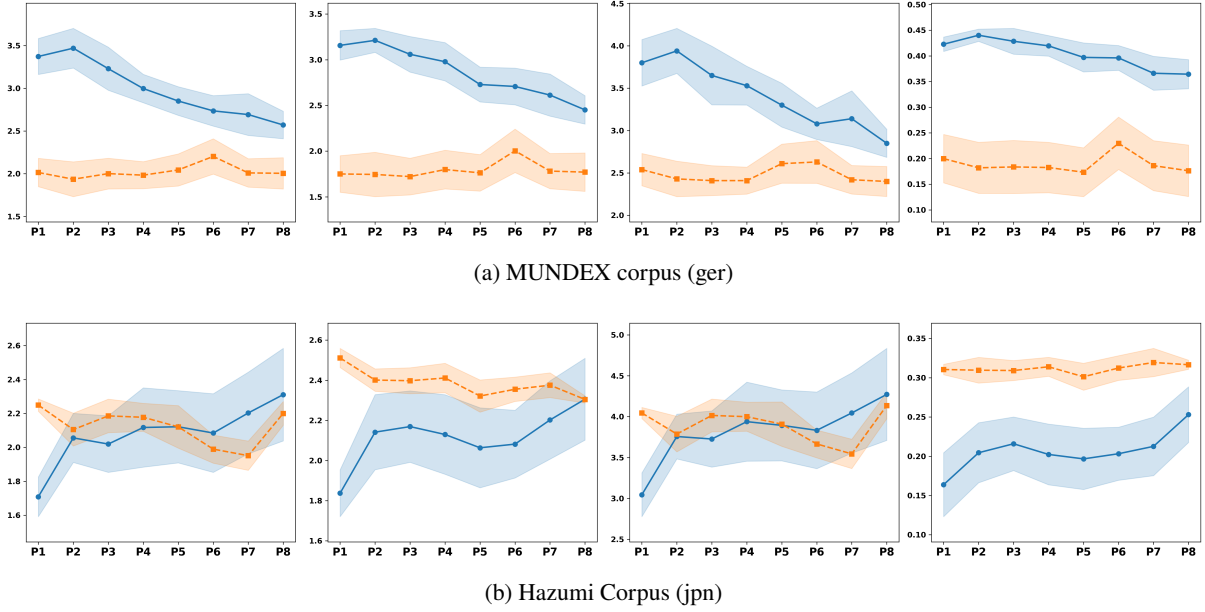


Figure 3: The plots show the syntactic complexity across phases for *MUNDEX* and *Hazumi*. The **blue line is the Dialogue Initiator**, the **orange line is the Dialogue Follower**. The four metrics in each sub-figure from left to right are: **syntactic complexity value**, **dependency length**, **tree depth**, and **complex Dependency Structure Ratio**. We calculate the mean value based on the eight phases (P1-P8) we pre-defined with 95% confidence interval.

the speakers converge to a point where one speaker’s syntactic complexity decreases while the other’s increases. *If any of the four complexity metrics shows such convergence with statistical significance, we treat it as evidence of syntactic complexity convergence in the dataset.* Based on this criterion, Pattern 2 appears more frequently than Pattern 1 (see Table 8 in the Appendix).

We then report the proportion of dialogues in each dataset that exhibit syntactic complexity convergence. To ensure that the ratios are not due to random effects, we generate null distributions by randomly shuffling the turns within each dialogue B times (scaled by dataset size; e.g., 1500 permutations for large datasets such as *BTSJ*, *CANDOR*, and *LLM_English*, and 1000 for smaller datasets such as *Hazumi* and *MUNDEX*). This procedure breaks the temporal structure while preserving turn and role information. If the values from the shuffled data are similar to or larger than the observed values, we conclude that convergence is not present.

The results of statistical testing, as summarized in Table 1 where p -values as well as Cohen’s d with 95% confidence intervals are reported, show that human task-oriented dialogue datasets exhibit statistically significant syntactic complexity convergence, whereas the human-LLM datasets do not.

Takeaway: Quantitative analysis shows that **convergence pattern 2** is dominant in all of the chosen datasets. Under the eight phases setting, syntactic complexity convergence can be statistically confirmed in four out of five human-human dialogue datasets, but not in the human-LLM dialogue datasets.

5.3 Potential Convergence among Different Metric Values

Furthermore, we hypothesize that syntactic complexity convergence is not univariate⁴ but rather multivariate given that: (i) the change of the average dependency length from speaker A can be aligned to the change of the tree depth from speaker B, and therefore the convergence may happen between two different metric values; (ii) the syntactic complexity from speaker A may not be strictly time-aligned with the equivalence from speaker B. To capture such dynamics, we consider the syntactic complexity of each phase as a matrix representation of the syntactic complexity: for each dialogue phase $p \in \{1, \dots, n\}$, we construct complexity matrices

⁴Based on the Interactive Alignment Model that priming can operate across representational dimensions, and evidence that although sharing similarities, priming in LLMs has differences from priming in humans (Sinclair et al., 2022; Jumelet et al., 2024), we hypothesize that human-LLM convergence may manifest cross-metrically rather than within individual metrics. See Appendix B for the full theoretical motivation.

Dataset (N Dialogues)	Dialogue Type	Metric	Observed	Random (Mean)	p -value	Cohen's d	95% CI
MUNDEX (50)	Task-oriented / Fixed topic	Syn. value	0.7600 ***	0.4775	<0.001	4.150	[0.1891, 0.3743]
		Dependency length	0.6600 **	0.4819	0.009	2.538	[0.0848, 0.2757]
		Tree depth	0.5600 ***	0.2657	<0.001	4.962	[0.1889, 0.4100]
		Complex dep. ratio	0.6000 *	0.4662	0.032	1.899	[0.0338, 0.2375]
Hazumi (33)	Task-oriented / Fixed topic	Syn. value	0.5152	0.4864	0.438	0.347	[-0.0951, 0.1636]
		Dependency length	0.7576 **	0.4937	0.002	3.036	[0.1563, 0.3758]
		Tree depth	0.3333	0.3723	0.750	-0.490	[-0.1617, 0.0823]
		Complex dep. ratio	0.6970 **	0.4942	0.010	2.376	[0.0940, 0.3392]
BTSJ (498)	Everyday conversation	Syn. value	0.4137	0.4705	0.997	-2.593	[-0.0900, -0.0264]
		Dependency length	0.4900	0.4873	0.473	0.116	[-0.0315, 0.0380]
		Tree depth	0.2691	0.2928	0.881	-1.189	[-0.0540, 0.0049]
		Complex dep. ratio	0.5321 *	0.4917	0.036	1.800	[0.0034, 0.0728]
Candor (1855)	Everyday conversation	Syn. value	0.3871	0.4817	1.000	-8.381	[-0.1113, -0.0773]
		Dependency length	0.4388	0.4828	0.999	-3.853	[-0.0614, -0.0274]
		Tree depth	0.2194	0.2103	0.161	1.024	[-0.0055, 0.0233]
		Complex dep. ratio	0.4642 †	0.4467	0.066	1.527	[0.0004, 0.0334]
Map Task (121)	Task-oriented / Fixed topic	Syn. value	0.4380	0.4875	0.899	-1.138	[-0.1140, 0.0188]
		Dependency length	0.4215	0.4896	0.945	-3.784	[-0.1306, 0.0037]
		Tree depth	0.3471	0.3446	0.503	0.057	[-0.0625, 0.0688]
		Complex dep. ratio	0.3967	0.4515	0.916	-1.237	[-0.1182, 0.0110]
LLM_German (234)	LLM-involved	Syn. value	0.4316	0.4822	0.945	-1.553	[-0.0971, -0.0025]
		Dependency length	0.3590	0.4818	1.000	-3.782	[-0.1715, -0.0749]
		Tree depth	0.3718	0.3826	0.659	-0.345	[-0.0626, 0.0347]
		Complex dep. ratio	0.4658	0.4723	0.601	-0.203	[-0.0534, 0.0443]
LLM_English (7850)	LLM-involved	Syn. value	0.4122	0.4726	1.000	-11.013	[-0.0681, -0.0516]
		Dependency length	0.4152	0.4722	1.000	-10.410	[-0.0657, -0.0483]
		Tree depth	0.3566	0.3789	1.000	-4.128	[-0.0304, -0.0147]
		Complex dep. ratio	0.4511	0.4596	0.935	-1.521	[-0.0158, -0.0004]

Table 1: Observed dialogue initiator–follower convergence compared to random baseline (8-phase analysis). N represents the number of dialogues meeting the ≥ 8 turns per speaker requirement. Bold indicates statistical significance ($p < 0.1$). Significance markers: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. CIs refer to the 95% confidence interval for Cohen's d .

X_p and $Y_p \in \mathbb{R}^{T \times 4}$, where T is the number of turns in the phase and 4 represents the integrated syntactic complexity metrics. We then calculate the alignment distance $d_p = d(X_p, Y_p)$ to track the convergence between the dialogue initiator and follower. We chose to use the Linear Centered Kernel Alignment (Linear-CKA; Kornblith et al., 2019), a distance function that was originally used to compare the similarity of different layers in language models' hidden representation (vector matrix) with the potential that it can address points (i) and (ii) mentioned above. Linear-CKA measures the representation distance by the similarity between representations' inter-sample similarity $\langle X_i^{p\top} X_i^p, Y_j^{p\top} Y_j^p \rangle$ in phase p from dialogue speaker i and j . Formally, for each dialogue phase $p \in \{1, \dots, n\}$, we construct complexity matrices X_i^p and $Y_j^p \in \mathbb{R}^{T \times 4}$. We then calculate the alignment distance using:

$$d_{\text{CKA}}(X_i^p, Y_j^p) = 1 - \frac{\|X_i^{p\top} Y_j^p\|_F^2}{\|X_i^{p\top} X_i^p\|_F \|Y_j^{p\top} Y_j^p\|_F}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Before the distance calculation, we first normalise the matrix values by applying MinMax normalisation. The smaller the distance value, the closer the syntactic

complexity between dialogue participants. In terms of linguistic interpretation, a small linear-CKA-based distance value can suggest that the dialogue participant (human or LLM) is adapting its overall structural complexity profile to match the other. For example, it may be the case that speaker A has less complex syntactic constructions in their utterances, and in response, speaker B produces utterances with larger average dependency length.

After obtaining the distance values from different phases, we fit the data points into a linear regression model to check if a statistically significant descending trend can be observed. We, again, set the phase number to eight, making it coherent with the previous experimental setting. As the results reported in Table 2 show, both human-LLM dialogue datasets have a strong tendency of syntactic complexity convergence under cross metric evaluation (negative slope value with statistical significance), which can only be confirmed for one of the human-human dialogue datasets.

Takeaway: By applying **Linear-CKA**, we verify a statistically significant tendency of syntactic complexity convergence in Human-LLM dialogue under cross metric setting.

Dataset	Slope	R^2	Std. Err	p -value
MUNDEX	-0.0015	0.0142	0.0050	0.7790
BTSJ	-0.0046	0.9077	0.0006	0.0003
Hazumi	0.0075	0.1701	0.0067	0.3099
Candor	-0.0004	0.0047	0.0026	0.8721
Map Task	-0.0011	0.1644	0.0010	0.3189
LLM (DE)	-0.0174	0.9099	0.0022	0.0002
LLM (EN)	-0.0156	0.9480	0.0015	<0.0001

Table 2: Illustration of the stacked comparison of 8-phase regression results. As we calculate the linear-CKA-based distance value between the syntactic complexity from both dialogue speakers under the cross-metrics setting, statistically significant negative slope values indicate the occurrence of syntactic complexity convergence.

5.4 Discussion

Our first finding is that syntactic complexity convergence is more evident in task-oriented or topic-focused dialogues than in daily conversations. Among the three such datasets (*MUNDEX*, *Hazumi*, *MAP TASK*), two show clear convergence tendencies (Figure 3). In contrast, this tendency is less visible in daily-conversation datasets (Figure 5a and 5b in the Appendix), and not observable in *MAP-TASK* and *Human-LLM* dialogues (Figure 5c to 5e in the Appendix). The statistical analysis in Table 1 supports this observation. No statistically significant syntactic complexity convergence is found in LLM-involved dialogues.

As syntactic complexity convergence can be interpreted as a form of linguistic alignment, these findings support the results of Reitter and Moore (2014), who report stronger syntactic alignment in task-oriented dialogues than in spontaneous conversations. Furthermore, human—human and human—LLM dialogues show different convergence patterns. Human—human dialogues exhibit convergence when measured within the same metric, whereas human—LLM dialogues do not. By analysing cross-metric convergence using linear-CKA, we see that clear convergence emerges in human—LLM dialogues (details in Table 2).

We further evaluate the impact of dialogue length by testing different pre-defined lengths in Appendix E. It appears that longer dialogues do not necessarily lead to stronger convergence, except in the *MUNDEX* dataset. When shorter dialogues are excluded in the cross-metric analysis, convergence becomes clearer in *MUNDEX* but disappears in *BTSJ* and *LLM_German*.

6 Application: LLM Dialogue Detection with Syntactic Complexity Features

As shown in Section 5, (i) human—human dialogues display clearer syntactic complexity convergence within the same metric, whereas (ii) human—LLM dialogues tend to converge across different metrics. These patterns suggest that syntactic complexity may serve as a useful feature for detecting LLM-involved dialogues.

Here, we investigate whether quantified syntactic complexity metric values can be used for human—LLM dialogue detection.⁵ We use our English dialogue datasets (*Candor* and *LLM_English*), whose relatively large sizes make them suitable for classification. The task is to distinguish human—human dialogue (*Candor*) from human—LLM dialogue (*LLM_English*). Our pipeline builds on Xu et al. (2024), who use a Discrete Fourier Transform (DFT) over surprisal values to detect LLM-generated text. Since syntactic complexity metric values also form fluctuating time series and, like surprisal, reflect cognitive load in language processing (Hale, 2001; Demberg and Keller, 2008; Demberg et al., 2012), it is reasonable to apply their DFT-based approach here. We therefore conduct DFT over syntactic complexity time series and use the resulting frequency spectra as features for classification.

We first feed the four syntactic-complexity time series into the FourierGPT pipeline to obtain their spectrum representations. After selecting the top-k peaks (set to 30), we concatenate them to form the feature vectors for classification. As a baseline, we report results using only the spectrum length to train a logistic regression model. In addition to the logistic regression model using spectrum representations, we also train a transformer-based encoder. We compare the original FourierGPT pipeline, which relies on surprisal estimates, with our version using quantified syntactic complexity as the main features, with results shown in Table 3. Further technical details and parameter settings are provided in Appendix D.

Using only the spectrum length, surprisal-based features reach 95.5% F1 score, while syntactic-complexity features reach 74.5%. This difference

⁵A related study by Bagdasarov and Alves (2025) uses syntactic features such as dependency length and tree depth to detect LLM-generated text. We follow the same idea but apply it to detecting LLM-involved dialogue. Li et al. (2025) also report large syntactic-tree differences between human-written and LLM-generated text and build a detector based on this.

Input	Model	F1	Accuracy	AUC
Syn Com	Spectrum Length (baseline)	0.745 (0.074)	0.753 (0.063)	0.757 (0.067)
	Linear Regression	0.962 (0.042)	0.960 (0.047)	1.000 (0.000)
	Transformer Encoder	0.947 (0.053)	0.947 (0.053)	0.993 (0.012)
Surprisal	Spectrum Length (baseline)	0.955 (0.012)	0.956 (0.011)	0.976 (0.016)
	Linear Regression	0.972 (0.014)	0.972 (0.014)	0.997 (0.002)
	Transformer Encoder	0.961 (0.017)	0.960 (0.018)	0.994 (0.004)

Table 3: Classification performance using the FourierGPT pipeline on the saved test set. We compare models using syntactic complexity (“Syn Com”) versus surprisal as input. Reported accuracies are means over 10 runs; standard deviations are shown in parentheses. Bold values indicate the best performance. Additional robustness analyses appear in Appendix G and Table 6.

stems from the fact that surprisal is computed over words and thus directly reflects dialogue length, whereas syntactic complexity is computed at the utterance level and does not. In our data, human–LLM dialogues compared to human–human dialogues have comparatively shorter length (less turns as shown in Table 4). When using spectral features, surprisal-based inputs still perform better overall (F1 score 97.2%), but syntactic-complexity features also yield competitive accuracy (F1 score 96.2%). Unlike surprisal-based detectors, such as *GLTR* (Gehrmann et al., 2019) and *GPT-who* (Venkattraman et al., 2024), which rely on language models to estimate likelihoods, syntactic-complexity computation requires no language model at all, making it a potentially valuable alternative.

7 Conclusion

In this paper, we revisit an under-investigated linguistic phenomenon, syntactic complexity convergence in dialogue, which is considered as an alignment behaviour indicating the development of mutual understanding along the change of the conversation. We extend prior work from topic-level analysis to full dialogues and examine five human–human dialogue datasets, as well as human–LLM dialogues. We review previous studies and integrate metrics, along with further evaluation, to quantify syntactic complexity. Our empirical analysis shows clear syntactic complexity convergence across the human–human dialogue datasets, particularly in task-oriented settings, and statistical tests confirm that this effect is not random. Although human–LLM dialogues show no significant convergence under the same metrics, linear-CKA reveals a different cross-metric convergence pattern. Given these differences, we use syntactic complexity as a feature for detecting LLM-involved dialogues. Classification results highlight its potential without relying on

language-model-based likelihood estimates (e.g., surprisal).

Limitations

Within the scope of this study, we consider the following four points as potential limitations.

First of all, although our study covers different languages, it appears that in dialogue datasets, English dialogues are still a dominant language resource in comparison to German and Japanese. The same situation holds for human–LLM dialogues. Although Don-Yehiya et al. (2025) introduce a larger collection of human–LLM dialogues with a more balanced language distribution, we use only Zheng et al. (2024) for human–LLM dialogue in this paper, as metadata details such as language type were missing in Don-Yehiya et al. (2025) during our last inspection of their dataset.

Secondly, the findings in this paper are highly dependent on the performance of the Stanza parser. As pointed out by Chen et al. (2026), experimental results on diachronic syntactic change can vary based on which dependency parser is used. That is to say, we can assume that, using a different parser (e.g., Spacy’s Dependency Parser, Montani et al., 2023), could provide us with slightly different results. However, the major reason for not including results from different existing parsers is that, to the best of our knowledge, Stanza is the most reliable off-the-shelf parser that can be used with dialogue data. One of the future works could be to train different parsers that can be used for parsing the dependency relationships of dialogue data and further verify the results reported in this paper.

Thirdly, the phase-based division of dialogues is rather suboptimal. Other elegant solutions can be, e.g., using a segmentation method to divide the dialogues based on, e.g., topic shifts. This, however, requires either a large human annotation effort to

train a classifier, or reliable off-the-shelf models to find topic shifts automatically. As we are uncertain about the quality of such models for our dataset, we leave this as future work.

Lastly, for the human-LLM dialogues, we didn't distinguish between different LLMs (GPT-4, LLama, etc) of various sizes. Therefore, in this study we cannot answer research questions such as, 'Do LLMs with larger scaling size show stronger tendency of syntactic complexity convergence?', or 'Do LLM-A as a dialogue participant show stronger complexity matching to the human speaker compared to LLM-B?'.

Ethics statement

For all of the selected dialogue datasets, we have checked that there is no revealing of participants' identity. The speaker names are all replaced with unique IDs. One of the potential ethical issues we can foresee for the syntactic complexity-based human-LLM dialogue detection is that such an application might be misused. Even if the classification result shows that an LLM is involved in the conversation, it may turn out that human users (e.g., students) are only using LLMs to assist with their writing in the chat. Therefore, such tool should be used to raise awareness that, 'LLMs are used to support the conversation/text-based chat'. Any judgement or punishment based on such classification results should be carefully evaluated.

In terms of our research topic, syntactic complexity convergence as a linguistic phenomenon, requires us to assume that dialogue participants have a mental model of language use in order to adapt to others. However, as noted in Kandra et al. (2025), LLMs as dialogue participants, are not expected to make conscious decisions to coordinate their conversation with interlocutors. We therefore need to be cautious when dealing with the question of whether LLMs do have a mental model of language use. Although LLMs as dialogue participants, under specific pre-defined evaluation metrics, can show the convergence of syntactic complexity to their communication partner (Section 5.3), it is still hard to answer if such convergence is more than the outcome of statistical learning or whether the LLM really starts to have a mental model like humans to adapt to its communication partner.

We acknowledge that, writing of the experimental code was supported by ChatGPT and GitHub Copilot. We used Gemini for grammar checking.

Supplementary Material

Code and data are available at <https://doi.org/10.5281/zenodo.22206634>, or, for your convenience, at https://github.com/owuQQQ/syntactic_complexity_in_dialogue.

Acknowledgement

The early structure of this paper benefited significantly from insightful discussions with Vera Demberg, particularly regarding the evaluation of syntactic complexity metrics and robustness tests on the convergence phenomenon. We are also grateful for her helpful proofreading of the original draft. This work is funded by the Deutsche Forschungsgemeinschaft (DFG): TRR 318/3 2026 – 438445824, A02 and B07. The research reported in this paper was also supported by the European Research Council (ERC), grant 101088098 – MEM-LANG. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The NLLG Lab gratefully acknowledges support from the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1. This work also acknowledges the funding from DFG project GRK 2853 'Neuroexplicit Models of Language, Vision, and Action' (project number 471607914).

References

- Drew H. Abney, Alexandra Paxton, Rick Dale, and Christopher T. Kello. 2014. *Complexity matching in dyadic conversation*. *Journal of Experimental Psychology: General*, 143(6):2304–2315.
- Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.
- Sergei Bagdasarov and Diego Alves. 2025. *Like a human? a linguistic analysis of human-written and machine-generated scientific texts*. In *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*, pages 38–47, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. *Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature*. In *The Twelfth*

International Conference on Learning Representations.

- Karen Beaman. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Deborah Tannen, editor, *Spoken and Written Language. Exploring Orality and Literacy*, pages 45–80. Ablex.
- Paras Bhatt and Anthony Rios. 2021. [Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3235–3247, Online. Association for Computational Linguistics.
- Terra Blevins, Susanne Schmalwieser, and Benjamin Roth. 2026. [Do language models accommodate their users? a study of linguistic convergence](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Hendrik Buschmeier and Stefan Kopp. 2018. [Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive](#). In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1213–1221, Stockholm, Sweden. IFAAMAS.
- Yanran Chen, Wei Zhao, Anne Breitbarth, Manuel Stoeckel, Alexander Mehler, Dominik Schlechtweg, and Steffen Eger. 2026. [Syntactic language change in english and german: Metrics, parsers, and convergences](#). *PLOS ONE*, 21(4):1–33.
- Gus Cooney and Andrew Reece. 2025. [Naturalturn: a method to segment speech into psychologically meaningful conversational turns](#). *Scientific Reports*, 15(1):39155.
- Vera Demberg and Frank Keller. 2008. [Data from eye-tracking corpora as evidence for theories of syntactic processing complexity](#). *Cognition*, 109(2):193–210.
- Vera Demberg, Asad Sayeed, Philip Gorinski, and Nikolaos Engonopoulos. 2012. [Syntactic surprisal affects spoken word duration in conversational contexts](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367, Jeju Island, Korea. Association for Computational Linguistics.
- Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2025. [The ShareLM collection and plugin: Contributing human-model chats for the benefit of the community](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 167–177, Vienna, Austria. Association for Computational Linguistics.
- Ivar Frisch and Mario Giulianelli. 2024. [LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models](#). In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.
- Simone Fuscone, Benoit Favre, and Laurent Prvot. 2020. [The contribution of dialogue act labels for convergence studies in natural conversations](#). In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. [Large-scale evidence of dependency length minimization in 37 languages](#). *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. [GLTR: Statistical detection and visualization of generated text](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.
- Mario Giulianelli and Raquel Fernández. 2021. [Analysing human strategies of information transmission as a function of discourse context](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.
- Talmy Givón. 1991. [Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure](#). *Studies in Language*, 15(2):335–370.
- John Hale. 2001. [A probabilistic earley parser as a psycholinguistic model](#). In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, NAACL '01*, page 1–8, USA. Association for Computational Linguistics.
- E.C. Holliman, J.J. Godfrey, and J. McDaniel. 1992. [SWITCHBOARD: telephone speech corpus for research and development](#). In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 2, pages 517–520, Los Alamitos, CA, USA. IEEE Computer Society.
- Kellogg W. Hunt. 1970. [Syntactic maturity in schoolchildren and adults](#). *Monographs of the Society for Research in Child Development*, 35(1):iii–67.
- Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. [Do language models exhibit human-like structural priming effects?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.
- Florian Kandra, Vera Demberg, and Alexander Koller. 2025. [LLMs syntactically adapt their language use to their conversational partner](#). In *Proceedings of the*

- 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 873–886, Vienna, Austria. Association for Computational Linguistics.
- Kazunori Komatani and Shogo Okada. 2021. [Multi-modal human-agent dialogue corpus with annotations at utterance and dialogue levels](#). In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. [Similarity of neural network representations revisited](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Xiang Li, Zhiyi Yin, Hexiang Tan, Shaoling Jing, Du Su, Yi Cheng, Huawei Shen, and Fei Sun. 2025. [PRDetect: Perturbation-robust LLM-generated text detection based on syntax tree](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8290–8301, Albuquerque, New Mexico. Association for Computational Linguistics.
- Dekang Lin. 1996. [On the structural complexity of natural language sentences](#). In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, pages 729–733, Copenhagen, Denmark.
- Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.
- Xiaofei Lu. 2010. [Automatic analysis of syntactic complexity in second language writing](#). *International journal of corpus linguistics*, 15(4):474–496.
- V. Mantha, J. Hamaker, Nakshatra Deshmukh, Aravind Ganapathiraju, and Joseph Picone. 1999. [Improved monosyllabic word modeling on SWITCHBOARD](#). Technical report, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, MS, USA.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. [DetectGPT: Zero-shot machine-generated text detection using probability curvature](#). In *International Conference on Machine Learning*, pages 24950–24962. PMLR.
- Aron Molnar, Jaap Jumelet, Mario Giulianelli, and Arabella Sinclair. 2023. [Attribution and alignment: Effects of local context repetition on utterance production and comprehension in dialogue](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 254–273, Singapore. Association for Computational Linguistics.
- Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. [explosion/spacy: v3.7.2: Fixes for apis and requirements](#).
- Joakim Nivre. 2010. [Dependency parsing](#). *Language and Linguistics Compass*, 4(3):138–152.
- Martin J. Pickering and Simon Garrod. 2004. [The interactive-alignment model: Developments and refinements](#). *Behavioral and Brain Sciences*, 27(2):212–225.
- Martin J. Pickering and Simon Garrod. 2006. [Alignment as the basis for successful communication](#). *Research on Language and Computation*, 4(2):203–228.
- Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. [Universal Dependency parsing from scratch](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Audio and Speech Processing (eess.AS)*, arXiv:2212.04356.
- Andrew Reece, Gus Cooney, Peter Bull, Arvin Chung, Benjamin Dawson, Katherine Mavor, and Richard Palumbo. 2023. [The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation](#). *Science Advances*, 9(12):eadf3197.
- David Reitter and Johanna D Moore. 2006. [Priming of syntactic rules in task-oriented dialogue and spontaneous conversation](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.
- David Reitter and Johanna D. Moore. 2007. [Predicting success in dialogue](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.
- David Reitter and Johanna D. Moore. 2014. [Alignment and task success in spoken dialogue](#). *Journal of Memory and Language*, 76:29–46.
- Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. [Structural persistence in language models: Priming as a window into abstract language representations](#). *Transactions of the Association for Computational Linguistics*, 10:1031–1050.
- Gabriel Skantze. 2021. [Turn-taking in conversational systems and human-robot interaction: A review](#). *Computer Speech & Language*, 67:101178.
- Mayank Soni, Brendan Spillane, Leo Muckley, Orla Cooney, Emer Gilmartin, Christian Saam, Benjamin Cowan, and Vincent Wade. 2022. [An empirical study of topic transition in dialogue](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 92–99, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

- Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. [DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.
- Benedikt M. Szmrecsányi. 2004. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, volume 2, pages 1031–1038, Louvain-la-Neuve, Belgium.
- Olcaç Türk, Petra Wagner, Hendrik Buschmeier, Angela Griminger, Yu Wang, and Stefan Lazarov. 2023. [Mundex: A multimodal corpus for the study of the understanding of explanations](#). In *Proceedings of the 1st International Multimodal Communication Symposium*, pages 63–64, Barcelona, Spain.
- Mayumi Usami, editor. 2023. *Building of a Japanese 1000 Person Natural Conversation Corpus for Pragmatic Analyses and Its Multilateral Studies, and NINJAL Institute-Based Projects: Multiple Approaches to Analyzing the Communication of Japanese Language Learners*. NINJAL.
- Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. [GPT-who: An information density-based machine-generated text detector](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115, Mexico City, Mexico. Association for Computational Linguistics.
- Yu Wang and Hendrik Buschmeier. 2024. [Revisiting the phenomenon of syntactic complexity convergence on German dialogue data](#). In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 75–80, Vienna, Austria. Association for Computational Linguistics.
- Yang Xu. 2021. [Global divergence and local convergence of utterance semantic representations in dialogue](#). In *Proceedings of the Society for Computation in Linguistics 2021*, pages 116–124, Online. Association for Computational Linguistics.
- Yang Xu and David Reitter. 2016. [Convergence of syntactic complexity in conversation](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448, Berlin, Germany.
- Yang Xu, Yu Wang, Hao An, Zhichen Liu, and Yongyuan Li. 2024. [Detecting subtle differences between human and model languages using spectrum of relative likelihood](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10121, Miami, Florida, USA. Association for Computational Linguistics.
- Zuhao Yang, Yingfang Yuan, Yang Xu, Shuo Zhan, Huijun Bai, and Kefan Chen. 2023. [FACE: Evaluating natural language generation with Fourier analysis of cross-entropy](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 17038–17056. Curran.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. [LMSYS-chat-1m: A large-scale real-world LLM conversation dataset](#). In *The Twelfth International Conference on Learning Representations*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Dialogue Dataset Description

German Dataset: **MUNDEX** is a German dialogue dataset and consists of around 87 task-oriented dialogues and focuses on explanation in interaction (Türk et al., 2023). Each dialogue is a explanation scenario involving a speaker (the explainer) explaining how to play a board game to a recipient (the explainee). The dataset in total consists of 87 dialogues between dyads of German native speakers, with each dialogue spans around 15-30 minutes. At its current stage, speech diarization was manually annotated with the help of Whisper ASR (Radford et al., 2022).

Japanese Dataset: **Hazumi Dataset** is a collection of 33 Japanese dialogues which are about 10 daily life topics (Komatani and Okada, 2021). One speaker (virtual agent controlled by human operator) plays the role that when the other shows no evidence of interest toward current conversation topic, the conversation will be guided toward the next topic. The setting is human agent interaction where the agent is a human operator located in a different room using Wizard-of-Oz method, each dialogue is around 15-20 minutes. Given that the agent’s utterances are still produced by the human operators based on participants’ utterance, we still include this dataset as human-human dialogue dataset. Another dataset is the **BTSJ** 1000 Person Japanese Natural Conversation Corpus (Usami, 2023), which consists of 541 spontaneous conversation with no specific pre-defined topics and each dialogue around 5-25 minutes approximately based

on our check given that there is no official statistics on that.

English Dataset: We use the open-sourced **CANDOR** dataset (Reece et al., 2023), which is a large-scale collection of nearly 1656 naturalistic, open-ended conversations between strangers through the video chat platform ⁶. **CANDOR** dataset has no specific topic and thus the content is closer to daily spontaneous conversation, average length of each dialogue is about 30 minutes. Additionally, we also include **Map Task** dataset (Anderson et al., 1991), which is a collection of 128 task-oriented dialogues. Despite its usage for confirming linguistic alignment phenomenon (e.g., Reitter and Moore, 2006, 2014), to the best of our knowledge, this corpus is not yet used for verifying the syntactic complexity convergence.

Human-LLM Dialogue dataset: we used the Human-LLM dialogue data provided by Zheng et al. (2024), which consists of one million real-world conversations between users and 25 large language models. These ‘dialogues’ ⁷ were collected “in the wild” through the Vicuna demo and the Chatbot Arena platform (Zheng et al., 2023). We selected the German and English dialogues for our study purpose, given that their sample size is comparatively larger while there is little Japanese data which is qualified for our study.

B Dialogue Dataset Pre-processing

Data Cleaning for Parsing: Since the calculation of syntactic complexity is mostly based on the syntactic structure of the given utterance, an important requirement for this purpose is that the parser must first be able to recognize the words in the transcribed spoken dialogue. In some of our dialogues we found that some of the transcriptions use informal word forms to reflect the original dialogue content instead of the corresponding correct word formats. This approach keeps the originality of the corpus, while adding difficulty for parsing. Another challenge is that special tokens such as ‘<p>’ and ‘<smile>’ are added to the transcriptions in order to reflect pauses and the paralinguistic

feature ‘<smile>’. This, again, is not processable by the parser. As our solution, when we pre-process such utterances, we replace all incorrect word formats with their correct ones in order to maximally increase the parsing accuracy, which is directly related to the accuracy of the calculation of syntactic complexity. For the special tokens, we remove them.

The Motivation of Setting the Phase Number to 8:

The motivation is firstly based on our review of the widely used dialogue corpus *Switchboard* (Holliman et al., 1992), which shows that within the 2438 dialogues, the duration of each one is around 3 to 10 minutes, with an average of 6 minutes (Mantha et al., 1999), each dialogue covers either 1 or 2 topics (they choose one topic at the beginning and later may shift to another topic). A follow-up study by Soni et al. (2022) shows that among the 215 selected dialogues from *Switchboard*, an average topic shift was observed to contain 12 turns and the average number of topics per dialogue is around 8; thus, around every 1 minute, a topic shift could happen.

A recent study in the **CANDOR** corpus proposed that a psychologically meaningful turn is around 12 seconds (Table 1 from Cooney and Reece, 2025). If a topic consists of 12 turns, then 12 turns multiplied by 12 seconds is around 2.4 minutes. Moreover, pauses, breathing, and backchanneling are also part of turns which take extra time (Skantze, 2021). Based on all of the above statistics from previous studies, our heuristic estimate for topic shifting is that within 4 minutes, a conversation topic can start and end. Given that the dialogues used in our study range from 15 minutes to 30 minutes in duration, while it is suboptimal, dividing each dialogue into 8 phases can ensure that each phase covers around 1 to 4 minutes.

The Motivation of Using Linear-CKA in Section 5.3:

The Interactive Alignment Model (Pickering and Garrod, 2004) proposes that, linguistic alignment arises through priming, with dialogue participants increasingly sharing lexical items or syntactic structures as the conversation unfolds. Because priming at one linguistic level, can trigger priming at others, alignment can thus occur across different representational dimensions. Recent work (e.g., Sinclair et al., 2022; Jumelet et al., 2024) shows that transformer-based models also exhibit priming effects, though they differ from

⁶While 1656 is the reported number, when we compiled the data for analysis, we obtained 1855 data points instead.

⁷It is worth noting that, human-LLM conversation is intrinsically different from natural conversation. In natural conversation, for example, interruption and backchanneling are expected to happen from time to time while this is not the case for human-LLM conversation where LLM agents wait human users to finish their turns.

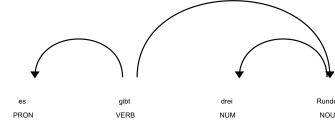
those observed in human communication. We believe that, these studies may overlook the possibility that such differences may produce qualitatively distinct alignment patterns. In the context of syntactic complexity, humans tend to converge within individual dimensions, whereas LLMs may converge across dimensions. We therefore use linear-CKA to capture these dynamics.

Handling Overlapped Speech: A key phenomenon distinguishing natural dialogue from monologue and human-LLM interaction is that speakers frequently interrupt, take turns, or backchannel while another speaks, resulting in overlapping speech. We observed that transcription strategies vary across the natural conversation datasets used in this study. For instance, in the BTSJ dataset, backchannels from one speaker are occasionally embedded directly within the other speaker’s utterance text. Conversely, in the MUNDEX dataset, overlapping speech is segmented chronologically based on which speaker spoke first. To ensure consistency, we parsed all utterances across all corpora; specifically for the BTSJ dataset, we stripped out backchannels embedded within the primary speaker’s turn.

Representation of the Syntactic Complexity:

One fact about the dialogue data we use is that a dialogue consists of many turns, and each turn can contain one or multiple utterances. Some utterances are complete sentences while others are not. In this paper, the measurement is conducted based on turns. Therefore, we use the averages of the syntactic complexity of the utterances to represent the syntactic complexity of each turn. That is to say, for each turn, we first measure how many sentences there are, and based on that, we calculate the syntactic complexity using our proposed metrics. After that, we take the average value from each metric, which is used to represent the syntactic complexity of the utterance in that turn. Here we also add more details by using two examples (dependency-based syntax trees of two sentences) shown in Figure 4 to explain how our proposed metric in Section 3 can capture utterances with different degree of syntactic complexity.

Cleaning the Human-LLM Dialogue Data: For the human-LLM dialogue, we notice that the dataset provided by [Zheng et al. \(2024\)](#) contains a lot of noise. For example, some of the dialogues are about asking the language agents to purify codes,



(a) Simple dependency structure (translation: “There are three rounds”), which corresponds to the example shown in Figure 2.



(b) Complex dependency structure (translation: “So I think I have explained all important things”).

Figure 4: Two examples showing the syntactic relationships according to UD. Edges are directed from heads to dependents. The root arc is omitted for the simplicity.

while others are about translating sentences in one language into another. These types of human-LLM conversation are not considered as ‘dialogue’ in this study. We conduct pre-processing and manual checks to exclude such noise.

Further Details of Dialogue: In Table 4, we report the details of the used dialogues, which include the dialogue type, the number of dialogues in each dataset, and turn information.

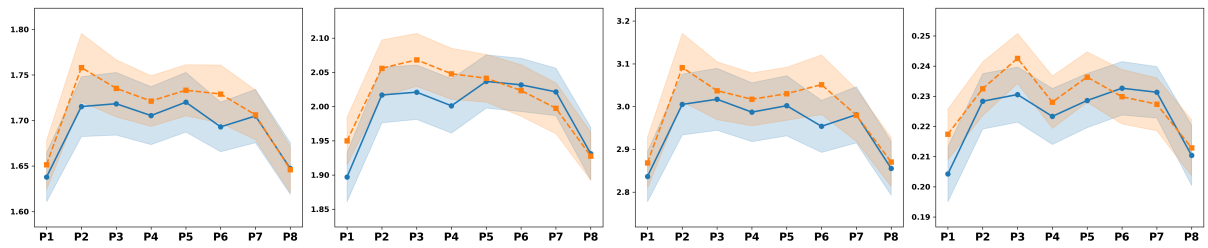
C Supplementary Plots for Figure 3 in in Section 5.1

Here we put supporting plots in Figure 5 for Figure 3 in Section 5.1 which includes the dialogue datasets showing minor or no syntactic convergence Tendency. We chose only to include plots that show obvious syntactic complexity convergence in the main text, while leaving the others in the Appendix.

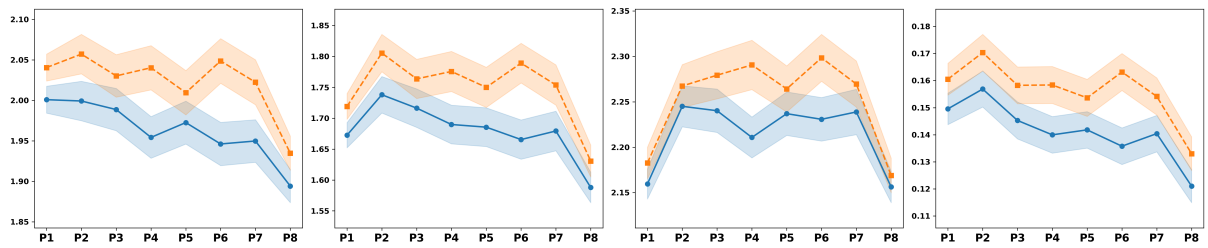
D Embedding Syntactic Complexity Features to FourierGPT Pipeline

In Section 6 we reported the classification results of using quantified syntactic complexity to detect LLM-involved dialogue. Here we report the technical details on how we integrate the syntactic complexity into the FourierGPT pipeline and use that to perform the classification task.

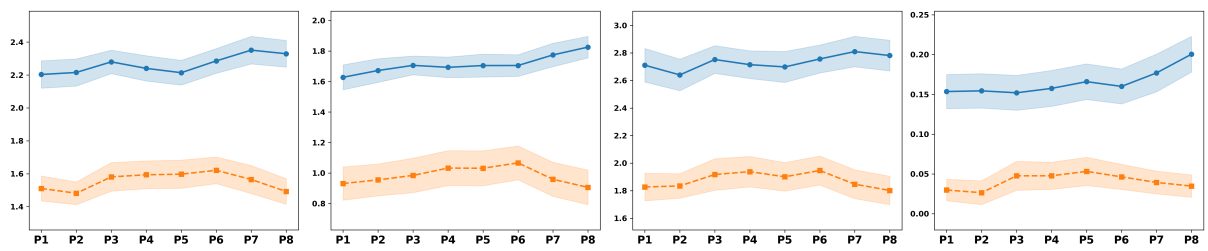
In the original FourierGPT, the detection pipeline works as follows: first, raw likelihood scores are estimated by running a forward pass on the input text. Taking an input of N tokens, $t_1, \dots, t_N$, the likelihood score of the i th token s_i is computed



(a) BTSJ Corpus (jpn)

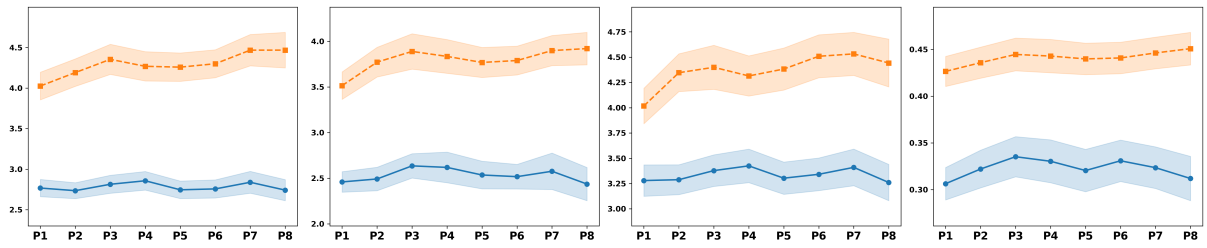


(b) Candor Corpus (eng)

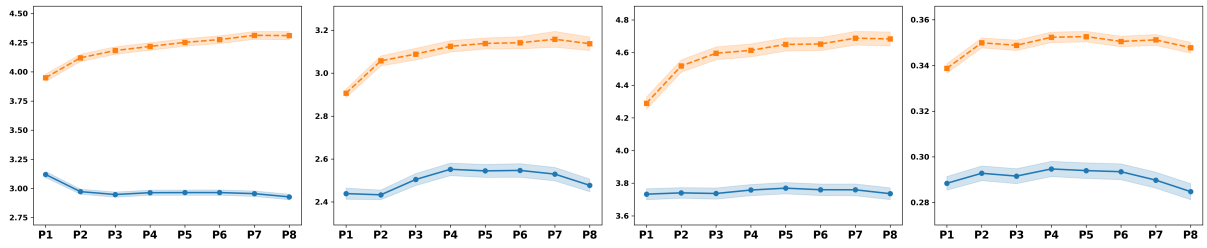


(c) Map Task (eng)

Figure 5: Syntactic complexity across phases across dialogue datasets (Continued on next page).



(d) LMSYS-Chat-1M (LLM ger)



(e) LMSYS-Chat-1M (LLM eng)

Figure 5: The plots here, as supplementary plots to Figure 3, show the syntactic complexity across phases for all dialogue datasets except *Mundex* and *Hazumi*. **Blue line is Dialogue Initiator, Yellow line is Dialogue Follower.** The four metric values in each sub-figure from left to right are: **syntactic complexity value, dependency length, tree depth, and complex Dependency Structure Ratio.** Syntactic complexity convergence is considered a progressive reduction in the syntactic complexity gap between the two interlocutors over the course of the interaction (from phase 1 to phase 8). We calculate the mean value based on the 8 phases we pre-defined with 95% confidence interval. We conduct pairwise t -tests for the value in each phase to make sure that the group of the values from current phase is statistically different from the group of the values from the next phase (except the last phase), the results, reported in Table 13 in the Appendix, indicate a strong degree of statistic significance.

Category	Dataset	Dialogue Type	Number of Dialogue	Avg. Turns (Init. / Fol.)	Min-Max Turns (Init. / Fol.)
Human-Human	MUNDEX	Task-oriented	87	121.07 / 134.62	2-426 / 3-771
	HAZUMI	Social/Open	33	74.30 / 83.61	57-95 / 63-120
	BTSJ	Social/Open	505	142.78 / 142.77	5-484 / 4-485
	CANDOR	Social/Open	1855	167.91 / 167.93	3-917 / 4-917
	MapTask	Task-oriented	121	166.32 / 130.69	24-535 / 18-432
Human-LLM	LLM_GERMAN	Instruction Mixed	9849	1.95 / 1.95	1-45 / 1-45
	LLM_ENGLISH	Instruction Mixed	476737	1.65 / 1.65	1-215 / 1-215

Table 4: Descriptive Statistics: Dialogue Types and Interaction Metrics per Dataset. Please note that the number of dialogue here is total dialogue number in the dataset, which is different from number of the dialogues used for analysis.

Dataset	Complexity Score	Dependency Length	Tree Depth	Dependency Ratio
Train: Original	2.9818	2.9826	4.0773	0.3721
Train: Simplification	2.9159	2.8459	3.8692	0.3725
Test: Original	2.9824	2.9887	4.0626	0.3734
Test: Simplification	2.9259	2.8621	3.8757	0.3733

Table 5: The average of the syntactic complexity metric values from sentences in train and test set of **DEplain** corpus. The higher the values, the more syntactically complex the sentences are. Our integrated metrics in general can capture the lower syntactic complexity in the simplified sentences in both training and testing datasets except *Dependency Ratio* in Training dataset.

by $s_i = -\log P(t_i | t_1, \dots, t_{i-1})$, i.e., the surprisal (negative log probability) returned by the language model (also called the *estimator* model). Then, the next step is to obtain the spectrum view of the likelihood sequence $\tilde{s}_0, \dots, \tilde{s}_N - 1$ as input. Discrete Fourier transform (**DFT**) is applied according to the following:

$$X(\omega_k) \triangleq \sum_{n=0}^{N-1} \tilde{s}_n e^{-j\omega_k n}. \quad (1)$$

The result is a set of complex numbers $\mathcal{F} = \{X(\omega_k)\}_{k=0, \dots, N-1}$ as the frequency-domain representation of the input time-domain signal (likelihood scores), in which ω_k is the k -th frequency component. The starting index of $\tilde{s}_i$ is set to 0 because **DFT** requires $k = 0$ as the lowest frequency component. $X(\omega_k)$ is a complex number made up of *real* and *imaginary* parts, $X(\omega_k) = \text{Re}(X(\omega_k)) + \text{Im}(X(\omega_k))$. The norm $\|X(\omega_k)\| = \sqrt{\text{Re}(X(\omega_k))^2 + \text{Im}(X(\omega_k))^2}$ represents the intensity of the k th component ω_k . Finally, the sequence $\{\|X(\omega_k)\|\}_{k=0, \dots, N-1}$ is used as the spectrum-view of likelihood, which then provides features for the classification step. The range of ω_k is $[0, \pi]$. Based on an intuitive interpretation provided by Yang et al. (2023), the likelihood score $\tilde{s}_i$ at the level of $\|X(\omega_k)\|$ tends to occur every $1/\omega_k$ tokens in the text data.

In our study, we take syntactic complexity as a

linguistic feature which can potentially describe the difference of human-human and human-LLM dialogue. As **quantified syntactic complexity metric values, similar to surprisal, fluctuate during the development of the dialogue**, it is rational to use **DFT** to represent the data. However different from the interpretation of periodicity of the likelihood (Xu et al., 2024), for the syntactic complexity we only assume the fluctuation of the value along the time step. The spectrum view of the four syntactic complexity metric values is thus rather a feature representation we use to capture the fluctuation of the syntactic complexity. Formally, for dialogue d , we have four syntactic-complexity-score based sequences:

$$\mathbf{s}^{(d,m)} = \{s_t^{(d,m)}\}_{t=0}^{T_{d,m}-1}, \quad m \in \{1, 2, 3, 4\}, \quad (2)$$

Where d denotes dialogue id, m denote metric index, T is the set of the time steps. We apply fast Fourier transform to each syntactic complexity metric value sequence and take the power spectrum p_k :

$$\begin{aligned} \hat{s}_k^{(d,m)} &= \sum_{t=0}^{T_{d,m}-1} s_t^{(d,m)} e^{-i2\pi kt/T_{d,m}} \\ p_k^{(d,m)} &= \left| \hat{s}_k^{(d,m)} \right|^2 \end{aligned} \quad (3)$$

We first compute the spectral representation of each dialogue’s syntactic complexity across metric m :

$$\mathbf{x}^{(d,m)} = [p_0^{(d,m)}, \dots, p_{T_{d,m}-1}^{(d,m)}]. \quad (4)$$

We then concatenate the four spectra into a single feature vector:

$$\mathbf{x}^{(d)} = \mathbf{x}^{(d,1)} \parallel \mathbf{x}^{(d,2)} \parallel \mathbf{x}^{(d,3)} \parallel \mathbf{x}^{(d,4)}. \quad (5)$$

Let $\ell_d = |\mathbf{x}^{(d)}|$ denote the length of the feature vector representing dialogue d , and let $y_d \in \{0, 1\}$ be its binary classification label.

As reported in Table 4, human–LLM dialogues generally contain significantly fewer turns than human–human dialogues. Consequently, a classifier might inadvertently rely on dialogue length as a shortcut feature. To evaluate how effectively quantified syntactic complexity metrics distinguish human–human from human–LLM dialogues compared to surprisal-based values, we introduce a length-balanced sampling strategy. After extracting the **DFT** spectral values and selecting the top- k frequencies, we construct B quantile bins over the dialogue length distribution $\{\ell_d\}$, ensuring dialogues of similar length fall into the same bin.

For each bin b , we define the set of class-specific dialogues as:

$$\mathcal{D}_{b,c} = \{d : \ell_d \in \text{bin } b, y_d = c\}, \quad c \in \{0, 1\}, \quad (6)$$

where c represents the ground-truth label ($c = 0$ for human–human dialogues and $c = 1$ for human–LLM dialogues). To prevent length bias and class imbalance, we uniformly sample without replacement exactly

$$n_b = \min\{|\mathcal{D}_{b,0}|, |\mathcal{D}_{b,1}|\} \quad (7)$$

dialogues from each class within bin b .

This procedure guarantees that each bin contains an equal number of human–human and human–LLM dialogues of comparable lengths. The resulting balanced dataset is then fed into the training pipeline, forcing the classifiers to rely strictly on syntactic complexity features during evaluation on the test set. To ensure fair comparability, we also apply this same quantile binning strategy to the surprisal-based FourierGPT baseline.

E Additional Experiment on Dialogue Length Effect

One of the primary issues we address here is whether longer dialogues exhibit a higher potential for syntactic complexity convergence than shorter ones. To answer this question, we conducted an experiment in which we incrementally excluded dialogues with fewer than 16, 24, 32, and 36 turns

from each dataset. We then recalculated the syntactic complexity convergence ratio following the methodology in Section 5.2; the results are summarized in Tables 9 through 12 in the Appendix.

When restricting the analysis to longer interactions by excluding short dialogues, MUNDEX displays an increasing ratio of syntactic complexity convergence. However, length-dependent effects cannot be confirmed for the Hazumi, BTSJ, and Candor datasets under our pre-defined settings, as all dialogues in these three corpora already exceed 36 turns. For human–LLM dialogues, no valid syntactic complexity convergence could be confirmed for longer interactions.

We conducted a similar experiment to test whether syntactic complexity convergence becomes more pronounced across different metrics in dialogues with more turns. The results, summarized in Table 14 in the Appendix, reveal that after excluding short dialogues, BTSJ and LLM German no longer display significant complexity convergence, whereas MUNDEX maintains a strong effect. For the LLM English dataset, complexity convergence holds under all length constraints. The lack of significant convergence in LLM German is likely attributable to sample size reduction, as very few dialogues remain after filtering (e.g., only 12 dialogues remain after excluding those with fewer than 24 turns), making meaningful correlation analysis unfeasible.

F Evaluation of Our Integrated Syntactic Complexity Metrics

To evaluate the validity of our integrated syntactic complexity metrics, we conducted two experiments. In the first experiment, we calculated the pairwise coefficient of determination (R^2) among the four syntactic complexity metrics across all dialogue datasets. Each dataset yields two data points (representing the syntactic complexity of the initiator and follower, respectively). The rationale is that weak or absent correlations between metric pairs indicate that each of the four metrics captures distinct, non-redundant aspects of syntactic complexity. The resulting heat map is shown in Figure 6.

We observed that across all dialogue datasets, Tree Depth and Complex Dependency Ratio consistently exhibit a weak correlation as a metric pair. Furthermore, among the 16 data points, 9 (from *Hazumi Initiator* downward) have more than three metric pairs showing weak correlations. Although

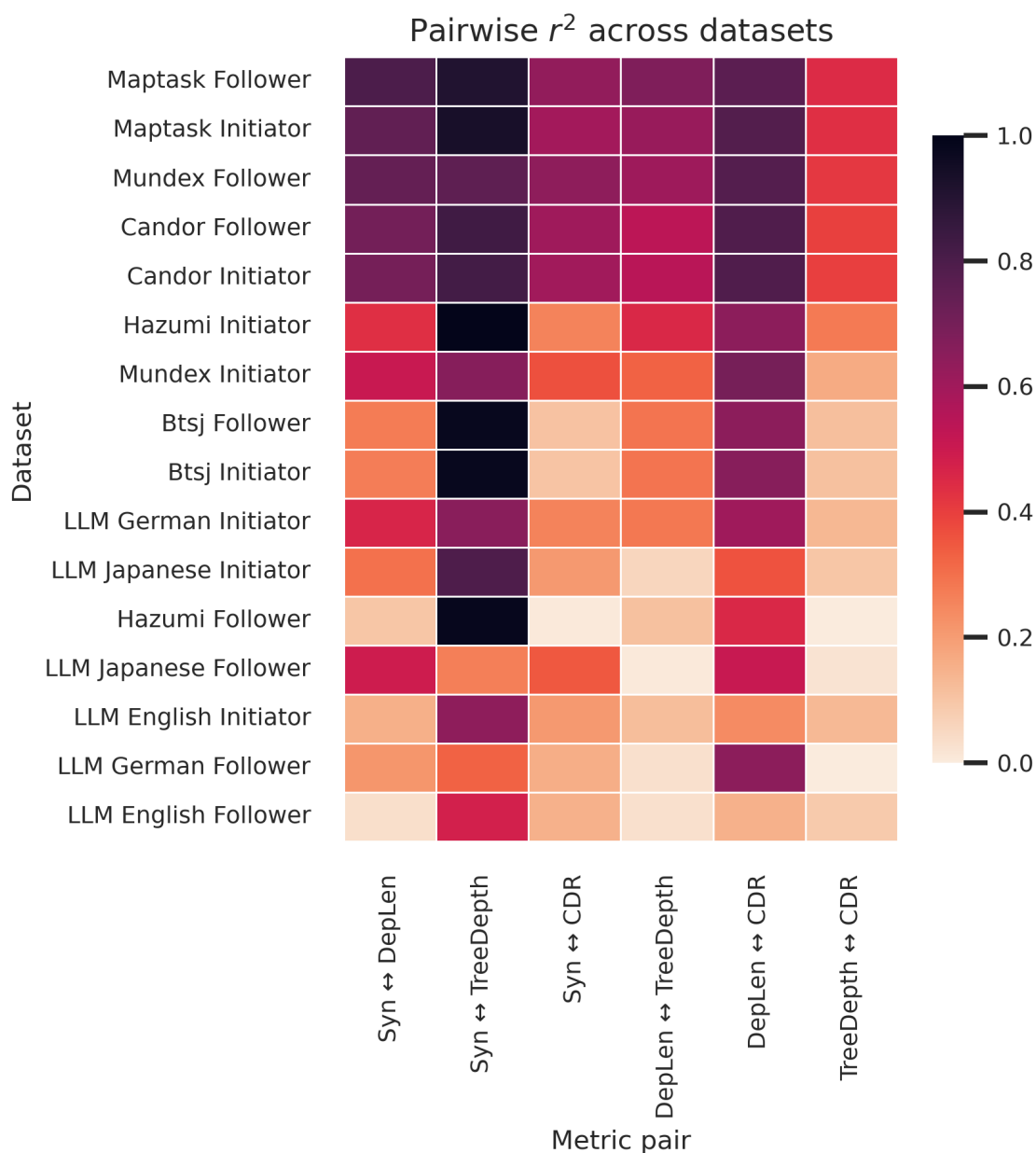


Figure 6: Inter-Correlation of the 4 Integrated Metrics in All Dialogue Datasets. 4 syntactic complexity metrics give us 6 distinct metric pairs. *syn*, *DepLen*, *CDR* denote syntactic complexity score, dependency length and complex dependency ratio. Darker colours indicate strong correlation, lighter colors indicate weak or no correlation.

Syntactic Complexity Score and Tree Depth generally display a strong correlation, they still show a weak correlation in *LLM Japanese Follower* and *LLM German Follower*. These findings justify using all four syntactic complexity metrics, as no single metric can sufficiently represent the others.

In our second experiment, we evaluated the validity of the integrated metrics on DEplain, an open-source parallel corpus. Its training set contains 10,661 sentence pairs, while its test set contains

1,232 sentence pairs. We applied our integrated syntactic complexity metrics to the dataset to obtain metric values for each sentence pair and calculated the average value for all sentences. As reported in Table 5, the simplified sentences generally exhibit lower syntactic complexity metric values, with the exception of the complex dependency ratio in the training set.

We acknowledge that in the DEplain dataset, simplified sentences may feature simpler lexical choices

rather than structural simplifications. Therefore, for our qualitative analysis, we sampled six sentence pairs where the simplified sentences were specifically selected for their lower syntactic complexity. Applying our proposed metrics to these sentences showed that our four metrics generally capture syntactic complexity effectively, as illustrated in Figure 7.

G GPU Usage, Parameters Used for Training, and Further Robustness Test for the Classification Shown in Section 6

To measure syntactic complexity metrics, we used an A6000 64GB GPU to accelerate calculations. The total computation for the human–human dialogues took around 26 hours, whereas the larger human–LLM dataset required approximately 3 days. We also used the GPU to train the transformer-based encoder for human–LLM dialogue detection, which took around 20 minutes. All hyperparameter settings are listed in Table 7 for reference. For classifier training, we sampled 20,000 human–LLM dialogues with varying turn counts from the English subset of LMSYS-Chat-1M. We used the Candor dataset (1,855 dialogues) as the human–human benchmark. Consequently, the majority-class baseline accuracy is $20,000/21,855 \approx 91.5\%$. The results in Table 3 demonstrate that our classifiers successfully learned to distinguish human–human from human–LLM dialogues, with both models achieving over 95.5% accuracy. Additionally, as reported in Table 4, the average turn count for LLM-involved conversations in the *LLM_English* dataset is relatively low (1.65 turns). To test the robustness of our syntactic complexity–based classification against turn-count bias, we evaluated whether performance holds after filtering out short human–LLM interactions. Removing dialogues with fewer than 8 turns per participant left 7,850 human–LLM conversations. We then merged these with the human–human dialogues and re-ran the classification task. As shown in Table 6, filtering out short dialogues did not significantly alter classification performance.

H The Algorithm Used for Recursive Tree Depth Measurement

Algorithm 1 Dependency Tree Depth Computation

Require: Sentence S , dependency parser P

Ensure: Tree depth D

```

1:  $doc \leftarrow P(S)$ 
2: if  $doc$  has no sentences then
3:   return 0
4: end if
5:  $sent \leftarrow$  first sentence in  $doc$ 
6: Initialize empty dictionary  $C$   $\triangleright C$  maps head
   IDs to lists of child IDs
7:  $R \leftarrow$  None  $\triangleright$  Root ID
8: for each word  $w$  in  $sent$  do
9:   if  $w.head = 0$  then
10:     $R \leftarrow w.id$   $\triangleright$  Found root
11:   else
12:     Append  $w.id$  to  $C[w.head]$ 
13:   end if
14: end for
15: if  $R$  is None then
16:   return 0
17: end if
18: function DEPTH( $node$ )
19:   if  $node \notin C$  then
20:     return 1
21:   else
22:     return  $1 + \max_{child \in C[node]} (DEPTH(child))$ 
23:   end if
24: end function
25:  $D \leftarrow DEPTH(R)$ 
26: return  $D$ 

```

So sollen die Treibhaus-Gase in der EU bis zum Jahr 2030 um mindestens 55 Prozent verringert werden.	Bis zum Jahr 2030 soll es in der EU um 55 Prozent weniger Treibhaus-Gase geben.
3.06, 4.1 , 4.0, 0.4	3.07, 3.94 , 4.0, 0.39
Die österreichische Fußball-Mannschaft Rapid Wien hat am Dienstag in der Qualifikation für die Champions League gegen Gent aus Belgien gespielt.	Rapid Wien ist eine von den besten Fußball-Mannschaften aus Österreich. Am Dienstag hat Rapid gegen Gent aus Belgien gespielt.
3.40 , 3.78 , 5.0 , 0.35	3.06 , 3.0 , 4.0 , 0.45
Wegen der vielen Neu-Infektionen haben Deutschland und Belgien die Stadt Wien zum Corona-Risikogebiet erklärt.	In Wien infizieren sich derzeit viele Menschen mit dem Corona-Virus. Deshalb haben Deutschland und Belgien eine Reise-Warnung für Wien herausgegeben.
2.77, 3.84 , 4.0, 0.42	3.11, 3.29 , 4.0, 0.33
Quarantäne bedeutet, dass man nicht mit anderen Menschen in Kontakt kommen darf. Man wird von allen anderen Menschen getrennt.	Quarantäne bedeutet, man wird von allen anderen Menschen getrennt.
3.35 , 4.05, 4.0, 0.5	3.13 , 4.3, 4.0, 0.6
Als sexueller Missbrauch gelten alle sexuellen Handlungen an Menschen, die das gar nicht wollen.	Sexueller Missbrauch sind sexuelle Handlungen an Menschen, die das gar nicht wollen.
3.5, 3.13 , 5.0, 0.4	3.5, 3.0. , 5.0, 0.46
Dann sind bei privaten Feiern und Veranstaltungen in Innenräumen nur noch 10 Personen erlaubt.	Bei privaten Feiern im Inneren dürfen dann nicht mehr als 10 Personen dabei sein.
3.0, 3.79 , 4.0, 0.36	3.5, 3.0 , 5.0, 0.27

Figure 7: This figure presents six sentence pairs, where the sentences in the green boxes represent the original inputs and those in the purple boxes represent the simplified versions, which we consider syntactically simple. The values in the blue boxes correspond to four syntactic complexity metrics, capturing complexity from distinct linguistic perspectives. Values in bold indicate instances where the change in syntactic complexity was successfully captured by a metric. The values are ordered from left to right as follows: **Syntactic Complexity Value**, **Dependency Length**, **Tree Depth**, **Complex Dependency Ratio**.

Input	Model	F1	Accuracy	AUC
Syn Com (Original Dialogues)	Spectrum Length (baseline)	0.745 (0.070)	0.753 (0.060)	0.757 (0.063)
	Linear Regression	0.962 (0.040)	0.960 (0.044)	1.000 (0.000)
	Transformer Encoder	0.933 (0.067)	0.927 (0.081)	0.982 (0.025)
Syn Com (Filtered Dialogues)	Spectrum Length (baseline)	0.836 (0.054)	0.832 (0.060)	0.832 (0.060)
	Linear Regression	0.979 (0.016)	0.978 (0.017)	0.999 (0.003)
	Transformer Encoder	0.964 (0.030)	0.962 (0.033)	0.997 (0.005)

Table 6: Comparison with dialogues all kept against dialogue with more than 8 turns kept

Component	Parameters / Values
Iterations	$n_iterations = 10$
LLM sampling	$llm_sample_size = 20000$
Length control	$length_match, length_match_bins = 20$
Feature selection	$fourier_percent \in \{100, 50, 25\}$ $fourier_top_k$ (optional), $fixed_len$ (optional)
Normalization	$power_norm \in \{none\}$
Splits (logistic Regression)	$test_size = 0.3$ (stratified)
Splits (Transformer Encoder)	$train/val = 0.8/0.2$, $test = 0.3$ if no ID split
Traditional model	StandardScaler; SMOTE (if enabled)
Logistic Regression	$max_iter = 1000$, $class_weight = balanced$
Transformer encoder	$D = 128$, $heads = 4$, $layers = 2$, $dropout = 0.1$
FFN / classifier	FFN dim = 256, classifier hidden = 64
Optimizer	Adam, $lr = 10^{-3}$
Scheduler	ReduceLROnPlateau (factor = 0.5, patience = 5)
Training	$epochs = 20$, $batch_size = 32$

Table 7: Training parameters used for, bin sampling and training the logistic regression, and transformer encoder.

Dataset	Metric	N	Both Down (%)	Converge (%)	Total (%)
MUNDEX	Syn Value		34.00	42.00	76.00
	Dependency Length		22.00	44.00	66.00
	Tree Depth		14.00	42.00	56.00
	Complex Dependency Ratio		4.00	56.00	60.00
	OVERALL (any)	50	38.00	48.00	86.00
HAZUMI	Syn Value		9.09	42.42	51.52
	Dependency Length		15.15	60.61	75.76
	Tree Depth		6.06	27.27	33.33
	Complex Dependency Ratio		0.00	69.70	69.70
	OVERALL (any)	33	21.21	75.76	96.97
BTSJ	Syn Value		7.03	34.34	41.37
	Dependency Length		10.64	38.35	49.00
	Tree Depth		2.61	24.30	26.91
	Complex Dependency Ratio		6.83	46.39	53.21
	OVERALL (any)	498	18.07	57.43	75.50
CANDOR	Syn Value		16.71	21.99	38.71
	Dependency Length		16.87	27.01	43.88
	Tree Depth		1.08	20.86	21.94
	Complex Dependency Ratio		12.61	33.80	46.42
	OVERALL (any)	1855	28.95	40.81	69.76
MAPTASK	Syn Value		5.79	38.02	43.80
	Dependency Length		4.96	37.19	42.15
	Tree Depth		4.13	30.58	34.71
	Complex Dependency Ratio		2.48	37.19	39.67
	OVERALL (any)	121	9.92	50.41	60.33
LLM_GERMAN	Syn Value		15.38	27.78	43.16
	Dependency Length		8.97	26.92	35.90
	Tree Depth		9.40	27.78	37.18
	Complex Dependency Ratio		5.56	41.03	46.58
	OVERALL (any)	234	23.50	51.28	74.79
LLM_ENGLISH	Syn Value		14.28	26.94	41.22
	Dependency Length		8.57	32.94	41.52
	Tree Depth		8.18	27.48	35.66
	Complex Dependency Ratio		8.08	37.03	45.11
	OVERALL (any)	7850	27.27	50.00	77.27

Table 8: Syntactic complexity convergence: Metric-specific and overall dialogue counts across datasets.

Dataset	N	Metric	Obs	Random Mean	NullSD	p-val	Cohen d	CI
MUNDEX	40	syn value	0.7750	0.4800	0.0783	0.0000***	3.7649	[0.1999, 0.3897]
		dependency length	0.7000	0.4930	0.0827	0.0150*	2.5046	[0.0996, 0.3147]
		tree depth	0.5500	0.2417	0.0649	0.0000***	4.7541	[0.1930, 0.4250]
		complex dependency ratio	0.6750	0.4675	0.0770	0.0067**	2.6925	[0.0925, 0.3151]
HAZUMI	33	syn value	0.5152	0.4864	0.0829	0.4384	0.3472	[-0.0951, 0.1636]
		dependency length	0.7576	0.4937	0.0869	0.0020**	3.0355	[0.1563, 0.3758]
		tree depth	0.3333	0.3723	0.0795	0.7495	-0.4898	[-0.1617, 0.0823]
		complex dependency ratio	0.6970	0.4942	0.0854	0.0101*	2.3762	[0.0940, 0.3392]
MAPTASK	121	syn value	0.4380	0.4875	0.0435	0.8987	-1.1376	[-0.1140, 0.0188]
		dependency length	0.4215	0.4896	0.0444	0.9453	-1.5342	[-0.1306, 0.0037]
		tree depth	0.3471	0.3446	0.0439	0.5027	0.0572	[-0.0625, 0.0688]
		complex dependency ratio	0.3967	0.4515	0.0443	0.9160	-1.2375	[-0.1182, 0.0110]
BTSJ	476	syn value	0.4055	0.4696	0.0230	0.9987	-2.7841	[-0.0982, -0.0288]
		dependency length	0.4895	0.4904	0.0226	0.5307	-0.0379	[-0.0343, 0.0303]
		tree depth	0.2584	0.2875	0.0201	0.9427	-1.4472	[-0.0595, -0.0003]
		complex dependency ratio	0.5315	0.4940	0.0217	0.0440*	1.7287	[0.0044, 0.0713]
CANDOR	1855	syn value	0.3871	0.4817	0.0113	1.0000	-8.4072	[-0.1109, -0.0793]
		dependency length	0.4388	0.4828	0.0115	0.9994	-3.8377	[-0.0614, -0.0265]
		tree depth	0.2194	0.2102	0.0090	0.1611	1.0234	[-0.0047, 0.0229]
		complex dependency ratio	0.4642	0.4467	0.0114	0.0651†	1.5279	[-0.0019, 0.0358]
LLM GERMAN	42	syn value	0.5000	0.4903	0.0790	0.5032	0.1225	[-0.1058, 0.1252]
		dependency length	0.3095	0.4985	0.0787	0.9952	-2.4001	[-0.2985, -0.0819]
		tree depth	0.4524	0.4044	0.0747	0.2921	0.6426	[-0.0711, 0.1557]
		complex dependency ratio	0.4524	0.4908	0.0800	0.7365	-0.4804	[-0.1575, 0.0744]
LLM ENGLISH	1313	syn value	0.3983	0.4921	0.0138	1.0000	-6.7929	[-0.1158, -0.0742]
		dependency length	0.4242	0.4905	0.0137	1.0000	-4.8344	[-0.0874, -0.0466]
		tree depth	0.3618	0.3978	0.0137	0.9960	-2.6348	[-0.0550, -0.0171]
		complex dependency ratio	0.4593	0.4766	0.0140	0.9000	-1.2385	[-0.0396, 0.0034]

Table 9: Convergence results in dialogues with at least 16 turns for each speaker.

Dataset	N	Metric	Obs	Random Mean	NullSD	p-val	Cohen d	CI
MUNDEX	37	syn value	0.7838	0.4850	0.0857	0.0000***	3.4861	[0.1968, 0.3950]
		dependency length	0.7297	0.4917	0.0792	0.0036**	3.0031	[0.1273, 0.3416]
		tree depth	0.5946	0.2284	0.0666	0.0000***	5.4975	[0.2477, 0.4916]
		complex dependency ratio	0.6757	0.4678	0.0803	0.0108*	2.5873	[0.0922, 0.3322]
HAZUMI	33	syn value	0.5152	0.4864	0.0829	0.4384	0.3472	[-0.0951, 0.1636]
		dependency length	0.7576	0.4937	0.0869	0.0020**	3.0355	[0.1563, 0.3758]
		tree depth	0.3333	0.3723	0.0795	0.7495	-0.4898	[-0.1617, 0.0823]
		complex dependency ratio	0.6970	0.4942	0.0854	0.0101*	2.3762	[0.0940, 0.3392]
MAPTASK	120	syn value	0.4417	0.4860	0.0452	0.8560	-0.9793	[-0.1110, 0.0209]
		dependency length	0.4250	0.4879	0.0455	0.9307	-1.3834	[-0.1279, -0.0007]
		tree depth	0.3417	0.3449	0.0452	0.5720	-0.0722	[-0.0672, 0.0661]
		complex dependency ratio	0.4000	0.4500	0.0438	0.8960	-1.1395	[-0.1166, 0.0158]
BTSJ	444	syn value	0.4054	0.4711	0.0224	0.9987	-2.9279	[-0.1002, -0.0318]
		dependency length	0.4752	0.4899	0.0242	0.7480	-0.6046	[-0.0481, 0.0191]
		tree depth	0.2500	0.2827	0.0209	0.9467	-1.5633	[-0.0641, -0.0027]
		complex dependency ratio	0.5293	0.4908	0.0228	0.0520†	1.6861	[0.0022, 0.0755]
CANDOR	1855	syn value	0.3871	0.4817	0.0113	1.0000	-8.4072	[-0.1109, -0.0793]
		dependency length	0.4388	0.4828	0.0115	0.9994	-3.8377	[-0.0614, -0.0265]
		tree depth	0.2194	0.2102	0.0090	0.1611	1.0234	[-0.0047, 0.0229]
		complex dependency ratio	0.4642	0.4467	0.0114	0.0651†	1.5279	[-0.0019, 0.0358]
LLM GERMAN	12	syn value	0.5000	0.4864	0.1350	0.5767	0.1008	[-0.2007, 0.2636]
		dependency length	0.4167	0.4842	0.1486	0.7700	-0.4544	[-0.3175, 0.1408]
		tree depth	0.5000	0.3819	0.1350	0.2700	0.8744	[-0.0962, 0.3323]
		complex dependency ratio	0.5000	0.4853	0.1588	0.5700	0.0927	[-0.1996, 0.2290]
LLM ENGLISH	428	syn value	0.3551	0.4819	0.0236	1.0000	-5.3727	[-0.1607, -0.0909]
		dependency length	0.4089	0.4830	0.0244	1.0000	-3.0372	[-0.1114, -0.0398]
		tree depth	0.3411	0.3702	0.0231	0.9093	-1.2577	[-0.0625, 0.0029]
		complex dependency ratio	0.4720	0.4714	0.0243	0.4987	0.0240	[-0.0344, 0.0363]

Table 10: Convergence results in dialogues with at least 24 turns for each speaker.

Dataset	N	Metric	Obs	Random Mean	NullSD	p-val	Cohen d	CI
MUNDEX	33	syn value	0.8485	0.4903	0.0863	0.0000***	4.1482	[0.2716, 0.4597]
		dependency length	0.7879	0.5011	0.0871	0.0000***	3.2930	[0.1831, 0.3989]
		tree depth	0.6061	0.2198	0.0697	0.0000***	5.5422	[0.2564, 0.5302]
		complex dependency ratio	0.6667	0.4694	0.0858	0.0162*	2.3004	[0.0761, 0.3306]
HAZUMI	33	syn value	0.5152	0.4864	0.0829	0.4384	0.3472	[-0.0951, 0.1636]
		dependency length	0.7576	0.4937	0.0869	0.0020**	3.0355	[0.1563, 0.3758]
		tree depth	0.3333	0.3723	0.0795	0.7495	-0.4898	[-0.1617, 0.0823]
		complex dependency ratio	0.6970	0.4942	0.0854	0.0101*	2.3762	[0.0940, 0.3392]
MAPTASK	119	syn value	0.4454	0.4840	0.0464	0.8213	-0.8321	[-0.1090, 0.0292]
		dependency length	0.4286	0.4876	0.0473	0.9027	-1.2474	[-0.1253, 0.0055]
		tree depth	0.3445	0.3425	0.0415	0.5253	0.0502	[-0.0625, 0.0685]
		complex dependency ratio	0.4034	0.4496	0.0433	0.8867	-1.0679	[-0.1118, 0.0170]
BTSJ	419	syn value	0.3986	0.4711	0.0234	1.0000	-3.1045	[-0.1089, -0.0393]
		dependency length	0.4558	0.4930	0.0244	0.9347	-1.5241	[-0.0719, -0.0019]
		tree depth	0.2411	0.2769	0.0199	0.9707	-1.8050	[-0.0668, -0.0049]
		complex dependency ratio	0.5155	0.4932	0.0242	0.1800	0.9210	[-0.0158, 0.0611]
CANDOR	1855	syn value	0.3871	0.4817	0.0113	1.0000	-8.4072	[-0.1109, -0.0793]
		dependency length	0.4388	0.4828	0.0115	0.9994	-3.8377	[-0.0614, -0.0265]
		tree depth	0.2194	0.2102	0.0090	0.1611	1.0234	[-0.0047, 0.0229]
		complex dependency ratio	0.4642	0.4467	0.0114	0.0651†	1.5279	[-0.0019, 0.0358]
LLM GERMAN	5	syn value	0.4000	0.4893	0.2170	0.8100	-0.4116	[-0.4893, 0.1773]
		dependency length	0.4000	0.5053	0.2239	0.8267	-0.4704	[-0.5053, 0.1613]
		tree depth	0.4000	0.3713	0.2203	0.5933	0.1301	[-0.3713, 0.2953]
		complex dependency ratio	0.4000	0.5340	0.2226	0.8267	-0.6020	[-0.5340, 0.1327]
LLM ENGLISH	184	syn value	0.3533	0.4888	0.0375	1.0000	-3.6128	[-0.1888, -0.0870]
		dependency length	0.4239	0.4910	0.0358	0.9720	-1.8731	[-0.1204, -0.0148]
		tree depth	0.3370	0.3711	0.0354	0.8627	-0.9626	[-0.0841, 0.0199]
		complex dependency ratio	0.4728	0.4835	0.0355	0.6240	-0.2997	[-0.0661, 0.0391]

Table 11: Convergence results in dialogues with at least 32 turns for each speaker.

Dataset	N	Metric	Obs	Random Mean	NullSD	p-val	Cohen d	CI
MUNDEX	32	syn value	0.8438	0.4866	0.0872	0.0000***	4.0981	[0.2634, 0.4634]
		dependency length	0.7812	0.5018	0.0870	0.0021**	3.2113	[0.1801, 0.3931]
		tree depth	0.5938	0.2150	0.0709	0.0000***	5.3408	[0.2555, 0.5122]
		complex dependency ratio	0.6562	0.4731	0.0885	0.0312*	2.0688	[0.0562, 0.3164]
HAZUMI	33	syn value	0.5152	0.4864	0.0829	0.4384	0.3472	[-0.0951, 0.1636]
		dependency length	0.7576	0.4937	0.0869	0.0020**	3.0355	[0.1563, 0.3758]
		tree depth	0.3333	0.3723	0.0795	0.7495	-0.4898	[-0.1617, 0.0823]
		complex dependency ratio	0.6970	0.4942	0.0854	0.0101*	2.3762	[0.0940, 0.3392]
MAPTASK	118	syn value	0.4407	0.4858	0.0420	0.8720	-1.0747	[-0.1092, 0.0277]
		dependency length	0.4237	0.4878	0.0445	0.9373	-1.4423	[-0.1242, 0.0051]
		tree depth	0.3390	0.3437	0.0413	0.5960	-0.1139	[-0.0653, 0.0617]
		complex dependency ratio	0.4068	0.4512	0.0441	0.8693	-1.0072	[-0.1051, 0.0163]
BTSJ	408	syn value	0.3897	0.4714	0.0252	1.0000	-3.2349	[-0.1139, -0.0428]
		dependency length	0.4485	0.4932	0.0245	0.9733	-1.8254	[-0.0785, -0.0067]
		tree depth	0.2353	0.2733	0.0215	0.9680	-1.7652	[-0.0677, -0.0045]
		complex dependency ratio	0.5049	0.4931	0.0245	0.3253	0.4837	[-0.0256, 0.0509]
CANDOR	1855	syn value	0.3871	0.4817	0.0113	1.0000	-8.4072	[-0.1109, -0.0793]
		dependency length	0.4388	0.4828	0.0115	0.9994	-3.8377	[-0.0614, -0.0265]
		tree depth	0.2194	0.2102	0.0090	0.1611	1.0234	[-0.0047, 0.0229]
		complex dependency ratio	0.4642	0.4467	0.0114	0.0651†	1.5279	[-0.0019, 0.0358]
LLM GERMAN	5	syn value	0.4000	0.4893	0.2170	0.8100	-0.4116	[-0.4893, 0.1773]
		dependency length	0.4000	0.5053	0.2239	0.8267	-0.4704	[-0.5053, 0.1613]
		tree depth	0.4000	0.3713	0.2203	0.5933	0.1301	[-0.3713, 0.2953]
		complex dependency ratio	0.4000	0.5340	0.2226	0.8267	-0.6020	[-0.5340, 0.1327]
LLM ENGLISH	134	syn value	0.3507	0.4884	0.0431	1.0000	-3.1923	[-0.1977, -0.0781]
		dependency length	0.4478	0.4885	0.0451	0.8373	-0.9032	[-0.1029, 0.0234]
		tree depth	0.3284	0.3680	0.0411	0.8520	-0.9633	[-0.1022, 0.0187]
		complex dependency ratio	0.4552	0.4813	0.0427	0.7493	-0.6111	[-0.0931, 0.0425]

Table 12: Convergence results in dialogues with at least 36 turns for each speaker.

Dataset	Role	syn_value	dep_length	tree_depth	complex_dep
MUNDEX	Initiator	150.33***	116.78***	111.79***	61.09***
	Follower	11.98 ^{ns}	10.84 ^{ns}	7.14 ^{ns}	15.02*
btsj	Initiator	35.24***	89.81***	38.83***	62.15***
	Follower	51.62***	70.94***	51.73***	67.24***
candor	Initiator	121.19***	84.53***	67.98***	105.23***
	Follower	110.36***	91.66***	101.82***	114.61***
Hazumi	Initiator	24.58***	19.61**	25.08***	21.42**
	Follower	41.61***	30.22***	33.23***	5.47 ^{ns}
maptask	Initiator	14.25*	16.68*	5.36 ^{ns}	17.97*
	Follower	10.87 ^{ns}	9.08 ^{ns}	7.27 ^{ns}	12.35 ^{ns}
llm_ger	Initiator	2.79 ^{ns}	8.96 ^{ns}	4.69 ^{ns}	8.11 ^{ns}
	Follower	21.65**	48.17***	22.57**	19.24**
llm_eng	Initiator	96.55***	60.48***	6.35 ^{ns}	38.40***
	Follower	837.49***	759.82***	438.42***	139.33***

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ^{ns} not significant.

Table 13: Friedman test results across datasets for initiator and follower roles across all of selected datasets.

Dataset	MinTurns>=16	MinTurns>=24	MinTurns>=32	MinTurns>=36
MUNDEX	0.0038 (0.111, 0.420)	0.0032 (0.129, 0.383)	0.0061 (0.596, 0.025)*	0.0059 (0.678, 0.012)*
HAZUMI	0.0075 (0.170, 0.310)	0.0075 (0.170, 0.310)	0.0075 (0.170, 0.310)	0.0075 (0.170, 0.310)
BTSJ	-0.0036 (0.875, 0.001)***	-0.0018 (0.604, 0.023)*	-0.0004 (0.079, 0.499)	-0.0000 (0.000, 0.990)
MAPTASK	-0.0011 (0.164, 0.319)	-0.0010 (0.084, 0.486)†	-0.0013 (0.133, 0.375)	-0.0011 (0.082, 0.491)†
CANDOR	-0.0004 (0.005, 0.872)	-0.0004 (0.005, 0.872)	-0.0004 (0.005, 0.872)	-0.0004 (0.005, 0.872)
LLM_ENGLISH	-0.0150 (0.916, 0.000)***	-0.0121 (0.963, 0.000)***	-0.0101 (0.844, 0.001)***	-0.0087 (0.567, 0.031)*
LLM_GERMAN	-0.0172 (0.437, 0.074)†	-0.0093 (0.100, 0.446)	-0.0087 (0.039, 0.638)	-0.0087 (0.039, 0.638)

Table 14: The syntactic complexity convergence across different metrics with longer dialogue length. The value outside of the parenthesis is the slope value. Parentheses show (correlation, p-value). Significance markers: † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.