How Much Does Nonverbal Communication Conform to Entropy Rate Constancy?: A Case Study on Listener Gaze in Interaction

Yu Wang[1]*, **Yang Xu**[2], **Gabriel Skantze**[3], **Hendrik Buschmeier**[1]
[1]Digital Linguistics Lab, Faculty of Linguistics and Literary Studies,
Bielefeld University, Bielefeld, Germany
[2]Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China
[3]Division of Speech, Music and Hearing (TMH),
KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

According to the Entropy Rate Constancy (ERC) principle, the information density of a text is approximately constant over its length. Whether this principle also applies to nonverbal communication signals is still under investigation. We perform empirical analyses of video-recorded dialogue data and investigate whether listener gaze, as an important nonverbal communication signal, adheres to the ERC principle. Results show (1) that the ERC principle holds for listener gaze; and (2) that the two linguistic factors syntactic complexity and turn transition potential are weakly correlated with local entropy of listener gaze.

1 Introduction

In human communication, information is conveyed via various channels. In addition to conveying meaning through linguistic units, interlocutors may express themselves through multimodal signals such as hand gestures, head movements, or prosody.

In face-to-face dialogue, interlocutors monitor signals from different modalities to decide when to take the turn, give feedback, or interject a brief response token (backchannel). Humans may also change their dialogue strategies based on their listeners' reactions, for example, repeating or explaining what has been said (Clark, 1996, p. 39, 378). By investigating such signals, we can gain insights about (1) how engaged dialogue participants are with the content of a conversation, or (2) how common ground (Clark, 1996) and mutual understanding evolve over the course of an interaction.

One important concept for measuring the growth of common ground is alignment of linguistic information (Pickering and Garrod, 2004). Paralinguistic information such as prosody, or nonverbal information such as body posture or gaze direction, showing different levels of convergence during interaction, can serve as evidence of linguistic alignment (Pickering and Garrod, 2004). Recent work introduces methods for investigating alignment of linguistic information through information theoretic measures such as entropy (Shannon, 1948). It is widely argued that linguistic information is uniformly distributed in language use, e.g., in the Uniform Information Density hypothesis (UID; Jaeger, 2010), or the Entropy Rate Constancy principle (ERC; Genzel and Charniak, 2002, 2003). Xu and Reitter (2017)'s analysis of spoken dialogue data shows that entropy of dialogue partners' speech, interactively evolves and converges as the discourse develops. This phenomenon indicates that mutual understanding has been consistently reached over the shifting of topics. Maës et al. (2022) shows that, while global entropy of dialogue remains constant, topic shift is a linguistic factor which leads to peaks in local entropy. Inspired by Xu et al. (2022) and Maës et al. (2022) this paper focuses on listener gaze during interaction.

In this paper, we quantify the information conveyed through gaze. Based upon and extending our previous work (Wang and Buschmeier, 2023), we first investigate gaze local entropy to add further evidence that non verbal communication conforms to the entropy rate constancy principle, complementing the finding of Xu et al. (2022). Our motivation for this is based on four points: (1) During interaction, interlocutors spend most of their time looking at each other (Rogers et al., 2018); (2) New gaze behaviour will emerge while the conversation develops; (3) Interlocutors are less likely to shift gaze from one corner of their visual field to the other when listening, as it can be perceived as being distracted (but see Goodwin 1985, p. 231); (4) If listener gaze targets the speaker, a gaze shift is less likely to happen, unless there is a cognitive processing-based need for gaze aversion (Argyle and Cook, 1976).

*Corresponding author: y.wang@uni-bielefeld.de

Findings of the Association for Computational Linguistics ACL 2024, pages 3533–3545

Based on these four points, if we consider gaze as meaningful communicative units (similar to lexical units), listeners' gaze behavior should be predictable to some extent. We also take a closer look at two high level linguistic factors: syntactic complexity (Xu and Reitter, 2016), and transition relevance places (potential turn-taking cues) (Sacks et al., 1974) in order to investigate if there is a correlation between gaze local entropy and these two linguistic factors. We do this as an initial attempt to explain why nonverbal communication (specifically listener gaze) may conform to the ERC principle. In this paper we specifically investigate the following two research questions:

- Does listener gaze, as a nonverbal signal, adhere to the Entropy Rate Constancy (ERC) principle?
- If so, why does listener gaze conform to ERC principle. Do syntactic complexity and transition relevance places correlate with this phenomenon?

The two main contributions of our work, compared to previous studies, are as follow:

- We add to the existing evidence that the ERC principle also holds for nonverbal signals (specifically listener gaze). We do this here for dyadic conversations in a controlled linguistic context, namely task oriented dialogue. This finding is based upon and extends our previous work (Wang and Buschmeier, 2023), improving its experiment setting.
- We investigate the effects of the two linguistic factors syntactic complexity and turn transition potential, which is based on the prediction of transition relevant places, on listener gaze entropy in conversational interaction.

2 Related Work

2.1 Studies on Gaze in Multimodal Communication

Gaze is an important social cue in human interaction that is used for indicating attention, eliciting feedback, or taking the turn (Kendon, 1967; Duncan, 1975; Goodwin and Goodwin, 1986). Brône et al. (2017) summarize these communicative functions of gaze as following two roles: a participation role, where interlocutors show attentiveness during interaction, and a regulation role, which coordinates the dynamics of speaker-listener role shifting, in addition to other cues, e.g., prosodic cues such as intonation, duration, loudness, and voice quality (Ward, 2019). A recent lab-based study of human interaction (Kendrick et al., 2023), in which gaze was recorded with professional eye trackers, found gaze direction not to be a significant predictor for turn shifts, while gaze aversion was a strong cue for turn holding (which is different from previous finding Nakano et al. 2003).

From the computational side, gaze has been thoroughly investigated for its communicative function in dialogue modeling tasks. For spoken dialogues systems, as well as for the design of human robot interaction, gaze modeling has been used to facilitate the grounding process (Nakano et al., 2003; Skantze et al., 2014). As a multimodal signal in interaction, gaze can also serve as an important feature for turn management (Jokinen et al. 2013; but see Kendrick et al. 2023, above). Onishi et al. (2023) combine gaze direction with other multimodal features (e.g., head rotation) to the baseline model of a self-supervised voice activity projection model (Ekstedt and Skantze, 2022b), and show that they further improve the accuracy of the model's performance on next speaker prediction. From a more theoretical perspective, Eberle et al. (2022) investigated transformer's self-attention mechanism and reveals that it has similar predictive power as task specific human gaze fixation pattern.

2.2 Measuring Information Content in Communication

In both written text and spoken dialog, information theory has been used extensively as a theory for uncovering properties of language. Let X_i denote a linguistic unit (e.g., a single lexical unit or a long utterance). As the relevant context grows, the predictability of the next linguistic unit will be higher, given that enough contextual cues are available as a prior, which increases the reliability of the posterior estimation. The information density of the random variable X_i is estimated as the entropy $H(X)$ (Shannon, 1948).

Genzel and Charniak (2002, 2003) initiated the idea of measuring the information density through an n-gram model on written text. In this study, we follow Xu and Reitter (2018) and Giulianelli et al. (2021) in calculating information content: For a linguistic unit, e.g., a sentence X which comprises of a sequence of smaller units: $\langle w_1, \ldots, w_i \rangle$, where $w_i \in \vartheta$, with ϑ being the set of vocabulary. The information content of the sentence is the aver-

age of the sum of negative logarithm conditional probability defined as:

$$H(X) = -\frac{1}{n} \sum_{w_i \in \vartheta} \log P(w_i \mid w_1, \ldots, w_{i-1})$$

Genzel and Charniak (2002, 2003) further proposed the so called Entropy Rate Constancy (ERC) principle, which stipulates that the rate of transmitted information of a given linguistic unit, from a global perspective of a written text, is roughly constant.

Here, we follow Genzel and Charniak (2002) to explain the ERC principle in details: Let $H(X_i \mid C_i, L_i)$ denote the conditional entropy of the word X_i, where $L_i = X_{i-n+1}, \ldots, X_{i-1}$ is the local n-gram context, and $C_i = X_0, X_1, \ldots, X_{i-1}$ the context which contains all of the words preceding the word X_i. The conditional entropy of the word X_i is then decomposed as:

$$\underbrace{H(X_i \mid C_i, L_i)}_{\text{global entropy}} = \underbrace{H(X_i \mid L_i)}_{\text{local entropy}} - \underbrace{I(X_i; C_i, L_i)}_{\text{mutual information}}$$

$H(X_i \mid C_i, L_i)$ is considered roughly constant based on following reasoning[1]: Given that $I(X_i; C_i, L_i)$ – as the mutual information between X_i under its local context L_i and its global context C_i – increases because the global context increases, the local entropy $H(X_i \mid L_i)$ would have to increase in order for $H(X_i \mid C_i, L_i)$ to remain roughly constant, though a variation of the value is possible.

A theory similar to the ERC principle, the Uniform Information Density (UID) Hypothesis (Jaeger, 2010), states that speakers tend to distribute information uniformly throughout an utterance so that less processing load is given to the listeners. Formally, the information expressed by a linguistic signal y (e.g., an utterance or, in this study, a sequence of gaze labels) as the so-called surprisal, is y's negative log probability: $s(y) = -\log_{p_\ell}(y)$, which can be further factorized. Recently, surprisal is commonly approximated using language models with learned parameters (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020). Besides constraining speaker choice of words, it has also been discussed how UID influences reading time (Meister et al., 2021), or speech duration (Pimentel et al., 2021). Moreover, recent studies investigated whether nonverbal signals adhere to the ERC principle as well. Xu et al. (2022), for example, encode co-speech gestures (in monological videos) into discrete labels. We used a similar approach in our previous work (Wang and Buschmeier, 2023) for listeners' gaze in interaction. These two studies show that gesture and gaze as nonverbal signal, conform to ERC principle as well.

In this study, we use surprisal (namely local entropy) to further explain how we transfer the ERC principle, which was originally developed for texts, to gaze. Here is an intuitive example: let us assume the surprisal of the word sequence ⟨a, good, day⟩ is computed as $-(\log P(\text{a}) + \log P(\text{good} \mid \text{a}) + \log P(\text{day} \mid \text{a, good}))/3$ (no n-gram model assumed here), the words are the representation of meaning. Analogously, gaze labels are representations of gaze (see Section 5 for details), which semantically represents the distance (distance features) of the listener's gaze toward speaker, e.g., close, far, left, right, attention, aversion, etc. Given an example gaze label sequence, e.g., ⟨41, 10, 12⟩, the surprisal is can thus be computed as $-(\log P(41) + \log P(10 \mid 41) + \log P(12 \mid 10, 41))/3$. The local context of gaze labels, e.g., ⟨41, 41, 41, 10, 10, . . . ⟩, is based on the input size, the global context of gaze labels, is a similar sequence, e.g., ⟨41, 41, 41, 10, 10, 41, 41, 41, 10, 10, . . .⟩, ahead of the current local context. Since mutual information is increasing with expansion of previous context (for both word context and gaze context), gaze can be treated analogously to words in research on written text. In this study, we thus investigate whether gaze local entropy increases to confirm our hypothesis.

3 Data Collection

Nonverbal communication takes place in parallel to verbal communication (Stivers and Sidnell, 2005). In order to control the variables in our study, we (roughly) control and constraint verbal information so that we can investigate the variation of nonverbal communication. The interactions we analyze are explanations of a board game (Deep Sea Adventure; Sasaki and Sasaki 2014) and were collected in an interaction study that resulted in the MUNDEX corpus (Türk et al., 2023). In each interaction an experienced explainer explains the games to an explainee who does not know the it yet (see Figure 1).

The interactions are organized into three phases: In the first phase, the game is explained without the game material being present. In a second phase, the game is put on the table and the explanation continues. In the third phase the two participants

[1]Note that Verma et al. (2023) re-evaluated the ERC principle in text with a neural sequential model and failed to find decisive evidence for it (with inconclusive results though).

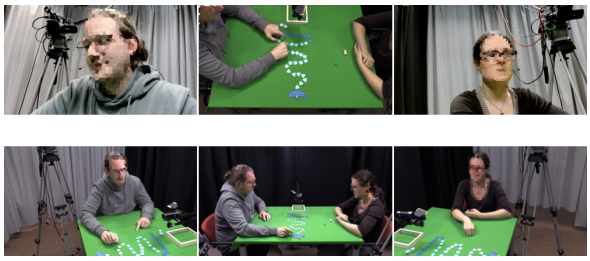


Figure 1: Study design for the data collections. The participant on the left (the explainer) explains the board game to the participant on the right (the explainee). The explainee's gaze behavior is captured by a camera behind the explainer.

start playing the game. The interactions can therefore be considered task-oriented dialogues. In this study we include the videos of 58 interactions from the corpus and extract the first phase (explanation without game). These explanations vary in length from 2:12 min to 17:36 min (mean length 7:04 min, standard deviation 3:15 min). All participants speak German.

4 Hypothesis

Our primary hypothesis is as follows: Similar to verbal information, which is considered to conform to the entropy rate constancy (ERC) principle, nonverbal communication signals such as the listener's gaze – which, similar to gestures, is usually coordinated with speech during interaction – also conform to it. The original hypothesis was posed in our previous work (Wang and Buschmeier, 2023), but the finding was not conclusive. In this study, we revisit the original hypothesis by refining the experimental setting, e.g., by detecting and removing local outliers and partitioning the video recordings into several groups in a more reasonable way.

We further test the ERC principle in nonverbal communication by extending the primary hypothesis as follows: the ERC principle in nonverbal communication in dialogue is based on the combination of multiple linguistic factors, of which we investigate two: syntactic complexity and turn transition potential of lexical items.

5 Processing Listener Gaze and Dialogue in Interaction

For each recorded video from the corpus, we extract the explainee's gaze information using the Openface framework (Baltrusaitis et al., 2018). Openface generates two types of gaze features (i) *gaze direction values* in radians averaged from both eyes (two dimensions) and (ii) *gaze vectors* in world coordinates (three dimensions for the left and right eye each). We integrate both representations so that the final gaze label does not only reflect gaze dynamics in three dimensions but also incorporates the effects from listener's head orientation during interaction. The computation of a gaze label follows these steps:

1. For the *gaze direction values*, we use the DBSCAN clustering algorithm (Ester et al., 1996) to find the spatial distribution of gaze and identify its "dense region" (Tran et al., 2020), both horizontally and vertically. The motivation of finding the "dense region" is based on the intuition that interlocutors spend a large proportion of time during the conversation looking at each other. Based on the minimal and maximal gaze direction values, a 3×3 grid is set up. The *gaze direction values* that are in this dense region are given the label '5', where the minimum and maximum gaze direction values among all gaze direction values in the dense region are used to decide on its boundary. The gaze direction values that are not in the dense region are assigned eight other number-based labels. The overall process follows the ICE algorithm (Tran et al., 2020) while the idea of using a 3×3 grid is based on Xu et al. (2022) and Xu and Cheng (2023).

2. For the *gaze vectors*, we only take the depth dimension into account and process it similarly to the previous step. DBSCAN-clustering is used to find the dense region where the gaze of the left and right eye are located in the depth dimension. The eye gaze vector inside this dense region is again given the label '5'. Based on how close the left eye vector or right eye vector are to the dense region, eight different labels are used. This process can sometimes fail, so that only one cluster is generated. We cope with these exceptional cases, by normalizing the depth value to a range $[0, 1]$ and consider values in the range $[0.25, 0.75]$ to be inside the 'dense region'.

3. Given the label $x_d \in [1, 9]$ for the gaze direction value and the label $x_v \in [1, 9]$ for the gaze vector, a combined label y, that represents the gaze information, is generated as $y = (x_d - 1) \cdot 9 + x_v$ (with $y \in [1, 81]$ representing the set of possible eye gaze labels). See Appendix A for further details on encoding eye gaze.

Figure 2 shows the distribution of the 15 most frequent gaze label in our data. The six most frequent

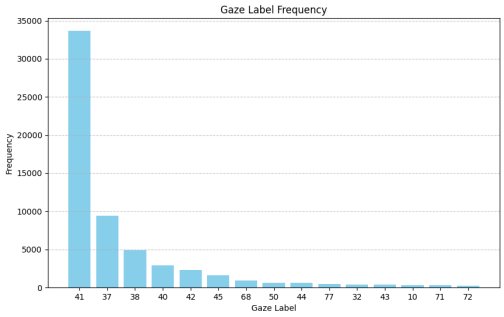


Figure 2: Distribution of the 15 most frequent gaze labels in the data set.

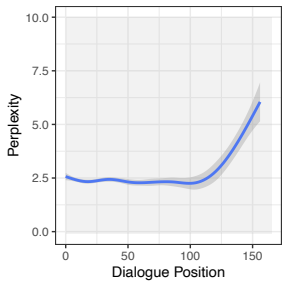


Figure 3: General trends in perplexity (local entropy) of explainee gaze gaze during explanations, without removing the effect from the length of dialogues.

Table 1: Statistics of gaze local entropy grouped by length of dialogues (see Figure 4 for labels **A, B, C**).

	A	**B**	**C**
Coefficient	5.98×10^{-3}	1.35×10^{-3}	3.49×10^{-7}
***p*-value**	< 0.05	< 0.01	< 0.01

labels (41, 37, 38 ,40, 42, 45) represent the explainees looking in the direction of explainers with various distance (closer or more distant). The less frequent labels (68, 50, . . .) instead start to show explainees averting their gaze from explainers.

Dialogues were automatically transcribed using 'Whisper' (Radford et al., 2022), which generates speech segments with a start and an end time. In order to calculate word timings, we approximated word onset times by calculating the duration of each speech segment, dividing it by its length in words, and approximating word duration (assuming, for this study, that words have uniform length). Eye gaze labels are then aligned with words based on the video's frame rate (50 fps).

We consider eye movements as sequences of gaze labels (Figure 9 shows an example of speech gaze alignment). To be able to compute the entropy for gaze, we train a transformer model (Vaswani et al., 2017) to estimate the underlying probability distribution of the gaze sequences.We have chosen a transformer model since it has a stronger psychometric predictive power compared to LSTM-RNNs models (Wilcox et al., 2020). To compute the local entropy of a gaze sequence, we first calculate its negative log probability:

$$\begin{aligned}&\mathrm{NLL}(e_1, e_2, \ldots, e_T)\\ &\quad = -\Big(\log P(e_1) + \sum_{i=2}^{T} \log P(e_i \mid e_1, \ldots, e_{i-1})\Big)\end{aligned}$$

where T is the maximum index of a given eye movement sequence and $e_i \in [1, 81]$ are the gaze labels. The local entropy $H(e_1, e_2, \ldots e_i)$ of the gaze sequence is then the exponential of NLL (i.e., perplexity). The learning task is thus to predict the next gaze label e_i based on the preceding sequence $\langle e_1, \ldots, e_{i-1} \rangle$ and minimize its negative log probability NLL.

6 Preliminary Result

After computing the local entropy, we first calculate statistical significance, removing outliers. We use a simple local outlier detection method by calculating the standard score (Z-score) of each data point. If the standard score is greater than a threshold (for the gaze local entropy, we define it as 4 so that only very extreme values are considered outliers), it is considered an outlier and replaced by the median of the local entropy.

Figure 3 shows the change of local entropy of the combined eye gaze sequences from all 58 interaction videos after removing outliers. The x-axis represents the dialogue position of the speech segments (with each dialogue position corresponding to about 7 seconds of speech) and the y-axis represents the local entropy of the gaze sequences. Although there seems to be a general increasing trend of the local entropy, we also have to consider the different length of the videos, which may bias the result. Therefore, we decided to divide the videos into three groups based on length to mitigate the length bias and replot the local entropy (see Figure 4).

We then use a linear model and conduct statistical significance tests between dialogue position and gaze local entropy for each of the three sub groups. The results show us that, coefficient values in all three sub groups are positive, with p-values smaller than 0.05. The increase of local entropy of gaze

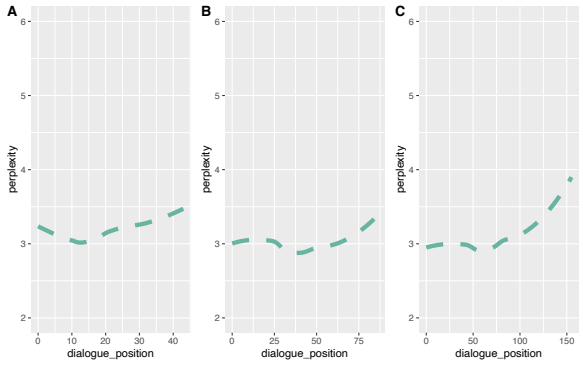


Figure 4: General trends in perplexity (local entropy) of explainees' eye gaze during explanations, grouped by length of dialogues (in dialogue positions): **A**: 27 dialogues shorter than 45; **B**: 24 dialogues between 46 to 90; **C**: 7 dialogues longer than 90. The reason for choosing 45 as the breaking point is that using this threshold, the data size in the first two groups is roughly equal, while dialog positions larger than 90 are comparatively rare.

(shown in Figure 4) is thus statistically significant, that is, the preliminary result supports our basic hypothesis. The general increasing trend of local entropy indicates that more new gaze labels appear over the course of the explanations.

7 Analysis of Linguistic Factors

The goal of this section is to answer the question why listener gaze, as a nonverbal signal in interaction, conforms to the ERC principle (see Section 6). Specifically, we focus on exploring linguistic factors which may influence the information amount conveyed by listeners' gaze during interaction. We select two metrics – *syntactic complexity* and *turn transition potential* – as the linguistic factors to investigate and perform statistical analyses to check for correlations between gaze local entropy and these two factors.

7.1 Syntactic Complexity and Gaze Local Entropy

Studies of human sentence processing show that when humans are exposed to written language that is structurally complex, processing difficulties accumulate, and this is reflected in the collected eye movement data, e.g., increasing eye fixation can be found in the syntactically complex area among all of the sentences given (cf. Clifton and Staub, 2011). That is, gaze behavior correlates with cognitive load, which can be increased by syntactic complexity. For example, it has been shown that cognitive load can cause gaze aversion during interaction (Argyle and Cook, 1976). This correlation between syntactic complexity and eye movement found in the literature leads us to the following conjecture: Is syntactic complexity also relevant to the increase of local entropy of listeners' gaze in face-to-face interaction?

7.1.1 Calculating Syntactic Complexity

Xu and Reitter (2016) show two metrics for calculating syntactic complexity: (1) tree depth of a syntactic tree, and (2) branching factor, i.e., the average number of children in non-leaf nodes when parsing a syntactic tree. We use dependency parsing (specifically 'Stanza'; Qi et al. 2018) to parse the automatically recognized speech and calculate its syntactic complexity. The advantage of using dependency parsing is that the input does not have to be a complete sentence to be parsed. We calculate syntactic complexity similarly to Xu and Reitter (2016). Given a speaker's utterance, we compute its length L and use dependency parsing to get the number of heads α as well as the maximum tree depth β. The syntactic complexity SC of the utterance is then computed as

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

λ is a tuning factor set to 0.5 by default.

First, an utterance which is lengthy is considered more complex, since it has a bigger tree depth value. Second, the more complex phrases an utterance contains, the more complex the utterance is. We take utterances from explainers as input to the linear model. Figure 5 shows a plot of the development of syntactic complexity of utterances over the course of the interactions. Within the defined topic we analyze (explanations of the board game), the syntactic complexity of utterances increases, as verified by a linear model which weakly correlates utterance position to syntactic complexity ($\beta = 0.0009$, $p < 0.001$). Although the beta coefficient is quite low, we consider that the effect size is still valid given that the value range of syntactic complexity is quite small (from 2.8 to 3.05 in Figure 5 [2] and 1.5 to 5.9 in the whole data).

7.1.2 Correlation with Gaze Local Entropy

Based on these results (Figure 5), we further investigate whether there is a correlation between the

[2]While the syntactic complexity is decreasing at the early stage, it is increasing after dialogue position 25.

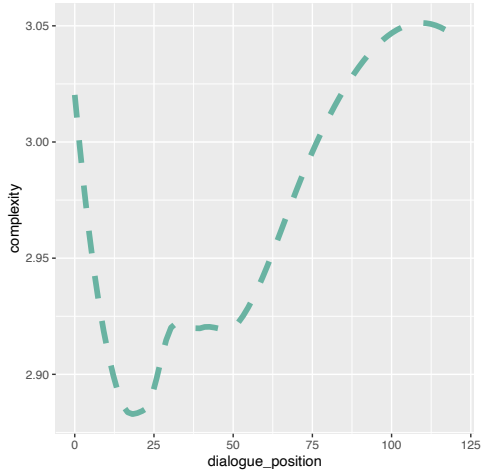


Figure 5: Syntactic complexity of explainers' speech. The x-axis shows dialogue positions, the y-axis scales the variation of measured syntactic complexity.

A …Es gibt drei Runden Und nach jeder Runde wenn man dann ertrinkt dann werden die Schätze am Ende…
41 41 41 41 41 41 41 41 71 71 71 71 71 71 71 41 41 41
Syntactic Complexity: 2.42 Perplexity: 1.928
(…there are three rounds and after each round if you drown then the treasures are at the end…)

B…Und dann versucht man wieder hoch zu kommen Also sprich es gibt unten Schätze und oben…
41 41 41 41 41 41 41 41 41 41 41 41 41 71 50 50
Syntactic Complexity: 2.9 Perplexity : 3.589
(…And then you try to get back up. So there are treasures below and above…)

C …dafür bezahlen Oder Sauerstoff dafür Das ist nicht ganz wie im echten Leben…
41 41 41 41 41 41 41 41 41 41 41 41 40
Syntactic Complexity: 2.74 Perplexity: 1.613
(…Pay for it or oxygen for it It's not quite like in real life…)

(a) Example 1

A… Ich glaube wir können jetzt anfangen oder Müssen wir uns jetzt was ausdenken oder was…
37 38 38 38 41 38 38 38 38 41 41 41 41 41 41
Syntactic Complexity: 3.2 Perplexity: 1.957
(... I think we can start now or do we have to think of something now or something…)

B…Also es gibt dann das U-Boot mit den verschiedenen Spielern und im U-Boot befindet sich ein Sauerstofftank Und ohne den Sauerstofftank…
41 41 41 41 41 41 41 41 41 41 41 41 41 41 41 38 38 41 41 41 41
Syntactic Complexity: 2.86 Perplexity: 2.721
(...So there is the submarine with the different players and in the submarine there is an oxygen tank and without the oxygen tank…)

C…Genau also Sauerstoff verringern ankündigen in welche Richtung du gehst Und dann wenn du gewürfelt hast….
41 41 41 41 41 41 41 41 41 41 41 62 62 62 62 62
Syntactic Complexity: 2.97 Perplexity: 3.59
(…Exactly so reduce oxygen announce in which direction you are going and then when you have rolled the dice….)

(b) Example 2

Figure 6: Sample dialogue segments from two videos. Utterances are followed by corresponding gaze label. The dialogues progress from A to C

local entropy of listeners' gaze and the syntactic complexity of speech [3]. A linear model shows a statistically significant relation between the two variables ($\beta = 0.02, p < 0.05$). A weak positive correlation between syntactic complexity and local entropy of gaze indicates that, as syntactic complexity is in general increasing, gaze local entropy increases accordingly. A possible explanation for the correlation between syntactic complexity and local entropy of gaze could be that, with the development of the board game explanation, the speech from explainers becomes more syntactically complex, the explainees instead, take more effort to process the information which may lead to, e.g., longer periods of gaze shift, and thus more variation in gaze labels.

Two illustrative examples from our data can be seen in Figure 6. In the chosen dialogue segments (from the two videos), an increase of local entropy is in general correlated with an increase of syntactic complexity (An exception is A of Example 2 in Figure 6b). Moreover, the chosen examples suggest that a higher local entropy value is generally caused by shifts between frequent gaze labels to less frequent ones and vice versa (e.g., shift from label 41 to 71 in A of example 1 (Figure 6a), shift from label 41 to 62 in C of example 2 (Figure 6b). This is in line with the UID theory: the appearance of less frequent units leads to a higher surprisal value (Jaeger, 2010).

7.2 Turn Transition Potential and Gaze Local Entropy

The interaction of gaze behavior and turn-taking have constantly attracted research attention (e.g., Skantze et al. 2014; Kendrick et al. 2023). For this study, we assume that, turn-taking signals are a linguistic factor which can influence listeners' gaze behavior and thus gaze local entropy. Our assumption is based on the intuition that listeners will usually wait until speakers finish speaking and then start to talk. And when listeners want to take a turn or give the turn back to previous speakers in interaction, they usually gaze towards the speakers. According to Sacks et al. (1974), in dyadic conversations, some lexical units, which are defined as transition relevant places (TRP), have higher probability that a turn shift can happen. These usually occur at syntactic or pragmatic completeness.

For our study, we did not use the real turn-taking occurrences from our data to test our hypothesis (because explainees are less likely to take the turn during the initial game explanation phase; Fisher et al. 2022), but instead analyze the transition rel-

[3]Here we use all of the utterances instead of only explainers' utterance since gaze labels are aligned with all of the utterances. Despite some noise, we still think it is the most optimal way based on our observation that explainee utterances are very limitted in our explanation data.

evance places as turn-taking potential in our data. The experimental setting is as follows: We train a neural sequential model (TurnGPT; Ekstedt and Skantze 2020) based on two available German dialogue data sets (VM2 and DialogueSUM, see below) to approximate human turn-taking behaviour. We aim at calculating the turn-taking potential of each lexical unit in dialogue, represented as a probability [4]. After getting the turn-taking potential for all of the lexical units, we analyze the interaction of turn-taking potential, dialogue position, and gaze local entropy.

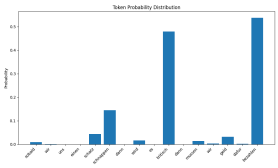
7.2.1 Estimating Turn Transition Potential using TurnGPT

We use TurnGPT (Ekstedt and Skantze, 2020) to calculate the turn transition potential for our speech data. TurnGPT extends the GPT-2 architecture (Radford et al., 2019) with additional speaker embeddings. The turn-taking ground truth is encoded by adding a token (<ts>) after lexical items where turn-taking occurred. To process German data, we replace the original GPT-2 model with German-GPT2 (Schweter, 2020) and fine-tune it on two datasets:

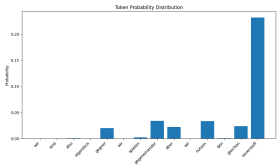
- VERBMOBIL (VM2; Kay, 1992): The dataset is based on recordings of various appointment scheduling scenario and consists of 30 800 utterances collected in face-to-face interactions. Utterances are annotated with dialogue acts and include a corresponding speaker ID.
- DialogueSum (Chen et al., 2021): DialogSum is a large-scale dialogue summarization dataset, consisting of 13 460 dialogues. Each utterance in the dataset is aligned with a speaker label. The original data is in English; we used a German translation of the data (Dialogsum-German) available on HuggingFace.

Figure 7 shows two examples of model predictions (with English translations). Lexical items are aligned with the turn transition probabilities predicted by our German TurnGPT model. Lexical items with higher probabilities indicate syntactic and pragmatic completeness, and are considered reasonable places for turns in real conversation. Additional examples can be found in the Appendix A.5.

[4] It is worth noting that a lexical item classified as a place with a high turn-taking potential probability, does not necessarily coincide with turn-taking in the real dialogue. We interpret the high probability as a reasonable point for taking the turn in a real conversation.

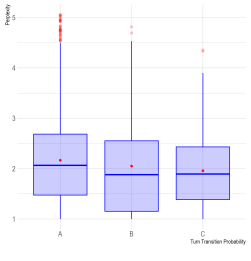


(a) Example 1: *Sobald wir uns einen **Schatz schnappen** dann wird es **kritisch** dann müssen wir Geld dafür **bezahlen*** ('As soon as we **grab** a **treasure**, things become **critical** and we have to **pay money** for it').

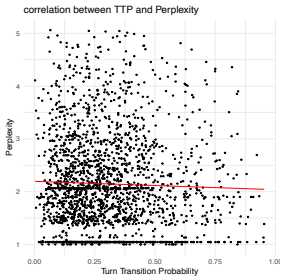


(b) Example 2: *wir sind also eigentlich **Gegner** wir spielen **gegeneinander** aber wir **nutzen** den gleichen **Sauerstoff*** ('So we are actually **opponents**, we play **[against each other]**, but we **use** the same **oxygen**').

Figure 7: Examples of turn transition potential probability predictions by TurnGPT, including English translations. Words in bold have higher turn transition probability.



(a) Turn transition potential probability in three bins (from left to right: A, B, C) and the relevant gaze local entropy. The red dot inside the boxes are the means of gaze local entropy.



(b) Regression model of turn transition potential probability and gaze local entropy. The red line shows the general trend of gaze local entropy with the increase of turn transition potential probability.

Figure 8: Statistics of turn transition potentials and gaze local entropy.

7.2.2 Correlation with Gaze Local Entropy

We first use a linear model to investigate whether turn transition probabilities (obtained from the process shown in Section 7.2.1) correlate with dialogue position. No statistically significant correlation is found between the two variables ($\beta = -3.28, p = 0.17, \alpha = 0.05$). However, the negative coefficient value indicates that, in general, higher turn transition probabilities are distributed more at the beginning of dialogues.

After feeding each utterance into the TurnGPT model to obtain the corresponding turn transition potential probability, we divide the data in to three groups by turn transition potential probability p_T. A/comparatively low: $p_T < 0.5$; B/comparatively high: $p_T \in [0.5, 0.75]$; C/high: $p_T > 0.75$. We then aggregate and analyze gaze local entropy by group (Figure 8a). In group B and C, where higher

turn transition probability is aggregated, the corresponding gaze local entropy generally has a slightly lower mean and median value compared to group A. A possible explanation for this result could be that listeners tend to look at the speaker's face when taking turns or returning the turn to the speaker, resulting in more uniform and predictable gaze labels defined in our study and thus a comparatively low gaze local entropy value.

Using a linear model, we find a weak, but statistically insignificant correlation ($\beta = -0.0064$, $p = 0.143$, $\alpha = 0.05$) between turn transition potential probability and gaze local entropy. Figure 8b shows this general decreasing linear trend (but data points are distributed quite sparsely).

8 Conclusion

In this study, we extracted listeners' gaze from video-recorded human interaction dialogues (specifically from explainees in explanatory dialogues) and found evidence that listener gaze, as an important nonverbal signal, adheres to the Entropy Constancy Rate principle – a property of human language use (Genzel and Charniak, 2002). This finding is supported by an increasing trend of the local entropy of listeners' gaze over the course of the dialogues we analyzed. We believe that this finding can provide future insight into the evaluation of co-speech gaze generation and interpretation in the field of human-agent interaction, e.g., in explanatory settings.

We further investigated two linguistic factors - syntactic complexity and turn-taking potential - that may contribute to this increase in gaze entropy, as they have been considered to interact with gaze behavior in previous studies. Our statistical analysis provides initial evidence that these two factors are correlated (weakly) with the local entropy value of gaze, which could potentially provide an explanation for why listeners' gaze may adhere to the ERC principle.

9 Limitations

At the current stage, our study has a number of limitations that need to be taken into account when interpreting its results.

A first limitation is that the calculation of gaze local entropy based on a model that takes only gaze labels as input is not optimal, since listener gaze is affected by other modalities, speech, speaker's prosody, etc. in the interaction. The neural sequential model (transformer) is autoregressive in that it predicts the current gaze label based only on previous gaze labels. This approach should be better than random guessing, but ignores the effects of the other modalities. This should be addressed in future work, e.g. by building a model that takes different modalities as input and calculates gaze local entropy based on them.

A second limitation concerns the the two weak correlations found in our study. They might be due to other linguistic factors that may influence the consolidation of the ERC principle for listener gaze (or nonverbal communication in general). For example, Giulianelli et al. (2022) found that the repetition of constructions (e.g., 'yes yeah I', 'uhm I think') in dialogue reduces its informational content. Similarly, Bowers et al. (2010) show that listeners pay more attention to fluent than to disfluent speech, suggesting that disfluencies influence listeners' gaze behavior. Possible question to investigate, then, might be whether repetitions of constructions or disfluencies influence the information content conveyed by listener gaze in dialogue.

A third limitation concerns the computation of the turn transition probabilities, which is based only on textual representations of the interactions. In practice, prosody is also an important factor that affects turn-taking behavior (Ekstedt and Skantze, 2022a) and such features should be taken into account as well.

A fourth limitation concerns the processing of the video data. The extraction of eye movements using OpenFace is suboptimal. For more reliable gaze estimation, dedicated eye-tracking hardware should be used instead – although this may affect the naturalness of the interactions.

10 Ethics Statement

The used data in the study is for research purposes only and commercial use is not allowed. The participants (i.e., dialogue participants) were mainly university students. They gave informed consent to participate in the study and to have their data used, and were paid 10 euros per hour. The study was approved by the university's internal ethics and data protection review boards.

Supplementary materials Code and data of analyses is available in the following data publication: https://doi.org/10.17605/OSF.IO/A4RHV

Acknowledgments This research was supported by the German Research Foundation (DFG) in the Collaborative Research Center TRR 318/1 2021 'Constructing Explainability' (438445824). Yang Xu is supported by National Science Foundation of the United States (CRII-HCC: 2105192).

References

Michael Argyle and Mark Cook. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, UK.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66, Xi'an, China.

Andrew L. Bowers, Stephen C. Crawcour, Tim Saltuklaroglu, and Joseph Kalinowski. 2010. Gaze aversion to stuttered speech: a pilot study investigating differential visual attention to stuttered and fluent speech. *International Journal of Language & Communication Disorders*, 45:133–144.

Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*, 28:449–483.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

Charles Clifton and Adrian Staub. 2011. Syntactic influences on eye movements during reading. In *The Oxford Handbook of Eye Movements*, pages 896–909. Oxford University Press.

Starkey Duncan. 1975. Interaction units during speaking turns in dyadic, face-to-face conversations. In Adam Kendon, Richard M. Harris, and Mary R. Key, editors, *Organization of Behavior in Face-to-Face Interaction*, pages 199–214. De Gruyter Mouton, Berlin, Germany.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4295–4309, Dublin, Ireland.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: A transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? Investigations using voice activity projection models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551, Edinburgh, UK.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice activity projection: Self-supervised learning of turn-taking events. In *Interspeech 2022*, pages 5190–5194, Incheon, Korea.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon.

Josephine B. Fisher, Vivien Lohmer, Friederike Kern, Winfried Barthlen, Sebastian Gaus, and Katharina J. Rohlfing. 2022. Exploring monological and dialogical phases in naturally occurring explanations. *KI – Künstliche Intelligenz*, 26:317–326.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, PA, USA.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction repetition reduces information rate in dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 665–682, Online.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, UT, USA.

Charles Goodwin. 1985. Notes on story structure and the organization of participation. In J. Maxwell Atkinson and John Heritage, editors, *Structures of Social Action*, pages 225–246. Cambridge University Press.

Marjorie Goodwin and Charles Goodwin. 1986. Gesture and coparticipation in the activity of searching for a word. *Semiotica*, 62:51–76.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61:23–62.

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems*, 3(2):12:1–30.

Martin Kay. 1992. *Verbmobil: A Translation System for Face-to-Face Dialog*. University of Chicago Press, Chicago, IL, USA.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Kobin H. Kendrick, Judith Holler, and Stephen C. Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210473.

Eliot Maës, Philippe Blache, and Leonor Becerra. 2022. Shared knowledge in natural conversations: can entropy metrics shed light on information transfers? In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 213–227, Abu Dhabi, UAE.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic.

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 553–561, Sapporo, Japan.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 13–21, Gothenburg, Sweden.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shane L. Rogers, Craig P. Speelman, Oliver Guidetti, and Melissa Longmuir. 2018. Using dual eye tracking to uncover personal gaze patterns during social interaction. *Scientific Reports*, 8:4271.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Jun Sasaki and Goro Sasaki. 2014. *Deep Sea Adventure (Tabletop Game)*. Oink Games, Tokyo, Japan.

Stefan Schweter. 2020. German GPT-2 model (dbmdz/german-gpt2).

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423.

Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65:50–66.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Tanya Stivers and Jack Sidnell. 2005. Introduction: Multimodal interaction. *Semiotica*, 2005(156):1–20.

Minh Tran, Taylan Sen, Kurtis Haut, Mohammad Rafayet Ali, and Ehsan Hoque. 2020. Are you really looking at me? a feature-extraction framework for estimating interpersonal eye gaze from conventional video. *IEEE Transactions on Affective Computing*, 13:912–925.

Olcay Türk, Petra Wagner, Hendrik Buschmeier, Angela Grimminger, 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Long Beach, CA, USA.

Vivek Verma, Nicholas Tomlin, and Dan Klein. 2023. Revisiting entropy rate constancy in text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15537–15549, Singapore.

Yu Wang and Hendrik Buschmeier. 2023. Does listener gaze in face-to-face interaction follow the entropy rate constancy principle: An empirical study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15372–15379, Singapore.

Nigel G. Ward. 2019. *Prosodic Patterns in English Conversation*. Cambridge University Press, Cambridge, UK.

Ethan G. Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713, Virtual.

Yang Xu and Yang Cheng. 2023. Spontaneous gestures encoded by hand positions improve language models: An information-theoretic motivated study. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9409–9424, Toronto, Canada.

Yang Xu, Yang Cheng, and Riya Bhatia. 2022. Gestures are used rationally: Information theoretic evidence from neural sequential models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 134–140, Gyeongju, Republic of Korea.

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 and David Reitter. 2017. Spectral analysis of information density in dialogue predicts collaborative task performance. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 623–633, Vancouver, Canada.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.

A Appendix

A.1 Model Parameter Settings

The following parameters were set for training the transformer model for gaze local entropy: batch size 35; input size 25; initial learning rate 0.05. The data was divided into 80% for training and 20% for testing. Both the training and testing data sets are used to compute the local entropy.

The following parameters were set for fine-tuning the German TurnGPT model: batch size 3; weight decay 0.01; dropout rate 0.1; learning rate 0.0001. A total of 15 epochs were used to finish the fine-tuning tasks. After each epoch, one checkpoint (model) was generated, saving the weight parameters gained during training. The model with minimal loss value (1.0083) was chosen as the final model for estimating the turn transition potential probability.

Das	*Problem*	*ist*	*wir*	*haben*	*nur*	*ein*	*billiges*	*U-Boot*	*wo*		
42	42	42	42	42	42	42	42	42	42		
wenig	*Sauerstoff*	*drin*	*ist*	*und*	*wir*	*müssen*	*uns*	*den*	*teilen*		
42	42	42	42	42	41	45	42	45	45		
Wir	*sind*	*also*	*eigentlich*	*Gegner*	*wir*	*spielen*	*gegeneinander*				
45	45	45	45	45	45	44	44				
aber	*wir*	*nutzen*	*den*	*gleichen*	*Sauerstoff*						
41	41	41	41	41	41						

Figure 9: An example of gaze-speech alignment (from Wang and Buschmeier (2023))

A.2 DBSCAN Algorithm Parameters

We set up a criterion to ensure that a point can only be considered a core point for a cluster if it is surrounded by at least ten samples.

A for-loop was used to find an optimal ϵ value (required by DBSCAN) in the range $[0.01, 0.1]$ such that the ratio between the two most frequent cluster labels is just below 15%. In this way, the cluster with the most frequent label is considered to be the region where an explainee's gaze is mostly located.

A.3 Example of Speech-Gaze Alignment

Figure 9 shows an example of gaze and speech alignment.

A.4 Examples ASR-Results

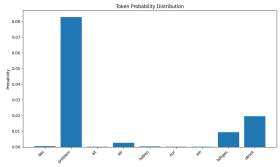
Figure 10 shows an example of a diarized speech transcript created automatically with the automatic speech recognition Whisper. We selected the model *large-v3*. Whisper's word error rate (WER) for German is given as 5.7% the in Common Voice 15 dataset and as 4.9% in the Fleurs datasets (see `https://github.com/openai/whisper/blob/main/README.md`). The transcripts were not corrected.

A.5 Additional Examples of Turn Transition Potential Estimation

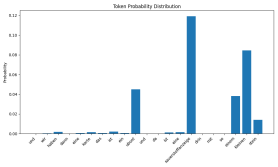
Four additional examples of model predictions for turn transition potential probability are shown in Figure 11.

68	00:06:54,072 -> 00:06:58,950 **Speaker 1:** *Du meintest doch vorher, wenn man wieder zurückgeht, dass man die Schätze dann da wieder ablegen kann, wo so ein leeres Feld ist.*
69	00:06:59,110 -> 00:07:02,208 **Speaker 0:** *Achso, das könntest du, wenn du merkst, ich habe zu viele Schätze mir gegaunert.*
70	00:07:03,493 -> 00:07:05,069 **Speaker 1:** *Ja, wenn der Schatz sowieso nichts bringt.*
71	00:07:05,650 -> 00:07:06,327 **Speaker 0:** *Das siehst du nicht.*
72	00:07:06,571 -> 00:07:07,186 **Speaker 0:** *Okay, das sage ich noch.*
73	00:07:08,331 -> 00:07:13,150 **Speaker 0:** *Unten drunter steht eine Zahl und die wird erst aufgedeckt, wenn du zurück im U-Boot bist und dir den Schatz anschaust.*

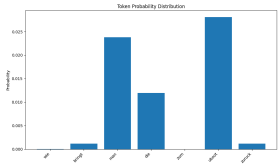
Figure 10: An example of an automatically generated speech transcript with speaker diarization.



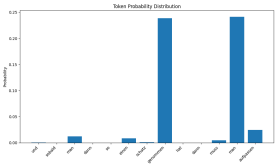
(a) *Das **Problem** ist wir haben nur ein **billiges Uboot*** (‘the **problem** is, we only have a **cheap submarine**’).



(b) *und wir haben dann eine Karte das ist ein **Uboot** und da ist eine **Sauerstoffanzeige** drin mit so **einem kleinen Stein*** (‘and then we have a card it’s a **submarine** and there’s an **oxygen** gauge with **a little stone** in it).



(c) *Wie **bringt man** die zum **Uboot** zurück* (‘How do **you** get **them** back to the **submarine**’).



(d) *und sobald man dann so einen Schatz **genommen** hat dann muss **man aufpassen*** (‘and once you have **taken** such a treasure, **you** have to be **careful**’).

Figure 11: Four additional examples of model predictions for turn transition potential probability, with translations in parentheses.