Predicting States of Understanding in Explanatory Interactions Using Cognitive Load-Related Linguistic Cues

Yu Wang[†‖] **Olcay Türk**[†‖] **Angela Grimminger**[‡‖] **Hendrik Buschmeier**[†‖]
[†]Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany
[‡]Faculty of Arts and Humanities, Paderborn University, Paderborn, Germany
[‖]SFB/Transregio 318 'Constructing Explainability', Bielefeld & Paderborn, Germany

Abstract

We investigate how verbal and nonverbal linguistic features, exhibited by speakers and listeners in dialogue, can contribute to predicting the listener's state of understanding in explanatory interactions on a moment-by-moment basis. Specifically, we examine three linguistic cues related to cognitive load and hypothesised to correlate with listener understanding: the information value (operationalised with surprisal) and syntactic complexity of the speaker's utterances, and the variation in the listener's interactive gaze behaviour. Based on statistical analyses of the MUNDEX corpus of face-to-face dialogic board game explanations, we find that individual cues vary with the listener's level of understanding. Listener states ('Understanding', 'Partial Understanding', 'Non-Understanding' and 'Misunderstanding') were self-annotated by the listeners using a retrospective video-recall method. The results of a subsequent classification experiment, involving two off-the-shelf classifiers and a fine-tuned German BERT-based multimodal classifier, demonstrate that prediction of these four states of understanding is generally possible and improves when the three linguistic cues are considered alongside textual features.

Keywords: dialogue, understanding, cognitive load, gaze, information value, syntactic complexity

1. Introduction

Explanatory interactions are a type of everyday communicative activity in which an 'explainer' tries to explain something to an 'explainee'. The explainee seeks to understand the explanation and common ground is continually built during the interaction (Clark and Brennan, 1991). In the grounding process, explainees frequently provide feedback, either verbally (e.g., in form of backchannels) or nonverbally (e.g., in form of head gestures), to explainers, and by that display or signal different states of understanding (e.g., Allwood et al., 1992, 2007): understanding might be signalled through nodding, partial understanding by more hesitant nodding, and non-understanding by providing feedback with decreased volume of the voice. Feedback allows explainers to monitor an explainee's state of understanding and potentially adapt their explanation such that mutual understanding can be co-constructed (Clark and Krych, 2004). The social practice of explaining entails processes such as monitoring understanding of the explainee and adapting to the explainee. These processes have recently been proposed as important components in building social explainable AI systems ('Social XAI'; Rohlfing et al., 2021, 2026).

Monitoring understanding moment-by-moment requires identification of the (non-)verbal cues that reflect different states of understanding throughout the interaction. Thus, understanding – and, more specifically, difficulties in understanding that might increase cognitive load – may be signalled by certain verbal and non-verbal behaviour. Previous studies show that backchannels such as 'mhm' or 'ja' are efficient vocal signals that communicate addressees' understanding (Allwood et al., 1992). Non-verbal signal such as gaze, can indicate cognitive processing activity, as supported by empirical experiments in which participants solve challenging tasks (Glenberg et al., 1998). As noted by Türk et al. (2024), variation in gaze direction demonstrates the extent of cognitive processing effort and can thus indicate an explainee's level of understanding – to a certain degree. Furthermore, surprisal theory (Hale, 2001) posits that the cognitive effort (cognitive load) required to process language depends on its contextual predictability, quantified as 'surprisal'. This was later substantiated numerically with empirical and theoretical evidence from reading studies (e.g., Levy, 2008; Smith and Levy, 2013; Shain, 2021).

In this paper, we hypothesise that explainee's understanding is related to cognitive load (e.g., Betz et al., 2023). Following this hypothesis, we review previous studies to identify potential verbal and nonverbal cues that can indicate cognitive load. From these, we select gaze variation, information value, and syntactic complexity as potential correlates. We use MUNDEX, a multimodal corpus of explanation dialogues (Türk et al., 2023), in which the above-mentioned cues are well-represented in annotated behaviour of explainees. We quantify these cues, first performing statistical analyses to verify their significance for explainees' understanding. Based on these analyses, we then use selected linguistic cues to perform a classification experiment on explainees' states of understanding.

Our main contributions in this paper are: (1) We investigate three linguistic (verbal and nonverbal)

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11368–11378
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

cues that previous linguistic studies considered as important indicators of cognitive load in language comprehension. Based on these cues that are also considered to be related to the explainees' understanding states, we obtain four different values to represent cognitive load, three of which we find to significantly vary with explainees' states of understanding. (2) We use two off-the-shelf classifiers and trained a BERT-based classifier to investigate the predictability of different states of understanding. Unlike previous work, which predicted understanding using a binary classification task (understanding vs. non-understanding) based on a series of multimodal signals (Kinoshita et al., 2023; Türk et al., 2024), we quantify linguistic cues potentially indicating cognitive load and perform multi-class classification of different understanding states. Results show that all of the classifiers perform better than chance, which indicates the feasibility of predicting different understandings states. Among the three classifiers, the BERT-based classifier is most robust. The variation in performance when predicting different understanding labels also suggests potential challenges for future work.

2. Background

2.1. Linguistic Cues for Cognitive Load

Cognitive load is considered as the amount of working memory dedicated to problem solving (Sweller, 1988; Paas et al., 2003). Language comprehension, like many other daily tasks, constantly requires working memory capacity (Just and Carpenter, 1992). From a psycholinguistic perspective, cognitive load, sometimes conceptualised as 'processing difficulty' (Levy, 2008; Mitchell et al., 2010; Betz et al., 2023), is relevant for various aspects of language processing. One example is syntactic complexity, which denotes the cognitive load to parse and process a sentence or an utterance (Szmrecsányi, 2004). Dependency locality theory Gibson (1998, 2000) suggests that language comprehension involves continuous integration cost for processing sentences with varied syntactic structures. Accordingly, the theory treats syntactic complexity as a potential indicator of cognitive load. This is supported by empirical evidence and analyses showing that sentences that rank as more syntactically complex are considered more difficult for humans to process (Lin, 1996), or that specific syntactic components, such as nouns or specific type of verbs, induce higher processing difficulty compared to other syntactic components (Demberg and Keller, 2008, 2009).

In language comprehension, cognitive load can also be estimated by the predictability of a word in its context. This is called 'surprisal' (Hale, 2001; Levy, 2008). Given a word w and its context, e.g., a sentence that comprises of a sequence of smaller units: $\langle w_1, \ldots, w_i \rangle$, where $w_i \in \vartheta$, with ϑ being the vocabulary, surprisal of the word is modelled as the negative log probability: $-\log P(w_i \mid w_1, \ldots, w_{i-1})$. The higher the surprisal value of a word, the more unpredictable it is, and the greater the cognitive load required to process it. Surprisal, as well as its related concept such as contextual entropy, are widely accepted as models of the effort for language comprehension, given their psychometric predictive power on cognitive load (measured, e.g., through reading time; see Frank, 2013; Wilcox et al., 2023). Focusing on dialogic interaction, previous studies show that surprisal based information values converge and diverge continuously while a dialogue unfolds (Xu and Reitter, 2018). This indicates a potential correlation between information value and interlocutors' development of understanding (Maës et al., 2022).

In addition to text based linguistic cues, nonverbal cues have also been discussed in terms of their relation to cognitive load. One example is gaze. In face to face interaction, the gaze behaviour of listeners serves as critical cues indicating their visual attention to the speaker's ongoing verbal content (Kendon, 1967). Moreover, gaze aversion is considered to be evidence of cognitive processing when answering questions (Glenberg et al., 1998), but has also been described as part of the so-called "thinking face" in interactional settings, displaying language processing (e.g., Bavelas and Chovil, 2018). Empirical studies further show that gaze aversion is more likely to occur when there is an increase of cognitive load (Morency et al., 2006; Glenberg et al., 1998). The variation of gaze during language comprehension in interaction can thus be considered a potential indicator of an interlocutors' cognitive load.

2.2. Computational Modelling for Understanding Evaluation

Evaluation of a human interlocutor's understanding is an emerging but important task in human interaction with adaptive conversational agents: conversational agents, such as embodied conversational agents, social robots, but also LLM-based chatbots, should be able to estimate the understanding of their interlocutor in order to adjust their utterances and by that, promote efficient communication (Reidsma et al., 2011; Buschmeier and Kopp, 2018; Axelsson and Skantze, 2022; Robrecht and Kopp, 2023; Mindlin et al., 2024). A competent conversation agent should also be able to correct any misunderstanding of the interlocutors when misunderstanding occurs. Regarding the evaluation and prediction of understanding in interaction, Howes

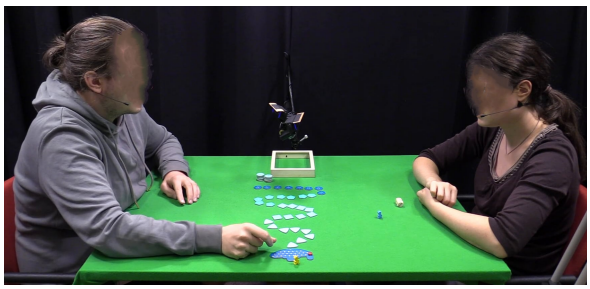

Figure 1: Explanation set-up in the MUNDEX corpus (screenshot from one camera-perspective). The person on the left is the explainer who explains a board game; the person on the right is the explainee.

and Eshghi (2021) model the effect of interlocutors' backchannels by incrementally tracking them as evidence of understanding. Similarly, Buschmeier and Kopp (2018) propose a probabilistic model in which an agent uses multimodal signals of the interlocutor to represent the current grounding state and update its belief about the listener's state of understanding during the interaction.

More recent work treats the evaluation of understanding as a machine learning task. Kinoshita et al. (2023), for example, built a dialogue corpus consisting of the listener's comprehension levels (a digital value from -2 to 2, annotated by third-party annotators) and the listener's multimodal information, and ran a regression model to predict the comprehension level. Similarly, Türk et al. (2024) investigate the predictability of understanding in the MUNDEX corpus (Türk et al., 2023), with understanding labels created based on participants' self-report of their understanding during the game explanation (Lazarov et al., 2025).

3. Dataset: MUNDEX Corpus

For our research, we used a the MUNDEX corpus (Türk et al., 2023) involving dyadic explanations of how to play a board game (Figure 1). MUNDEX contains manual and automatic multimodal annotations (Buschmeier et al., 2025) of the acoustic signal (e.g., voice quality), textual descriptors (e.g., discourse functions), and nonverbal behaviour (e.g., gaze, head gestures, and adaptors). The corpus was created to study how different states of understanding of explanations are multimodally signalled. These states were annotated using 'retrospective video-recall', a self-annotation method (Lazarov et al., 2025) in which explainees watched a recording of their own interaction (immediately after the interaction was completed) and commented on their state of understanding (into four levels). These comments were grouped under two labels: understanding and non-understanding. Inter-annotator agreement regarding understanding annotations was high (Cohen's κ = 0.90; Türk et al., 2024).

Label	N	%
Understanding	176	27.4
Partial Understanding	162	25.2
Non-Understanding	191	29.8
Misunderstanding	113	17.6

Table 1: Distribution of labels for different states of understanding extracted from annotations during the game explanation phase.

In the current study, we expand on the analysis of Türk et al. (2024), using these four levels of understanding. The additional understanding labels reflect the interaction dynamics more deeply, such as when the explainee only understood part of the content (i.e., Partial Understanding), or they understood what was explained in a different way (i.e., Misunderstanding). This expansion enables a more comprehensive investigation of understanding, acknowledging its gradual nature. Here, we selected a total of 21 explanatory dialogues from MUNDEX and calculated the distribution of different understanding labels (Table 1): Understanding (29.8%) and Non-Understanding (27.4%) have slightly higher proportions compared to Partial Understanding (25.2%) and Misunderstanding (17.6%).

4. Methods

In Section 2.1, we discussed potential linguistic cues, indicating cognitive load during language comprehension, to analyse. Based on the survey, we choose the following three linguistic aspects in order to see how effective they are for predicting different understanding states: (i) semantic information conveyed in the utterances (measured using information value); (ii) structural complexity of the utterances (measured using syntactic complexity score and dependency length); and (iii) variation in the listener's gaze behaviour (measured using gaze entropy). In the following, we introduce the methods for cognitive load quantification.

Information Value Quantification We quantified the information value of utterances based on the work of Xu and Reitter (2018) and Giulianelli et al. (2021), where the information value of an utterance, $H(X)$, is defined as the average surprisal of each word in the utterance. Given an utterance of words $\langle w_1, \ldots, w_N \rangle$, average gaze entropy is defined as:

$$H(X) = -\frac{1}{N} \sum_{i=1}^{N} \log P(w_i \mid w_{<i})$$

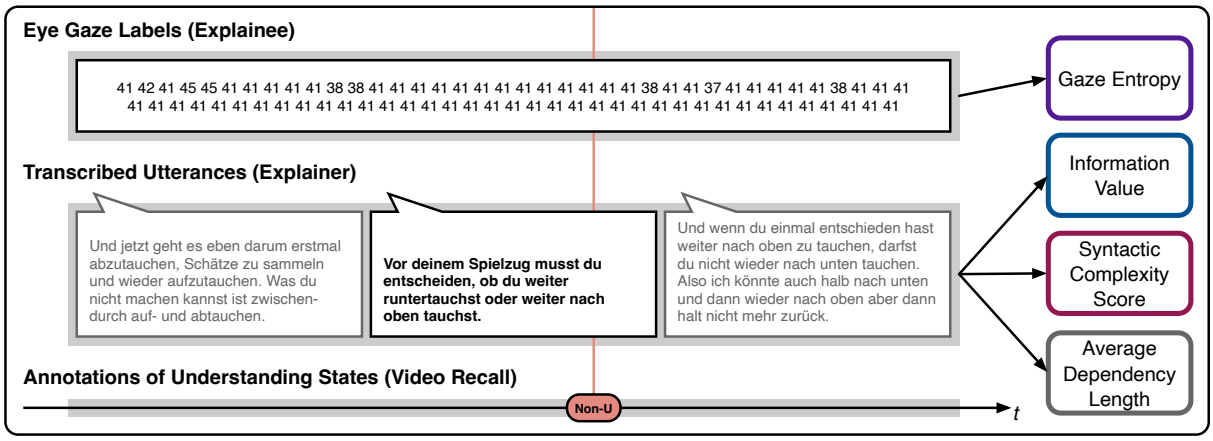

Figure 2: The general quantification pipeline for getting average information value, average gaze entropy, average syntactic complexity score, and average dependency length.

The greater the information value, the more unpredictable the content. In our study, being exposed to utterances with higher information value is expected to increase the explainee's cognitive load. We measure utterance surprisal by estimating the information content with GPT-2 (specifically, dbmdz/german-gpt2; Schweter, 2020).

For our approach to syntactic complexity quantification of utterances, we now introduce Syntactic Complexity Score and Average Dependency Length as two alternative metrics.

Syntactic Complexity Score Given a speaker's utterance, we calculate its length L and use dependency parsing (Qi et al., 2018) to obtain the number of heads α as well as the maximum tree depth β (see Wang and Buschmeier, 2024). The syntactic complexity SC of the utterance is then computed as follows:

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

where λ is a tuning factor (0.5 by default). The formula above takes the following aspects into consideration: Generally (i) an utterance with longer length (L) requires more working memory, and thus a higher cognitive load; (ii) an utterance with more heads (α) in its syntactic structure is harder to process; (iii) an utterance with larger tree depth (β) tends to have more embedded clauses, which is also a sign of syntactic complexity.

Average Dependency Length An alternative approach for quantifying syntactic complexity relies on the average dependency length. Dependency length minimization theory states that speakers will minimise dependency length in order to decrease processing difficulty (Futrell et al., 2015), which indicates that the dependency length is a decent measurement for syntactic complexity. We follow Liu (2008) and Futrell et al. (2015) to formally define dependency length as follows: We regard an utterance as a large syntactic structure. Let an utterance of length n be a syntactic structure where each node will have one arc. Its representation can then have a set of dependencies $D = \{(h_i, d_i)\}_{i=1}^{n}$, where each (h_i, d_i) denotes a syntactic dependency between a head word at position h_i and its dependent at position d_i. The dependency length for a word and its dependent is the number of words in between. The average dependency length of an utterance then represents its syntactic complexity:

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

In general, a larger average dependency length value can indicate that the utterance is more lengthy and has more embedded syntactic structure, and thus, requires more working memory to process.

Average Gaze Entropy Gaze variation during interaction is considered an important nonverbal cue to estimate a listener's understanding. To quantify gaze variation, we use the method proposed in Wang and Buschmeier (2023) and Wang et al. (2024). It quantizes the speech aligned gaze vector, generated from explainee videos using OpenFace (Baltrusaitis et al., 2018), into a discrete number ranging from 1 to 81, where 41 is the most frequent label indicating explainee's gaze direction is on the explainer (attention on the explainer). The other values suggest gaze aversion to varying degrees. Figure 2 illustrates how the explainee's gaze is aligned with the words of the explainer and quantized into a gaze labels (the numbers). Next, we apply a transformer model that has been trained using gaze label data in order to obtain gaze entropy values.

Given a sequence of gaze labels $\langle e_1, \ldots, e_T \rangle$, the average gaze entropy is defined as:

$$\text{NLL}(e_1, e_2, \ldots, e_T) = -\frac{1}{T} \sum_{i=1}^{T} \log P(e_i \mid e_{<i})$$

The higher the average gaze entropy value, the more unpredictable the gaze label becomes. This can indicate different degrees of gaze variation. Variation in gaze, such as for example gaze aversion – considered a cue for memory search (Glenberg et al., 1998) – is expected when potentially high cognitive load is present on the side of the explainee.

Quantification Pipeline The schema in Figure 2 illustrates the pipeline we use to obtain the quantifications for our statistical analysis and understanding state classification. For each annotated understanding state in our corpus ('Non-Understanding' in the example shown in Figure 2), we first identify its corresponding utterance. This utterance is then extended with its immediate context in form of its preceding and its succeeding utterance. In Figure 2, the utterance corresponding with the annotated understanding state is shown in the second box (in bold), while the preceding and succeeding utterances (in the first and third boxes, in grey) are prepended and appended to form its context. This processing is based on our observation that understanding states can be confirmed with greater confidence when contextual factors are taken into consideration. The combined utterances are then used to compute the average information value, syntactic complexity score, and average dependency length value. In addition, they are aligned with the corresponding gaze labels of the explainee (gaze behaviour while these utterances were spoken) to compute the average gaze entropy value.

5. Results and Discussion

In our survey of previous studies, we discussed linguistic cues which may be related to the development of understanding during interaction. Here, we examine selected linguistic cues from the MUNDEX corpus and quantify the linguistic information based on the approach proposed in Section 4. We first present a statistical analysis of the individual cues, which is followed by classification experiments involving the cues.[1]

[1]All values used in the statistical analyses and classification tasks have been normalised using the function *MinMaxScaler* from scikit-learn (Pedregosa et al., 2011).

Predictor	**H**	***p***	η^2
Information Value	9.029	0.0289	0.0094
Gaze Entropy	9.035	0.0288	0.0095
Syntactic Complexity	8.853	0.0313	0.0092
Dependency Length	1.344	0.7187	−0.0025

Table 2: Kruskal-Wallis test results showing associations between different understanding states and each predictor variable.

5.1. Statistical Analysis

With our statistical analysis we aim to establish whether the selected linguistic cues individually vary among the four different states of understanding. As can be seen by comparing the boxplot diagrams in Figure 3, the median *information value* is lower for misunderstanding and non-understanding, the median *gaze entropy* is lower for misunderstanding, and the median *syntactic complexity* score is higher in understanding and non understanding. For dependency length, the medians are almost identical across the four different understanding states.

For our statistical analysis, we first conduct Kruskal-Wallis tests for each linguistic cue, assuming no difference in median value among the understanding states as the null hypothesis. The results, as shown in Table 2, indicate statistically significantly different medians between understanding states for information value, gaze entropy, and syntactic complexity scores, but not for average dependency length (α = 0.05). It should be noted though that effect sizes (η^2) are very small.

These analyses are then followed up by pairwise post-hoc tests between understanding states (Dunn's test with Bonferroni correction for multiple comparison). Results show statistically significant differences between two states for each linguistic cue (see bars in Figure 3, *alpha* = 0.05). Median information value differs between Partial Understanding and Misunderstanding (p = 0.032), median gaze entropy differs between Non-Understanding and Misunderstanding (p = 0.036), and median syntactic complexity score differs between Understanding and Partial Understanding (p = 0.042).

It is worth noting that we do not assume a direct link between potentially higher cognitive load, as measured by average information value, and misunderstanding or non-understanding. However, based on our observations from Figure 3, when explainees are exposed to speech which takes a higher cognitive load to process (i.e., a higher average information value and higher average gaze entropy value), they are more likely to report, in the subsequent video recall task, that they understand, or partially understand the explanation. Specifically, Figure 3 suggests that Understanding and Partial

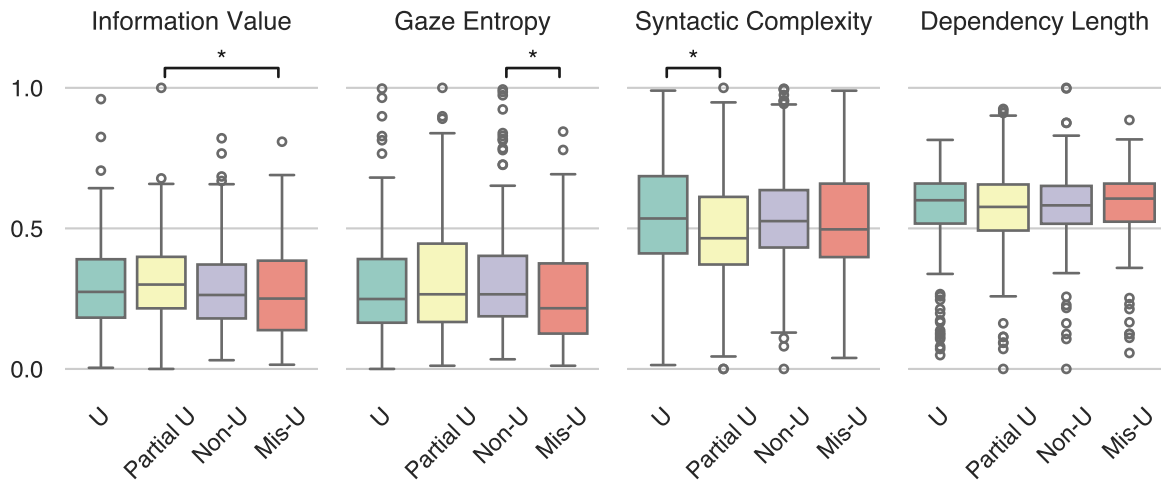

Figure 3: Variation of the quantified linguistic cues under different states of understanding ('U'). Horizontal bars show statistically significant Dunn's post-hoc tests (Bonferroni-corrected, α = 0.05).

Understanding tendencies are slightly more prevalent when information value is higher. This may indicate that listeners have a higher probability of understanding the content when exposed to speech with a higher information value (high cognitive load). One possible explanation is that listeners are more attentive when the information value is higher which aligns with previous studies assuming that high information density may help to maintain listener's attention (see Tsipidi et al. 2024, citing Bjare et al. 2024 for findings on music: listener engagement can be influenced by modulating surprisal). However, this assumption still requires further experimentation. Additionally, average gaze entropy also shows a similar pattern, with higher average values for Understanding and Partial Understanding compared to the Non-Understanding and Misunderstanding. We hypothesise that higher average gaze entropy values correspond to greater variation in gaze labels, which is considered a cue to cognitive load in some empirical analyses (e.g., Morency et al., 2006; Glenberg et al., 1998).

In the following analysis, we go beyond individual linguistic cues and will analyse how they can be jointly used in a classification task.

5.2. Classification Experiments

We conduct classification experiments to investigate, based on the analysis of the classification result, how predictable different understanding labels. Based on the results of the statistical analysis in Section 5, we decided to only use average information value, average gaze entropy value, and syntactic complexity score, excluding average dependency length.

We employ two commonly used classifiers, Random Forest from the scikit-learn toolkit (Pedregosa et al., 2011) and XGBoost (Chen and Guestrin, 2016). In addition, we also fine-tune a German BERT model (Devlin et al., 2019; Minixhofer et al., 2019) with the utterance data and then fuse the three selected linguistic cues into the model as a third classifier (see Figure 4). Based on earlier empirical findings that BERT tends to encode semantic and co-reference information (which is potentially related to understanding in our study) in the higher layers (see Tenney et al., 2019), we only used the last four hidden layers in the fusion step. The three linguistic cues are concatenated with the textual features represented in the hidden layer, and further passed through the linear layer to perform the classification task.[2] To make the comparison fair, we test the performance of the three classifiers under two different settings: (1) using only textual features to predict understanding states and (2) combining textual features and the selected linguistic cues. The textual features are encoded using scikit-learn's TF-IDF for the Random Forest and XGBoost classifiers, whereas the custom BERT model encodes them as a language model. For model evaluation, the training and testing data are split 7:3. Evaluation is performed using only the test data. In addition, 10-fold cross-validation was used as another evaluation method. The classification results are reported in Table 3. We report precision, recall, F1 score, as well as average accuracy and Macro F1 for each model accordingly.

Since the classification task contains four labels, a random classifier would achieve a chance baseline accuracy of 0.25. First, all of the classifiers under both settings perform better than chance.

[2]The training employed 10-fold stratified cross-validation with AdamW: 15 epochs using Cross Entropy loss; learning rate: 2*e*-5; batch size: 8; learning dropout rate: 0.2.

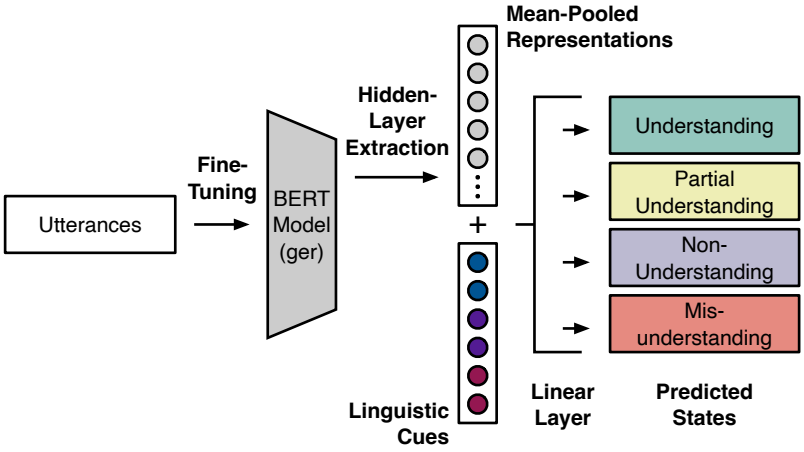

Figure 4: Understanding state classification by fusing linguistic cues to a fine-tuned BERT model. We first fine-tuned a (German) BERT model with the dialogue data from the MUNDEX corpus in order to learn potential textual features related to understanding states. We then focused on the last four hidden layers, fusing them with the three significant linguistic cues identified in Section 5.1 (average information value, average gaze entropy, syntactic complexity score) .

		Random Forest			XGBoost			German BERT		
	Class	Precis.	Recall	F1	Precis.	Recall	F1	Precis.	Recall	F1
Textual only	Understanding	0.74	0.70	0.72	0.59	**0.80**	0.68	**0.76**	**0.80**	**0.78**
	Partial Understanding	0.75	**0.67**	0.71	0.71	0.59	0.65	**0.80**	**0.67**	**0.73**
	Non-Understanding	0.55	0.77	0.64	**0.79**	0.65	0.71	0.71	**0.91**	**0.80**
	Misunderstanding	0.67	0.33	0.44	0.54	0.54	0.54	**0.88**	**0.58**	**0.70**
	10-CV Acc. (mean ± SD)	0.762 ± 0.050			0.695 ± 0.049			**0.768** ± 0.139		
	10-CV Macro F1 (mean ± SD)	0.756 ± 0.058			0.693 ± 0.052			**0.758** ± 0.158		
Textual & Linguistic	Understanding	0.71	0.67	0.69	0.50	0.72	0.59	**0.74**	**0.85**	**0.79**
	Partial Understanding	0.87	**0.81**	**0.84**	0.77	0.62	0.69	**0.88**	0.78	0.82
	Non-Understanding	0.70	**0.84**	0.76	0.74	0.74	0.74	**0.82**	0.82	**0.82**
	Misunderstanding	0.78	0.64	0.70	**0.83**	0.45	0.59	0.82	**0.75**	**0.78**
	10-CV Acc. (mean ± SD)	0.790 ± 0.032			0.709 ± 0.051			**0.816** ± 0.038		
	10-CV Macro F1 (mean ± SD)	0.793 ± 0.031			0.708 ± 0.050			**0.812** ± 0.042		

Table 3: Comparison of the three classifiers with textual features only (upper half) and textual features and linguistic cues combined (lower half). Best scores in bold.

Secondly, although the linguistic cues' effect size is small (Table 2 in Section 5.1), classification performance is better when the they are used in addition to the textual cues – which indicates the combined effectiveness of the linguistic cues in predicting different states of understanding. Third, it appears that the labels Partial Understanding and Misunderstanding are harder to predict. Although for the BERT-based classifier, the F1 score for Misunderstanding is 0.78 when we use both textual features and linguistic cues, it is still lower than predictions of the other three labels. Fourth, Non-Understanding appears to be the easiest label to predict. This is most apparent for the BERT based classifiers, where F1 scores are 0.8 and 0.82 respectively, 10 pp and 4 pp higher than for the label Misunderstanding.

We now consider why Misunderstanding seems to be harder to predict correctly. First, we believe the weak performance to be related to the class imbalance of understanding labels. There is less training data for Misunderstanding and more training data for Non-Understanding (see Table 1). Secondly, zooming in on the predictions made the German BERT model with textual and linguistic cues with the help of a confusion matrix (see Figure 5), it can be observed that for Misunderstanding, the model assigns uniform probability to the three other labels. In face-to-face interaction, interlocutors may not be aware at the time that their understanding is impaired, and may only later realise that they misunderstood what was said. Similar model behaviour can be seen for Partial Understanding, which tends to be predicted as Understanding.

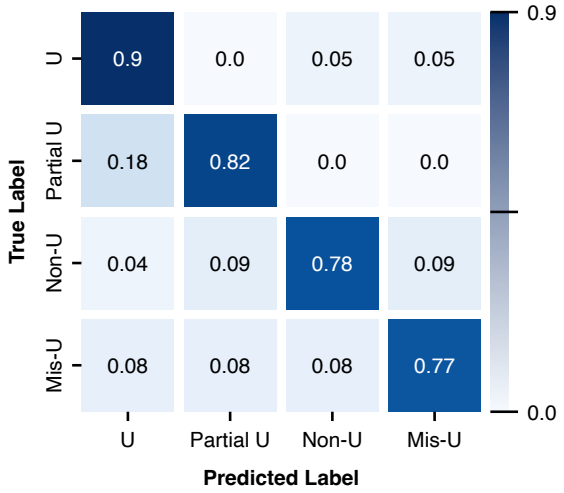

Figure 5: Confusion matrix for the German BERT model for classifying understanding states ('U') with textual features and linguistic cues.

6. Conclusions and Future Work

We aimed at investigating the predictability of different understanding states of explainees in explanatory interaction using different cognitive load related verbal and non-verbal linguistic cues. This builds on previous work of Türk et al. (2024) that analysed the predictability of different understanding states based on two classes: Understanding and Non-Understanding. For the analyses presented here, we used an updated version of the MUNDEX corpus with four different understanding states. Based on our survey of the literature, we hypothesised that the cognitive activity of understanding should be correlated with cognitive load. We chose linguistic cues potentially related to cognitive load: information value, syntactic complexity, and gaze. We used 21 explanatory dialogues of the MUNDEX corpus and quantified these three linguistic cues into four different values which may indicate cognitive load.

Our statistical analysis shows that the information value and the syntactic complexity score of explainer's utterances, as wells as explainee's gaze variation, differ significantly between listener's states of understanding. We then used these three measures in two off-the-shelf classifiers as well a fine-tuned German BERT-based classifier to conduct classification of understanding states. The results show that the BERT-based classifier generally performs much better than the two off-the-shelf classifiers, which further hints at the predictability of different states of understanding. Model performance also reveals potential challenges in predicting different understanding states, especially states labelled Misunderstanding and Partial Understanding.

As predicting states of understanding requires feature engineering of different multimodal signals, we believe there is scope to improve our current work. We consider the following to be important for interpreting different states of understanding: speech signals such as pitch and voice quality; vocal backchannels; as well as hand and head gestures. In future work, we will focus on these as additional factors. We plan to incorporate these multimodal features to see if we can improve the classification results. Another area for future research will be to try different models besides BERT to see if better classification results can be achieved.

7. Ethical Considerations and Limitations

The publicly available MUNDEX corpus does not contain personal data of the study participants. The corpus was collected with the protection of personal data in mind, and was approved by our institutional review board.

Within the scope of this study, we consider the following three limitations which we would like to leave for future work: Firstly, although MUNDEX is the only publicly available multimodal corpus which provides richly annotated understanding labels, the data is still quite limited from a machine learning perspective. In the future, we hope that comparable data will be made available in order to build more reliable and robust classifiers for tracking listeners' states of understanding in human–human (and human–agent) interaction. Secondly, the MUNDEX corpus contains data from German speakers. The results and analyses reported here therefore lack linguistic generality. It is possible that the tendency to express different states of understanding (e.g., Understanding/Non-Understanding) varies cross-culturally. In future work, we hope to use multimodal corpora of other languages in order to further generalise our findings. Thirdly, to obtain the gaze labels, we relied on OpenFace (Baltrusaitis et al., 2018), which could lead to the gaze labels not capturing the gaze movement with state-of-the-art accuracy (e.g., above 95%). In future work, if more advanced gaze tracking software becomes available, we hope to create gaze labels with greater accuracy.

8. Supplementary Material

Code and data are available on Zenodo: https://doi.org/10.5281/zenodo.19003190

9. Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/3 2026 – 438445824, project A02. We thank the participants and our colleagues and student assistants

for their support in data collection, transcription, and annotation. We would also like to thank the anonymous LREC 2026 reviewers who provided constructive feedback to help us improve this work.

10. Bibliographical References

Jens Allwood, Stefan Kopp, Karl Grammer, Elisabeth Ahlsén, Elisabeth Oberzaucher, and Markus Koppensteiner. 2007. The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behaviour simulation. *Language Resources and Evaluation*, 41:255–272.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Agnes Axelsson and Gabriel Skantze. 2022. Multimodal user feedback during adaptive robot-human presentations. *Frontiers in Computer Science*, 3:741148:22.

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.

Janet Bavelas and Nicole Chovil. 2018. Some pragmatic functions of conversational facial gestures. *Gesture*, 17(1):98–127.

Simon Betz, Nataliya Bryhadyr, Olcay Türk, and Petra Wagner. 2023. Cognitive load increases spoken and gestural hesitation frequency. *Languages*, 8(1).

Mathias Rose Bjare, Stefan Lattner, and Gerhard Widmer. 2024. Controlling surprisal in music generation via information content curve matching. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pages 922–929, San Francisco, CA, USA. ISMIR.

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.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA, USA. ACM.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 222–233. American Psychological Association, Washington, DC, USA.

Herbert H. Clark and Meredyth Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81.

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 and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN, USA. ACL.

Stefan L. Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494.

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.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. The MIT Press.

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.

Arthur M. Glenberg, Jennifer L. Schroeder, and David A. Robertson. 1998. Averting the gaze

disengages the environment and facilitates remembering. *Memory & Cognition*, 26:651–658.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, PA, USA. ACL.

Christine Howes and Arash Eshghi. 2021. Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *Journal of Logic, Language and Information*, 30:331–362.

Marcel A. Just and Patricia A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1):122–149.

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

Shunichi Kinoshita, Toshiki Onishi, Naoki Azuma, Ryo Ishii, Atsushi Fukayama, Takao Nakamura, and Akihiro Miyata. 2023. A study of prediction of listener's comprehension based on multimodal information. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 30:1–4, Würzburg, Germany. ACM.

Stefan Lazarov, Michael Schaffer, Viviane Gladow, Hendrik Buschmeier, Angela Grimminger, and Heike M. Buhl. 2025. Applications of video-recall for the assessment of understanding and knowledge in explanatory contexts. OSFPreprints.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Dekang Lin. 1996. On the structural complexity of natural language sentences. In *Proceedings of 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.

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. ACL.

Dimitry. Mindlin, Amelie S. Robrecht, Michael Morasch, and Philipp Cimiano. 2024. Measuring user understanding in dialogue-based xAI systems. In *ECAI 2024: 27th European Conference on Artificial Intelligence*, pages 1148–1155, Santiago de Compostela, Spain. IOS Press.

Benjamin Minixhofer, Thomas Winkler, Martin Schmitt, Hinrich Schütze, and Markus Leuck. 2019. German BERT models (bert-base-german-dbmdz-uncased). Hugging Face.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Uppsala, Sweden. ACL.

Louis-Philippe Morency, C. Mario Christoudias, and Trevor Darrell. 2006. Recognizing gaze aversion gestures in embodied conversational discourse. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, page 287–294, Banff, Canada.

Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Dennis Reidsma, Iwan de Kok, Daniel Neiberg, Sathish Pammi, Bart van Straalen, Khiet Truong, and Herwin van Welbergen. 2011. Continuous interaction with a virtual human. *Journal on Multimodal User Interfaces*, 4:97–118.

Amelie Robrecht and Stefan Kopp. 2023. SNAPE: A sequential non-stationary decision process model for adaptive explanation generation. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 48–58, Lisbon, Portugal. SciTePress.

Katharina Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike Buhl, Hendrik Buschmeier, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath,

Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. 2021. Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13:717–728.

Katharina Rohlfing, Kary Främling, Brian Lim, Kirsten Thommes, and Suzana Alpsancar, editors. 2026. *Social Explainable AI*. Communications of NII Shonan Meetings. Springer, Singapore.

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

Cory Shain. 2021. CDRNN: Discovering complex dynamics in human language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3718–3734, Online. ACL.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

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.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18820–18836, Miami, Florida, USA. ACL.

Olcay Türk, Stefan Lazarov, Yu Wang, Hendrik Buschmeier, Angela Grimminger, and Petra Wagner. 2024. Predictability of understanding in explanatory interactions based on multimodal cues. In *Proceedings of the 26th ACM International Conference on Multimodal Interaction*, pages 449–458, San José, Costa Rica. ACM.

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. ACL.

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*, pages 75–80, Vienna, Austria. ACL.

Yu Wang, Yang Xu, Gabriel Skantze, and Hendrik Buschmeier. 2024. How much does nonverbal communication conform to entropy rate constancy?: A case study on listener gaze in interaction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3533–3545, Bangkok, Thailand. ACL.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

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

11. Language Resource References

Hendrik Buschmeier, Angela Grimminger, Petra Wagner, Stefan Lazarov, Olcay Türk, and Yu Wang. 2025. MUNDEX annotations (version 0.7). Zenodo.

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.