Exploring Content Predictability in Turn-Taking Through Different Computer-Mediated Communications

Wanqing He, Calen C. MacDonald, Yejoon Yoo, Marcos Eizayaga, Ryun Shim, Lev D. Katreczko, Susan R. Fussell
Cornell University, United States
{wh385, ccm262, yy937, me437, rs2279, ldk62, sfussell}@cornell.edu

Abstract

Previous studies of face-to-face (f2f) communication have suggested that speakers rely heavily on a variety of multi-modal cues to make real-time predictions about upcoming words in rapid turn-taking. To understand how computer-mediated communication (CMC) differs from f2f communication in terms of the prediction mechanism, this study assessed how the loss of multi-modal cues would affect word predictability in turn-taking. Participants watched videos, listened to audio, or read a transcript of f2f conversations. Across these three conditions, they predicted the same set of omitted words with different levels of predictability and semantic relatedness to other words in the context. Results showed that words of higher predictability were more accurately predicted regardless of CMC types. Higher response accuracy but longer response time were observed in conditions with richer cues, and for participants with more positive and less negative self-emotions. Meanwhile, semantic relatedness did not affect predictability. These results confirmed the key role of prediction in language processing and conversation smoothness, especially its importance in CMC.

1 Introduction

Effective communication depends on the predictability of conversational dynamics, enabling smooth turn-taking and facilitating the overall flow of communication (Torreira et al., 2015). Studies on face-to-face (f2f) communication have highlighted that speakers rely on a range of verbal and non-verbal cues to make real-time predictions about upcoming words (Corps et al., 2018). However, computer-mediated conversation (CMC) introduces distinctive challenges stemming from the variations in the availability and effectiveness of these cues across diverse communication channels, encompassing video, audio, and text-only (Kalman et al., 2006). In CMC, the absence or limited presence of these cues impairs the prediction process and frequently engenders disruption in coherence and fluidity in speech flow.

More notably, turn-taking fundamentally depends on cooperative information sharing and common ground (Tomasello, 2008). The degree of interpersonal alignment is subject to the level of intentionality, the specific communicative goals and tasks at hand, attention and affective states and emotion (Hall et al., 2019; Rasenberg et al., 2022). These sociopsychological aspects are intertwined to affect the prediction process, thus influencing overall speech comprehension in conversational turns (Verhees et al., 2015; Hinojosa et al., 2019).

Despite extensive research on the implications of predictability for turn-taking in f2f communication, a comprehensive understanding of the nuanced variations of predictability across different communication channels is limited. The inherent disparities regarding the availability and effectiveness of verbal and non-verbal cues in modalities characteristic of CMC engender a gap in knowledge regarding the prediction and grounding mechanisms in these contexts.

As such, this study aims to fill this gap and address the open topic of how the predictability of incoming material impacts the dynamic of turn-taking in CMC as opposed to f2f communication. Additionally, this research aims to shed light on the

Proceedings of the 31st International Conference on Computational Linguistics, pages 7949–7962
January 19–24, 2025.

intricate relationship between predictability and turn-taking interruptions in CMC.

To achieve these objectives, we compared predictability in video, audio, and text formats in a behavioral study in which participants were asked to complete a word prediction task with a prerecorded conversation. Accuracy, semantic relatedness, and time of responses were measured.

Our results showed that higher response accuracy, but longer response time was observed in conditions with richer cues. Semantic relatedness or attention did not affect predictability. Participants with more positive emotions showed slower responses and those with less negative emotions showed lower response accuracy. We further discussed implications of these findings in language processing and conversation smoothness, and how they can inform the processing mechanisms in CMC with varying availability of multi-modal cues.

2 Related work and hypotheses

Prediction in Language Processing Disfluency is often observed in real-time interactions (Brennan and Williams, 1995; Shriberg, 1994; Smith and Clark, 1993), not restricted to f2f conversations but also extensively in CMC, which further induces miscomprehension (Duan et al., 2021; Lim et al. 2022; Walther, 2011). Disfluency can be attributed to multiple aspects: age, gender, speed, clarity, and individual differences in language experience, memory, and cognition among other factors (Bortfeld et al., 2001; Li et al., 1995). One prominent yet often overlooked source of the problem is content prediction in conversation. In psycholinguistics models, prediction functions as a central mechanism of comprehension and production in language processing. An increasingly popular hypothesis about the fundamental workings of the language system in the human mind and brain is that people are generally able to keep up with language input by predicting what comes next – by activating the meaning and potentially other aspects of words ahead of time (see Ryskin and Nieuwland, 2023). These predictions rely on conversational cues, especially the preceding sentence context, and global information arisen from the context.

Prominent Cues in Prediction Multi-modal communication is shown to have an outstanding efficacy compared to unimodal communication (Froehlich et al., 2019) with a temporal advantage of simultaneous information transmission in two modalities, a spatial advantage in the visual complementarity to auditory information (Gergle et al., 2013), and a pragmatic advantage of allowing for multi-modal interpersonal adaptive behaviors (Toma et al., 2014). It increased the likelihood and efficiency of successful dyadic, f2f social contacts amongst conversational partners.

For instance, people predict what the other interactant is about to say based on content (Corps et al., 2018), when the turn will end based on semantic, syntactic, and prosodic features (Bögels and Torreira, 2015; De Ruiter et al., 2012; Hadley and Culling, 2022) and use these cues to early preparation for what they need to say in the next turn (Magyari and De Ruiter, 2012). This online processing in comprehension and prediction from both parties is continuous, simultaneous, and interactive that draws on highly compressed and demanding cognitive resources (Levinson, 2016).

Emotion and Attention Social factors such as emotional states and attention have shown to influence different aspects of language processing (Fredrickson and Branigan, 2005; Hinojosa et al., 2019; Verhees et al., 2015). Positive emotion has been linked to cognitive flexibility, a more global, category-level processing style with a broad attentional focus, while negative emotion, contrastively, is associated with a more localized, bottom-up, analytic, and systematic approach with a narrower attentional scope. (Gasper and Clore, 2002). A substantial body of research has demonstrated that emotion plays a significant role in influencing attentional processing (see Kaspar and König, 2012 for a comprehensive review). Therefore, emotion and attention are critical factors to consider when comprehensively examining the prediction mechanisms underlying CMC.

Prediction Challenges in CMC Turn-taking is inherently complex, posing significant cognitive demands as it requires interactive and simultaneous processing of a vast array of cues. The absence of any cue hinders the smooth comprehension and prediction of the upcoming content. This challenge is particularly pronounced in CMC, where varying degrees of cue loss—depending on the modality, such as video meetings, phone calls, or text messaging—impede the efficiency of rapid turn-taking (Trujillo et al., 2021; Levinson, 2016).

CMC mainly differs in which multi-modal cues are available: text cues in messaging, verbal cues in voice calls, and both verbal and non-verbal cues in

video meetings. In all forms of CMC, there is a great level of miscommunication incurred by the absence of certain multi-modal cues. Interlocutors face varying degree of cue losses and technical difficulties that obstruct comprehension, prediction, and speech planning processes. Therefore, it is important to understand how interlocutors utilize multi-modal cues with different conversational strategies in response to cognitive, emotional, and environmental constraints.

Based on the objectives above, the following hypotheses have guided this study:

H1: There will be more accurate responses for words with greater sequential predictability in all types of CMC.

H1a: The video condition (V) will show the highest response accuracy, followed by the audio condition (A) and followed by the text condition (T), given the gradual loss in informative multimodal cues.

H1b: Participants with more positive self-emotion and higher level of attention will provide more accurate responses for all types of CMC.

H2: We do not expect the semantic relatedness of responses to differ as much across CMC types, as this feature is heavily based on context and should be less sensitive to the availability of cues.

H3: There will be faster responses for words with higher sequential predictability in all CMC types.

H3a: Participants will show fastest response in V, followed by the A, and followed by T, given the gradual loss in informative multimodal cues.

H3b: Participants with more positive emotions and higher level of attention during the task will provide faster responses for all types of CMC.

By investigating these hypotheses, this research seeks to shed light on the relationship between predictability and turn-taking disruptions in different types of CMC. The findings contribute to a deeper understanding of the unique challenges posed by CMC and inform the development of strategies to improve communication effectiveness in various online communication contexts.

3 Method

3.1 Participants

Participants (n = 191; age = 18-28; M = 54, F = 122, non-binary = 1) were undergraduate students recruited via the SONA platform in exchange for 0.5 course credit or a $5 Amazon gift card if they completed the study online or in-lab. In total, 13 participants were excluded due to incompleteness of the task or survey. Participants have no known visual or auditory impairments.

3.2 Materials

Detailed study materials can be found on Open Science Framework (https://osf.io/35z6a).

Conversation materials Two conversation clips were conducted by two separate pairs of volunteers who agreed to hold a 10-minute recorded conversation on the prompt: "What do you think of opening up a new dairy bar somewhere on campus? Where would be a good choice? What are the pros and cons that you think of?" This prompt ensures high topic familiarity for all potential participants and constrains the array of context to be homogenous enough between two conversations. The recordings were modified into video, audio, and text-only versions. Six counter-balanced task conditions were accumulated from three different CMC (V, A, T), each with two different conversations. This study design ensures that the observed results were not due to any specific aspects of the conversation recordings. An example trial for each condition can be found in appendix.

Prediction Task Individual words were omitted and replaced by an underline in the middle or towards the end of the utterances for each conversation. The stream of video, audio, or text immediately paused before the selected words. Participants were asked to fill out these individual word slots based with their own guesses by typing in a popped-up textbox.

Word Selection We assessed the word's predictability from two aspects, 1) sequential predictability (in three levels: high-mid-low) and 2) semantic relatedness (also high-mid-low). In dyadic turn-taking, the predictability of turns by one interlocutor is based on both the previous turns of this interlocutor and the previous turns of the other interlocutor. Sequential predictability can capture the predictability of words within a turn well but performs poorly across turns from different interlocutors. Semantic relatedness is used to evaluate the predictability of content across interlocutors based on contextual fit within three consecutive turns. Three consecutive turns were chosen given the memory limit in turn-taking.

We defined sequential predictability as the likelihood for a word to occur given its context. It was computed as the word predictability using the state-of-art language model – generative pre-training 2 (GPT-2; Radford et al., 2019).

We defined semantic relatedness as how closely related the words are in terms of the taxonomic and thematic relation in context. It is computed using GloVe word embeddings (Pennington et al., 2014), where every word in turn *N* is compared to the entire paragraph which contains all the preceding words in the same turn *N* and all the words in turn *N* – 1 and *N* - 2. For example, the semantic score for the word "apple" was obtained by comparing "apple" to the preceding context "I like __." from the same turn by speaker 1, the previous turn by speaker 2, and the previous turn by speaker 1. This score was adopted as a measure of the contextual fit of any individual word (Luke and Christianson, 2018).

For all words in each conversation, both the predictability score and semantic score were divided into three levels (high, mid, and low), each with a 20% quantile. Paired *t*-test ensured no multicollinearity across three levels. Six words were selected from each of the nine groups (high-high, high-mid, etc.), giving a total of 54 words omitted from the clips. Note that this nine-way grouping is for counterbalancing purpose only. In the data analysis, both scores were coded as continuous variables. The stimuli from these two conversation recordings were counterbalanced in their predictability scores and semantic scores throughout the task.

Post-test Questionnaire A short survey was designed to collect a brief summary of the main points of the conversation and the emotional states of participants towards the overall conversation and the interlocutors (with a 1- to 10-point rating for each emotion word that spans from positive to negative categories). Volunteers and participants both filled out the same post-test questionnaire for comparison purpose.

3.3 Experiment Set-up

Participants used their personal computers or research assistants' computers to access the study link on SONA, regardless of online or in-person study. For in-person studies, participants were placed alone in an isolated testing room on campus.

3.4 Procedure

This study was hosted on Pavlovia and run online via Qualtrics survey and PsychoPy software. The study design is fully between subjects: participants were randomly assigned to the prediction task in any of the three CMC types (V, A, T) of either conversation 1 or 2. Participants clicked on the screen to start the prediction task in PsychoPy. They were informed that they need to type down the next upcoming word when a text box pops up immediately after the video or the audio pauses, or when they see an underline in the text. The instruction varied based on the condition they were assigned to. Participants completed a demonstration trial to get a better understanding of the prediction task before the task officially started. After completion of all trials in the prediction task, they were redirected to Qualtrics to complete a post-test questionnaire to fill out their summaries, perception and experience, emotion ratings of the conversation. Demographic information such as age, gender, and major were collected at the end of the questionnaire.

4 Measures

4.1 Experiment measures

Performance in three CMC groups was compared against each other. We measured 1) *response accuracy* – how related the response is to the actual word of the slot. It was computed as the cosine similarity between the embeddings of the actual word and the response; 2) the *semantic score* of each response – how related the response is to any possible words for the slot. It was computed as the weighted average of the embeddings of the first 500 possible words where the weight is the normalized sequential predictability; 3) *response time* – the time taken after the video/audio/text stopped playing and participants typed down the response in the textbox and hit the "continue" button. In light of the rapid growth and significant improvement of computational models, we updated the word sequential predictability with Llama 3.0 (Dubley et al., 2024) in our data analysis.

All the measures of text responses only considered the first full word response, no matter how many words the rest of the response had or how accurate they were. This ensures fairness across trials and across participants per study instruction. Misspellings and typos of symbols and apostrophes were corrected by Speller from Autocorrect (Sondej, 2022) and then manually revised by research assistants to avoid penalty.

4.2 Survey measures

We assessed how the level of comprehension, engagement, attention, and different emotional

states may contribute to the willingness of active prediction in the conversation, which in turn affects the task performance.

Conversation Summary In the post-test questionnaire, we asked participants to summarize the pros and cons mentioned by each speaker in the conversation. We computed 1) the similarity between their responses and each of the volunteers' responses, and 2) the cosine similarity between their responses and the conversation transcript using spaCy (Honnibal and Montani, 2017). Given the abstractness of volunteers' responses, we decided to use the second approach as a more accurate measure of participants' attention. The averaged similarity formed a reliable scale (Cronback's $\alpha = .80$) and was included as a fixed effect in the linear mixed-effect (LME) model as the *attention level* in the conversation.

Perceptions and Experience Participants rated their perceptions of the emotions and relationships of both speakers in the conversation on 10-point Likert scale for 9 descriptive adjective words. The ratings of each word between the actual speakers and the participants as observers were compared. Given that the speakers did not respond fully to each item, the comparison was inconsistent within the question and cannot form a reliable scale. Therefore, this measure was removed.

Self-Emotion Participants also rated their own emotional states towards the conversation on 10-point Likert scale for each emotion word. These words were reliably loaded onto two categorical factors: positive self-emotion ($n = 5$; $\alpha = .90$) and negative self-emotion ($n = 3$; $\alpha = .69$). The average ratings within each category were included two predictors in LME model as *self-emotion* and *negative self-emotion*.

5 Results

5.1 Linear mixed-effect regression analysis

Response times longer than 30000 ms were removed, which resulted in 1.1% of data loss. We analyzed the results using LME models in R version 3.4.0 (R Core Team, 2017), with packages LME4 version 1.1.19 (Bates et al., 2015) and lmerTest version 2.0.33 (Kuznetsova et al., 2017). Empty responses were removed by models.

The *lmer* package was used to define the linear mixed-effect model, which included *participant ID*, *trial ID*, and *conversation type* as random effects. Fixed effects included *CMC type*, *overall score group of target words*, *self-emotion, negative self-emotion*, and *attention level* as the basic model.

Using the above-described LME model as a starting point, we conducted model selection using Least Absolute Shrinkage and Selection Operator (LASSO) to reduce the number of fixed effects and the number of interactions between fixed effects. This model selection is for better theoretical interpretability model fit due to the large number of non-significant effects presented in LME model. LASSO regression was deployed by *cv.glmnet* package version 1.6.1 (Friedman et al., 2010). With post-hoc models, we used the *emmeans* package (Searle et al., 1980) to gauge all the pairwise comparisons of *CMC type*. The corrections of *p* values based on three-way comparisons offer a better measure for multi-way comparison across CMC types with smaller chances of false positives.

5.2 Experiment results

Response Accuracy (RA) As predicted, reaction accuracy (range = [-.28, 1.00], M = .60, SD = .29) was better in V compared to T, and negative self-emotion shows a detrimental effect on RA.

As shown in table 1, overall score group was removed by model selection. H1 was not supported. However, it is worth noted that in the baseline model, participants' performance in the prediction task was affected by word predictability: compared to words with high predictability and high semantic relatedness (the baseline), RA of words with mid or low predictability was significantly lower regardless of their semantic relatedness within the context except words of low predictability but mid-level semantic relatedness. H1 was supported in the baseline model.

CMC type showed a significant main effect: participants had similar RA in A compared to V ($\beta = -.01$, $p = .53$), but significantly worse RA in T compared to V ($\beta = -.03$, $p = .009$) as shown by the decrease in cosine similarity of the response. With a pairwise comparison in CMC type, there was no significant change comparing A to V ($\beta = -.006$, $p = .80$), or comparing T to A ($\beta = -.03$, $p = .12$), but only significantly better RA in T compared to V ($\beta = -.03$, $p = .02$). H1a was partially supported.

Negative self-emotion had a significant main effect on RA: Participants predicted significantly worse ($\beta = -.01$, $p = .006$) if their negative self-emotion ratings were higher. As seen in figure 1, RA shows stronger negative correlation with negative self-emotion in both A and T, but less in

Fixed effects	Estimate	*SE*	df	*t* value	*p* value
(Intercept)	0.64	0.03	2.17	19.15	0.002**
CMC type: audio	-0.01	0.01	173.00	-0.63	0.53
CMC type: text	-0.03	0.01	172.63	-2.62	0.009**
Negative self-emotion	-0.01	0.002	172.96	-2.79	0.006**

Table 1: Post-hoc LME model for Response Accuracy (*Overall Score Group, Self-emotion* & *Attention* removed by model selection)

V. Neither did self-emotion (β = -.003, p = .24) or attention (β = -.04, p = .38) enter the final model. H1b was partially supported.

Response Semantic Relatedness As predicted, semantic relatedness (range = [-.33, 1.00], M = .37, SD = .33) was not affected by CMC type. As expected, CMC type did not enter the final model selection (A: β = -.002, p = .74; T: β = -.007, p = .34 in baseline LME model)**).** In table 2, none of the predictors showed significant effect on the semantic relatedness of responses except the overall score of the words: all groups except high-mid group showed different degrees of significant decrease in semantic relatedness comparing to the baseline high-high group. The lower the predictability of word slot and the lower the original level of semantic relatedness of word slot the lower the semantic relatedness of responses. Neither self-emotion (β = .001, p = .51) or negative self-emotion (β = -.001, p = .67) showed significant main effect. Attention (β = .03, p = .32) did not reach significance, either. H2 was supported.

Response Time (RT) Expectedly, RT (range = [3.40, 29415.10], M = 2933.8, SD = 3155.02) did not show the hypothesized pattern. The maximum value was set at 30000ms given that the required task was word typing. Although the textbox allows multi-word entries, we deemed it exceptionally long to spend over 30000ms on a single trial with sufficient attention. No minimum cutoff value was set for RT (and RT per character thereafter) given that all the empty responses associated with short RT were removed by LME models.

As shown in table 3, for the overall score group of words, all groups except high-mid and low-high groups showed different degrees of significant decrease in RT comparing to the baseline high-high group. Words with low semantic relatedness generally showed significantly shorter RT, but predictability did not correlate with longer RT. H3 was partially supported.

Participants did not show significant longer RT in A compared to V (β = 203.33, p = .43), but significantly shorter RT in T compared to V (β = -1750.77, p < .001). With a pairwise comparison in CMC type, there was no significant difference comparing A to V (β = 203.00, p = .70), but significantly shorter RT in T compared to V (β = -1751.00, p < .0001) and compared to A (β = -1954.00, p < .0001). H3a was not supported from this reversed pattern observed.

Self-emotion (β = 148.97, p = .01) had a marginal effect on RT: participants with more positive self-ratings of emotion spent significantly

Fixed effects	Estimate	*SE*	df	*t* value	*p* value
(Intercept)	0.56	0.15	1.04	3.64	0.16
Score group: high-low	-0.35	0.01	9193.13	-32.61	<.001***
Score group: high-mid	-0.01	0.01	8911.25	-1.38	0.17
Score group: low-high	-0.09	0.01	8559.95	-7.27	<.001***
Score group: low-low	-0.18	0.01	9222.02	-16.66	<.001***
Score group: low-mid	-0.19	0.01	8447.67	-15.22	<.001***
Score group: mid-high	-0.26	0.01	8977.64	-23.39	<.001***
Score group: mid-low	-0.39	0.01	8797.61	-27.49	<.001***
Score group: mid-mid	-0.36	0.01	8755.87	-26.67	<.001***
Self-emotion	0.001	0.002	172.70	0.70	0.49
Negative self-emotion	-0.001	0.002	172.19	-0.59	0.56
Attention level	0.03	0.03	171.89	1.00	0.32

Table 2: Post-hoc model for Response Semantic Relatedness (*CMC type* removed by model selection)

Fixed effects	**Estimate**	***SE***	**df**	***t* value**	***p* value**
(Intercept)	2626.48	615.84	38.44	4.26	<.001***
Score group: high-low	659.90	150.63	2766.28	4.38	<.001***
Score group: high-mid	22.33	147.58	6204.86	0.15	0.88
Score group: low-high	39.03	180.33	5127.26	0.22	0.83
Score group: low-low	367.66	152.71	4549.87	2.41	0.02*
Score group: low-mid	775.39	177.25	4238.16	4.37	<.001***
Score group: mid-high	472.06	156.59	5656.55	3.01	0.003**
Score group: mid-low	989.80	201.21	2313.64	4.92	<.001***
Score group: mid-mid	733.74	189.98	2641.42	3.86	<.001***
CMC type: audio	203.33	255.06	170.60	0.80	0.43
CMC type: text	-1750.77	262.55	170.17	-6.67	<.001***
Self-emotion	148.97	59.11	171.65	2.52	0.01*
Negative self-emotion	93.87	59.63	170.75	1.57	0.12
Attention level	778.32	1259.35	170.46	0.62	0.54

Table 3: Post-hoc model for Response Time (All variables were included after model selection)

longer in typing responses. However, negative self-emotion did not show significant effect. As seen in figure 2, the change in RT was large in V and A, but reversely in T. Lastly, attention level (β = 778.32, p = .54) did not show any significant main effect. H3b was not supported.

RT Per Character Given the varying lengths in text response while we only considered the first full-word entry, we performed the same analysis with *RT per character* to avoid misinterpretation using total RT. RT per character (range = [31.54, 19225.60], M = 665.33, SD = 1041.86) showed a different pattern from total RT.

First, overall word score group did not entered the full model. H3 was not supported. It is worth noted that even in the basic LME model, only words with high or mid predictability showed significant increase in RT per character compared to the baseline high-high group, regardless of their level of semantic relatedness.

The result for CMC type was similar: there was no significant difference between V and A (β = 22.73, p = .68), but significantly shorter RT in T compared to V (β = -627.14, p < .001). With a pairwise comparison in CMC type, there was no significant difference comparing A to V (β = 22.90, p = .91), but significantly much shorter RT in both T (β = -627.00, p < .001) and A (β = -649.90, p < .001) comparing to V. H3a was still not supported.

Self-emotion ratings and attention also entered the post-hoc model. However, unlike in RT, neither self-emotion (β = 17.09, p = .17) or negative self-emotion (β = -8.30, p = .52) showed any significant effect in RT per character. Similarly, attention did not significantly increase RT per character (β = 118.34, p = .66). H3b was not supported.

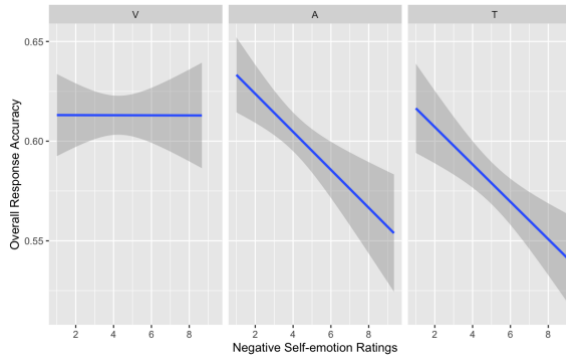

Figure 1: The difference of response accuracy with the change in negative self-emotion ratings (on a 1- to 10-point scale) across CMC types

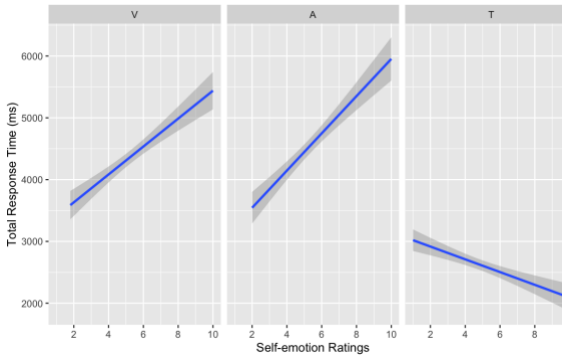

Figure 2: The difference of response time with the change in self-emotion ratings (on a 1- to 10-point scale) across CMC types

6 Discussion

In this study, we explored the effect of three CMC types (V, A, T) on word predictability. Our findings indicated that there were more accurate responses for words with higher sequential predictability in all CMC types. Higher RA but longer RT were observed in conditions with richer cues, and for participants with less negative emotions and higher attention level. Semantic relatedness did not affect predictability. These results confirmed the key role of prediction in language processing and conversation smoothness, especially its importance in CMC. In the rest of this section, we discuss these quantitative findings considering the implications.

Word Differences Our results suggested that prediction ability was sensitive to both the sequential predictability of the word itself and its semantic relatedness to the context. However, the sensitivity was limited: it emerged only if there was a salient change.

Results of RA did not show the expected trend: The accuracy of word prediction was consistent across word groups - it was not affected by the predictability or semantic relatedness of words.

As expected, results of RT suggested that participants responded more slowly when the predictability or semantic relatedness of selected word slots decreased. The lower the word predictability or semantic relatedness, the slower the response. However, when there is a decrease in word predictability with an equal level of semantic relatedness, no difference was observed. Similarly, when there is a drop in semantic relatedness with an equal level of predictability, no difference was observed, either. But when both levels of predictability and semantic relatedness changed, a deteriorating effect on RT was observed.

Unexpectedly, results of RT per character were in the opposite direction of our hypotheses. The time spent for word prediction was consistent across word groups - it was not affected by the predictability or semantic relatedness of words. These findings of RT and RT per character had similar implications as those of RA: prediction rate may be slightly sensitive to word predictability, but the sensitivity emerged only if the relative change in predictability and semantic relatedness was salient enough.

CMC Differences As expected, participants had a lower RA in word prediction if they read pure text compared to watching videos or listening to audios. Their performance was not critically different with the loss of visual cues, but auditory cues played the central role in affecting their predictability level.

Regarding RT and RT per character, both V and A showed slower responses than T. RT increases as there are more multimodal cues available. This pattern indicated that participants may have experienced heavier cognitive load in processing A and V, which may have required more time for prediction. One possible theoretical explanation lies in reading ability. Text condition may have measured different cognitive abilities in audio and text processing. There are individual differences in reading speed, working memory, and short-term memory capability (Freed et al., 2017; Just and Carpenter, 1992) that may have contributed to the difference observed between T and other conditions. Another possible explanation is related to the study design, where texts were displayed with full sentences in short paragraphs so that readers were not forced to spend a certain amount of time to process the content as they would do in A and V. They may skim through the text and respond to finish the task sooner. A sliding window design to display the text or incorporating eye-tracking method would be valuable to further gauge the behavioral differences across conditions. A within-subject study with text-audio or text-video comparisons could also address this issue with a future study.

When comparing CMC types for both measures, no difference in visual cue loss was observed, but

Fixed effects	**Estimate**	***SE***	**df**	***t* value**	***p* value**
(Intercept)	1098.90	122.15	189.70	9.00	<.001***
CMC type: audio	22.73	54.42	171.90	0.42	0.68
CMC type: text	-627.14	56.02	171.02	-11.19	<.001***
Self-emotion	17.09	12.52	172.93	1.37	0.17
Negative self-emotion	-8.30	12.75	172.85	-0.65	0.52
Attention level	118.34	268.46	171.30	0.44	0.66

Table 4: Post-hoc model for Response Time Per Character (*Overall Score Group* removed by model selection)

only a significant difference incurred by auditory cue loss. It provides substantial evidence for the deterministic role of auditory cues in the cognitive process of prediction compared to visual non-verbal cues. This finding also validates the role of prediction in processing difficulties of CMC: speakers need to be scaffolded with richer multimodal cues to maintain prediction accuracy.

Emotion and Perception RT and RA provided complementary findings: self-emotion showed marginal facilitatory effect on RT, while negative self-emotion had deteriorating effect on RA.

For RA, figure 1 indicates that the larger change in negative self-emotion in A and T may be due to the unavailability of visual cues from the conversation. The similarity between A and T may show a floor effect: Without visual cues, the availability of auditory cues could not compensate for the increase in negative self-emotions.

As shown in figure 2, self-emotion increased RT for V and A. There are two possible reasonings: participants with more positive emotions would 1) invest more in thinking, or 2) invest more in typing more words. Since self-emotion did not significantly increase RT per character, it indicates that self-emotion leads to slower RT regardless of response length. Therefore, there may be a higher level of willingness to contribute more to the task incurred by more positive emotional states when auditory cues were present. However, from the opposite direction observed in T, where self-emotion decreased RT, it is difficult to gauge whether it was due to participants with more positive emotions 1) invested less in thinking or 2) invested less in typing fewer words. Meanwhile, negative self-emotion did not show deteriorating effect to lengthening RT or RT per character. Participants may not be disadvantaged in their willingness to invest more in response typing by negative self-emotions. Given the study design where participants needed to click on "continue" to proceed to response typing in T, RT measure may not be consistent across participants due to individual differences in the clicking habit: participants may click before, during, and after their thinking process. Again, incorporating a sliding window design to display the text or using eye-tracking method would help further validate the current findings of the prediction mechanism.

Attention level did not boost prediction accuracy or show any promoting effect on RT. The willingness or capability of information processing in the prediction task were not influenced by the attention level to the task. This finding did not align with the body of literature claiming the relationship between attention and language processing (e.g., Hinojosa et al., 2019; Verhees et al., 2015). Either the measure of attention in the present study may not have accurately captured online attentional processes, or there may be fundamental distinctions between online and offline attentional processing that underpin online prediction. The word prediction task employed in the study allowed participants to engage with the conversation as observers rather than as speakers. This distinction suggests that the prediction mechanism of an observer may not fully align with that of a speaker, given the methodological constraints of this behavioral study. While this study provides substantial support for the current findings, future research could further elucidate these effects through computational modeling and simulations or by measuring neural signals of speakers using EEG during real-time CMC.

Consistent observations suggested the intricate link among predictability, attention, emotion, and processing rate: those who had more positive self-emotions may have processed information more slowly and willingly, and those who had more negative self-emotions may be more vulnerable to cue loss and produce prediction with lower accuracy. Notably, online attention level in the task was not critical for the cognitive process.

Implications The findings underscore the critical roles of multimodal cues, especially auditory cues, and social factors such as positive and negative emotional states have their unique roles in influencing information processing strategies and predictive accuracy in CMC as in f2f communication. More attention is needed to evaluate the relative reliance on these multi-modal cues and the relative reliability of social cues given the difference of modality in different CMC. Given that predictability is largely contingent on individual's general cognitive ability and language experience, within-subject studies and comparison studies should be conducted to gauge how speakers use different strategies to cope with cue deficits. This study also provides valuable insights for constructing multi-modal models, highlighting the primacy of auditory cues and sociopsychological aspects in conveying critical information to sustain smooth comprehension and prediction in real-time conversation.

Limitations This study used GPT-2 in word selection during design phase and switched to a more updated model Llama 3.0 during data analysis. More up-to-date models will offer more accurate measures of predictability. A design flaw existed in RT measure, where entries of more than one word should be disabled on the platform, providing a more consistent measure of response behaviors. Other limitations of study design and materials were included in the main paper.

Acknowledgments

We thank our advisors Susan R. Fussell and Morten H. Christiansen, and anonymous reviewers for their valuable input and feedback. We are also grateful of all the participants who contributed to the study, and our research assistants Harim Hahn, Kate Hahnenberg, Jay Huang, and Bryan Kongnyu for their contribution to this project.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67.

Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44:123-147.

Susan E. Brennan and Maurice Williams. 1995. The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34:383–398

Ruth E. Corps, Abigail Crossley, Chiara Gambi, and Martin J. Pickering. 2018. Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition,* 175:77–95.

Wen Duan, Naomi Yamashita, Yoshinari Shirai, and Susan R. Fussell. 2021. Bridging Fluency Disparity between Native and Nonnative Speakers in Multilingual Multiparty Collaboration Using a Clarification Agent. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1-31.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783v2. Version 2.

Fernanda Ferreira and Karl G.D. Bailey. 2004. Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, 8:231-237.

Barbara L. Fredrickson and Christine Branigan. 2005. Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion*, 19:313-332.

Erin M. Freed, Stephen T. Hamilton, and Debra L. Long. 2017. Comprehension in proficient readers: The nature of individual variation. *Journal of Memory and Language*, 97:135-153

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software,* 33.

Marlen Fröhlich, Christine Sievers, Simon W. Townsend, Thibaud Gruber, and Carel P. van Schaik. 2019. Multimodal Communication and language origins: Integrating gestures and vocalizations. *Biological Reviews,* 94:1809–1829.

Karen Gasper and Gerald L. Clore. 2002. Attending to the big picture: mood and global vs. local processing of visual information. *Psychol Sci*, 13: 43-40.

Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2013. Using visual information for grounding and awareness in collaborative tasks. *Human-Computer Interaction,* 28:1–39.

Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. 2019. Nonverbal communication. *Annual review of psychology* 70:271-294.

Lauren V. Hadley and John F. Culling. 2022. Timing of head turns to upcoming talkers in triadic conversation: Evidence for prediction of turn ends and interruptions. *Front Psychol*, 13:1061582.

José A Hinojosa, Eva María Moreno Montes, and Pilar Ferré. 2019. Affective neurolinguistics: towards a framework for reconciling language and emotion. *Language, Cognition and Neuroscience*, 35:813 - 839.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7:411-420.

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

Yoram M. Kalman, Gilad Ravid, Daphne Raban, Sheizaf Rafaeli. 2006. Speak *now* or forever hold your peace: Power law chronemics of turn-taking and response in asynchronous CMC. *In Annual Conference of the International Communication Association, Dresden*.

Kai Kaspar and Peter König. 2012. Emotions and personality traits as high-level factors in visual

attention: a review. *Frontiers in Human Neuroscience*, 6:321.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. Christensen. 2017. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software,* 82.

Stephen C. Levinson. 2016. Turn-taking in human communication – origins and implications for Language Processing. *Trends in Cognitive Sciences*, 20:6–14.

Edith Chin Li, Sarah E. Williams, and Angela Della Volpe. 1995. The effects of topic and listener familiarity on discourse variables in procedural and narrative discourse tasks. *Journal of communication disorders*, 28:39-55.

Hajin Lim, Dan Cosley, and Susan R. Fussell. 2022. Understanding Cross-lingual Pragmatic Misunderstandings in Email Communication. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1-32.

Steven G. Luke and Kiel Christianson. 2017. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50:826–833.

Lilla Magyari and J.P. de Ruiter. 2012. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

R Core Team. 2017. R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria.*

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*.

Marlou Rasenberg, Asli Özyürek, Sara Bögels, and Mark Dingemanse. 2022. The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, 59:209-236.

Jan P. de Ruiter, Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science,* 4:232-248.

Rachel Ryskin and Mante S. Nieuwland. 2023. Prediction during language comprehension: What is next? *Trends in Cognitive Sciences,* 27:1032-1052.

Shayle R. Searle, Fred M. Speed, and George A. Milliken. 1980. Population marginal means in the linear model: An alternative to least squares means. *The American Statistician,* 34:216-221.

Elizabeth Ellen Shriberg. 1994. Preliminaries to a Theory of Speech Disfluencies.

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of memory and language*, 32:25–38.

Filip Sondej. 2022. Filyp/autocorrect: Spelling Corrector in python.

Catalina L. Toma. 2014. Towards conceptual convergence: An examination of interpersonal adaptation. *Communication Quarterly,* 62:155-178.

Michael Tomasello. 2008. *Origins of human communication*. MIT Press, Cambridge, United States.

Francisco Torreira, Sara Bögels, and Stephen C. Levinson. 2015. Breathing for answering: The time course of Response Planning in conversation. *Frontiers in Psychology,* 6:284.

James P. Trujillo, Stephen C. Levinson, and Judith Holler. 2021. Visual information in computer-mediated interaction matters: Investigating the association between the availability of gesture and turn transition timing in conversation. *Lecture Notes in Computer Science*, pages 643-657.

Martine W. F. T. Verhees, Dorothee J. Chwilla, Johanne Tromp, and Constance T. W. M. Vissers. 2015. Contributions of emotional state and attention to the processing of syntactic agreement errors: evidence from P600. *Front Psychol*, 6:388.

Joseph B. Walther. 2011. Theories of Computer-Mediated Communication and Interpersonal Relations. *The SAGE Handbook of Interpersonal Communication/SAGE Publications, Inc*.

A Study Materials and Data

All the study materials and data analyses were included in supplementary material archive files for data transparency and reproducibility purpose on Open Science Framework (https://osf.io/35z6a). Illustrations of prediction task can be found in figures 6 to 8. For audio condition, participants saw a blank page when the conversation audio played to minimize disruption.

B Additional Supporting Materials

One of the assumptions of LME models is normal distribution of the variable. In terms of response accuracy, its distribution was indeed not normal as

shown in figure 3: there were many perfect responses (i.e., participants guessed the exact same words) which yielded perfect scores in predictability. However, the residuals are normally distributed as shown in figures 4 and 5, so response accuracy still fits LME model's assumption.

Another concern may arise for the cut-off point of response time at 30000 ms. We did not use standard deviation (SD) to exclude the outliers of response time, given that SD was over 460000 ms due to unreasonably long response time for some trials that skewed the entire raw data. Therefore, we decided to set the cut-off point at 30000 ms since most of the data scattered below 10000 ms. We also reasoned from the nature of the word prediction task that it was not valuable to have responses longer than 30000 ms where the attention was not sustained sufficiently to the task.

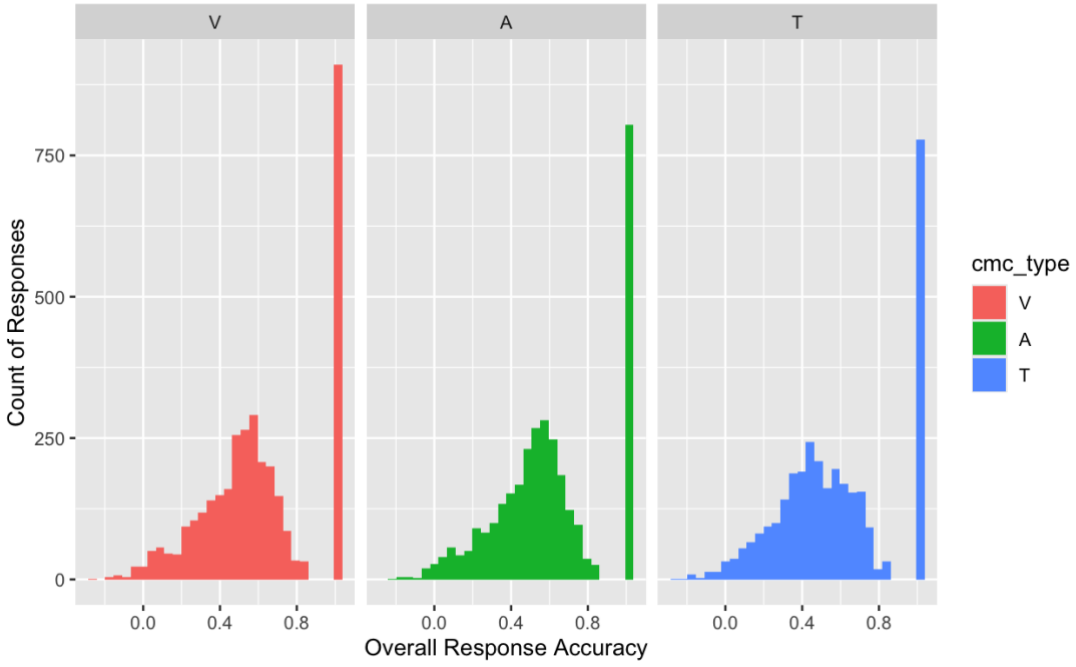

Figure 3: The overall distribution of response accuracy in each CMC type

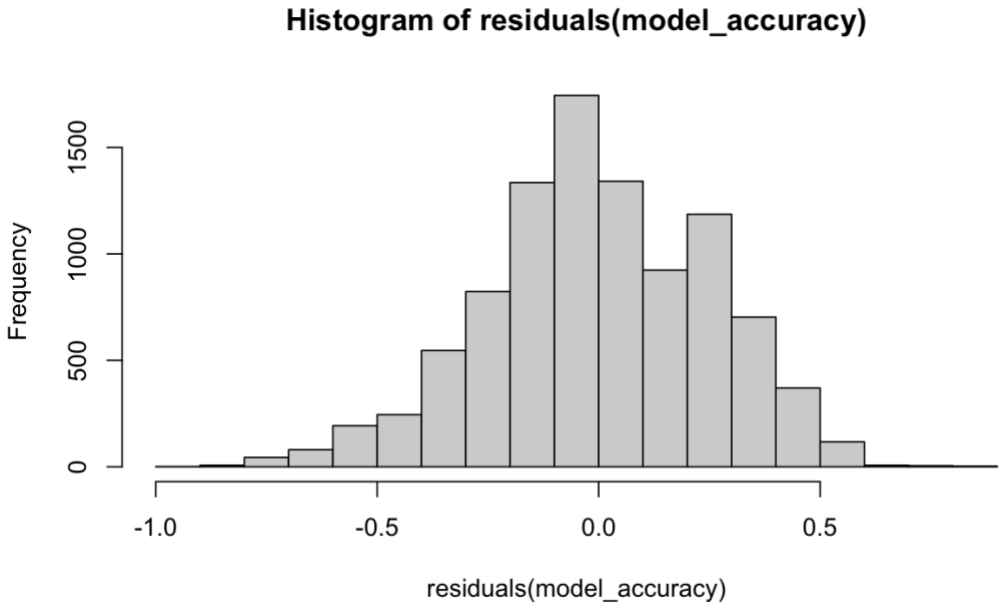

Figure 4: The distribution of residuals of response accuracy

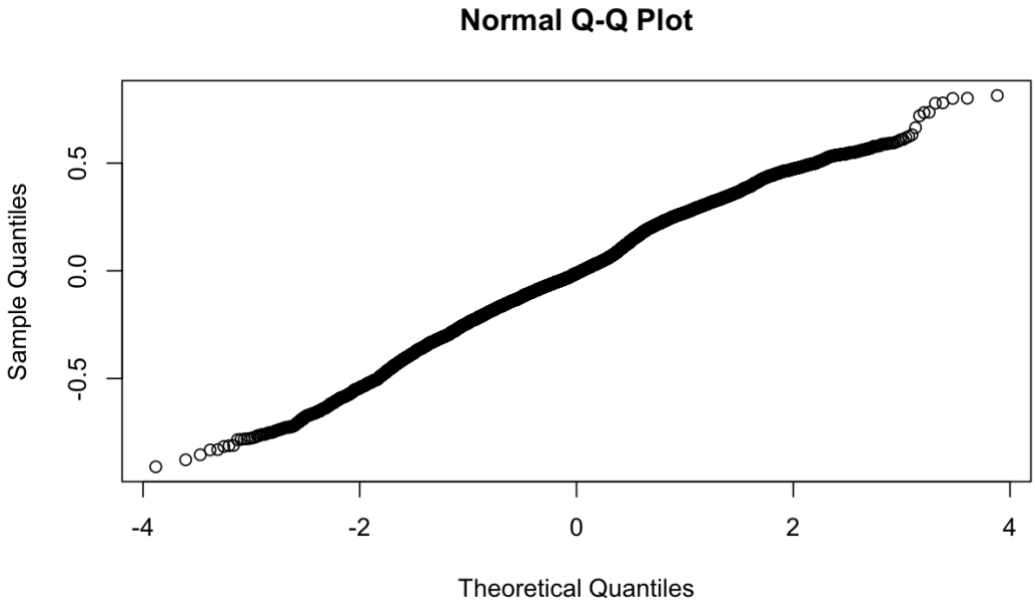

Figure 5: The results of normality test for residuals

Figure 8: Prediction task interface for Video condition

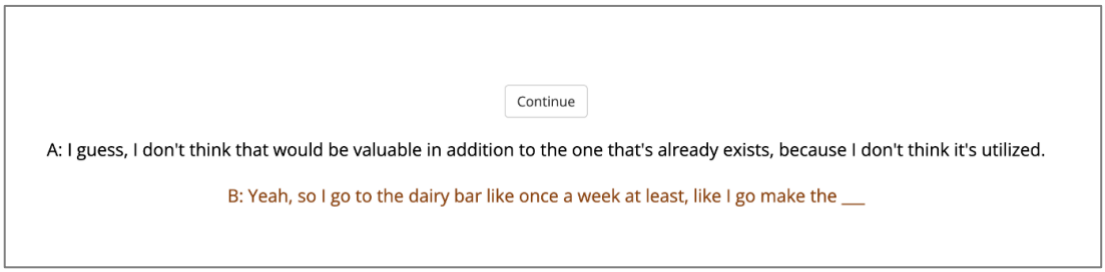

Figure 7: Prediction task interface for Text condition

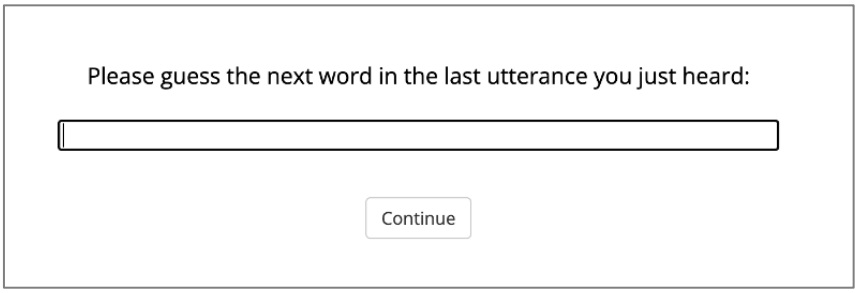

Figure 6: Prediction task interface for participants entering response