Speech Act Prediction Across Turn Boundaries in Conversation

A thesis submitted by

Lena Warnke

In partial fulfilment for the degree of

Master of Science

in

Psychology

Tufts University

May 2022

Advisor: J.P. de Ruiter

Abstract

In conversation, every utterance performs a *speech act*, such as an invitation, a complaint or a greeting. Recognizing an utterance's speech act is essential for the listener in planning a relevant response yet presents a non-trivial challenge: there is no one-to-one correspondence between a turn's form and its speech act, and speech acts must be recognized very quickly. Despite these challenges, we typically identify the speech acts of utterances efficiently. One explanation is that interlocutors *predict* the speech act of the upcoming turn. In this study, we explicitly investigate whether listeners form expectations about upcoming speech acts based on the preceding turn in conversation. The results show that listeners draw on prior context to anticipate upcoming speech acts in conversation, leading to the same utterance being processed differently depending on the action it performs in conversation. This sheds light onto the fact that we have a cognitive architecture oriented to speech acts.

Table of Contents

Speech Act Prediction Across Turn Boundaries in Conversation

Conversation is a ubiquitous form of human communication in which social interactions are conducted through language. Successful conversation not only requires comprehension at lower levels of language representation, such syntactic and semantic levels, but also at a higher, pragmatic level of language representation. In everyday conversations, interlocutors use language to perform actions, such as requests, invitations, or complaints. The underlying social function of an utterance is known as its *speech act* (Austin, 1975; Searle, 1969), and the ability to extract the speech act from an utterance is a fundamental skill for successful conversation. In fact, recognizing speech acts is an essential condition of conversation because it is only once the listeners has extracted the utterance's underlying function that she can begin planning a relevant response.

Speech act recognition is fundamental to successful conversation, yet presents a non-trivial challenge for listeners. First, there is no one-to-one mapping between an utterance's linguistic form and the action it performs – what is meant cannot necessarily be extracted from what is actually said. For example, the utterance "*Are you kidding?*" is formally interrogative, yet functionally does not request information (de Ruiter, 2012). Moreover, the same speech act can be expressed with many different utterances, and the same utterance can perform multiple actions depending on the context. The statement "*I'm cold*" could be the answer to a question, a complaint, or a request to turn up the heat. Given that the form and semantic content of an utterance is underspecified for the action that it performs, listeners must rely on the context to extract speech acts in conversation.

The rapid timing of natural conversation further augments the challenge of speech act recognition. Across languages, the modal gap between turns in a conversation is 200 ms, with gaps of 0 ms or even negative gaps (overlapping speech) occurring frequently (de Ruiter et al., 2006; Stivers et al., 2009). Within these tight time constraints, a listener must process the incoming speech, extract the underlying speech act, and then plan an appropriate response. Research shows that it takes at least around 600 ms to prepare a one-word utterance (Indefrey & Levelt, 2004), implying that interlocutors in a conversation must recognize the speech act and begin preparing their response in the middle of an incoming utterance.

Despite the many-to-many mapping between an utterance's form and its function, the fast turn transitions in conversation, and the noise in our communicative environments, humans typically extract speech acts from utterances very efficiently. Currently, very little is known about the cognitive mechanisms that underlie the critical and challenging task of speech act recognition in conversation.

Literature in Conversation Analysis, a sub-discipline of sociology dedicated to studying natural social interaction, indicates that conversation is organized systematically. One essential component of conversational organization is the *adjacency pair*, consisting of paired action sequences (Schegloff & Sacks, 1973). Adjacency pairs function such that the speech act of a given turn in a conversation places strong constraints on the possible speech acts of the following turn (Schegloff, 2007; Schegloff & Sacks, 1973). An invitation, for example, is typically followed by an acceptance or a declination, and a question is typically followed by an answer. Given that action sequences in conversation follow a structure, and that actions in conversation have implications for how we should respond, the preceding context of a conversation provides a basis for interpreting upcoming utterances. In the face of ambiguity, listeners likely draw on their knowledge of this structure in order to recognize the speech act of an incoming turn.

To date, only a handful of studies have experimentally investigated speech act recognition. Using written dialogue, Gisladottir et al. (2012) showed that the same sentence is categorized as different speech acts depending on the preceding context, even when this context is limited. For the same sentence, behavioral reading times differed depending on the action that it performs, diverging already at the first word. Participants mobilized their top-down knowledge of the preceding context when interpreting the speech act of the current turn, affecting their interpretation of the input very early in the turn. Further work by the same authors using naturalistic spoken dialogues found event-related potential (ERP) evidence that speech acts are recognized early in the turn, before the utterance has completely unfolded (Gisladottir et al., 2015). The same study also showed that the time course of speech act recognition depends on the utterance's relationship to the preceding context. When the context was highly constraining, the final word of the utterance required little processing as evidenced by the ERP signal, but when the context was less constraining, processing at the final word of the utterance was increased. The latter finding highlights the role of context in speech act recognition, and provides preliminary evidence that listeners may employ *predictive* mechanisms to recognize speech acts early in the utterance.

Indeed, prediction provides a powerful explanation for the efficiency with which humans understand speech acts in conversation given the cognitive complexity of the task. If the speech act is recognized early in the turn through predictive mechanisms, production planning can begin before the end of the incoming utterance, allowing interlocutors to respond within the 200 ms window that is characteristic of conversation.

Converging neural and behavioral evidence from sentence processing research suggests that comprehenders are able to probabilistically predict upcoming linguistic input. Eye-tracking studies consistently report that readers fixate less on predictable compared to unpredictable words (Ehrlich & Rayner, 1981; Rayner, Slattery, & Drieghe, 2011; see Staub, 2015 for a review). Similarly, evidence suggests that predictable words are processed more quickly (Schwanenflugel & Shoben, 1985) and more accurately (Jordan & Thomas, 2002) than unpredictable words. Importantly, a series of eye-tracking studies using the visual-world paradigm showed that when the sentential context constrained for a particular word (i.e. the word was predictable), participants move their eyes towards images of objects related to that word *before* the onset of the word (Altmann & Kamide, 1999; Kamide et al., 2003; Tanenhaus et al., 1995), reflecting anticipatory cognitive processes in advance of any bottom-up linguistic input.

ERPs provide some of the strongest evidence that the brain engages in predictive cognitive processes during language comprehension. The N400 ERP component, a negative-going waveform that peaks around 400 ms after a word's onset, varies inversely in amplitude with the semantic predictability of the incoming word. Predictable words show a smaller N400 than unpredictable words (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984). More direct evidence for predictive preactivation of upcoming linguistic input come from ERP studies showing differential brain activity before the onset of a predictable versus an unpredictable word. In these studies, function words that differ depending on the next word elicited an N400 ERP component when incongruent with the predicted next word (DeLong et al., 2005; Van Berkum et al., 2005). For example, the sentence context "*the day was breezy so the boy went out to fly* ***a*** ____" is highly constraining for the next word to be *kite*. When participants instead saw "*the day was breezy so the boy went out to fly* ***an*** ____", larger N400s were observed to the word "*an*" compared to "*a*" because it is incongruent with "*kite*", the predictable next word (DeLong et al., 2005). This suggests that prediction during language comprehension operates not only at the semantic level, but also at the orthographic and phonological level.

Comprehenders are able to predict upcoming linguistic input at many other levels of representation in language. Top-down contextual information can facilitate the bottom-up processing of information at the event-structure level (Altmann & Kamide, 1999; Xiang & Kuperberg, 2015), the semantic level (Federmeier & Kutas, 1999), the syntactic level (Kamide et al., 2003; Strijkers et al., 2019), and the orthographic level (Laszlo & Federmeier, 2009). ERP evidence suggests that the brain distinguishes between prediction error at these different levels of representation. A strong lexical prediction violation evokes an N400 response in addition to a late anterior positivity (Federmeier et al., 2007), while a syntactic-semantic prediction violation elicits an N400 and a late posterior positivity (P600) (Kuperberg, 2007). A recent study by Kuperberg et al. (2020) replicated these findings in a within-participants design, providing further support that there are spatially and temporally dissociable neural networks involved in prediction at different levels of representation in language.

Taken together, the psycholinguistic literature provides convincing support for comprehenders' ability to anticipate upcoming linguistic input. The evidence, however, comes almost exclusively from studies of sentence processing, which differs from conversation comprehension along a number of dimensions. Most importantly, conversation is characterized by rapid communicative interactions across multiple speakers and turns, requiring pragmatic inferences that are not necessary for isolated sentence comprehension. Furthermore, language in conversation unfolds rapidly in the auditory domain. Most studies investigating language prediction, however, have employed a relatively slow visual presentation of words. Participants in a conversation have to do *more* processing in *less* time, so it is not clear how the findings of predictive language comprehension from the sentence processing literature translate to language comprehension in the context of interactive conversation.

A limited set of studies have explicitly investigated the role of predictive cognition in conversation. De Ruiter et al. (2006) presented participants with turns taken from natural conversations and asked them to press a button at the moment they thought the turn would end. The average response was approximately 200 ms *before* the end of the turn, indicating that listeners anticipated turn-ends. Two eye-gaze studies also found that children as young as 1 and 2 years old are able to anticipate when a turn-switch will occur (Casillas & Frank, 2013; Keitel et al., 2013). De Ruiter et al. (2006) further report that lexico-syntactic content rather than intonational contour is a necessary (and perhaps even sufficient) cue for turn-end prediction. Specifically, while both semantic and syntactic cues are needed to accurately anticipate the end of a turn, the semantic content may be a more important anticipation cue (Riest et al., 2015). Lastly, evidence shows that when listeners are more accurate at estimating the end of a turn they are also more accurate at guessing how a turn ends as measured using a gating task (Magyari & de Ruiter, 2012). Taken together, these findings suggest that interlocutors in a conversation predict the upcoming linguistic content of a turn, and use this prediction to estimate the duration of the turn to plan their response accordingly within the time constraints of the turn-taking system.

While turn-taking research provides substantial evidence that prediction underlies the rapid timing of natural conversation, these studies investigate the cognitive processes of language comprehension in single-speaker single-turn environments. An important question that remains to be addressed is whether humans recruit anticipatory mechanisms to predict across turn boundaries and speaker switches in natural conversation. We know that sequential conversational context influences comprehension of an incoming turn (Gisladottir et al., 2012; Gisladottir et al., 2015), but whether an incoming turn is predicted, and at what level of linguistic representation, remains unclear.

To date, only a handful of studies have investigated prediction across turn boundaries in dialogue comprehension. Goregliad Fjaellingsdal, Ruigendijk, Scherbaum, & Bleichner (2016) showed for the first time that the N400 ERP component can index semantic expectations across a speaker-switch. The generalizability of this result, however, is not clear, as the stimuli consisted of spoken sentences with a speaker-switch occurring prior to the last word – one speaker was essentially finishing the other speaker's turn. While this does occur in natural conversations, it does not constitute a true turn boundary. More convincing evidence of prediction across turn boundaries comes from another ERP study by Bögels, Kendrick, & Levinson (2015). Here, listeners were able to draw on their knowledge of the timing of preferred versus dispreferred answers to initiating actions (e.g. invitations) to update their semantic predictions about a speaker's answer to a question. Importantly, natural turns from a corpus of telephone recordings were used as the stimuli for this study.

These two studies provide evidence for prediction across turn boundaries at the semantic level. While people in a conversation must process an incoming turn at this level of representation of language, a critical task of the listener is to comprehend the pragmatic level of the utterance – its speech act. It is only once the speech act of the turn has been recognized that a listener can begin to plan an appropriate response to their partner's utterance. As previously discussed, Gisladottir et al. (2015) found neural activity indicating that speech acts are recognized early in a turn, possibly through predictive mechanisms. In a follow-up study, the authors explicitly investigated this claim, asking whether listeners draw on preceding conversational context to predict the speech act of the next turn. The experiment showed that oscillatory EEG activity differs depending on the speech act, with more predictable speech acts (declinations) eliciting reduced power in alpha/beta bands, relative to unpredictable speech acts (answers and pre-offers) (Gisladottir et al., 2018). Based on the previously observed role of alpha and beta desynchronization in anticipatory processing, the authors conclude that prediction plays a role in speech act recognition. While this study provides preliminary evidence for prediction at the speech act level, there a number of issues. Firstly, the authors assume that the utterances in their stimuli constrain for specific speech acts in the following turn, such that some of their speech act stimuli are more predictable than others. However, they do not provide any concrete evidence for this assumption, such as cloze or plausibility norming data. Without empirical evidence that certain speech acts are in fact more predictable than others given the context, it is difficult to make the claim that any observed difference in brain activity can be attributed to predictive processing. Secondly, the authors' finding relies on the idea that anticipatory oscillatory neural activity across domains of cognition is identical, which is not necessarily the case. At the moment, very little is understood about the role of neural oscillations in the language domain, and it is unclear whether a particular frequency band observed in one paradigm reflects the same cognitive mechanisms as one observed in another (Hauk et al., 2017). Further investigation is necessary in order to confirm the role of predictive cognition in speech act recognition.

Taken together, the sentence processing and turn-taking literature suggests that listeners engage predictive mechanisms during language comprehension. Prediction provides a powerful explanation for the speed and accuracy with which humans understand language in the face of noise, ambiguity, and the speed of the unfolding input. This is especially pertinent to conversations, in which listeners reliably extract the underlying speech act from an underspecified utterance and plan a relevant response within a remarkably tight time frame. Currently, the cognitive mechanisms underlying this capability are poorly understood. Though prior studies have shown that listeners engage predictive mechanisms to pre-activate upcoming linguistic input across turn-boundaries and speaker switches, it is unclear what type of information comprehenders are able to predict across turn-boundaries. Building on previous work, the current study investigates the role of anticipatory processing in facilitating speech act recognition. Given the considerable body of behavioral and neural evidence for prediction at multiple levels of representation of language, and the systematic organization of conversation, we hypothesize that listeners draw on prior context to probabilistically predict the speech act of the next utterance in a conversation.

The current study

In the current study, we consider the fact that turn duration in conversation is highly variable, and unknown to the listener in advance (de Ruiter, 2019). A turn construction unit (TCU) is the smallest interactionally complete linguistic unit that can make up a turn in conversation; turns can either consist of one or several TCUs (Sacks et al., 1974). Thus, at the end of one TCU in conversation, the same speaker could either utter a second TCU, or another speaker could take the next TCU. Given that speaker switches are not predictable, and that the interpretation of an utterance depends on who says it (Van Berkum et al., 2008), we hypothesize that listeners anticipate speaker-specific speech acts across turn boundaries. That is, at the end of one TCU in a conversation, listeners probabilistically pre-activate one set of plausible speech acts if the next TCU is spoken by the same speaker, and another set of plausible speech acts if the next TCU is spoken by a different speaker. See Figure 1 for an illustration of the listener's presumed communication model.

Study Design

To examine the nature of speech act anticipation across turn boundaries, we created a set of naturalistic-sounding conversational scenarios consisting of two turn TCUs. The first TCU in our stimuli serves as the context utterance, spoken by one speaker, and the second TCU serves as the target utterance, either spoken by the same speaker or a different speaker. Target utterances fall into one of three conditions. In *congruent* trials, the speech act of the critical turn matches the previous speech act, thereby forming a congruent pair of TCUs. In *speech act violation* trials, the speech act does not match the previous one, thereby forming an incongruent pair of TCUs. To control for lexico-semantic content, for the *speech act violation* condition we took the TCUs in the *congruent* condition and switched the speaker identity (same or different) such that the two speech acts did not match anymore, but would match if spoken by the other speaker identity pairing. In *speaker-independent violation* trials, the speech act and lexico-semantic content of the utterance do not match in either of the speaker identity assignments (see Table 1 for examples).

The strength of this design is twofold. Firstly, we are able to manipulate congruency of the speech act while fully controlling for lexico-semantic content, allowing us to measure comprehension purely at the speech act level. Secondly, unlike previous studies of speech act comprehension, we include stimuli that violate the constraints of the context, such that we can detect effects of prediction violation.

Methods

Development of stimuli

The stimulus materials for this study are auditory recordings of conversational scenarios consisting of two TCUs. Participants listened to six types of scenarios (see Table 1), in which the second TCU served as the target utterance. To construct these, we generated 125 unique utterances that are not constraining for a particular speaker or utterance in the next TCU. For each of these, we then constructed four version of a second TCU that corresponded to each of the four experimental conditions as follows. First, we constructed a second TCU such that the first TCU followed by the second TCU created a naturalistic, congruent conversational scenario when the second TCU was spoken by the same speaker as the first TCU (*same speaker, congruent*), and an incongruent conversational scenario when spoken by a different speaker as the first TCU (*different speaker, speech act violation*). Similarly, for each first TCU, we also constructed a second TCU such that the first TCU followed by that second TCU created a congruent conversational scenario when the second TCU was spoken by a different speaker as the first TCU (*different speaker, congruent*), and an incongruent conversational scenario when spoken by the same speaker (*same speaker, speech act violation*). To create the *speaker-independent violation* control conditions, each *same speaker, speech act violation* and *different speaker, speech act violation* target utterance (second TCU) was paired with a first TCU such that the resulting conversational scenario was incongruent regardless of whether the same speaker or a different speaker spoke the target utterance. This resulted in a total of 750 stimuli, with 125 items per condition. All target utterances were either one or two syllables long.

To ensure that participants would encounter more congruent than incongruent utterances, in order to induce naturalistic comprehension processes, we included a set of 120 congruent fillers. Identical to the critical stimuli, filler stimuli consisted of two TCUs, half of which included a speaker switch with the second TCU. All second TCUs were also maximally two syllables long. All filler stimuli were taken from recordings of natural conversations from the Tufts University Human Interaction Lab's InConversation Corpus, and were then transcribed and re-recorded as described below.

Stimuli were recorded in two soundproof rooms separated by a glass pane at 44 kHz sampling rate using Shure MX153T/O-TQG Omnidirectional Earset Headworn Microphones. Six native speakers of American English (three women, three men) were recorded in male-female pairs to maximally distinguish their voices. Speakers were seated across from each other in the separate rooms, and could see each other and hear each other through Sennheiser PX200 headphones. The native speakers were instructed to act out the written conversational scenarios as naturally and informally as possible, without reading. For the *speaker switch* scenarios, the partners of each pair switched off acting out the first and the second TCU. All TCUs were recorded in their congruent form and spliced together after recording to create incongruent scenarios as described below.

After recording, the stimuli were processed using the software Praat (Boersma & Weenink, 2019), version 6.0.48. Each stimulus was first cut into two TCUs, and then spliced together again with a pause of 300 ms, a naturalistic gap between turns in natural conversations (de Ruiter et al., 2006; Stivers et al., 2009). The overall sound intensity of the recordings was normalized to 65 dB to prevent loudness differences between TCUs within items, and between items.

Plausibility norming

In order to characterize our stimuli and to confirm that the *speech act violation* as well as the *speaker-independent violation* conversational scenarios are perceived as more implausible than the *congruent* conversational scenarios, we collected plausibility ratings for all stimuli. We divided our stimuli into 8 balanced and randomized lists, containing an equal proportion of items from each condition, and collected ratings from 160 participants (20 per list). Participants were recruited from Amazon's Mechanical Turk (AMT), and screened on the basis of the following criteria: first language learned was English, no self-reported psychiatric or neurological disorders, no use of psychoactive medication at the time of the survey, and aged between 18 and 35. Informed consent was obtained from all participants, and they were compensated for their time. Protocols were approved by the Tufts University Social, Behavioral, and Educational Research Institutional Review Board.

Participants were asked to listen to each stimulus, and rate how plausible it is that they would hear it in a conversation on a scale of 1 to 6 (1 for highly implausible and 6 for highly plausible). Each stimulus could be played no more than twice by participants. Prior to carrying out the norming task, participants completed a guided practice with examples to familiarize themselves with the task. In addition, "catch" questions were included to identify and omit bots, and participants were manually excluded if their responses indicated a failure to comply with instructions.

The plausibility rating data was analyzed by fitting a series of linear mixed effect regression models (Baayen et al., 2008) using R version 4.11 (R Core Team, 2021). We chose to perform a Bayesian analysis rather than a traditional frequentist analysis because this approach provides information about the strength of the evidence in favor of either the alternative or the null hypothesis. Frequentist significance tests, on the other hand, compute the probability of the data or more extreme data under a null hypothesis, and are therefore designed to reject the null hypothesis. In the current Bayesian analysis, we report Bayes factors, representing the relative probability of the observed data under model *a* over model *b,* to determine the model under which the data are the most likely. In this paper, we interpreted Bayes factor values using evidence categories from Wetzels (2011), adapted from Jeffreys (1961).

For the Bayesian linear mixed effects models, we used glmer from the rstan (Stan Development Team, 2020) and rstanarm (Goodrich et al., 2020) packages, and then used the bridgesampling package (Gronau et al., 2020) to obtain Bayes factors for model comparisons. The rstanarm package estimates multilevel models using full Bayesian Inference via Markov Chain Motel Carlo (MCMC) estimation. In the current analysis, we used the default (weakly informative) priors from rstan. We added fixed and random effects incrementally to a minimal model and tested if the inclusion of the additional term was justified by comparing the likelihood of the data under both models.

Descriptive statistics for plausibility ratings are shown in Table 2. The mean plausibility rating for *congruent* trials was 5.28 for different speakers and 5.09 for same speaker. For the *speech act violation condition*, the mean plausibility rating was 3.35 for different speaker trials and 4.31 for same speaker trials. Lastly, in the *speaker-independent violation* condition, different speaker trials had a mean of 2.07 and same speaker trials had a mean of 2.14.

The model under which the data were most likely was the one that contained condition, speaker switch, and the interaction between condition and speaker switch as switch as fixed factors, and random intercepts for both participants and items. The data were more than 100 times more likely under this model than under a null model, as well as a model that included speaker switch and condition without the interaction.

As expected, both the *speech act violation* and *speaker-independent violation* stimuli were rated as less plausible than the *congruent* stimuli. Interestingly, the *speech act violation* condition was rated as more plausible than the *speaker-independent violation* condition, and in the *speech act violation* condition, stimuli spoken by the same speaker were rated as more plausible than stimuli spoken by different speakers. One explanation for this finding is that participants listening to the stimuli spoken by the same speaker in the *speech act violation* condition are actually hearing different speakers rather than the same speaker, leading to a plausible pair of TCUs (Warnke & De Ruiter, in prep.). This phenomenon is discussed further in the discussion. We compensated for potential effects of a speaker illusion effect by pictorially showing participants if they would hear one speaker or different speakers, as described below.

Experimental lists

The stimuli were initially divided into 6 lists such that the first TCU of the stimuli occurred once in every list, and the second TCUs, the target utterances, was fully counterbalanced across the lists and across the three levels of congruency. After pilot testing, we then further divided each list into two lists to ensure that the length of the experiment was manageable for participants, resulting in a total of 12 lists. Each list contained an approximately equal number of stimuli from each condition (with some lists containing one additional stimulus in one of the conditions), and an equal number of same speaker and different speaker scenarios. No stimuli were repeated within lists. In addition, the six speakers appeared as equally as possible in each condition within each list. The order of the stimuli in each list was randomized.

Experimental Procedures

The current experiment employed the same behavioral paradigm as de Ruiter et al. (2006). The experiment was implemented online with PsychPy 2020.2.10 (Peirce et al., 2019) and Pavlovia, which provide reasonable timing precision for presenting audio and visual stimuli and receiving responses (Bridges et al., 2020). Participants could only participate on desktop or laptop devices, and were asked to wear headphones during the study. They were instructed to listen to conversations consisting of two turns, and to press a button on the keyboard at the precise moment that they think the speaker of the second turn would be finished speaking. The instructions ask participants to *anticipate* this moment, and not to wait until the speaker finished speaking. At the beginning of each session, participants performed a short practice session consisting of six trials, followed by three experimental blocks. Each trial began with a visual countdown from 3 to 1 presented on the screen, followed by an image of an audio signal and an image indicating whether they would hear one or two speakers. The acoustic presentation of the stimulus began 1 second after the countdown ended. The structure of each trial is shown in Figure 2. As soon as the participant pressed the button, the sound file cut out so that participants didn't receive feedback on their accuracy, as this may have preferenced them to become more conservative in their turn-end estimations. The duration between the end of a turn and the button-press (called *bias*) was recorded by the computer. The next trial began as soon as the button was pressed. If a participant didn't press the button within 4000 ms of the stimulus offset, a time-out was recorded and the next trial began automatically. In order to ensure that participants were paying attention throughout the experiment, attention check trials occurred at random intervals throughout the experiment. The visual presentation of these trials was identical to the experimental trials, and the audio instructed the participants to press a specific number on the keyboard. The whole experiment lasted approximately 25 minutes.

The strength of our paradigm is two-fold: first, it does not disclose the experimental manipulation to participants and second, the task we are asking participants to perform – estimating the end of a turn – is one that listeners must engage in in everyday natural conversation. This provides a controlled yet ecologically valid language comprehension environment for participants.

Participants

We report data from 120 participants (age = 18-35 years, mean = 24.6, SD = 5.0, 70 women) who were recruited from the Tufts University community as well as via Prolific (Palan & Schitter, 2018). Originally, 137 participants were recruited, but 10 failed to complete the session, 5 failed one or more attention checks, and the data from two participants was removed because they consistently pressed the button before the end of the first turn. Participants were screened on the basis of the following exclusion criteria: significant exposure to any language other than English before the age of 5, history of psychiatric or neurological diagnoses, and use of psychoactive medication within the preceding 6 months. Participants had corrected or corrected-to-normal vision. All participants provided written informed consent and were compensated for their time. Protocols were approved by the Tufts University Social, Behavioral, and Educational Research Institutional Review Board.

Predictions

The dependent variable for these analyses is the duration between the end of the second TCU and the button-press (*bias*). If listeners are generating speech act predictions across turn boundaries, we predict that their *bias* will be smaller in the *congruent* compared to both the *speech act violation* and the *speaker-independent violation* conditions since both violate speech act congruency. If listeners are only sensitive to the lexico-semantic content of the utterance but not its function in context, then we predict that the *bias* will be smaller in both the *congruent* and the *speech act violation* conditions compared to the *speaker-independent violation* condition.

Results

A total of 7500 trials were recorded. After screening the data for deviations, we removed any trials in which the timer expired (i.e. no button press was recorded), and any trials in which participants responded before the onset of the second turn. As a result, 340 trials (4.53% of the data) were excluded from analyses.

Analysis strategy

The *bias* data was analyzed using the same procedures as the plausibility rating data analysis. The *bias* data is visualized in Figures 3 and 4. The descriptive statistics are presented in Table 3. The mean *bias* for *congruent* trials was 306 ms for different speakers and 308 ms for same speaker. For the *speech act violation condition*, the mean *bias* was 341 ms for different speaker trials and 323 ms for same speaker trials. Lastly, in the *speaker-independent violation* condition, different speaker trials had a mean of 372 ms and same speaker trials had a mean of 363 ms.

All models included the duration of the second TCU as a covariate. Though the content of the second TCU was counterbalanced across all levels of congruency, the duration varied slightly because the stimuli were recorded by different speakers. Previous research using this paradigm found effects of turn duration on turn-end estimation (de Ruiter et al., 2006), so including duration as a covariate allowed us to isolate any effects of our experimental manipulation. In our models, the duration of the second TCU was mean centered, and the speaker switch factor was contrast coded.

The model under which the data were most likely was the one that contained condition as a fixed factor, and random intercepts for both participants and items. The Bayes factor for the data under this model was 2770, providing decisive evidence for this model over the null model (intercept only). The data were about 24 times more likely under this model than under a model that also included speaker switch as a factor, and about 2212 times more likely under this model than under a model that included speaker switch as well as the interaction between speaker switch and condition. This provides very strong evidence that that whether the critical turn was spoken by the same or a different speaker was not a factor that affected *bias*.

The summary for the final model for the data is shown in Table 4 (with *congruent* as the reference level) and Table 5 (with *speaker-independent violation* as the reference level). Looking at the effects in the model, it can be seen that expected *bias* values for the *speech act violation* condition are 18.8 ms larger than in the *congruent* condition, with a model estimated probability of 0.98. The expected *bias* values for the *speaker-independent condition* are 57 ms larger than in the *congruent* condition, with a model estimated probability of 1. The model expected *bias* was also 38.1 ms smaller in the *speech act violation* condition compared to the *speaker-independent violation* condition, with a model estimated probability of 1.

Discussion

The aim of the current study was to investigate whether listeners draw on preceding context to anticipate speech acts in conversation. We used a set of well-controlled stimuli and an established, ecologically valid paradigm to examine turn-end estimation to turns that confirmed and violated speech act expectations. Our data show that listeners were more accurate at estimating the end of a turn in conversation when the speech act was congruent with the preceding turn, even when we controlled for lexico-semantic content. The exact same utterance was processed differently depending on its speech act. We also found an unexpected three-way distinction in turn-end estimation times between turns that are congruent with the context, turns that violate speech act constraints of the context, and turns that violate both the speech act and lexico-semantic constraints of the context. Taken together, these findings suggest that comprehenders draw on both their pragmatic speech act expectations as well as their lexico-semantic expectations while comprehending an unfolding utterance.

Communication is a crucial aspect of social behavior and involves performing actions. The ability to extract the speech act from an utterance is a fundamental skill for successful conversation: it is only once the action of an utterance has been recognized that the listener can begin to plan a response. Natural conversation is a rapid affair in which interlocutors not only have to extract speech acts before the end of the turn, but also must predict when the turn will end in order to produce a relevant and temporally appropriate response. The task from the current study thus approximates the task that we perform every day in our interactions with others. We know that context shapes our understanding of others' intentions: the same utterance can perform many speech acts, and the same speech act can be performed by many utterances. Here we show evidence that listeners draw on context to interpret the underlying action of an utterance, and that this affects their estimation of turn ends. The discourse context helps the listener comprehend the action of an utterance, reducing the time needed to respond. This sheds light onto the fact that we have a cognitive architecture that is oriented to speech acts.

The fact that we found an effect of speech act expectation on the timing of turn end estimation is particularly remarkable given that 1) the context participants heard was only one TCU and 2) our target utterances were very short, consisting of only one or two words. Even in such a limited context environment with short turns, listeners integrated the action of the unfolding utterance with their expectations online. When the speech act matched the context, the turn-end estimation was more accurate than when the speech act violated the context, requiring additional processing time as reflected by larger *bias* values. It is important to note that even though the size of the effect we found is substantial, the effect we found is relatively small, with *bias* values differing, on average, less than 74 ms between conditions. The next step to deepening our understanding of speech act anticipation across turn-boundaries is to investigate this ability with turns taken from natural conversation with richer contexts and turns of varying content and length that better represent the richness of natural conversation.

We found evidence that speech act level expectations are not the only cue that participants attended to when estimating the ends of turns. Our data suggests that mismatching lexico-semantic context leads to less accurate turn end estimation compared to just mismatching action content. There are a number of possible explanations for this finding. First, the second TCU in the *speech act violation* condition would have been congruent had they been spoken by a different speaker. As comprehenders, we use all contextual cues in the environment that are available to us to make sense of an utterance. In a separate experiment using the same stimuli as in the current study, we have found evidence that in the *speech act violation* condition spoken by the same speaker, listeners often hear two different speakers (Warnke & De Ruiter, in prep.), resulting in a congruent pair of TCUs. Thus, it is possible that listeners were actually inferring whether there was a speaker switch based on the lexico-semantic content of the utterances. This could explain the smaller *bias* values in the *speech act violation* condition compared to the *speaker-independent violation* condition.

A second possible explanation for difference in *bias* values between the *speech act violation* condition and the *speaker-independent violation* condition is that predictions at the speech act level interact with lexico-semantic information at lower levels. Generative models of language comprehension posit that listeners generate top-down predictions (Kuperberg & Jaeger, 2016). Hypotheses at higher levels of representation generate predictions at lower levels of representation prior to the bottom-up input arriving. Within this framework, speech act level predictions would lead to probabilistic predictions of possible upcoming words that make sense for the speech act. In the current study, the content of the utterance at these lower levels is identical in the *congruent* and the *speech act violation* conditions. It is thus possible that the content has been pre-activated in the *speech act violation* condition by predictions at the speech act level. This may explain the difference in *bias* between the *speech act violation* condition and the *speaker-independent violation* condition, the content of which would not be preactivated by speech act level predictions.

It is important to note that the *bias* values in the current study are mostly positive, indicating that participants, for the most part, estimated the end of the turn after its actual end. This was not the case in other studies that used the same paradigm (de Ruiter et al., 2006; Riest et al., 2015), where *bias* was mostly negative. This obscures the anticipatory nature of speech act comprehension, as the positive *bias* values could reflect the process of integration rather than prediction. However, if participants are faster at processing a speech act as indicated by a shorter *bias* (even if it is positive), the cognitive system must have been in a state to more readily integrate that speech act. While the debate surrounding prediction versus integration in language is still very much active, it has recently been argued that they are "two sides of the same coin" (Ferreira & Chantavarin, 2018): integration builds representations of already processed input, which in turn prepares the listener's system to receive new information (Ferreira & Chantavarin, 2018). Further research employing experimental designs and methods that explicitly tap into predictive cognitive mechanisms, such as EEG/ERP and eye tracking, is needed to investigate predictive pre-activation of speech acts.

In sum, we have presented evidence showing that listeners draw on context to anticipate speech acts, and this anticipatory processing affects turn-end estimation. The results of this study provide evidence that humans efficiently extract speech acts in the face of the many-to-many mapping between function and form by anticipating upcoming language at the highest level of language: pragmatic intention. This sheds light onto the cognitive processes underlying the non-trivial process of intention recognition that we engage in everyday communication. Further experimental and theoretical work is needed to integrate exactly how speech act prediction and comprehension meshes with prediction and comprehension at lower levels of representation in language.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer*. (6.0.48) [Computer software]. http://www.praat.org/

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no. . . How the brain interprets the pregnant pause in conversation. *PLoS ONE*, *10*(12), 1–15. https://doi.org/10.1371/journal.pone.0145474

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Casillas, M., & Frank, M. C. (2013). The development of predictive processes in children's discourse understanding. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 299–304.

de Ruiter, J. P. (2012). Introduction: Questions are what they do. In J. P. De Ruiter (Ed.), *Questions: Formal, functional and interactional perspectives*. Cambridge University Press.

de Ruiter, J. P. (2019). Turn-Taking. In C. Cummins & N. Katsos (Eds.), *Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford University Press.

de Ruiter, J. P., Mitterer, Holger., & Enfield, N. J. (2006). Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, *82*(3), 515–535. https://doi.org/10.1353/lan.2006.0130

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat Neurosci*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. https://doi.org/10.1016/S0022-5371(81)90220-6

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Ferreira, F., & Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. *Current Directions in Psychological Science*, *27*(6), 443–448. https://doi.org/10.1177/0963721418794491

Gisladottir, R., Chwila, D., Schriefers, H., & Levinson, S. (2012). Speech Act Recognition in Conversation: Experimental Evidence. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). https://escholarship.org/uc/item/94n80472

Gisladottir, R. S., Bögels, S., & Levinson, S. C. (2018). Oscillatory Brain Responses Reflect Anticipation during Comprehension of Speech Acts in Spoken Dialog. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00034

Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS One*, *10*(3), e0120068. https://doi.org/10.1371/journal.pone.0120068

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). *Rstanarm: Bayesian Applied Regression Modeling via Stan. R Package v. 2.19. 2.*

Goregliad Fjaellingsdal, T., Ruigendijk, E., Scherbaum, S., & Bleichner, M. G. (2016). The N400 Effect during Speaker-Switch-Towards a Conversational Approach of Measuring Neural Correlates of Language. *Front Psychol*, *7*, 1854. https://doi.org/10.3389/fpsyg.2016.01854

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Hauk, O., Giraud, A.-L., & Clarke, A. (2017). Brain oscillations in language comprehension. *Language, Cognition and Neuroscience*, *32*(5), 533–535. https://doi.org/10.1080/23273798.2017.1297842

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Jordan, T. R., & Thomas, S. M. (2002). In search of perceptual influences of sentence context on word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 34–45. https://doi.org/10.1037//0278-7393.28.1.34

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of*

Memory and Language, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Keitel, A., Prinz, W., Friederici, A. D., Hofsten, C. von, & Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *Journal of Experimental Child Psychology*, *116*(2), 264–277. https://doi.org/10.1016/j.jecp.2013.06.005

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Res*, *1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension. *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Magyari, L., & de Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Front Psychol*, *3*, 376. https://doi.org/10.3389/fpsyg.2012.00376

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rayner, K., Slattery, T. J., & Drieghe, D. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514–528. http://dx.doi.org/10.1037/a0020990

Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00089

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.1353/lan.1974.0010

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I* (Vol. 1). Cambridge University Press.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*(4), 289–327.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*(2), 232–252. https://doi.org/10.1016/0749-596X(85)90026-9

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Stan Development Team. (2020). *RStan: The R interface to Stan*. http://mc-stan.org/

Staub, A. (2015). The Effect of Lexical Predictability on Eye Movements in Reading: Critical Review and Theoretical Interpretation. *Language and Linguistics Compass*, *9*(8), 311–327. https://doi.org/10.1111/lnc3.12151

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Strijkers, K., Chanoine, V., Munding, D., Dubarry, A.-S., Trébuchon, A., Badier, J.-M., & Alario, F.-X. (2019). Grammatical class modulates the (left) inferior frontal gyrus within 100 milliseconds when syntactic context is predictive. *Scientific Reports*, *9*(1), 4830. https://doi.org/10.1038/s41598-019-41376-x

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Van Berkum, J. J. A., Van Den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591. https://doi.org/10.1162/jocn.2008.20054

Warnke, L., & De Ruiter, J. P. (in prep.). *The Role of Top-down Pragmatic Knowledge in Speaker Change Detection*.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. https://doi.org/10.1080/23273798.2014.995679

Table 1

Conditions of the experiment, 3 (congruency: congruent, speech act violation, speaker-independent violation) x 2 (speaker: same speaker, different speaker)

Context Utterance (TCU 1): "*We just moved into a new house*."

Target Utterance (TCU 2):

	Congruent	Speech act violation	Speaker-independent violation
Different speaker	"*Where?*"	"*Come by*"	"*I'm lost*"
Same speaker	"*Come by*"	"*Where?*"	"*You sure?*"

Table 2

Means and standard deviations for plausibility ratings as a function of congruency and speaker

	Congruency					
	Congruent		Speech act violation		Speaker-independent violation	
Speaker	*M*	*SD*	*M*	*SD*	*M*	*SD*
Different speaker	5.28	1.15	3.35	1.67	2.07	1.43
Same speaker	5.09	1.30	4.31	1.78	2.14	1.45

Table 3

Means and standard deviations for bias in milliseconds as a function of congruency and speaker

	Congruency					
	Congruent		Speech act violation		Speaker-independent violation	
Speaker	*M*	*SD*	*M*	*SD*	*M*	*SD*
Different speaker	306	385	341	405	372	419
Same speaker	308	450	323	406	362	440

Table 4

Summary of Linear Mixed Effects Model for bias for different levels of congruency with congruent as the reference level

Fixed effects	Mean	95% CI	P(b>0)	P(b<0)
Speech act violation	18.8	0.5 – 37	0.98	0.02
Speaker-independent violation	57	38.6 – 75.2	1	0

Note. Mean represents the posterior mean unstandardized beta coefficient, 95% CI represents the credible interval around the mean, P(b>0) represents the probability that the coefficient is greater than zero, and P(b<0) represents the probability that the coefficient is less than zero.

Table 5

Summary of Linear Mixed Effects Model for bias for different levels of congruency with speaker-independent violation as the reference level

Fixed effects	Mean	95% CI	P(b>0)	P(b<0)
Congruent	-57	-75.3 – -38.8	0	1
Speech act violation	-38.1	-56.6 – -19.6	0	1

Note. Mean represents the posterior mean of the unstandardized beta coefficient, 95% CI represents the credible interval around the mean, P(b>0) represents the probability that the coefficient is greater than zero, and P(b<0) represents the probability that the coefficient is less than zero.

Figure 1

Overhearer's communication model

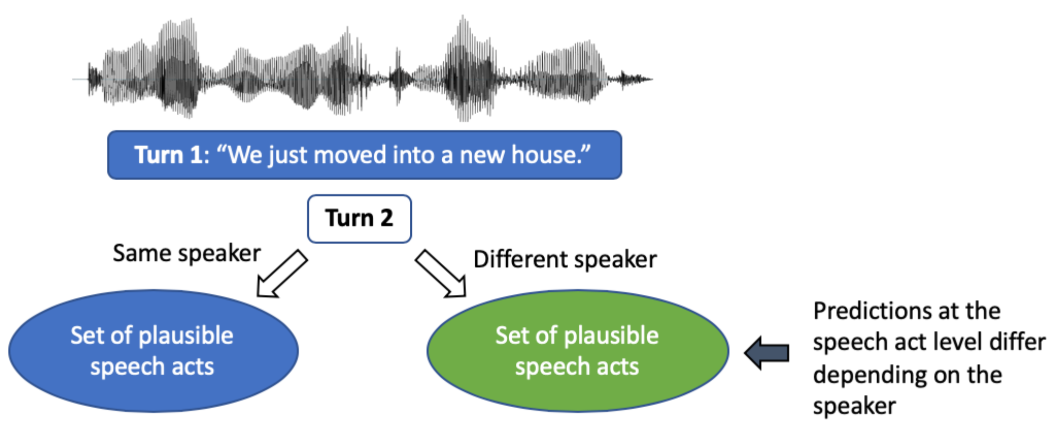

Figure 2

Presentation of each stimulus in a trial of the experiment.

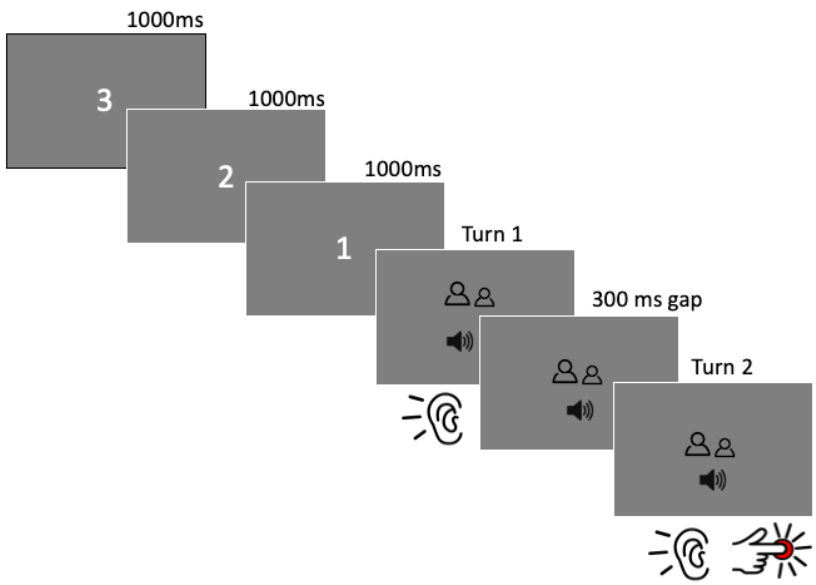

Figure 3

Mean bias scores and standard errors separated by congruency and speaker

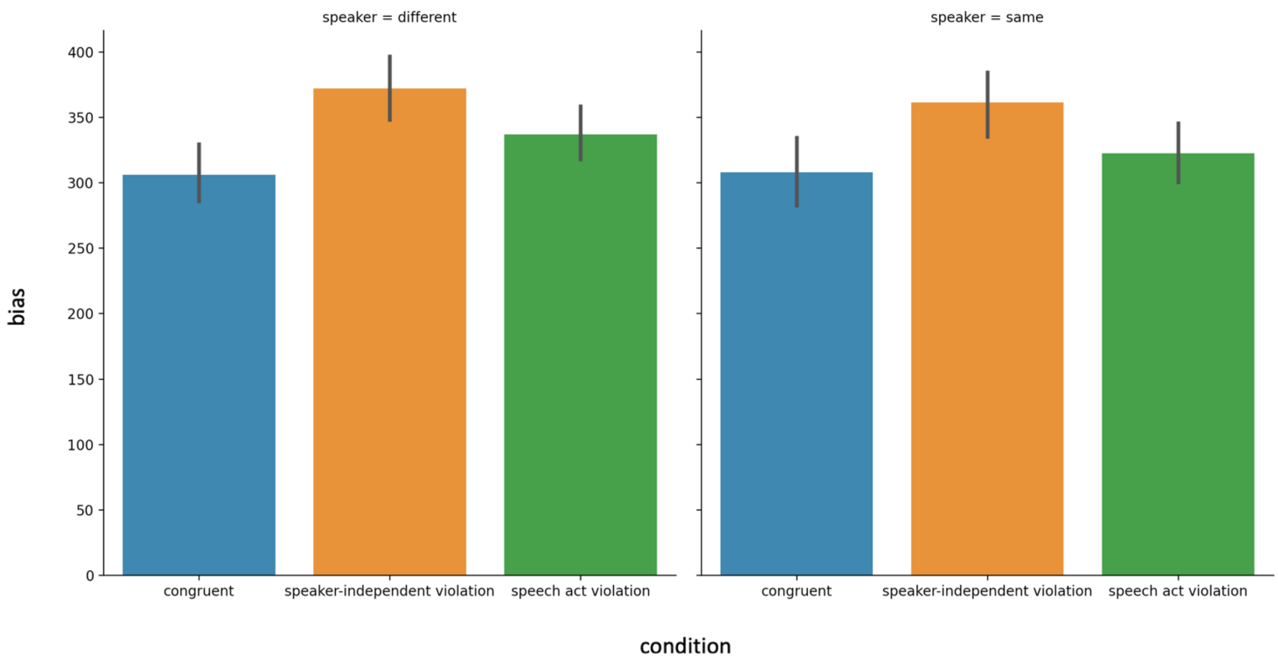

Figure 4

Probability density of bias separated by congruency