The Communication of Social Meaning in Conversation

A dissertation submitted by

Lena Warnke

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Psychology and Cognitive Science

Tufts University

February 2024

Advisor: Dr. Jan P. de Ruiter

Abstract

Language is a social tool and talking to one another entails a deep knowledge and consideration of the broader social context. In this dissertation, I present a series of studies that examine how high-level social knowledge of conversational sequence organization shape lower-level conversational mechanics. In Chapter 1, I introduce the rich structure of conversation, and the social preferences that shape how we interact with one another. Chapter 2 presents experimental evidence showing that top-down knowledge of sequence organization in conversation informs how and when we perceive a speaker switch. In chapter 3, I show that turns in conversation are predicted at the speech act level, extending top-down prediction in language comprehension beyond the sentence level and across turn and speaker boundaries in conversation. Chapter 4 directly compares direct vs. indirect speech acts and their respective cognitive processing demands in natural dialogue: speech acts are not recognized *by* their sentence type, but *in spite of* their sentence type. Higher-level pragmatic expectations inform how incoming linguistic input is incrementally integrated at lower levels. Lastly, in chapter 5, I explore the relationship between turn length and turn timing in experimental versus natural social settings and find that longer turns are followed by longer gaps in natural conversation but not in laboratory experiments**.** Together, the work presented here suggests that the social dynamics of interaction shape how we understand speech and language in conversation.

Acknowledgments

First and foremost, I'd like to thank JP de Ruiter for being an incredibly supportive advisor since day one at Tufts. Thank you for your guidance, for all of our insightful discussions, and for giving me the space and trust to see my ideas through. Most importantly, thank you for always advocating for me, and for being there for me not only as an advisor but also a person. Second, I'd like to thank my dissertation committee, Gina Kuperberg, Ari Goldberg and Kristen Bottema-Beutel, for their help and insights in developing this dissertation and for supporting me throughout the process. I am especially grateful to Gina Kuperberg for adopting me into the Kuperberg lab back in 2016 and shaping me as a researcher.

Thank you to my colleagues and friends at Tufts, especially Samer Nour Eddine, Raea Rasmussen, Julia Mertens, Chas Threlkeld and Nick Rabb, for being there alongside me through the last five plus years. I'm so glad to have had you in my corner. Thanks for letting me bounce ideas off you, for commiserating with me, and for letting loose sometimes. Doing a PhD is, frankly, a bizarre experience, and having a community of people who understand this journey has kept me grounded and sane.

Thank you to my parents Gilla and Peter and my incredible siblings Louisa and Jonas. Thank you for making sure I kept things in perspective, for making me laugh, for keeping me humble, for pushing me, and for teaching me everything I know about being a person, which, it turns out, is a very important component of being a researcher. I could not have done this without your love and support.

Finally, I'd like to thank my incredible friends for making my life so rich. Thank you to my bandmates Hava Horowitz, Cara Giaimo and Martha Schnee for giving me a space to dream, to forget and remember, to play. Thank you especially to Martha for always asking questions, for

our endless walking and talking, for being a home for my ideas. To Cecilia MacArthur and Patricia Noto, for making my literal home a place of joy and refuge. To Amulya Mandava, for your wisdom and humor. To Jessy Reed, for your steadfast support, curiosity and laughter. Last but not least, thank you to Lucie March, for everything, but especially for thinking and looking with me, and for answering on my behalf when someone at a party asks what my dissertation is about. I would never have survived this PhD without you all, so thank you very much, for everything.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1: Introduction

1. Language as a social phenomenon

Language is a social phenomenon. Its origins are social: language developed out of humans' cooperative sociality, a milieu in which language became advantageous for communication (Pagel, 2016; Smit, 2014). Children learn language through social interaction in conversation (Kuhl, 2007, 2011), and throughout our lifetimes we use language to engage with others, communicating our goals, ideas, needs and desires with those around us. Language evolved out of humans' social nature, and continues to sustain the sophisticated social basis of human society.

Though the primary ecological niche of language is social interaction, cognitive science research has largely focused on studying language at the level of the individual, neglecting its social communicative function. The field has instead focused on language's hierarchical, recursive, and generative structure, for example, and the complex cognitive computations that give rise to these phenomena. Studying language processing in this limited and unnatural context is problematic because it ignores the complex social dynamics at play in everyday conversation and their influence on how we process language. Most language comprehension studies investigate the cognitive processes underlying reading or listening to isolated, slow sentences that are unrelated to each other, a setting that is starkly different from everyday conversation where several interlocutors in a social environment rapidly exchange utterances. This discrepancy between how language is often studied and how it is actually used poses a threat to ecological validity, making it difficult to generalize conclusions from sentence processing studies to real-world communication.

The literature suggests that context plays a critical role in language comprehension – our understanding of words and even whole utterances depends on their linguistic and non-linguistic

environment. Preceding language in a sentence shapes the interpretation of words, for example (Kutas & Federmeier, 2011; Kutas & Hillyard, 1984), as does the physical environment in which said words occur (Altmann & Kamide, 1999; Huettig et al., 2011). Given that most linguistic research does not take into account how language comprehension functions in the context of social interaction, the role of top-down social context on bottom-up language processing has been overlooked. In this chapter I will examine the prosocial structure of conversation and review existing evidence for the importance of social context in the processes of language comprehension.

2. Conversation is structured to support social interaction

Conversation is a socially collaborative endeavor: several people have to coordinate with each other to perform *participatory actions* (Clark, 1996). Spontaneous conversation follows a structure, both in terms of timing and content, that is shaped by social context. This structure both derives from and supports the social dynamics of human interaction.

Interlocutors in conversation use language to *do things*, i.e. to accomplish social actions such as requests, invitations, or promises. An essential component of conversational organization is the *speech act* (Austin, 1962; Searle, 1969): a verbal action that performs a social function. Understanding the speech act of your interlocutor is a fundamental prerequisite of conversation: it is only once you have extracted the underlying function of the utterance that you can begin to plan a relevant response. Understanding an utterance's speech act, however, is not a trivial process. There is no one-to-one mapping between at utterance's linguistic form (the words in the utterance) and the social action it performs (Cummins & De Ruiter, 2014). For example, the utterance "*Have you lost it?"* is formally interrogative, i.e. it is grammatically a question, but, depending on the context, it could communicate a criticism or an insult rather than request

information, as questions purport to do. Similarly, the same speech act can be expressed by many different utterances taking varying grammatical forms. One could ask if you are attending a function by saying "*Are you going to the party later?*" or *"Perhaps I'll see you at the party later*". Consequently, the semantic and syntactic contents of an utterance are underspecified with respect to the action that it performs. Social context determines our interpretation of what is being communicated in conversation – meaning is derived interactively.

Consecutive speech acts form paired action sequences known as *adjacency pairs* (Schegloff & Sacks, 1973). Adjacency pairs function such that the social action of a given turn in a conversation places strong constraints on the possible speech acts of the following turn; turns in conversation are structured to relevantly respond to the preceding social act (Levinson, 1983; 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 we generally prefer to communicate rationally and cooperatively (Davies, 2007; Grice, 1989), actions in conversations have implications for how we should respond, and the preceding social act in the conversation provides a basis for interpreting upcoming utterances. This structure of action sequences that informs pragmatic meaning in conversation is cross-cultural and cross-linguistic (Enfield et al., 2010; Kendrick et al., 2020).

Conversation is structured by communication at the pragmatic level: we have to understand the speech act of a turn in order to respond appropriately. Conversational structure is also constrained by people's preference to maintain social affiliation in conversation (Pillet-Shore, 2017). Conversation Analysis, a sub-discipline of sociology, conceptualizes this phenomenon as *preference organization*. The systematic construction of people's actions with the goal of maintaining social solidarity is both highly generalized across speakers and

communicative environments (Pillet-Shore, 2017). Perhaps the most central preference supporting social affiliation that interlocutors demonstrate in everyday conversation is *progressivity* – forward movement in conversation (Stivers & Robinson, 2006). People talking to each other work together to maximize progressivity in their interactions.

Interlocutors are progressive in their timing – they typically transition rapidly between speaker and listener roles with minimal gaps or overlaps. The lack of pauses between turns appears to be universal: the modal gap between conversations across languages is approximately 200 ms, with gaps of 0 ms or even negative gaps (overlapping speech) occurring frequently (De Ruiter et al., 2006; Sacks et al., 1974; Stivers et al., 2009). Interlocutors work together to minimize gaps. If a speaker is having difficulty finding a word in conversation, another person will complete their turn to progress the interaction (Goodwin & Goodwin, 1986). Similarly, when one person asks a question and allocates the next turn to a specific person, but that person does not respond within the typical time frame, someone else in the conversation will speak in their place to keep the conversation moving forward and to avoid gaps between turns (Stivers & Robinson, 2006). This latter finding is particularly pertinent in highlighting the importance of progressivity. In conversation, there is a preference for the selected next speaker to take the turn. Stivers & Robinson (2006) demonstrate that when this preference competes with the preference for progressivity, the preference for progressivity prevails. Non-selected interlocutors consistently take the subsequent turn so that the interaction proceeds.

Speakers in a conversation are also progressive in their content production. Speakers minimize the length of their phrases while structuring them such that references are recognizable with minimal need for clarification (Heritage, 2007). Interlocutors limit the number of clarification requests in conversation (Schegloff, 1979). Should a speaker miscommunicate, there

is a preference for self-repair, meaning speakers will try to clarify their miscommunication within the same turn (Schegloff et al., 1977), rather than waiting for their interlocutor to insert an additional turn to flag the misunderstanding. Lastly, speakers respond to the preceding speaker's turn with socially relevant speech acts, facilitating the listener's interpretation of the speaker's utterance (Levinson, 1983). Furthermore, content and timing are closely linked. The speech act of the utterance influences its timing (Bögels et al., 2015; Kendrick & Torreira, 2015; Pomerantz & Heritage, 2013; Schegloff, 2007): speakers respond faster when their turn is socially preferred, e.g. accepting rather than declining an invitation. Listeners use this timing information in their interpretation of the turn's meaning (Bögels, Kendrick, et al., 2015).

Conversations are progressive both in terms of timing and content, allowing conversations to move forward smoothly and rapidly. This universal preference in conversation is a social phenomenon. By working together and maximizing smoothness in conversation, interlocutors exhibit their preference to "go on" with each other (Shotter, 1996). Progressive interactions preserve *face*, i.e. the positive social value that is revealed in people's actions with each other (Goffman, 1967). As opposed to a slow or irrelevant response, which indicates troubled understanding, a fast and relevant response indicates that the listener understood what the speaker intended with ease. Participants' smooth progress from one turn to the next also signals an absence of misunderstanding: by transitioning between turns quickly and providing relevant responses, participants communicate that they have achieved shared understanding (Albert & De Ruiter, 2018). Sharing understanding, in turn, is indicative of social closeness. Interactions between people from similar backgrounds are generally easier, as evidenced by more overlapping common ground (Clark & Wilkes-Gibbs, 1992), and more accurate turn-end prediction (Hadley et al., 2020). Thus, when a conversation is fast and smooth, participants can

assume that they are “on the same page”, sharing understand and grounded in the same knowledge.

3. Language is hierarchical: Top-down processing matters

Conversation is a highly structured system organized around social components, and language lies at its center. To understand how higher-level social context in conversations influences language comprehension at lower levels, we must first examine the structure of language and our cognitive architecture for comprehending language. Critical to this understanding is language’s hierarchical organization: we comprehend (and produce) language at several different levels of representation. Understanding spoken language involves transforming a sound wave entering our ear into a hierarchy of representations. We understand utterances in terms of the speech sounds (phonemes) we hear, how those speech sounds together form words (lexical units), the meaning of those words and the sentence they form (semantics), the systematic arrangement of those words and phrases (syntax), and finally, at the highest level, what the intended social action of the utterance is (pragmatics).

The cognitive mechanisms underlying language comprehension mirror this hierarchical nature. Comprehenders use bottom-up processing to come to an understanding of an utterance: we process language at the lowest levels of representation (phonemes), and move up the hierarchy until multiple levels or representation come together to inform the pragmatic, message-level understanding of the utterance. Listeners also employ top-down cognitive processes in language comprehension where higher levels of representation inform our understanding of language at lower levels. For example, our understanding of a word at the lexical level shapes the phonemes we hear (Ganong, 1980; Gwilliams et al., 2018) and syntactic structure informs our semantic comprehension of words, for example (Ferreira & Henderson, 1991; Frazier & Rayner,

1982). Preceding linguistic context also shapes how we understand words at the semantic level (DeLong et al., 2005; Kutas & Hillyard, 1980, 1984; Staub, 2015). Bottom-up processing derives from perceptual sources, while top-down processing derives from contextual knowledge. The two processes are interdependent, acting in parallel to inform our understanding of language (Field, 2004; Tsui & Fullilove, 1998).

There are several reasons why top-down processing during language comprehension is beneficial. Our environment is noisy and uncertain, and neural processing itself is noisy (Shadlen & Newsome, 1994). Top-down comprehension reduces ambiguity in the bottom-up perceptual input: if a loud noise occurs while your interlocutor is speaking, you can use the preceding and subsequent linguistic content to understand the word in the face of noise, for example. Given the ambiguity and noisiness of linguistic input that our brains need to process, and the rapid speed at which these inputs are received, top-down processing facilitates integration and provides information that bottom-up processing alone cannot provide.

Our understanding of how top-down processing at higher levels of representation shapes processing at lower levels of representations stems largely from the sentence processing literature. This work has led to fruitful findings that have significantly advanced our understanding of the cognitive mechanisms underlying language comprehension. A significant gap in this literature, however, is that sentence processing research does not represent the complex social context within which we communicate with each other. Interactions in conversation involve a higher social (pragmatic) level of language representation that goes beyond the level of the sentence. How top-down social context influences language processing at lower levels of representation is thus not well understood, presenting a gap in our understanding of language in its most natural context: social interaction.

4. Top-down social expectations and preferences influence how we perceive spoken language in conversation

Why would social context influence how we perceive language at lower levels of representation, and what evidence do we have for this phenomenon? First, we know that social context affects how we perceive our surroundings across other domains. We look at images differently when we know that another person is looking at the same image (Richardson et al., 2012), and visual attention is guided by how socially relevant a visual cue is perceived to be (Gobel et al., 2018). Social relevance also enhances memory for impressions, particularly in older adults (Cassidy & Gutchess, 2012), and social context modifies brain responses and subsequent behavior to touch (Saarinen et al., 2021). Our social environments substantially shape how we perceive the world around us. Second, we know that language comprehension and production mechanisms are influenced by many different contexts including visual information (Mcgurk & Macdonald, 1976), the concurrent visual world (Huettig et al., 2011), the spatial location of objects (Altmann, 2004), and even tactile information (Gick & Derrick, 2009). Given that social factors influence perception and cognition across domains, that language comprehension is contextual, and that language is a socially situated phenomenon, language comprehension is doubtlessly shaped by social contexts. Surprisingly, models of language processing and psycholinguistic theories have largely disregarded the social environment of communication.

The effect of social context on language processing is evident in early development. Newborn babies, for example, already express a preference for their mother's voice compared to the voice of a stranger (DeCasper & Fifer, 1980). Infants prefer to look at faces which were previously paired with their native language vs. a foreign language (Kinzler et al., 2007) and

8-month-olds learn faster from facial cues than other attentional cues (Yurovsky et al., 2011). Lastly, social interaction between a caregiver and their child plays a fundamental role in language acquisition (Kuhl, 2007; Kuhl et al., 2003). Communicative development is dependent on and shaped by social relationships, suggesting an early close link between social context and language.

Psycholinguistic research suggest that social context has a top-down effect on real-time language processing in adults as well. Social cues about speakers such as eye gaze (Hanna & Brennan, 2007), gestures (Holle et al., 2012; Wu & Coulson, 2005) and facial expressions (Carminati & Knoeferle, 2013) are rapidly and incrementally integrated into language processing. Listeners draw on information about a speaker's ethnicity and speaking style to resolve ambiguity in their speech (Casasanto, 2008, 2010), and social expectations improve speech perception in noise (McGowan, 2015). Event-related brain responses show that the brain incorporates information about a speaker's identity when processing a word's meaning (Van Berkum et al., 2008). What these findings suggest is that language does not exist in a vacuum and comprehension is not a local, context-independent process. Rather, the wider social communicative context directly shapes how we hear and understand an utterance's meaning.

While the vast majority of psycholinguistic models do not include non-linguistic social representations, a limited number of psycholinguistic models of language processing have attempted to incorporate social context. Giles at al.'s (Giles et al., 1991) Communication Adaptation Theory (CAT) proposes that interlocutors adapt their speech, gestures, vocal patterns and accents to each other through social norms. The greater the linguistic resemblance between interlocutors, the more they sympathize with one another. The linguistic style matching (LSM) theory mirrors this proposal: listers are unconsciously primed by speakers in their subsequent

response to said speaker (Niederhoffer & Pennebaker, 2002). Both of these accounts propose linguistic coordination during communication, but do not account for how this process occurs in real time during comprehension. Pickering and Garrod's (2004) Interactive Alignment Theory (IAT) argues that linguistic representations become automatically aligned at many levels across interlocutors in dialogue, facilitating both the comprehension and production of language as joint activity between interlocutors (Garrod & Pickering, 2009). The IAT, however, does not include any explicit representations of social context in the mechanism of language processing. The Coordinated Interplay Account (CIA) (Knoeferle et al., 2014; Knoeferle & Crocker, 2006) includes at least a representation of visual context, and a link between visual context and real-time language processing. The same authors' social Coordinated Interplay Account (sCIA) (Münster & Knoeferle, 2018) builds on this, incorporating socially relevant speaker and comprehender characteristics into real time language processing.

What is missing from both theoretical and experimental approaches to studying the role of social context in language processing is interactive conversation is the incorporation of naturalistic conversation. Most language use occurs in social interactions, namely conversations, between two or more parties. As discussed earlier, conversation is structured for social interaction and shaped by social preferences. This structure provides social context for the interpretation of utterances in conversation: the ultimate meaning of an utterance is not necessarily contained in its linguistic form, but rather inferred through pragmatic social context. How exactly the social infrastructure of conversation influences language comprehension at lower levels of representation remains an important open question.

5. Spotlight on prediction

A fruitful domain for examining top-down processing in language is prediction. In the context of cognitive science, prediction refers to the idea that top-down context influences the state of the processing system before any bottom-up input is received (Kuperberg & Jaeger, 2016). Comprehenders predict upcoming language at multiple levels of representation and make use of various aspects of context. Top-down contextual information facilitates the bottom-up processing of information at the event-structure level (knowledge of events or states) (Altmann & Kamide, 1999, p. 199; Xiang & Kuperberg, 2015), the semantic level (Federmeier & Kutas, 1999; Kutas & Hillyard, 1984), the syntactic level (Kamide et al., 2003; Strijkers et al., 2019), and the orthographic level (the written representation of language) (Laszlo & Federmeier, 2009). Listeners draw on various aspects of the context to generate predictions, including preceding semantic and syntactic information (DeLong et al., 2005; Kuperberg et al., 2020; Wicha et al., 2004), social context such as speaker identity (Lattner & Friederici, 2003; Van Berkum et al., 2008), and knowledge about the world (Hagoort et al., 2004). Visual context also shapes linguistic prediction: eye-movement studies, for example, show that participants who are listening to sentences while looking at a visual scene will look at objects that are uniquely identifiable from the visual context before the onset of the word for that object (Tanenhaus et al., 1995).

Again, the vast majority of research investigating the role of top-down prediction in language comes form sentence processing studies. Even the limited evidence that listeners draw on social context such as speaker identity to predict upcoming linguistic input comes from studies of monologue. Dialogue, or conversation, is inherently an interactive process, providing a rich and complex social context for language. How does top-down prediction facilitate language

processing in the context of social interaction? What role does top-down social information play in predicting upcoming linguistic input in conversation?

As discussed early, conversation is structured by and for social interaction. This social structure of conversation provides a contextual cue for top-down prediction at lower levels of representation in language. The timing of turns in conversation differs depending on the action that they perform, giving listeners a social cue for the speech act of the subsequent turn. Socially dispreferred actions such as declining an invitation are delayed compared to preferred actions such as accepting an invitation (Kendrick & Torreira, 2015). Listeners draw on their knowledge of the timing of preferred versus dispreferred answers to initiating actions such as invitations to update their semantic predictions about a speaker's answer to a question. Neural evidence shows that listeners update their predictions of the word in the next turn in conversation depending on the length of the gap between turns (Bögels, Kendrick, et al., 2015): as the gap lengthens, people begin to change their predictions to a declination rather an acceptance to an invitation.

The speech act or social action of a turn in conversation has implications for how we should respond, and action sequences in conversation follow a structure. Listeners use this predictable structure to recognize the social of action of a turn before the end of the turn though predictive mechanisms (Gisladottir et al., 2012; Gisladottir et al., 2015, 2018). The social act of the preceding turn shapes how we interpret the meaning of the incoming utterance, before the whole utterance has even unfolded, through top-down prediction. What this means is that the same string of words can be interpreted differently depending on the social context.

The advantage of top-down prediction in conversation is clear. Conversation is complex and time-critical. Listeners must begin to plan their response while the speaker is still speaking, meaning that language comprehension and production processes must work in parallel, drawing

on overlapping cognitive resources. Predictive comprehension facilitates this simultaneous process: when a word, for example, is pre-activated, it is easier to incorporate. It also furthers the social preference of progressivity: prediction encourages smooth, forward-moving conversation. In conversation, listeners draw on conversational structure as context for top-down prediction at lower levels of language representation. This prediction, mediated by social context, then further facilitates the interaction by advancing mutual understanding and pacing, i.e. progressivity. The evidence reviewed here, showing that social conversation context shapes top-down processing through prediction, thus provides important insight into social effects in language processing.

6. Summary

Language is a communicative tool that we use for social interaction. Our knowledge of how we process language, however, stems from a line of research investigating language in the context of isolated senteces. How language has, for the most part, been studied does not correspond to how it is actually used. The primary goal of the current work is to further our understanding of socially situatied language. Specifically, I aim to uncover how the social dynamics of interaction shape how we understand speech and language in everyday conversation. The second goal is to emphasize the importance of ecological validity in psycholinguistic research. Conversation is where language developed, where humans learn language, and where most social exchanges are conducted. If we want to make claims about the cognitive mechanisms underlying language comprehension, this process needs to be studied in the context of conversation as much as possible.

Lastly, the link between language comprehension and prosocial structural aspects of conversation highlights the importance of an interdisciplinary approach to the study of language and interaction. Part of the reason why cognitive scientists have largely neglected to study

language in the context of conversation is because experimental control is so difficult. Experimental control is critical for establishing causation in scientific research, and the spontaneous, rich and multifaceted nature of conversation is incredibly difficult to control. Conversation analysis, on the other hand, studies language as it occurs naturally in conversation, but does not make claims about the cognitive mechanisms underlying the observable phenomena. The two approaches have opposing benefits and drawbacks: the inference of causation at the cost of ecological validity on the one hand, and the natural structure and properties of interaction without cognitive explanations, on the other. An interdisciplinary approach drawing on these complementary methods is essential to paint a full picture of how we understand language.

Chapter 2: Top-Down Effect of Dialogue Coherence on Perceived Speaker Identity

n.b. this work has been previously published in *Scientific Reports*: Warnke, L., & De Ruiter, J. P. (2023). Top-down effect of dialogue coherence on perceived speaker identity. Scientific Reports, 13(1), Article 1. https://doi.org/10.1038/s41598-023-30435-z

Introduction

A critical skill in conversation is recognizing when a speaker switch occurs. Knowing who is speaking is important for interpreting the meaning of what is being said. Knowledge of a speaker provides important social context such as identity, ethnicity and speaking style, which helps listeners resolve ambiguity and improve speech perception in noise (Casasanto, 2008, 2010; McGowan, 2015). Given that the same utterance can mean different things depending on who is speaking (Krauss & Fussell, 1991; Metzing & Brennan, 2003; Van Berkum et al., 2008), listeners' ability to detect speaker changes in conversation is a fundamental prerequisite for successfully understanding utterances in conversation.

In the absence of visual cues (e.g. while talking on the phone or listening to a podcast), listeners must identify speaker changes based on the auditory signal of the speakers' voices. Humans are generally very good at this. In fact, the ability to recognize individuals from their vocalizations is an adaptive trait that many social animals demonstrate (e.g. Balcombe & McCracken, 1992; Wiley, 2005). A specific aspect of human voice identification, however, is that listeners have access to both the *sound* of the talker as well as the *linguistic representations* that the utterance entails. That is, an utterances in conversation contains both a sound with acoustic information (e.g. frequency, amplitude, and pitch) as well as linguistic representations including phonetics, phonology, syntax, semantics, and pragmatics. How exactly these different sources of

information interact in the process of speaker change detection is not clear from the existing literature.

Acoustic-focused accounts of speaker change detection suggest that listeners attend to bottom-up acoustic features in the auditory signal; when those features change, a speaker change is perceived. Kuwabara and Takagi (1991), for example, found that formant shifts significantly affect voice-individuality perception. Differences in vocal tract length also inform speaker identity perception (Gaudrain et al., 2009), with vocal tract features playing a more important role than glottal source features (Lavner et al., 2000). The reliability of acoustic features as cues for speaker change detection is, however, inconsistent. Evidence shows that voice information is not continuously monitored at a fine-grain level of acoustic representation (Fenn et al., 2011; Sell et al., 2015), and listeners can identify speakers even when these acoustic features are removed from the auditory signal (Lavner et al., 2000; Sheffert et al., 2002). Furthermore, an acoustic-focused account of speaker change detection does not sufficiently explain the phenomena of speaker change deafness, in which listeners fail to detect a speaker change (Fenn et al., 2011; Vitevitch, 2003). These results suggest that bottom-up processing of acoustic information contained in the auditory speech signal alone does not suffice in explaining how listeners recognize when a speaker change occurs.

An alternative proposal is that bottom-up processing of acoustic features interacts with top-down representations in the process of speaker change detection. Listeners are better at recognizing who is speaking when they can access the phonological representations of the linguistic signal: listeners are more accurate at identifying who is speaking in their native language compared to an unfamiliar language (Perrachione & Wong, 2007), and speaker identity recognition is compromised when abstract linguistic representations of words are impaired, such

as in dyslexia (Perrachione et al., 2011). In line with these findings, one study found that when listeners heard a story in both their native language and an unfamiliar language in which a speaker change occurred halfway through the story, they were better at detecting the speaker change in the unfamiliar language (Neuhoff et al., 2014). The authors explain this effect through listeners' linguistic knowledge in their native language: when they have access to lexical and semantic representations and are not cued to listen for changes, their linguistic expertise overrides their detection of a change in acoustic features of the voice. This finding supports the idea that top-down linguistic knowledge shapes the way we direct attention and detect speaker changes in conversations (Clarke et al., 2014; Gaudrain et al., 2009), aligning with a rich literature showing that speech processing involves top-down processes (see, for instance, Ganong, 1980; Norris et al., 2003).

While the aforementioned literature suggests that listeners draw on top-down linguistic representations to identify speakers, most of this research does not do justice to representing the rich social communicative environment that humans converse in every day. Many of the studies cited here used only single vowels or consonant-vowel (CV) pairings in their stimuli, or asked participants listen to isolated sentences or stories to identify who is speaking and when a speaker switch occurs. Conversation, however, is characterized by rapid communicative interactions across multiple speakers and turns. Importantly, this involves processing at the pragmatic level of language representation: the knowledge of how language is used in social communication. Conversations follow a normative social structure. For example, turns often occur in paired action sequences called *adjacency pairs* (Schegloff & Sacks, 1973) such that the speech act of a given turn places strong constraints on the possible speech acts of the following turn (Schegloff, 2007). A question is typically followed by an answer, and a greeting is typically responded to

with a subsequent greeting - sequences in conversation are pragmatically ordered. To date, no studies that we know of have investigated how speaker change detection is influenced by top-down pragmatic expectations. Given that 1) conversations follow a pragmatic structure (Levinson, 1983; Schegloff, 2007; Schegloff & Sacks, 1973), 2) pragmatics plays an important top-down role in comprehension (Van Berkum et al., 2008), and 3) speaker change detection is influenced by top-down processing at other levels of representation, we hypothesize that top-down pragmatic expectations bias listeners' perception of speaker changes in conversation.

Experiment 1

To investigate whether listeners' pragmatic expectations influence their perception of who is speaking, we ran a randomized controlled experiment. We created a set of naturalistic conversational scenarios consisting of two turn construction units (TCUs) (Sacks et al., 1974), the smallest interactionally complete linguistic unit that can make up a turn in conversation. Both TCUs were spoken by the same speaker. The first TCU served as the context utterance, and the second TCU served as the target utterance. The experiment consisted of three experimental conditions. In the *congruent* condition, the second (target) utterance made sense for the same speaker to say, forming a plausible sequence of TCUs. In the *speaker violation* condition, the second (target) utterance did not make sense for the same speaker to say, but, importantly, would have formed a pragmatically coherent sequence of TCUs if it were spoken by a different speaker. In the *full violation* condition, the second (target) utterance formed a pragmatically incoherent sequence of TCUs regardless of whether it was spoken by the same speaker or different speakers. See Table 1 for examples. The task of the participants was to listen to the pairs of TCUs and indicate whether they heard the same speaker or different speakers. We hypothesize that listeners perceive a greater proportion of stimuli as spoken by different speakers in the *speaker violation*

condition compared to both the *congruent* and *full violation* condition because the *speaker violation* condition is pragmatically more plausible when spoken by two different speakers. Because the *full violation* condition is pragmatically incoherent, we hypothesize that listeners rely on bottom-up acoustic features (as opposed to top-down pragmatic inferences) and thus perceive these stimuli as spoken by the same speaker. We therefore predict no difference between the *congruent* and *full violation* condition.

Method

Participants

We report data from 60 participants who were recruited through the online platform Amazon Mechanical Turk; 20 men and 40 women. Participants' ages ranged from 23 to 35 (*Median* = 31; *M* = 30.7; *SD* = 3.36). Originally, 68 participants were recruited, but 5 failed to complete the session and 3 failed one or more attention checks. All participants indicated having normal or corrected-to-normal vision and hearing and used headphones during the study. Participants were pre-screened to meet the following criteria: native English speaker, living in the USA, no history of psychiatric or neurological diagnoses, and no use of psychoactive medication within the preceding 6 months. All participants in this study provided written informed consent and were financially compensated for their time. Protocols were approved by the Tufts University Social, Behavioral, and Educational Research Institutional Review Board, and all research was performed in accordance with their guidelines and recommendations.

Stimuli

The stimulus materials consisted of audio recordings of conversations. Stimuli were recorded in soundproof rooms at a 44 kHz sampling rate using Shure MX153T/O-TQG Omnidirectional Earset Headworn Microphones. Six native speakers of American English (three

men, three women) acted out the written conversational scenarios as naturally and informally as possible, without reading, to emulate natural conversation. All TCUs were recorded separately and processed using the software Praat (Boersma & Weenink, 2019), version 6.0.48. The overall sound intensity of the recordings was normalized to 65 dB-SPL to prevent loudness differences between TCUs within the same stimulus and between stimuli. The relevant TCUs were then spliced together with a pause of 300 ms (a frequent gap between TCUs in natural conversation) in Python (Van Rossum & Fred, 2009), using the package Parselmouth, version 0.4.1 (Jadoul et al., 2018). Both TCUs in each stimulus were always spoken by the same speaker. The average length of the context TCUs was 1.71 seconds, while the average length of the target TCUs was 0.52 seconds. The average length of the stimuli (both TCUs separated by 300ms of silence) was 2.53 seconds. A total of 375 stimuli were created, with 125 items per conditions. Each target utterance was fully counterbalanced across conditions: the exact same second TCU appeared in each of the three conditions. The stimuli were then divided into three experimental lists such that no TCUs were repeated. Each list contained an approximately equal proportion of items per condition (*congruent*, *speaker violation* and *full violation*): 42 stimuli from two of the conditions, and 41 stimuli from the remaining condition. The condition with one less stimulus item varied between the three lists. Similarly, the six recorded speakers appeared as equally as possible in each condition and within each list. The order of the stimuli was randomized.

Procedure

The study was conducted online via the Qualtrics platform (*Qualtrics*, 2022). Participants completed a pre-screening survey to ensure that they met the eligibility criteria to participate in the study. They were instructed to listen carefully to the conversations and to indicate whether they heard the same speaker or different speakers in the first and second TCU by selecting either

"same" or "different" from a multiple-choice menu using a mouse click after hearing the stimuli. Stimuli from the three conditions were presented in a randomized order. Prior to carrying out the task, participants completed a guided practice with examples to familiarize themselves with the task. Five "catch" trials were included at random intervals to ensure that participants were paying attention throughout the experiment and to identify and omit data from bots. These trials consisted of audio recordings of a speaker instructing the participant to either select "same" or "different" from the two possible response options. Participants listened to a total of 125 conversations and took approximately 20 minutes to complete the study.

Statistical Analysis

We analyzed the data with logistic linear mixed-effects regression models (Baayen et al., 2008) with R version 4.11 (R Core Team, 2021). We performed both a Bayesian and a traditional frequentist analysis. Bayesian analyses provide information about the strength of the evidence in favor of either the alternative or the null hypothesis, while frequentist tests compute the probability of the data or more extreme data under a null hypothesis. For the Bayesian analyses, 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 interpret Bayes factor values using evidence categories from Wetzels (2011), adapted from Jeffreys (1961).

For the Bayesian 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 our fixed effect (condition) and random effects (participants and items) 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.

For the frequentist models, we used glmer from the R package lme4[34] and used the lmerTest package (Kuznetsova et al., 2017) to calculate p-values for models. Again, we added the same fixed and random effects incrementally to a minimal model and then used the likelihood ratio test for model comparison to test if the inclusion of an additional term was justified (Pinheiro & Bates, 2000).

Results and Discussion

In the congruent condition, 82.44% of trials were perceived as being spoken by the same speaker. In the speaker violation condition, only 61.4% of trials were heard as the same speaker, and in the full violation trials, 70.6% of trials were heard as the same speaker. The data are visualized in Figure 1. The distribution of participants' responses across the 6 speakers of the stimuli are reported in Table 2.

According to the Bayesian analysis, the model under which the data were most likely was 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 6.23×10^{74}, providing decisive evidence for this model over the null model (intercept only). This indicates a main effect of condition: whether participants heard the same or different speakers was affected by our experimental conditions.

The summary for the final model for the data is shown in Table 3, with *congruent* as the reference level. Participants heard the second TCU in the *congruent* stimuli as the same speaker more often than in both the *speaker violation* and the *full violation* condition, with a model

estimated probability of 1. Participants also heard the second TCU in the *speaker violation* condition as different speakers more often than in the *full violation* condition, with an estimated model probability of 1.

The frequentist analysis was consistent with these findings. The final model with *congruent* as the reference level is shown in Table 4, and with *full violation* as the reference level in Table 5, and contains condition as a fixed factor with random intercepts for participants and items. We found a significant main effect of condition ($X^2(2) = 335.58$, $p<.001$). Participants heard different speakers more frequently in both the *speaker violation* ($\beta_{\text{speaker violation}} = 1.41$, $p <.001$) and the *full violation* ($\beta_{\text{full violation}} = 0.87$, $p <.001$) condition compared to the *congruent* condition. Participants also heard different speakers more frequently in the *speaker violation* condition compared to the *full violation* condition ($\beta_{\text{speaker violation}} = 0.55$, $p <.001$).

The aim of this study was to investigate whether listeners' pragmatic expectations in conversation have a top-down influence on their perception of who is speaking. Our data confirm our first hypothesis: when two segments in conversation make more sense for different speakers to say, participants report hearing different speakers, even in the absence of a speaker change. We explain this as a top-down effect of pragmatic representation: listeners draw on their knowledge of normative conversational structure such that when an utterance makes sense for a different speaker to say, they infer that a speaker change occurred and perceive a different voice. Interestingly, participants also reported hearing different speakers more frequently when the two segments did not make sense for a different speaker *or* the same speaker to utter. Our hypothesis that listeners draw more strongly on bottom-up acoustic features when an utterance is pragmatically incoherent and unpredictable was therefore not supported. We also explain this

result as an effect of pragmatic inference: incoherent adjacent turns are more likely to be uttered by two different speakers than by the same speaker.

A surprising result of Experiment 1 is that in the *congruent* condition, only 82% of the stimuli were perceived as spoken by the same speaker. This is unexpected because both the acoustic features as well as the pragmatic context should reinforce the coherence of these stimuli, i.e. we would expect 100% to be perceived as spoken by the same speaker. This finding could be driven by several factors. First, participants in web-based experiments tend to be less attentive than those participating in in-person experiments[37], so our result could be explained through noise in the data. However, "catch questions" and large sample sizes such as ours should reduce and compensate for this noise (Paolacci et al., 2010). Second, our experimental design could have influenced participants' selection of answers. Given that all our stimuli were always spoken by the same speaker but that participants are, in general, not inclined to keep giving the same answer, participants could have been biased to select having heard different speakers. Since our effect of condition is very large, this possibility does not alter our conclusion: the pragmatic relationship between two TCUs plays a role in speaker change detection. The potential of our experimental design biasing participants' answers does, however, suggest that this effect may be more subtle in real, everyday conversations.

Though our results strongly implicate pragmatic context in speaker change detection, our experimental design does not allow us so strictly rule out any influence of acoustic cues. Research suggests a strong interaction between top-down lexical and sentential content and indexical (talker-specific) content (Koelewijn et al., 2021; Zaltz et al., 2018): the presence of linguistic context at the word and sentence levels affects participants' discrimination of voice cues. In order to test the causal role of pragmatic context in speaker change detection and to

further rule out any effect of acoustic cues in Experiment 1, we ran a follow up experiment in which we removed all pragmatic information from the same stimuli, leaving acoustic information intact.

Experiment 2

In Experiment 1, participants listened to utterances in conversation that were either pragmatically coherent or incoherent, and our results suggest that pragmatic coherence influences listeners' speaker change detection. To investigate the causal nature of this top-down pragmatic effect on speaker change detection, we ran a control experiment in which we took the same conversational stimuli and removed all pragmatic information while preserving the acoustic information. In our second experiment, we tested this by asking a new set of participants to partake in the same task using the same stimuli with the exception that the word order in the stimuli was scrambled, removing any pragmatic information from the utterances. We hypothesized that the rate of speaker change detection would now be identical across conditions.

Method

Participants

Sixty participants were recruited from Amazon Mechanical Turk, including 33 men and 27 women whose ages ranged from 23 to 35 (*Median* = 32; *M* = 31.15; *SD* = 3.43). Participants were compensated and screened according to the same procedures are Experiment 1. All participants in this study provided written informed consent and were financially compensated for their time. Protocols were approved by the Tufts University Social, Behavioral, and Educational Research Institutional Review Board, and all research was performed in accordance with their guidelines and recommendations. None of the participants in this experiment had

participated in Experiment 1.

Stimuli and Procedure

To remove the pragmatic information from the stimuli, we randomly shuffled the order of the words. For each TCU of the stimulus, we first found the boundary between each word using timing information from Gailbot (Umair et al., 2022), an automated transcription software that calculates word-by-word timing. All word boundaries were then manually inspected and corrected and moved to zero crossings in the waveform using Praat (Boersma & Weenink, 2019). Sound files were cut at the word boundaries and spliced back together in a random order. Because randomizing the word order made it difficult to recognize the boundary between TCUs, and our task asked participants if they hear the same or different speakers in the first and second TCU, we included a beep sound between the two TCUs. The beep was 300ms long, the same length as the gap between turns in Experiment 1. To ensure that no pragmatic information was retained in the stimuli, no more than two words from the original word order in Experiment 1 were adjacent to each other, and all stimuli were manually checked to confirm unintelligibility. See Table 6 for example stimuli.

The experimental procedure was identical to that of Experiment 1. Participants were instructed to listen to the scrambled conversation segments and to indicate whether they heard the same or different speakers before and after the beep.

Results and Discussion

We analyzed the data following the same procedures as Experiment 1. In the congruent condition, 62.24% of trials were perceived as being spoken by the same speaker. In the speaker violation condition, only 62.2% of trials were heard as the same speaker, and in the full violation trials, 60.5% of trials were heard as the same speaker. The distribution of participants' responses

across the 6 speakers of the stimuli are reported in Table 7. As Figure 2 confirms, there was no observable difference in speaker ratings between our three experimental conditions when the pragmatic information was removed from the stimuli. Both the Bayesian and frequentist analyses confirm this finding. According to the Bayesian analysis, the model under which the data were most likely was one that contained only random intercepts for both participants and items. The data were 1938 times more likely under this null model compared to a model that also contained condition as a fixed effect, indicating that there was no main effect of condition. For the frequentist model, the final model was the same. There was no main effect of condition ($X^2(2) = 2.82$, p=.24), and none of the conditions significantly predicted whether participants heard the same or different speakers. See Table 8 for the model with *congruent* as the reference level and Table 9 for the model with *full violation* as the reference level. The results from the current study thus strongly indicate that the effect found in Experiment 1 can be explained by pragmatic context rather than acoustic features in the auditory signal, which remained, for the most part, identical across experiments. We discuss implications and possible limitations to this conclusion below.

General Discussion

Our experiments demonstrate that listeners use their expectations of dialogue coherence to infer who is speaking in conversation. We find that when an utterance makes sense for a different speaker to say, listeners experience the illusion that a speaker change has occurred and hear a different voice, even in the absence of a speaker change. Pragmatic knowledge thus has a top-down influence on how speaker identity is perceived by listeners in conversation.

Our study is experimentally well-controlled, supporting our conclusions, and has a high level of ecological relevance: speaker change detection is a process that almost all of us engage

in every day. That said, there are several limitations that warrant discussion. First, some of the experimental manipulations in Experiment 2 that were designed to remove pragmatic intelligibility may also have altered acoustic features that could be relevant to speaker identification. Specifically, scrambling the order of the words impacts the rhythmic structure underpinning the speech in the utterance. While the literature on the role of speech rhythm in speaker change detection is limited, research suggest consistent between-speaker variability of speech rhythm (Dellwo et al., 2015), which, according to one study, contributes to speaker recognition (Van Dommelen, 1987). The acoustic features of signal may therefore not have been identical across Experiments 1 and 2. However, the acoustic voice cues that have been shown to be most salient for the discrimination between voices of different talkers – fundamental frequency (F0) (Baumann et al., 2010; Chhabra et al., 2012; Darwin et al., 2003; Koelewijn et al., 2021) and vocal tract length (Hillenbrand et al., 1994) – were not altered by Experiment 2. The long term average spectrum (Löfqvist, 1986), another cue used by listeners to differentiate between voices (Durlach, 2006), also remained consistent across our experiments.
Further research is needed to understand the effect of scrambled word order on acoustic features that are relevant for talker discrimination.

Second, it is possible that acoustic features in the signal were attended to differently across Experiments 1 and 2, which may have influenced participants' auditory perceptions. Research shows top-down effects of linguistic content on the differentiation of voice cues that are relevant to speaker differentiation. The presence of linguistic content at the word level, for example, increases participants' ability to differentiate between vocal tract lengths of voices (Koelewijn et al., 2021), and both vocal tract length and F0 differentiation increases in the context of sentences (Sinnett et al., 2006). These findings indicate a top-down effect of linguistic

content on how we integrate acoustic information in speaker voice discrimination processes, meaning that participants in our experiment may have actually been more sensitive to differences in acoustic features in Experiment 1 compared to Experiment 2. While this insight does not alter our conclusions, it does highlight a possible difference in acoustic feature processing between our two experiments.

A third possibly significant difference between our experiments is the presence of the beep tone separating the TCUs in Experiment 2, which could have disturbed the evaluation of the voices. The beep could be acting as a masking stimulus, affecting and compromising the perception of the speech (Eramudugolla et al., 2005) and rendering it more difficult to compare the speaker voices of the first and second TCU in Experiment 2 compared to Experiment 1. The literature suggests that the ability to detect information in the speech stream is reduced when attention is directed away from speech (Neuhoff & Bochtler, 2018), and work on change deafness specifically indicates that distractors and the way in which we direct attention have robust effects on change detection (Fenn et al., 2011; Vitevitch, 2003). The presence of the beep in Experiment 2 thus reduces the comparability of our two experiments.

Our primary interest in the current study was the difference in the effect of condition between experiments 1 and 2, which shows that pragmatic information influences listeners' perception of when a speaker change occurs. It is, however, also worth comparing the raw proportions of speaker perceptions between the two experiments. In experiment 2 listeners perceived the same speaker in approximately 62% of the stimuli across all three conditions. Given that both TCUs were spoken by the same speaker in all stimuli and that no pragmatic information was available in this experiment, it is surprising that participants reported hearing such a large number of different speakers. One possible explanation for this finding is that our

binary choice experimental design biased speakers to select hearing different speakers more than they actually did. Another possibility is that the beep tone, as discussed above, had a masking effect, reducing the comparability of the voices in the two TCUs. Paired with our experimental design, listeners may have therefore been swayed to perceive two different speakers more than if there had been no beep tone. Finally, it is possible that the destruction of the rhythmic structure of the utterances in Experiment 2, which, as discussed above, may be a possible cue for speaker change detection, affected participants' perception of speaker changes. It is particularly interesting that the response rate in all conditions of Experiment 2 is identical to that of the *speaker violation* condition in Experiment 1, indicating that the high proportion of "different speaker" responses may be an artifact of our experimental design rather than any acoustic processing differences across the two experiments.

Though the limitations discussed above impact the comparability of experiments 1 and 2 and should be considered in future research on speaker change detection, we do not think that they significantly alter our conclusion that listeners use their pragmatic expectations of dialogue coherence to infer who is speaking in conversation. We therefore now turn to the practical and theoretical implications of this finding. First, our work provides important insight for computational models of language processing. Automatic analysis of conversation audio remains a challenge with multiple talkers, intonational variation and overlapping speech. The results from this study provide further evidence that speaker change detection relies on linguistic representations at multiple levels, including the pragmatic level. Understanding the cognitive basis of how humans detect speaker changes in conversation is fundamental to implementing this process in machines.

Second, and most importantly, the findings from this study have important theoretical implications. They show that prior talk in conversation contains a rich structure that proactively affects interpretations of something as low-level as acoustic information. This aligns with a rich literature across many modalities demonstrating top-down effects of cognition on the interpretation of a percept. Furthermore, our results suggests that listeners assume cooperativity in conversation: when an utterance makes more sense for one speaker to say than another, listeners hear the speaker that makes the most sense. Given that the world is noisy and the bottom-up signal is often unreliable, we suggest that this top-down mechanisms of speaker change detection has evolved for optimal social communication.

Chapter 3: Speech Act Prediction Across Turn Boundaries in Conversation

n.b. the data in for this chapter were drawn from my master's thesis: Warnke, L. (2022). *Speech Act Prediction* [Master's thesis, Tufts University]. ProQuest.

Introduction

Understanding the speech act of an utterance in conversation is essential for pragmatic comprehension and thus for the conversation to progress. Despite its critical role, speech act comprehension is quite challenging because the speech act of an utterance is not necessarily contained in its linguistic form. What is meant, in the sense of Grice's "non-natural meaning" (Grice, 1957), cannot in general be extracted from what is literally said. Moreover, the same speech act can be expressed by 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 with respect to the action it performs, listeners must rely on 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 approximately 200 ms, with gaps of 0 ms or even negative gaps (overlapping speech) occurring frequently (De Ruiter et al., 2006; Heldner & Edlund, 2010; Stivers et al., 2009)Within these tight time constraints, a listener must process the incoming speech, extract the underlying speech act, and plan an appropriate response. Earlier research has shown 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 before the end of the current 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 quickly and reliably (Gisladottir et al., 2012; Levinson, 1983). Currently, very little is known about the cognitive mechanisms that underlie the critical and challenging task of speech act recognition in conversation. Research from conversation analysis, however, may provide clues for the role of conversational context in speech act recognition. We know that conversation is organized systematically in paired speech act sequences (*adjacency pairs*) (Schegloff & Sacks, 1973). Questions are typically followed by answers, and greetings are followed by response greetings, for example. Speech act sequences in conversation are predictably structured - the speech act of an utterance has implications for how we should respond. Listeners probably draw on their knowledge of this structure in the process of recognizing speech acts as a turn unfolds.

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 less processing as evidenced by the ERP signal compared to when the context was less constraining. 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 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 required for isolated sentence comprehension. Furthermore, language in conversation unfolds rapidly in the auditory modality. 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 accurate 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. 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 caveats. First, 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; Threlkeld & De Ruiter, 2022).

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 3 for an illustration of the listener's presumed communication model.

Study Design

To investigate whether people predict and anticipate speech acts 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 10 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.

Method

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 10), 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 versions of a second TCU that corresponded to each of the four experimental conditions as follows. First, we constructed a target TCU such that the context TCU followed by the target TCU created a naturalistic, congruent conversational scenario when the target TCU was spoken by the same speaker as the context TCU (*same speaker, congruent*), and an incongruent conversational scenario when spoken by a different speaker as the context TCU (*different speaker, speech act violation*). Similarly, for each context TCU, we also constructed a target TCU such that the context TCU followed by that target TCU created a congruent conversational scenario when the target TCU was spoken by a different speaker as the context 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 context utterance (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 TCUs 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 target (second) TCU. All target 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 at 44 kHz sampling rate using Shure MX153T/O-TQG Omnidirectional Earset Headworn Microphones in two soundproof rooms separated by a glass pane. Six American English native speakers (three women, three men) were recorded in women-men 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 context and the target 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 (Heldner & Edlund, 2010; 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' data 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. In the current Bayesian analysis, we report Bayes factors which represent the relative probability of the observed data under model *a* over model *b,* in order to determine the model under which the data are the most likely. In this paper, we used evidence categories from Wetzels (2011), adapted from Jeffreys (1961), to interpret Bayes factor values.

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

Descriptive statistics for plausibility ratings are shown in Table 11. 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, 2023). This phenomenon is discussed further in the discussion section below. We compensated for potential effects of this "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 strength of this paradigm is three-fold: first, it does not disclose the experimental manipulation to participants, 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, and third, the paradigm has successfully captured effects of anticipatory processing in conversation. This provides a controlled yet ecologically valid language comprehension environment for participants.

The experiment was implemented online with PsychoPy 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 computers, 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 4. 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.

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). 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. 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.

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 target turn. As a result, 340 trials (4.53% of the data) were excluded from analyses.

Analysis

The *bias* data was analyzed using the same procedures as the plausibility rating data analysis. The *bias* data is visualized in Figures 5 and 6. The descriptive statistics are presented in Table 12. 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 13 (with *congruent* as the reference level) and Table 14 (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.

The ability to extract the speech act from an utterance is a fundamental skill for successful conversation: it is only once the social function 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 precisely when the turn will end in order to produce a relevant and temporally appropriate response. The results of our study highlight that the prediction of speech acts provides a plausible explanation for the ease with which we perform the challenging task of speech act recognition. We show 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

speech act of an utterance, reducing the time needed to respond. This is evidence for a language processing architecture that is oriented to speech acts.

Though our study has a relatively high ecological validity, and the button-press task approximates the turn-end estimation task that we perform every day in our interactions with others, our experiment has several limitations. First, we use an overhearer paradigm and extrapolate that the cognitive processes of listening to conversation are the same as those of participating in the conversation. While they are likely similar, participating in a conversation requires some multitasking and additional social processing that simply listening to conversations does not. When we interact through talk, we must plan our response to a turn while said turn is still coming it, requiring simultaneous comprehension and production. Our experiment approximates the onset of a response turn with a button press, but not does ask participants to respond verbally or to engage socially with an interlocutor. Second and relatedly, our study does not account for the role of social factors in the turn timing. In conversation, socially dispreferred turns such as declinations to invitations, requests or offers are likely to occur with a delay (Pomerantz, 1984). Furthermore, listeners draw on the length of the gap preceding a turn to predict the speech act of the following turn (Bögels, Kendrick, et al., 2015). While gaps between turns in conversation vary with preference, our study had a stable gap between turns of 300 ms with no variation, possibly biasing participants to only expect socially preferred speech acts, which tend to follow this shorter gap. Given that both cognitive and social delays factor into to turn timing (Mertens & De Ruiter, 2021), future research investigating speech act prediction should control for the relationship between turn timing and preference organization, employing real participatory conversation tasks to increase ecological validity.

It is also 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. It is possible that these positive *bias* values could reflect the process of integration rather than prediction. That said, the modal *bias* in the *congruent* condition was on the order of 61ms, which is generally too fast to reflect reactive cognitive processing and far more likely reflects anticipation. Furthermore, 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. 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.

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 effect size is substantial, the absolute 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 expectations at the speech act level are not the only cue that participants used to estimate the ends of turns. When turns were always incongruent regardless of the speaker, i.e. they violated both the speech act constraints as well as the lexico-semantic constraints of the context, participants were less accurate at estimating the turn's end compared to turns with just mismatching speech act content. That is, when the words *and* the speech act of the turn were contextually incongruent, participants' turn end estimation was slower compared to when just the speech act was incongruent. 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, 2023), 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 speech acts 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.

In sum, we have presented evidence 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 has important theoretical implications. The study of linguistic prediction has to date only been studied in the sentence processing domain. We show, for the first time, that language prediction operates beyond the boundaries speaker change. This means that prediction in language processing is not only facilitating our internal cognitive processing, but also contributes to intra-personal social processing. When we predict speech acts, we predict at a higher, social level that transcends yet interacts with lower language-only predictions. This reveals how the non-trivial process of intention recognition in human interaction is deeply intertwined with our lower-level cognitive processing.

Chapter 4: Indirect Speech Acts Do Not Change FTOs in Conversation

N.b. this work is in collaboration with Charles Threlkeld.

Introduction

We use language to 'do things with words' (Austin, 1962) – utterances in conversation serve as a social tool that we use to accomplish things such as ask for information, answer a question, or offer a compliment. In fact, speech acts are the currency of communication – they are the primary units of meaning in natural dialogue (Vanderveken, 1990). Conversation's central function is to communicate social actions, a dimension that does not exist in non-dialogic contexts such as isolated sentences or texts.

Understanding the speech act of an utterance is critical because all actions in conversations have implications for how we should respond (Schegloff, 2007), i.e. we cannot respond appropriately to our interlocutor's turn until we have deduced its speech act. Speech act recognition, however, is a complex cognitive challenge, not only because of the tight time constraints of natural dialogue, but also, and crucially, because there is not necessarily a one-to-one correspondence between at utterance's speech act and its sentence level form. As Grice (1957) argued, communication involves a process of intention recognition, and identical utterances can correspond to different intentions. That is, the signal itself is ambiguous because it does not necessarily "contain" the speech act – intentions are not contained within the external signal of language (Reddy, 1979). Context is therefore essential in determining the intention, or speech act, that a speaker intends to convey.

This many-to-many mapping challenge is known as the *pragmatics problem* (Cummins & De Ruiter, 2014). Consider the following example. You're sitting on the couch with a friend about to start a movie, and you both realize that the overhead light is still on. You might say to

your friend "*You're closest to the light switch*". The speech act of this utterance is a demand to your friend to get up and turn off the light, but it is expressed in the declarative rather than the imperative form. Similarly, "*Can you pass the salt?"* is formally interrogative, but functions as a request rather than a question. Given that speech acts cannot necessarily be deduced from their grammatical sentence type (interrogative, declarative, imperative) (Sadock & Zwicky, 1985), a key open question is how listeners so successfully map speech acts onto their interlocutors' utterances.

When an utterance's sentence type and function match, it is a direct speech act. According to the Literal Force Hypothesis (Gazdar, 1981 via Levinson, 1983), declaratives function as statements, interrogatives as questions, and imperatives as commands. The sentence type provides a basis for deducing the speech act – function can be inferred from form. When an utterance's function does not match its sentence type, it is known an indirect speech act (Searle, 1979). How exactly indirect speech acts are processed cognitively is unclear, given that function cannot be inferred from surface level form. The goal of the work presented in this chapter is to investigate whether indirect speech act comprehension involves different cognitive processes compared to direct speech act comprehension.

Two competing theories propose how we process indirect speech acts. According to a traditional language-philosophical view, comprehending indirect speech acts requires Gricean reasoning, first assessing the sentence type and then determining whether the corresponding speech act makes sense in the conversational context (Searle, 1976). Grice (1975) proposed the *cooperative principle*, suggesting that people in conversations generally act cooperatively and assume their interlocutor's cooperativity to achieve effective communication in conversation. Grice breaks this principle down into four maxims of conversation – quantity, quality, relation

and manner (Grice, 1975). These rational principles provide a guide for effective and socially cooperative production and comprehension of language in conversation. The flouting of Grice's maxim of manner, specifically, provides a framework for indirect speech comprehension. Gordon and Lakoff (1975), for instance, utilize Gricean maxims to analyze the following example (via Asher & Lascarides, 2001):

A: Let's go to the movies tonight.
B: I have to study for an exam.

Speaker B rejects speaker's A's proposal by asserting something. According to a Gricean inference account, speaker B is violating, or *flouting,* the maxim of manner: to be as clear and orderly as possible, and to avoid obscurity and ambiguity (Grice, 1975). To reach the understanding that speaker B's utterance is a rejection, speaker A first interprets speaker B's utterance literally (as an assertion), concludes, given the context, that this interpretation violates the maxim of manner and the cooperativity principal, and then reinterprets speaker B's utterance as an indirect speech act performing a rejection. The Gricean reanalysis model has been adopted by Gordon & Lakoff (1975), and successfully implemented in a computational model by Sarathy et al. (2020), for example.

The Gricean inference account proposes a reanalysis process in comprehending indirect speech acts. Experimental research, however, shows that we recognize speech acts early in the turn, before the entire utterance has unfolded (Bögels et al., 2015; Gisladottir et al., 2012; Gisladottir et al., 2015, 2018). In fact, turn-taking research more broadly demonstrates that listeners in conversation plan their responses as early as possible in an incoming turn, as soon as sufficient information is available, which is often midway through a turn (Barthel et al., 2016, 2017; Bögels, 2020; Bögels et al., 2018; Corps et al., 2018). Since speech act recognition is a prerequisite for planning an appropriate response, listeners must be extracting the speech act

from an utterance before it has fully unfolded at the sentence level. Early speech act recognition and response planning thus suggests that indirect speech acts are recognized not *by* their sentence type, but *in spite of* their sentence type. Behavioral experiments have demonstrated that indirect speech act recognition is an automatic process for both participants and observers (Holtgraves, 2008a), and that people identify and remember the speech acts that people perform with their utterances, forming one of the organizing principles of conversation memory (Holtgraves, 2008b). Gisladottir et al. (2012) show that the exact same sentence is processed differently depending on its speech act, which was always indirect, with reading times already differing at the first word. Neuroscientific experiments using spoken dialogue stimuli support this finding: neural responses to the same utterance performing different indirect speech acts differed as early as 200ms after the onset of the utterance (Gisladottir et al., 2015, 2018). This aligns with data showing that response planning in conversation starts midway through an incoming turn, implying that the speech act must already be known before the turn is over (Bögels, Magyari, et al., 2015). In sum, experimental research provides evidence for an alternative theory to the Gricean account of indirect speech act comprehension, in which the speech act is extracted as soon as possible before the utterance's sentence-level form has unfolded fully, without reinterpretation.

While experimental work seems to support an account of indirect speech act processing in which the speech act is extracted early in the turn without reanalyzing the literal meaning, the conclusions drawn from the above-mentioned work contains several caveats. Holtgraves (2008b, 2008a), for example, uses artificial tasks such as lexical decision and recognition probe that are quite far removed from the ecological setting of natural dialogue. Similarly, Gisaldottir et al. (2012) use written dialogue and a self-paced reading paradigm in their study. Further, all of the

studies mentioned above draw on overhearer paradigms in which participants are listening to other people speak or interact. Conversation is a rich socially interactive milieu, and if we want to understand the cognitive processes underlying indirect speech act comprehension, we should ideally study the phenomenon in real life, natural interaction. Lastly, no research that we know of has directly compared the processing of direct versus indirect speech acts.

In the current study, we investigate the *timing* of natural dialogue from a large conversational corpus. The two theories discussed above predict different cognitive processes in speech act recognition. In the philosophical (i.e. Gricean) view, indirect speech act recognition is reactive: we first extract a direct speech act according to sentence type and then reinterpret the speech act given context. The alternative model does not suggest reinterpretation, but rather early integration: we recognize speech acts early in the turn before the sentence level signal has fully unfolded, by way of context. According to this model, speech act recognition operates beyond the sentence level, i.e. without sentence-level reanalysis. We predict that if indirect speech acts require extra cognitive processing, interlocutors in conversation should take longer to respond to indirect vs. direct speech acts. If, however, the processing of indirect speech acts does not require reanalysis, i.e. no extra cognitive processing, then turns following indirect vs. direct speech acts should be similar in their timing.

Method

We operationalize the cognitive processing of the previous speaker's turn as the *floor transfer offset* (FTO) in conversation. FTO is the time between the end of one speaker's utterance and the start of their interlocutor's utterance (De Ruiter et al., 2006), and it has been shown that longer FTOs index cognitive demand in conversation (Mertens & De Ruiter., 2021; Walczyk et al., 2003).

Conversation data came from the Switchboard corpus (Godfrey et al., 1992), which contains 2,400 two-sided telephone conversations among 543 speakers of American English (302 men, 241 women), resulting in a total of 260 hours of conversational data. We retrieved precise word by word timing information in each utterance in the corpus from the MSU transcriptions of the Switchboard Corpus (*OSLR*, 2023) and used these to construct the start and end times of each utterance and thus the timing between utterances. We divided utterances into turn construction units (TCUs) – the smallest interactionally complete part of a turn – and retrieved the speech act, sentence type and subsequent FTO for all TCUs that were followed by a speaker switch.

Turn construction unit (TCU) segmentation and speech act tagging was obtained from the Stanford University transcriptions (Jurafsky, Shriberg, et al., 1997). These transcriptions contain 42 speech act types. We reduced the set to three broader speech act categories: question, statement, and commands, as motivated by Gazdar (1981), and excluded utterances that could not be categorized into the above three speech act categories. We did this by examining the syntax of the existing 42 types and breaking them down into their overarching speech act category. Opinions, appreciations and apologies were all categorized as statements, for example. See appendix A for the full set of the original speech acts and how we mapped them into questions, statements and commands.

Our final dataset included only questions and statements because the data did not contain a large enough number of commands. Next, we extracted the grammatical sentence type for each TCU in the corpus and assigned a probability of each TCU being either interrogative, declarative, or imperative using fine-tuning of DistilBERT (Sanh et al., 2020). To do so, we took the pre-trained DistilBERT model and fine-tuned it with a hand-annotated utterance-syntax label set. We chose DistilBERT because it is trained on a word-masking task that makes it suitable to

syntax recognition. Fine-tuning is a form of artificial transfer learning where we freeze the original weights in the word-mask trained model and add a final layer to predict (in our case) utterance syntax.

We performed two analyses. In the first analysis, we used only TCUs assigned at least 50% chance of some sentence type, i.e. the probability value for either interrogative, declarative or imperative had to at least 0.5 for one of the sentence type categories. We then included sentence type as a categorical predictor. This analysis included 1930 TCUs and FTOs. We used R (R Core Team, 2021) with the rstanarm (Goodrich et al., 2020) and bridgesampling (Gronau et al., 2020) packages to build several Bayesian linear mixed effects models. We included random intercepts for conversation ID in all models, and sentence type (interrogative or declarative), speech act (question or statement) and the interaction between sentence type and speech act as fixed effects to predict FTOs. Sentence type and speech act were both contrast coded. We added fixed effects incrementally to a null model, testing if the inclusion of the additional term was justified by comparing the marginal likelihood of the data under each model.

In the second analysis, we included all 2465 available turn-end TCUs and their FTOs and included the probability of directness of speech act as a continuous predictor. The probability of directness is the probability (from DistilBERT) associated with the direct sentence type of the tagged speech act. We ran Bayesian linear mixed effects regression models including directness probability as a fixed factor and conversation ID as a random factor to predict FTOs.

Results

In our first analysis, the mean FTO following all turns was 75 ms. The mean FTO was 171 ms for grammatically interrogative turns and 64 ms for grammatically declarative turns. In

terms of speech acts, the mean FTO was 206 ms for questions and 60 ms for statements. Our dataset contained a total of. The descriptive statistics are presented in Table 15.

For our first analysis, the final model under which the data were most likely included speech act as a fixed effect. The summary for the final model is show in Table 16. The data were 522 times more likely under this model compared to the null model and 71 times more likely under this model compared to a model that included both sentence type and speech acts as fixed effects. See Figure 7 for a probability density distribution of FTOs following direct vs. indirect speech acts. The data under the speech act only model were also 4270 times more likely compared to a model that included the interaction between speech act and sentence type. See Figure 8 for a probability density distribution of FTOs following statements vs. questions. This provides decisive evidence that speech acts predict FTOs and that the interaction between sentence type and speech act does not. We also find very strong evidence that sentence type does not predict FTOs in conversation (Wetzels et al., 2011).

For our second analysis, the mean probability of directness of the speech act was 0.62, and the mean FTO was 63.35 ms. Descriptive statistics are shown in Table 17. The data were 5.7 times more likely under the null model, which included only random intercepts for conversation ID, than the model including the probability of directness. This constitutes moderate evidence in favor of the null model, replicating the above findings that speech act directness does not predict FTOs.

Discussion

The aim of the current study was to investigate whether comprehending indirect speech acts, whose function does not align with grammatical sentence type, requires additional cognitive processing compared to direct speech acts, whose function can be derived directly from sentence

type. We used timing data from natural conversations from a large corpus, predicting that if indirect speech acts require extra cognitive processing, interlocutors in conversation should take longer to respond to indirect vs. direct speech acts. Our data suggest that FTOs in natural conversation are sensitive to the preceding turn's speech act regardless of that turn's grammatical structure. These findings are not compatible with cognitive models positing that indirect speech act recognition involves reinterpretation based on Grice's maxim of manner. There is no delay in response to indirect speech acts compared to direct speech acts performing the same act in our data. Instead, we find that turn timing in conversation is sensitive to speech act independent of syntax, indicating that conversational structure orients to social speech act rather than linguistic structure. This finding highlights that conversation is structured by communication at the pragmatic level, and further constitutes evidence for a cognitive architecture oriented to speech acts.

Our primary interest in the current study was to investigate the difference in timing of indirect versus direct speech acts. It is, however, also worth comparing the timing differences between the speech acts analyzed in our data. Specifically, we found that people in conversations take longer to reply to speech acts that are questions compared to statements, regardless of directness. Other studies have found similar results: Gisladottir (2012), for example, found that offers take longer to process compared to answers. ERP data confirms this finding: offers have a later neural processing onset compared to answers, which may be explained by a questions having a more complex action sequence structure (Gisladottir et al., 2015). Questions may demand a more specific speech act response, such as an answer, compared to statements, requiring the listener to engage in more reasoning and processing. Furthermore, some questions contain critical information for answering said question only late in the turn or even at the very

end. Listeners may have deduced the speech act early on in the turn, but may not have received all of the information to answer until later in the turn, delaying the onset of their response planning (Bögels, Magyari, et al., 2015). Further research is needed to understand the relationship between different speech acts, the cognitive processing required to comprehend them, and the relationship between cognitive processing and FTOs in conversation.

While our study has high ecological validity, there are several limitations that warrant discussion. First, FTO may not be a sufficiently sensitive measure to capture any cognitive reanalysis after a turn end – incremental reanalysis of indirect speech acts may occur as the turn unfolds. Second, we do not control for other factors, such as preference organization, that affect timing in conversation: socially dispreferred actions (e.g. rejections or declination), for example, are delayed compared to preferred actions (e.g. acceptances) (Bögels, Kendrick, et al., 2015; Kendrick & Torreira, 2015; Mertens & De Ruiter, 2021; Pillet-Shore, 2017). Thirdly, it is possible that certain indirect speech acts are conventionalized, while others require Gricean reasoning for their interpretation, as proposed by Asher and Lascarides' model (2001). While we do not have a way for controlling for conventionalization, we find no evidence for directness of speech act affecting turn timing in conversation, suggesting that even indirect speech acts that are not conventionalized are processed within the same time frame as direct speech acts. Lastly, our data consists of telephone conversation, which may differ in systematic but hidden ways from face-to-face conversation, possibly affecting the timing of conversations.

We now turn to the theoretical implications of the current work. The fact that the timing of indirect speech acts and direct speech acts is the same in natural conversations suggests that they are processed similarly. Our findings do not support a cognitive model in which indirect speech acts are recognized through Gricean reinterpretation. Rather, our work suggests that interlocutors

orient directly and immediately to the social speech acts, rather than grammatical sentence type. This supports experimental evidence that speech acts are recognized both early in the turn and through a process that operates independently of syntax-level sentence comprehension. Speech acts form the social currency of communication, and listeners primarily orient to this social action dimension in conversation, rather than sentence form. The work presented here suggests that we have a cognitive architecture that is primarily oriented to the social action level in everyday conversation.

Chapter 5: The Relationship between Turn Length and Turn Timing in Conversation

The timing of one's turns is a critical matter in conversation. Across cultures and languages, turns in conversation are very tightly timed – speakers and listeners exhibit a strong preference for minimal gaps and overlaps (Sacks et al., 1974). This is particularly remarkable given the cognitive complexity of turn taking. Speakers must simultaneously perceive and process the incoming speech of their interlocutor's turn, prepare their own utterance in advance, and launch articulation at the correct time, all while monitoring hand gestures and other bodily actions linked to the speech. Turn timing is not only precise in that gaps are minimized; the length of these minimal gaps in conversation can also be a social signal in and of itself (Clayman, 2002), with longer gaps indicating an upcoming dispreferred turn structure, for example (Kendrick & Torreira, 2015). In the preceding chapter, I argued that the timing of turns in conversation is sensitive to speech acts independent of syntax. This indicates that the social actions of utterances inform conversational structure, and that our cognitive architecture is tailored towards processing speech acts beyond just sentence level meaning. Timing in conversation, according to these results, may therefore depend on social factors in interaction. In this chapter, I further explore this topic by investigating the extent to which turn timing in conversation is a consequence of cognitive processing of linguistic information versus a social signal.

The conversation analysis literature indicates that social factors such as conversational sequence structure and preference organization influence both the minimization of gaps as well as gap duration variation. According to this view, the highly precise timing of turn taking is not determined by the timing of cognitive processing demands, but rather by social norms. Interrupting is considered rude (Goldberg, 1990) and less sociable (Robinson & Reis, 1989), and

so people don't begin their turn as soon as they have planned their response, but wait until the current speaker has finished their turn (Bögels, 2020; Bögels, Magyari, et al., 2015) to begin speaking. Similarly, minimizing gaps is socially consequential: people perceive speakers who leave long pauses before beginning their turn as less willing to comply with requests (Roberts et al., 2006), and orient to gaps longer than 1 second as trouble indicative (Jefferson, 1989). Again, speakers time their turns not according to the timing of their cognitive processes, but according to the socially normative gap length. Speakers often begin their turns without having fully comprehended the preceding turn or planned their own current turn by using fillers such as "uh" or "um" to avoid silence. Lastly, precise turn timing is driven by our preference for progressivity (Stivers & Robinson, 2006) – the forward movement in conversation – bolstering social affiliation between interlocutors. Taken together, these findings suggest that timing in conversation is at least partially determined by socially normative turn timing.

Another facet of conversation that is pertinent to turn timing is the role of preference organization in speech act structure across turns. Utterances are often paired together in what are known as adjacency pairs (Levinson, 1983) – a first turn performs an initiating action that is then followed by a second turn performing a responding action. A question is typically followed by an answer, an invitation is followed by an acceptance or a declination, and a greeting is followed by a greeting, for example (Schegloff, 2007). With regard to timing, turns that are socially dispreferred, such as declining an offer or rejecting a proposal, are often delayed (Kendrick & Torreira, 2015) compared to socially preferred turns. This delay performs the social function of alerting the interlocutor to the valence of the responding action (Bögels, Kendrick, et al., 2015). Turn-timing, according to the conversation analytic perspective, is not just a by-product of cognitive processing, but an intentional signal (see e.g., Clark & Fox Tree, 2002).

While social factors clearly play a role in the timing of turns in conversation, it is likely that there is also a relationship between cognitive processing demands and the duration of turn transitions. Conversation is cognitively complex in that interlocutors must engage in many cognitive processes simultaneously, and processing takes time (Donders, 1969; Zhang et al., 2018). Psycholinguistics research shows that several variables present in conversation increase language processing times. Words that are more frequent, for example, take less time to process and respond to across a variety of experimental tasks (Monsell et al., 1989; see Brysbaert et al., 2018 for a review), words that are more concrete correlate with faster lexical decision times (Schwanenflugel et al., 1988), and syntactically more complex sentences are more difficult to produce and understand (Kemper et al., 1989). Turns that are more surprising, i.e. carry more information, are more difficult to understand and more likely to be targeted by repair in conversation (Mertens, 2022). Lastly, sentences that are longer require more cognitive effort to process, as measured with pupillometry (Piquado et al., 2010). Taken together, these findings suggest that demands on processing such as those listed above should influence turn transition times: more cognitively demanding turns should be followed by a longer gap. The duration of gaps between turns may therefore not only serve as a social signal but may also reflect the amount of processing required to comprehend the previous turn and plan the upcoming turn.

Teasing apart social effects and cognitive processing effects in turn transition timing is not trivial. A handful of corpus studies have, however, directly compared a variety of cognitive and social variables, and have found mixed results. Mertens and de Ruiter (2021) examined the effects of cognitive versus social factors in the timing of other initiated repair (OIR) – when a listener notices and signals a comprehension problem (e.g. “What?”). The authors included four cognitive variables (turn duration, word frequency, syllable rate and availability of discourse

context), two social variables (the duration ratio of the repair solution to the trouble source turn and OIR specificity (open versus closed)) and one variable that could be both cognitive and social: overlapping talk. Both cognitive and social factors predicted Floor Transfer Offset (FTO) timing, in particular turn duration and OIR specificity. FTOs were longer when turns were shorter and when OIRs were less specific. Roberts et al (2015) conducted a similar analysis investigating the role of sequence organization and a battery of processing variables known to affect language processing using a large corpus of conversation. Using a random forest analysis, the authors calculated the relative contributions of cognitive planning and comprehension factors such as turn duration, speech rate, concreteness, frequency, surprisal and information density to FTO timing. They compared these cognitive factors to a series of social sequence organization factors including indexical properties of the speaker, laughter, adjacency pairs and backchannels. The results show that turn duration, speech rate, and speaker sex were the most important factors in predicting FTOs. Together, these studies suggest that both processing constraints and social interactional constraints are important for the timing of turn taking.

Of note across the two corpus studies mentioned above is the relationship between turn length and FTO. Both Mertens and de Ruiter (2021) and Roberts et al. (2015) found that turn duration, above and beyond a host of other variables included in their analyses, predicted the subsequent FTO, but found opposite effects. Mertens and de Ruiter (2021) found an inverse relationship: the longer the turn the shorter the FTO, while Roberts et al. (2015) found a non-linear relationship: overall, longer turns are followed by longer FTOs, with the exception of very short turns such as backchannels and agreements, which are followed by longer FTOs. Mertens and de Ruiter (2021) measured turn length in duration while Roberts et al. (2015) measured turn length in number of syllables. Interestingly, in an experimental lab setting in which participants

have to press a button at the end of a turn rather than respond verbally, the longer the turn is, the faster participants respond (Corps et al., 2019; De Ruiter et al., 2006). Based on this literature, which includes both experimental and corpus data, the relationship between turn length and turn timing is not clear.

From a cognitive perspective, turn length could have two effects on turn timing. Longer utterances may result in longer subsequent FTOs because a longer turn is likely to be more complex, containing more information, and therefore requiring more cognitive processing to comprehend (Roberts et al., 2015). This perspective is supported by pupillometry data showing that participants have increased pupil size, indicating greater cognitive load (see Beatty, 1982, for a review), while comprehending longer sentences (Piquado et al., 2010). On the other hand, a longer turn may be followed by shorter FTOs for two reasons. First, longer turns gives the listener more time to plan their own turn, allowing listeners to begin speaking sooner. Second, longer turns typically contain more linguistic information to draw from when generating predictions about the turn-end. Words in an utterance become more predictable as the turn unfolds (Van Petten & Kutas, 1990), thus words at the end of longer turns are most predictable, contributing to more accurate turn-end estimation. In sum, the fact that longer turns contain more information may require increased processing, leading to longer FTOs, or may decrease processing because more information allows for increased turn-end prediction.

The aim of the work presented in this chapter is twofold. The primary goal is to clarify the relationship between turn length and turn timing by examining multiple measures of length in a large corpus of conversation. Specifically, I am interested in whether turn length predicts turn timing, or whether turn timing operates independently of turn length, indicating social processing rather than information processing. The second goal is to compare the relationship between turn

length and turn timing across experimental and natural dialogue contexts to evaluate how well turn-taking behavior generalizes from the lab to real world conversation.

Study 1

Method

The data for this study came from de Ruiter et al. (2006), in which the authors conducted an experiment to measure the timing of participants' end-of-turn estimation to natural turns. Participants listened to utterances taken from natural telephone conversations and were instructed to *anticipate* the moment at which the turn would end using a button press. The original study focused on comparing the turn-end estimation times to stimuli that were modified to reduce/remove lexical information. In this particular reanalysis, I focus on de Ruiter et al.'s natural condition, in which stimuli were not subject to any acoustic alterations. This dataset closely mimics the constraints of real conversation, in which interactants must anticipate the end of their partner's turn in order to plan their response and respond within the appropriate time constraints of natural dialogue. I use the same measure of turn timing as in chapter 3: *bias*, i.e. the time between the end of the turn and the button press. A negative *bias* indicates a button press before the end of the turn, and a positive *bias* indicates a button press after the end of the turn. I am specifically interested in comparing the relationship between *bias* and stimulus duration, as these measurements are analogous to FTO and turn length in natural conversation.

In this study, I also compare the relationship between turn length and response time between experimental and natural settings. In order to directly compare these two environments, I obtained the FTOs from the natural conversations that de Ruiter et al.'s stimuli were extracted from. That is, for the exact same turns, I analyzed participants' *bias* from the experiment as well as the time it took the interlocutor to respond to that turn in the original (natural) telephone

conversations, allowing for a direct comparison between turn length and turn timing across experimental and natural settings.

Data Cleaning and Analysis

Any *bias* and FTO values greater than 2000ms and less than -2000ms were excluded. The rationale for this criteria is that gaps greater than 1000 ms are very rare in conversation (Sacks et al., 1974), and people orient to gaps that are approximately 1 second or longer as trouble indicative (Jefferson, 1989). It is therefore reasonable to assume that FTOs and *bias* values greater than 2000ms no longer constitute typical turn-taking behavior. Similarly, overlaps greater than -2000ms likely reflect a false start rather than a response to a previous turn. For the experimental *bias* data, 131 data points were excluded, leaving 2129 data points total. For the natural FTO data, 60 data points were excluded, leaving 48 data points total. We conducted a Bayesian linear mixed effects regression for the experimental data, with *bias* as the outcome variable and turn length as the predictor. We used JASP (JASP Team, 2021) to calculate the model and R version 4.11 (R Core Team, 2021) to calculate Bayes factors using the rstanarm (Goodrich et al., 2020) and bridgesampling (Gronau et al., 2020) packages. Random intercepts were included for both participants and items. For the natural FTO data, we conducted a Bayesian regression with FTO as the outcome variable and turn length as the predictor using JASP (JASP Team, 2021). All evidence categories for Bayes Factors in this chapter are taken from Wetzels et al. (2011).

Results

Descriptive statistics are shown in Table 18. Overall, the average turn length was 2644.74 ms. The mean FTO in the natural data was 18.68 ms and the mean *bias* from the experimental data was 20.37. The experimental *bias* data were 1.78×10^{17} more likely under a model that

included turn length as a predictor compared to the null model, providing decisive evidence that turn length predicts *bias.* The relationship between turn length and *bias* was negative such that a 1 millisecond increase in turn duration was associated with a 0.08 millisecond decrease in *bias*. The longer the turn, the more accurate participants were at estimating its end. A scatterplot of the data are shown in Figure 9; a summary of the model is shown in Table 19. For the natural data taken from the original telephone conversations, we regressed FTO on turn duration and compared this model to a null model and found only anecdotal evidence that turn length is associated with FTO ($R^2 = 0.003$, $BF_{inclusion} = 1.459$). A scatterplot of the data is shown in Figure 10. See Table 20 for the regression model summary.

Discussion

The results presented here indicate different effects of turn length on turn timing in experimental versus natural settings even when the utterances are identical. In an experimental setting, longer turns are associated with shorter *bias* (time elapsed between the end of the turn and participants' button press). In natural dialogue, the relationship between turn length and FTO, i.e. when the next speaker initiated their turn, is weaker. The plotted data in Figures 9 and 10 illustrates this difference.

One explanation for the finding that longer turns are followed by shorter FTOs in the experimental setting is that longer turns contain more linguistic information to draw from when generating predictions about the turn-end. Words in an utterance become more predictable as the turn unfolds (Van Petten & Kutas, 1990), thus words at the end of longer turns are more predictable, contributing to more accurate turn-end estimation. Participants in the experiment were explicitly tasked with *anticipating* the turn end, so may rely on cues in the turn that would allow them to do so. Participants in the telephone conversations, however, were not explicitly

tasked with anything besides chatting with each other. While the task of turn-end-estimation approximates a cognitive task that we perform in informal everyday conversation, explicitly asking participants to anticipate a turn end with a button press is quite different from timing one's utterance implicitly in a conversation. Importantly, people in the natural conversations are interacting socially – they are conversing with an interlocutor, introducing a set of social expectations, preferences and dynamics that are not present in the experimental setting. Getting one's timing right in this setting is socially consequential, as opposed to a laboratory setting.

It is important to note that the dataset for the FTOs from the telephone conversations is small: while there are many *bias* data points for the same turn due to the repeated measures paradigm, there is only one FTO value for the same turn in the natural data. To more closely investigate how these findings translate to natural dialogue Study 2, as described below, examines the relationship between turn length and FTO in a much larger dataset and includes additional measures of length beyond raw duration: number of syllables and number of words. While turns that are longer in duration are likely to have more words and more syllables, speakers sometimes pause such that utterance duration increases while the number of syllables and words stays the same. Similarly, a turn composed few slow syllables may have the same duration as a turn consisting of many fast syllables. Furthermore, under the hypothesis that interlocutors anticipate the other speaker's end of turn, turns that contain a large number of syllables but relatively few words would have larger FTOs than turns that contain as many syllables but more words for the same duration. Since I am interested in turn length as a proxy for the amount of information contained in a turn, it is critical to accurately capture turn length using a variety of measures.

Study 2

Method

Conversation data came from the Switchboard corpus (Godfrey et al., 1992) containing 260 hours of telephone conversations between speakers of American English. Using an approach similar to Chapter 4, we segmented each turn into turn construction units (TCU) based on the Stanford University transcriptions (Jurafsky, Coccaro, et al., 1997) and included only turn final TCUs in our dataset. FTO duration was derived from the MSU transcriptions (*OSLR*, 2023).

Turn length was calculated in three ways: number of words, number of syllables, and duration in milliseconds. Turn duration was extracted from MSU transcriptions (*OSLR*, 2023) of the Switchboard corpus by subtracting the end time of the last word of the TCU from the start time of the first word of the TCU. In order to calculate the number of words per TCU, we identified word boundaries as strings of space-separated characters. Contractions such as “I’m”, “it’s” and “wasn’t” were each counted as two words. To count syllables, we looked up each word in the turn in the Carnegie Mellon Pronouncing Dictionary (*The CMU Pronouncing Dictionary*, n.d.), an open source machine-readable dictionary of North American English containing over 134,000 words and their pronunciations, via the Python cmudict package (Day, 2021). If the word was not contained in the dictionary, we attempted to match it against the alternative pronunciation contained in the Open Speech and Language Resources (OSLR) MSU Switchboard transcripts (*OSLR*, 2023). In 0.3% of cases a syllable count was not found in either of these sources, so we estimated the number of syllables through a vowel-pattern counting heuristic. Finally, duration was measured as the time elapsed (in milliseconds) between the onset of the first word and the offset of the final word in the TCU.

Data Cleaning and Analysis

As in Study 1, all FTOs greater than 2000ms and less than -2000ms were removed from the dataset. We also excluded backchannels, which speakers use to signal that they have understood their interlocutor's turn and that they do not wish to take a full turn (Schegloff, 1982). Using these criteria we excluded 1872 data points, leaving us with a total of 3992 TCUs and their subsequent FTOs in our analysis. We analyzed the data with Bayesian linear mixed effects regression models (Baayen et al., 2008). We used JASP (JASP Team, 2021) to calculate intercepts and R version 4.11 (R Core Team, 2021) to calculate Bayes factors using the rstanarm (Goodrich et al., 2020) and bridgesampling (Gronau et al., 2020) packages. We built a separate model for each fixed effect predictor (number of words, number of syllables, and duration in milliseconds) and compared that model to a null model. We then tested if the inclusion of the predictor was justified by comparing the marginal likelihood of the data under each model. All models included random intercepts for conversation ID, i.e. the conversation the turns were taken from.

Results and Discussion

Descriptive statistics are shown in Table 21. Overall, the average turn length was 2280.769 ms, the average number of words per turn was 8.26, and the average number of syllables was 10.16. The mean FTO was 8.39 ms; the distribution of FTOs is shown in Figure 11. The data were 5.16×10^{30} times more likely under a model that included number of words as a predictor compared to the null model, providing decisive evidence that number of words predicts FTOs. Each additional word in a TCU is associated with a 4.26 millisecond increase in FTO. See Figure 12 for a scatterplot of the data. A summary of the model is shown in Table 22. The data were 9.46×10^{27} times more likely under a model that included number of syllables as a predictor compared to the null model, again providing decisive evidence that number of syllables

predicts FTOs. Each additional syllable in a TCU is associated with a 3.51 millisecond increase in FTO. See Figure 13 for a scatterplot of the data and Table 23 for a summary of the model. Lastly, the data were infinitely more likely under a model that included TCU duration (in milliseconds) compared to the null model. This provides decisive evidence that duration predicts FTOs. A one millisecond increase in TCU duration is associated with a 0.01 millisecond increase in FTO. See Figure 14 for a scatterplot and Table 24 for a summary of the model. Evidence categories for all Bayes factors were, again, taken from Wetzels et al. (2011).

In the current analysis of turn length and turn timing in the Switchboard corpus, we find a positive relationship between a turn's length and the amount of time it takes for the next speaker to respond. Longer TCUs, as measured in milliseconds, number of words, and number of syllables, are followed by longer FTOs. We hypothesized that longer turns contain more information and could therefore have two effects on FTO timing: longer turns could increase FTOs due to increased processing, or could decrease FTOs because longer turns would allow for better turn-end estimation due to facilitated prediction, allowing listeners more time to plan their turn. Our results show that longer turns are followed by longer FTOs, supporting an account that longer turns contain more information, which takes longer to process cognitively. We discuss implications and possible limitations of this conclusion below.

General Discussion

This chapter has examined the relationship between turn length and turn timing across various contexts and various measures of length. Longer turns, especially when measured not just by their duration but also their number of syllables and words, contain more information. In Study 1 we found that in an experimental lab setting, participants are better at estimating the end of a turn when the turn is longer: longer turns lead to faster responses. We also found that for

those same turns in the context of their original conversational context, turn length was not correlated with FTO. We attribute these differences to the different contexts (experiment versus natural dialogue). When participants are explicitly asked to *anticipate* the end of a turn, i.e. anticipation is the goal, more information in the longer turns allows participants to predict the turn's ending more accurately. In Study 2, which investigated turn-taking behavior in a large corpus of natural conversations, we find the opposite effect: longer turns are followed by longer FTOs. We interpret this finding as an effect of cognitive processing: in the context of real conversation that doesn't involve explicit anticipation task but does involve planning a socially relevant turn, more information takes longer to process and respond to, so listeners initiate their turns later after a longer turn.

One limitation of the current research is the suitability of the measures used. We used timing (FTO) as an indicator of cognitive processing. Timing in conversation is not a direct measure of cognitive activity, but a measure that is known to correlate with cognitive processing: the longer participants take to initiate a response, the more cognitive processing they are likely engaging in (Donders, 1969). A more direct on-line measure cognitive activity, such as pupil dilation or neural activity, would provide more conclusive evidence, and would allow us to directly see if participants do indeed demonstrate increased cognitive processing when FTOs are longer. A drawback of such an approach is that the high level of experimental control would likely compromise the high ecological validity that the current approach prioritizes. Finding a balance between ecological validity and experimental control is one of the biggest challenges in conversation research, and striking a balance between the two is critical to future research.

We used turn length as a measure of information content of the turn. Again, turn length is not a direct measure of the amount of information contained in a turn, but we believe it is a

suitable proxy. Corpus analyses have found that shorter words are correlated with lower Shannon information content than longer words (Piantadosi et al., 2011), even when controlling for word meaning (Mahowald et al., 2013). Behavioral work also shows that people use short forms of longer words (e.g. *chimp* vs. *chimpanzee*) more often in predictive contexts in order to be information-theoretically efficient (Mahowald et al., 2013). Our syllable measure in particular should capture this relationship between word length and information content. Other studies have explicitly measured information in conversation by measuring surprisal values for each word in a turn and averaging across words go attain information for the whole turn (e.g. Roberts et al. 2015). Word by word surprisal values can be calculated from large corpora of English using large language models such as GPT. Recent work, however, demonstrates that surprisal as calculated by GPT, for example, does not necessarily translate accurately to behavior observed in spoken dialogue. Large language models are primarily trained on written text, and do not capture dialogic behavior (Mertens et al., submitted). We use language differently when writing compared to when speaking in conversation, making it difficult to generalize measures of information content from such models to behavior demonstrated in everyday conversation.

Lastly, it is important to note here that in Study 2, though our measures of turn length (duration, words and syllables) all predict FTO. A number of variables are known to affect the speed of responses in conversation that we did not take into account including syntactic complexity (Piquado et al., 2010; Roberts et al., 2015), the identity of the speaker (Roberts et al., 2015), preference structure (Kendrick & Torreira, 2015; Mertens & De Ruiter, 2021) and dialogue act type (Bögels, Kendrick, et al., 2015). While we found a relationship between turn length and FTO, our results contribute to a body of literature suggesting that variation in gap length is not determined by any one variable, let alone by exclusively cognitive or social factors.

In fact, constraints on processing and social factors such as preference organization may not be entirely distinct mechanisms. It could be the case, for example, that socially preferred turns are both faster in their onset in conversation due to social signaling, but also because they are easier to produce. Further ecologically valid experimental and corpus research is needed to finely tease apart social and cognitive variables and their relative contributions to FTO variation in dialogue.

In conclusion, we have shown that longer turns in conversation lead to longer FTOs in natural conversation. Previous research has found different effects of turn lengths on FTO, partially due to measuring length in inconsistent ways and only examining the relationship between length and FTO in highly specific contexts such as repair or turn-taking experiments. By measuring length across a series of variables in a large corpus and directly comparing response times to the same turns in experiments versus natural dialogue, we show that turn length has a relationship FTOs, but that is effect may not be discernible in task-based turn-taking experiments. These findings warrant two conclusions. First, the temporal patterns of dialogue can, at least partially, be accounted by the length of a turn, which is most likely an effect of cognitive processing demands. Second, experimental findings from laboratory studies do not necessarily represent mental processes or behavior observed in real world conversation. Verifying lab findings through corpus work and thus optimizing internal and ecological validity is critical.

Chapter 6: Concluding Remarks

Language is a social tool. Interacting with other people through language entails a deep consideration of the broader social context. The overarching goal of this dissertation was to examine how high-level social processing impacts lower-level conversational mechanics. The introductory chapter outlines the social nature of language learning and use. I review how the rich structure of conversation is shaped by social preferences and facilitates socially affiliative interaction. Here I lay the groundwork for conceptualizing language use in the context of social interaction, the ecological niche of language. Chapter 2 presents experimental evidence showing that top-down knowledge of sequence organization in conversation informs how and when we perceive a speaker switch, a critical skill in everyday conversation. In chapter 3, I show that turns in conversation are predicted at the speech act level, extending top-down prediction in language comprehension beyond the sentence level and across turn and speaker boundaries in conversation. Listeners anticipate the social content of utterances. Chapter 4 further supports this point by directly comparing direct vs. indirect speech acts and their respective cognitive processing demands in natural dialogue. Speech acts are not recognized *by* their sentence type, but *in spite of* their sentence type. Again, higher-level pragmatic expectations inform how incoming linguistic input is incrementally integrated at lower levels. Lastly, in chapter 5, I explore the relationship between turn length and turn timing in experimental versus natural social settings and find that longer turns are followed by longer gaps in natural conversation but not in laboratory experiments**.** Together, the work presented here suggests that the social dynamics of interaction shape how we understand speech and language in conversation.

Language in dialogue entails a many-to-many mapping system between form and meaning. Questions are not always interrogative, and statements are not always declarative.

Understanding the meaning of an utterance in conversation is critical for conversation, yet the ambiguous link between form and function renders the process of meaning extraction difficult. The work I have presented here, particularly in chapters 2 and 3, suggests that listeners draw on their knowledge of the normative social organization of conversation to anticipate upcoming speech acts in dialogue. These predictions, in turn, shape how the incoming turn is processed. The findings of my dissertation contribute supporting evidence to the hypothesis that predictive processes facilitate speech act recognition, providing a plausible explanation for smooth and effortless conversation in the face of the vast ambiguity of language. The work in these chapters also suggests that listeners assume cooperativity in conversation: when a turn is socially marked given the context, listeners interpret the utterance in such a way that it still fits the social context. Top-down pragmatic knowledge shapes how and what we hear, streamlining optimal social communication.

Traditionally, both language and our language comprehension systems have been described in terms of levels of structure that are represented mentally. This dissertation contributes to understanding the relationship between higher levels of representation, specifically the relationship between sentences and speech acts. Every turn in a conversation performs a social action, and but that action cannot necessarily be derived from information at the sentence level. My findings in chapter 3 suggests that we mentally represent speech acts – they are not just inferred from lower levels of representation, and, in fact, often can't be. My findings in chapter 4 support this: indirect speech acts are processed within the same time frame as direct speech acts. Listeners likely process the social action right away, rather than first extracting the literal sentence level meaning of utterances then inferring the social action. This has broader implications for understanding the cognitive underpinnings of language comprehension: our

cognitive system represents speech acts and uses those representations to facilitate conversation comprehension.

If we want to understand how language is used and understood, we must study it in ecologically valid environments. The bulk of language research is conducted in relatively contrived and artificial experimental contexts using isolated, slowly presented and often unrelated linguistic stimuli. While these contexts allow for tight experimental control and thus high internal validity, this setting is quite different from natural conversation, where interlocutors in a social environment rapidly exchange utterances. These differences make it difficult to generalize findings to real-world communication. In this dissertation, I combine tightly controlled experimental designs that emulate everyday conversation with corpus-based research from large datasets of natural dialogue. Chapter 5 finds differences in conversational timing across experimental and natural settings, underscoring the importance of studying language processing in the context of spontaneous conversation. My work combines experiments with data from corpora of natural conversations, merging the strengths of both methods to optimize internal and ecological validity.

Two broader limitations exist within the work presented here. First, the experiments employ an overhearer paradigm where participants listen to conversations rather than participate in them. By necessity, my research assumes that the cognitive processes of someone listening to a conversation are similar to those involved in the conversation, but this may not be the case. Social cognition is different when we are interacting with others compared to merely observing interactions (Schilbach et al., 2013) and the very process of understanding is different for overhearers compared to those being addressed in conversation (Schober & Clark, 1989), impacting the generalizability of these experiments. Future work should expand on preliminary

research such as Bögels et al. (2020) measuring cognitive activity and behavioral responses of participants actively engaged in a conversation. Second, the corpus analyses included in the last two chapters of my dissertation are limited in terms of the size and homogeneity of the data. The Switchboard Corpus only contains telephone conversations recorded in North America, and our subset of the data in Chapter 4 specifically is small. This work should be expanded to include more diverse data from different languages and cultures, and different types of conversations, especially in-person conversations.

Lastly, I turn to the broader applications of the research presented in this dissertation. First, my dissertation provides important insight for computational models of language processing. Some models, particularly Large Language Models (LLMs) such as GPT, are now widely available and routinely used. While LLMs can generally predict human-like writing, they are unable to predict or produce the dynamics of dialogue (Mertens et al., submitted). LLMs only have feedforward connections, but producing and comprehending language, especially dialogue, involves both feedforward and feedback connections. Predictions at higher levels affect lower-level activity. Understanding the meaning of a turn in conversation entails realizing its speech act, which humans do through top-down anticipation. Determining when a speaker changes in conversation also requires drawing on higher level representations of normative sequence organization. Without linguistic representations at multiple levels, including the pragmatic level, it may not be possible to emulate human-like conversation. Understanding the cognitive basis of how humans process conversation can create better informed computational models of language. Secondly, the findings of my dissertation have clinical applications. Autism spectrum disorder (ASD) is *often* characterized by impaired social interaction and communication. Children with ASD, for example, show different use and processing of speech acts compared to typically

developing children (Bauminger-Zviely et al., 2017). Language abnormalities in people with schizophrenia also include atypical building of higher-order meaning (Kuperberg, 2010). If we can understand the cognitive (and neural) architecture of the normal communication system, particularly as it pertains to social communication, we can compare it to the communication systems of those with atypical social communication. This can inform both diagnosis and treatment.

The work included in this dissertation suggests that the social dynamics of interaction play a significant role in language comprehension in conversation. Our social cognition has a top-down influence on lower-level processing in conversation. We are social animals, and this has consequences for our cognition. Language is a tool we use for social communication.

References

Albert, S., & De Ruiter, J. P. (2018). Repair: The Interface Between Interaction and Cognition. *Topics in Cognitive Science*, *10*(2), 279–313. https://doi.org/10.1111/tops.12339

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm.' *Cognition*, *93*(2), B79–B87. https://doi.org/10.1016/j.cognition.2004.02.005

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

Asher, N., & Lascarides, A. (2001). Indirect Speech Acts. *Synthese*, *128*(1), 183–228. https://doi.org/10.1023/A:1010340508140

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

Balcombe, J. P., & McCracken, G. F. (1992). Vocal recognition in mexican free-tailed bats: Do pups recognize mothers? *Animal Behaviour*, *43*(1), 79–87. https://doi.org/10.1016/S0003-3472(05)80073-9

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next Speakers Plan Their Turn Early and Speak after Turn-Final "Go-Signals." *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00393

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The Timing of Utterance Planning in Task-Oriented Dialogue: Evidence from a Novel List-Completion Paradigm. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01858

Baumann, O., Belin, Æ. P., Baumann, O., Belin, P., & Baumann, O. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110.

Bauminger-Zviely, N., Golan-Itshaky, A., & Tubul-Lavy, G. (2017). Speech Acts During Friends' and Non-friends' Spontaneous Conversations in Preschool Dyads with High-Functioning Autism Spectrum Disorder versus Typical Development. *Journal of Autism and Developmental Disorders*, *47*(5), 1380–1390. https://doi.org/10.1007/s10803-017-3064-x

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

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

Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition*, *203*, 104347. https://doi.org/10.1016/j.cognition.2020.104347

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia*, *109*, 295–310. https://doi.org/10.1016/j.neuropsychologia.2017.12.028

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

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Sci Rep*, *5*, 12881. https://doi.org/10.1038/srep12881

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

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Carminati, M. N., & Knoeferle, P. (2013). Effects of Speaker Emotional Facial Expression and Listener Age on Incremental Sentence Processing. *PLOS ONE*, *8*(9), e72559. https://doi.org/10.1371/journal.pone.0072559

Casasanto, L. S. (2008). Does Social Information Influence Sentence Processing? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*.

Casasanto, L. S. (2010). What do Listeners Know about Sociolinguistic Variation? *University of Pennsylvania Working Papers in Linguistics*, *15*(2).

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.

Cassidy, B. S., & Gutchess, A. H. (2012). Social relevance enhances memory for impressions in older adults. *Memory*, *20*(4), 332–345. https://doi.org/10.1080/09658211.2012.660956

Chhabra, S., Badcock, J. C., Maybery, M. T., & Leung, D. (2012). Voice identity discrimination in schizophrenia. *Neuropsychologia*, *50*(12), 2730–2735. https://doi.org/10.1016/j.neuropsychologia.2012.08.006

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, H., & Wilkes-Gibbs, D. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 183–194.

Clarke, J., Gaudrain, E., Chatterjee, M., & Başkent, D. (2014). T'ain't the way you say it, it's what you say – Perceptual continuity of voice and top–down restoration of speech. *Hearing Research*, *315*, 80–87. https://doi.org/10.1016/j.heares.2014.07.002

Clayman, S. E. (2002). Sequence and solidarity. In *Advances in Group Processes* (Vol. 19, pp. 229–253). Emerald Group Publishing Limited. https://doi.org/10.1016/S0882-6145(02)19009-6

Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating Utterances During Turn-Taking: The Role of Prediction, Response Preparation, and Articulation. *Discourse Processes*, *55*(2), 230–240. https://doi.org/10.1080/0163853X.2017.1330031

Corps, R. E., Pickering, M. J., & Gambi, C. (2019). Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience*, *34*(5), 615–627. https://doi.org/10.1080/23273798.2018.1552008

Cummins, C., & De Ruiter, J. P. (2014). Computational Approaches to the Pragmatics Problem. *Linguistics and Language Compass*, *8*(4), 133–143. https://doi.org/10.1111/lnc3.12072

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913–2922. https://doi.org/10.1121/1.1616924

Davies, B. L. (2007). Grice's Cooperative Principle: Meaning and rationality. *Journal of Pragmatics*, *39*(12), 2308–2331. https://doi.org/10.1016/j.pragma.2007.09.002

Day, D. L. (2021). *cmudict: "A versioned python wrapper package for The CMU Pronouncing Dictionary data files."* [Computer software]. https://github.com/prosegrinder/python-cmudict

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

DeCasper, A. J., & Fifer, W. P. (1980). Of Human Bonding: Newborns Prefer Their Mothers' Voices. *Science*, *208*(4448), 1174–1176. https://doi.org/10.1126/science.7375928

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513–1528. https://doi.org/10.1121/1.4906837

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

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Durlach, N. (2006). Auditory masking: Need for improved conceptual structure. *The Journal of the Acoustical Society of America*, *120*(4), 1787–1790. https://doi.org/10.1121/1.2335426

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

Enfield, N. J., Stivers, T., & Levinson, S. C. (2010). Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, *42*(10), 2615–2619. https://doi.org/10.1016/j.pragma.2010.04.001

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed Attention Eliminates 'Change Deafness' in Complex Auditory Scenes. *Current Biology*, *15*(12), 1108–1113. https://doi.org/10.1016/j.cub.2005.05.051

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

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

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When Less is Heard than Meets the Ear: Change Deafness in a Telephone Conversation. *Quarterly Journal of Experimental Psychology*, *64*(7), 1442–1456. https://doi.org/10.1080/17470218.2011.570353

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

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*(6), 725–745. https://doi.org/10.1016/0749-596X(91)90034-H

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *System*, *32*(3), 363–377. https://doi.org/10.1016/j.system.2004.05.002

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Garrod, S., & Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, *1*(2), 292–304. https://doi.org/10.1111/j.1756-8765.2009.01020.x

Gaudrain, E., Li, S., Shen Ban, V., & Patterson, R. D. (2009, September). The role of glottal pulse rate and vocal tract length in the perception of speaker identity. *Interspeech 2009*. https://hal.archives-ouvertes.fr/hal-02144510

Gazdar, G. (1981). Speech Act Assignment. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of Discourse Understanding*. Cambridge University Press.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*(7272), Article 7272. https://doi.org/10.1038/nature08572

Giles, H., Coupland, J., & Coupland, N. (Eds.). (1991). *Contexts of Accommodation: Developments in Applied Sociolinguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511663673

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

Gobel, M. S., Tufft, M. R. A., & Richardson, D. C. (2018). Social Beliefs and Visual Attention: How the Social Relevance of a Cue Influences Spatial Orienting. *Cognitive Science*, *42*(S1), 161–185. https://doi.org/10.1111/cogs.12529

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 517–520 vol.1. https://doi.org/10.1109/ICASSP.1992.225858

Goffman, E. (1967). On face-work. *Interaction Ritual*, 5–45.

Goldberg, J. A. (1990). Interrupting the discourse on interruptions. *Journal of Pragmatics*, *14*(6), 883–903. https://doi.org/10.1016/0378-2166(90)90045-F

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

Gordon, D., & Lakoff, G. (1975). Conversational Postulates. In Cole, Peter & Morgan, Jerry L. (Eds.), *Syntax and Semantics Volume 3: Speech Acts* (pp. 83–106). Academic Press. https://doi.org/10.1163/9789004368811_005

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

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388. https://doi.org/10.2307/2182440

Grice, H. P. (1975). Logic and Conversation. *Speech Acts*, 41–58. https://doi.org/10.1163/9789004368811_003

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

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

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. *The Journal of Neuroscience*, *38*(35), 7585–7599. https://doi.org/10.1523/JNEUROSCI.0065-18.2018

Hadley, L. V., Fisher, N. K., & Pickering, M. J. (2020). Listeners are better at predicting speakers similar to themselves. *Acta Psychologica*, *208*, 103094. https://doi.org/10.1016/j.actpsy.2020.103094

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, *304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615. https://doi.org/10.1016/j.jml.2007.01.008

Harness Goodwin, M., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica*, *62*(1–2). https://doi.org/10.1515/semi.1986.62.1-2.51

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

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Heritage, J. (2007). Intersubjectivity and Progressivity in References to Persons (and Places). In *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives* (pp. 255–280). Cambridge University Press.

Hillenbrand, J., Getty, L. A., Wheeler, K., & Clark, M. J. (1994). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *95*(5), 2875–2875. https://doi.org/10.1121/1.409456

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A., Ward, J., & Gunter, T. (2012). Gesture Facilitates the Syntactic Analysis of Speech. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00074

Holtgraves, T. (2008a). Automatic intention recognition in conversation processing. *Journal of Memory and Language*, *58*(3), 627–645. https://doi.org/10.1016/j.jml.2007.06.001

Holtgraves, T. (2008b). Conversation, speech acts, and memory. *Memory & Cognition*, *36*(2), 361–374. https://doi.org/10.3758/MC.36.2.361

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

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

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

JASP Team. (2021). *JASP* (0.16) [Computer software].

Jefferson, G. (1989). Notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An interdisciplinary perspective* (Vol. 3, pp. 166–196). Multilingual Matters.

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

Jurafsky, D., Coccaro, N., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., & Ess-Dykema, C. V. (1997). *Johns Hopkins LVCSR Workshop-97 Switchboard Discourse Language Modeling Project Final Report*. 50.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *WS-97 Switchboard DAMSL Coders Manual*. https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

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

Kemper, S., Kynette, D., Rash, S., O'Brien, K., & Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, *10*(1), 49–66. https://doi.org/10.1017/S0142716400008419

Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., Hoey, E., Hoymann, G., Manrique, E., Rossi, G., & Levinson, S. C. (2020). Sequence organization: A universal infrastructure for social action. *Journal of Pragmatics*, *168*, 119–138. https://doi.org/10.1016/j.pragma.2020.06.009

Kendrick, K. H., & Torreira, F. (2015). The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes*, *52*(4), 255–289. https://doi.org/10.1080/0163853X.2014.955997

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580. https://doi.org/10.1073/pnas.0705345104

Knoeferle, P., & Crocker, M. W. (2006). The Coordinated Interplay of Scene, Utterance, and World Knowledge: Evidence From Eye Tracking. *Cognitive Science*, *30*(3), 481–529. https://doi.org/10.1207/s15516709cog0000_65

Knoeferle, P., Urbach, T. P., & Kutas, M. (2014). Different mechanisms for role relations versus verb–action congruence effects: Evidence from ERPs in picture–sentence verification. *Acta Psychologica*, *152*, 133–148. https://doi.org/10.1016/j.actpsy.2014.08.004

Koelewijn, T., Gaudrain, E., Tamati, T., & Başkent, D. (2021). The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception. *The Journal of the Acoustical Society of America*, *150*(3), 1620–1634. https://doi.org/10.1121/10.0005938

Krauss, R. M., & Fussell, S. R. (1991). Perspective-Taking in Communication: Representations of Others' Knowledge in Reference. *Social Cognition*, *9*(1), 2–24. https://doi.org/10.1521/soco.1991.9.1.2

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K. (2011). Early Language Learning and Literacy: Neuroscience Implications for Education. *Mind, Brain, and Education*, *5*(3), 128–142. https://doi.org/10.1111/j.1751-228X.2011.01121.x

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

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. (2010). Language in Schizophrenia Part 2: What Can Psycholinguistics Bring to the Study of Schizophreniaand Vice Versa? *Language and Linguistics Compass*, *4*(8), 590–604. https://doi.org/10.1111/j.1749-818X.2010.00217.x

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. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

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

Kuwabara, H., & Takagi, T. (1991). Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Communication*, *10*(5–6), 491–495. https://doi.org/10.1016/0167-6393(91)90052-U

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

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

Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing—Evidence from event-related brain potentials. *Neuroscience Letters*, *339*(3), 191–194. https://doi.org/10.1016/S0304-3940(03)00027-2

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9–26. https://doi.org/10.1016/S0167-6393(99)00028-X

Levinson, S. (1983). *Pragmatics*. Cambridge University Press.

Levinson, S. C. (1983). Conversational structure. In *Pragmatics* (pp. 284–370).

Löfqvist, A. (1986). The long-time-average spectrum as a tool in voice research. *Journal of Phonetics*, *14*(3–4), 471–475. https://doi.org/10.1016/S0095-4470(19)30692-8

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

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010

McGowan, K. B. (2015). Social Expectation Improves Speech Perception in Noise. *Language and Speech*, *58*(4), 502–521. https://doi.org/10.1177/0023830914565191

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), Article 5588. https://doi.org/10.1038/264746a0

Mertens, J. B. (2022). Miscommunication: How Prediction and Egocentricity Increase Progressivity but Decrease Intersubjectivity [Ph.D., Tufts University]. In *ProQuest Dissertations and Theses*. https://www.proquest.com/docview/2712775900/abstract/1F31937F937549F3PQ/1

Mertens, J. B. & De Ruiter, Jan P. (2021). Cognitive and social delays in the initiation of conversational repair. *Dialogue & Discourse*, *12*(1), Article 1. https://doi.org/10.5210/dad.2021.102

Mertens, J. B., Umair, M., Warnke, L., & De Ruiter, J. P. (submitted). *Can Models Trained on Written Monologue Learn to Predict Spoken Dialogue?*

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213. https://doi.org/10.1016/S0749-596X(03)00028-7

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, *118*(1), 43–71. https://doi.org/10.1037/0096-3445.118.1.43

Münster, K., & Knoeferle, P. (2018). Extending Situated Language Comprehension (Accounts) with Speaker and Comprehender Characteristics: Toward Socially Situated Interpretation. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02267

Neuhoff, J. G., & Bochtler, K. S. (2018). Change deafness, dual-task performance, and domain-specific expertise. *Quarterly Journal of Experimental Psychology*, *71*(5), 1100–1111. https://doi.org/10.1080/17470218.2017.1310266

Neuhoff, J. G., Schott, S. A., Kropf, A. J., & Neuhoff, E. M. (2014). Familiarity, Expertise, and Change Detection: Change Deafness is Worse in Your Native Language. *Perception*, *43*(2–3), 219–222. https://doi.org/10.1068/p7665

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360. https://doi.org/10.1177/026192702237953

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Openslr.org. (n.d.). Open Speech and Language Resources. Retrieved March 18, 2022, from http://www.openslr.org/5/

Pagel, M. (2016). Language: Why Us and Only Us? *Trends in Ecology & Evolution*, *31*(4), 258–259. https://doi.org/10.1016/j.tree.2016.01.009

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

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419. https://doi.org/10.1017/S1930297500002205

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

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human Voice Recognition Depends on Language Ability. *Science*, *333*(6042), 595–595. https://doi.org/10.1126/science.1207327

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910. https://doi.org/10.1016/j.neuropsychologia.2006.11.015

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). https://doi.org/10.1017/S0140525X04000056

Pillet-Shore, D. (2017, March 29). *Preference Organization*. Oxford Research Encyclopedia of Communication. https://doi.org/10.1093/acrefore/9780190228613.013.132

Pinheiro, J. C., & Bates, D. M. (Eds.). (2000). Linear Mixed-Effects Models: Basic Concepts and Examples. In *Mixed-Effects Models in S and S-PLUS* (pp. 3–56). Springer New York. https://doi.org/10.1007/0-387-22747-4_1

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In Atkinson, JM & Heritage, J (Eds.), *Structures of Social Action: Studies in Conversation Analysis* (pp. 57–101). Cambridge University Press.

Pomerantz, A & Heritage, J. (2013). Preference. In Sidnell, J & Stivers, T (Eds.), *Handbook of conversation analysis* (pp. 210–228). Cambridge University Press.

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

Reddy, Michael. (1979). The Conduit Metaphor. *Metaphor and Thought*, *2*, 285–324.

Richardson, D. C., Street, C. N. H., Tan, J. Y. M., Kirkham, N. Z., Hoover, M. A., & Ghane Cavanaugh, A. (2012). Joint perception: Gaze and social context. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00194

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

Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of “trouble” in conversation. *Speech Communication*, *48*(9), 1079–1093. https://doi.org/10.1016/j.specom.2006.02.001

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, *6*, 509.

Robinson, L. F., & Reis, H. T. (1989). The effects of interruption, gender, and status on interpersonal perceptions. *Journal of Nonverbal Behavior*, *13*(3), 141–153. https://doi.org/10.1007/BF00987046

Saarinen, A., Harjunen, V., Jasinskaja-Lahti, I., Jääskeläinen, I. P., & Ravaja, N. (2021). Social touch experience in different contexts: A review. *Neuroscience & Biobehavioral Reviews*, *131*, 360–372. https://doi.org/10.1016/j.neubiorev.2021.09.027

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

Sadock, J., & Zwicky, A. (1985). Speech act distinctions in syntax. In *Language typology and syntactic description* (pp. 155–196). Cambridge University Press.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv:1910.01108 [Cs]*. http://arxiv.org/abs/1910.01108

Sarathy, V., Tsuetaki, A., Roque, A., & Scheutz, M. (2020). Reasoning Requirements for Indirect Speech Act Interpretation. *Proceedings of the 28th International Conference on Computational Linguistics*, 4937–4948. https://doi.org/10.18653/v1/2020.coling-main.433

Schegloff, E. (1979). The Relevance of Repair to Syntax-for-Conversation. In *Syntax and Semantics* (Vol. 12, pp. 261–286). https://doi.org/10.1163/9789004368897_012

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

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, *53*(2), 361–382.

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

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. https://doi.org/10.1016/0010-0285(89)90008-X

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499–520. https://doi.org/10.1016/0749-596X(88)90022-8

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.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, *5*(01), 1. https://doi.org/10.1017/S0047404500006837

Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

Sell, G., Suied, C., Elhilali, M., & Shamma, S. (2015). Perceptual susceptibility to acoustic manipulations in speaker discrimination. *The Journal of the Acoustical Society of America*, *137*(2), 911–922. https://doi.org/10.1121/1.4906826

Shadlen, M. N., & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, *4*(4), 569–579. https://doi.org/10.1016/0959-4388(94)90059-0

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1447–1469. http://dx.doi.org/10.1037/0096-1523.28.6.1447

Shotter, J. (1996). 'Now I can go on:' Wittgenstein and our embodied embeddedness in the 'Hurly-Burly' of life. *Human Studies*, *19*(4), 385–407. https://doi.org/10.1007/BF00188850

Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattentional blindness within and across sensory modalities. *Quarterly Journal of Experimental Psychology*, *59*(8), 1425–1442. https://doi.org/10.1080/17470210500298948

Smit, H. (2014). *The social evolution of human nature: From biology to language*. 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

Stivers, T., & Robinson, J. D. (2006). A preference for progressivity in interaction. *Language in Society*, *35*(03). https://doi.org/10.1017/S0047404506060179

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

The CMU Pronouncing Dictionary. (n.d.). Retrieved July 21, 2023, from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Threlkeld, C., & De Ruiter, J. (2022). The Duration of a Turn Cannot be Used to Predict When It Ends. *Proceedings of the SIGdial 2022 Conference*, 361–367.

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or Top-down Processing as a Discriminator of L2 Listening Performance. *Applied Linguistics*, *19*(4), 432–451. https://doi.org/10.1093/applin/19.4.432

Umair, M., Mertens, J. B., Albert, S., & De Ruiter, J. P. (2022). GailBot: An automatic transcription system for Conversation Analysis. *Dialogue & Discourse*, *13*(1), Article 1. https://doi.org/10.5210/dad.2022.103

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

Van Dommelen, W. A. (1987). The Contribution of Speech Rhythm and Pitch to Speaker Recognition. *Language and Speech*, *30*(4), 325–338. https://doi.org/10.1177/002383098703000403

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequencyinevent-related brainpotentials. *Memory & Cognition*, *4*(18), 380–393. https://doi.org/10.3758/BF03197127

Van Rossum, G., & Fred, D. (2009). *Python 3 Reference Manual*. CreateSpace.

Vanderveken, D. (1990). *Meaning and Speech Acts: Volume 1, Principles of Language Use*. Cambridge University Press.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance; Washington*, *29*(2), 333–342. http://dx.doi.org.ezproxy.library.tufts.edu/10.1037/0096-1523.29.2.333

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology*, *17*(7), 755–774. https://doi.org/10.1002/acp.914

Warnke, L., & De Ruiter, J. P. (2023). Top-down effect of dialogue coherence on perceived speaker identity. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-30435-z

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

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating Words and Their Gender: An Event-related Brain Potential Study of Semantic Integration, Gender Expectancy, and Gender Agreement in Spanish Sentence Reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Wiley, R. H. (2005). Individuality in songs of Acadian flycatchers and recognition of neighbours. *Animal Behaviour*, *70*(1), 237–247. https://doi.org/10.1016/j.anbehav.2004.09.027

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, *42*(6), 654–667. https://doi.org/10.1111/j.1469-8986.2005.00356.x

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

Yurovsky, D., Wu, R., Yu, C., Kirkham, N. Z., & Smith, L. B. (2011). Model selection for eye movements: Assessing the role of attentional cues in infant learning. *Connectionist Models of Neurocognition and Emergent Behavior*, 58–75. https://doi.org/10.1142/9789814340359_0005

Zaltz, Y., Goldsworthy, R. L., Kishon-Rabin, L., & Eisenberg, L. S. (2018). Voice Discrimination by Adults with Cochlear Implants: The Benefits of Early Implantation for Vocal-Tract Length Perception. *Journal of the Association for Research in Otolaryngology*, *19*(2), 193–209. https://doi.org/10.1007/s10162-017-0653-5

Zhang, Q., Walsh, M. M., & Anderson, J. R. (2018). The Impact of Inserting an Additional Mental Process. *Computational Brain & Behavior*, *1*(1), 22–35. https://doi.org/10.1007/s42113-018-0002-8

Tables

Table 1

Example stimulus from Experiment 1 in the congruent, speaker violation, and full violation conditions

Context Utterance (TCU 1): "*We just moved into a new house.*"			
Target Utterance (TCU 2):			
	Congruent	**Speaker violation**	**Full violation**
Same speaker	"*Come by*"	"*Where?*"	"*You sure?*"

Table 2

Frequencies of speaker ratings in Experiment 1 separated by speakers of the stimuli

	Recorded Speakers					
	1	**2**	**3**	**4**	**5**	**6**
Same Speaker	0.70	0.66	0.75	0.74	0.59	0.86
Different Speaker	0.30	0.34	0.25	0.26	0.41	0.14

Table 3

Summary of Bayesian Logistic Linear Mixed Effects Model for Experiment 1 with congruent as the reference level

Fixed effects	Mean	95% CI	P(b>0)	P(b<0)
(Intercept)	-2.05	-2.40 – -1.71	0	1
Speaker violation	1.41	1.26 – 1.56	1	0
Full violation	0.87	0.71 – 1.02	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 4

Summary of Frequentist Logistic Linear Mixed Effects Model for Experiment 1 with congruent as the reference level

Fixed effects	Estimate	Std. error	Z value	Pr(> |z|)
(Intercept)	-2.05	0.17	-12.00	**<.001**
Speaker violation	1.41	0.08	18.32	**<.001**
Full violation	0.87	0.08	11.14	**<.001**

Table 5

Summary of Frequentist Logistic Linear Mixed Effects Model for Experiment 1 with full violation as the reference level

Fixed effects	Estimate	Std. error	Z value	Pr(> |z|)
(Intercept)	-1.19	.17	-7.10	**<.001**
Congruent	-.87	.08	-11.14	**<.001**
Speaker violation	.55	.07	7.89	**<.001**

Table 6

Example stimulus from Experiment 2 in the congruent, speaker violation, and full violation conditions

Context Utterance (TCU 1): "*moved we new just a into house"* Target Utterance (TCU 2):			
	Congruent	**Speaker violation**	**Full violation**
Same speaker	"*by come"*	"w*here"*	"*sure you"*

Table 7

Frequencies of speaker ratings in Experiment 2 separated by speakers of the stimuli

	Recorded Speakers					
	1	**2**	**3**	**4**	**5**	**6**
Same Speaker	0.57	0.63	0.57	0.63	0.58	0.73
Different Speaker	0.43	0.37	0.43	0.37	0.42	0.27

Table 8

Summary of Frequentist Logistic Linear Mixed Effects Model for Experiment 2 with congruent as the reference level

Fixed effects	Estimate	Std. error	Z value	Pr(> |z|)
(Intercept)	-0.61	0.14	-4.49	**<.001**
Speaker violation	0.01	0.05	0.102	0.92
Full violation	0.09	0.06	1.50	0.13

Table 9

Summary of Frequentist Logistic Linear Mixed Effects Model for Experiment 2 with full violation as the reference level

Fixed effects	Estimate	Std. error	Z value	Pr(> |z|)
(Intercept)	-0.52	.14	-3.80	**<.001**
Congruent	-.010	.06	-1.50	0.13
Speaker violation	-0.09	.06	-1.40	0.16

Table 10

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 11

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 12

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 13

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 14

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.

Table 15

Means, standard deviations and modes for FTO in milliseconds as a function of grammatical sentence type and speech act

	FTO			
	Interrogative	**Declarative**	**Question**	**Statement**
N	201	1729	200	1730
Mean	170.99	64.12	64.12	60.15
Mode	-74	-137	-137	-74
SD	391.14	373.49	391.53	372.09

Table 16

Summary of Linear Mixed Effects Model for FTOs for statement versus question speech acts.

Fixed effects	Mean	95% CI	P(b>0)	P(b<0)
Intercept	140.32	104.2 – 176.84	1	0
Speech Act	130.2266	75.64 – 184.74	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. Since factors were contrast coded, the coefficient for Speech Act is the difference in FTO between statements and questions.

Table 17

Means, standard deviations and modes for FTO in milliseconds and probability of speech act directness

	Directness Probability	**FTO**
Mean	0.62	63.348
Mode	0.03	207
SD	0.27	375.53

Table 18

Descriptive statistics for turn length, FTO and BIAS data for Study 1. Values shown in milliseconds.

	Duration	**FTO**	**BIAS**
Mean	2644.74	18.68	20.37
Std. Deviation	1693.69	572.20	435.05
Minimum	926.23	-1880.22	-1976.44
Maximum	9951.79	1360.38	1791.00

Table 19

Summary of Linear Mixed Effects Model for bias with duration as predictor (Study 1)

Fixed effects	Mean	SE	95% CI
Intercept	233.02	51.31	133.95 – 335.76
Duration	-0.08	0.01	-0.11– -0.05

Note. Mean represents the posterior mean unstandardized beta coefficient, 95% CI represents the credible interval around the mean**.**

Table 20

Regression coefficients for FTO from natural data regressed on Duration added to a null model (Study 1)

Coefficient	Mean	SD	95% CI	P(incl|data)	P(excl|data)	$BF_{inclusion}$
Intercept	18.68	12.84	-5.57 – 44.23	1.00	0.00	1.00
Duration	-0.01	0.01	-0.03 - 0.00	0.59	0.41	1.46

Note. Mean represents the posterior mean unstandardized beta coefficient, 95% CI represents the credible interval around the mean**.**

Table 21

Descriptive statistics for turn length, number of words, number of syllables and FTO in Study 2

	Duration	**Words**	**Syllables**	**FTO**
Mean	2280.69	8.26	10.16	8.35
Std. Deviation	2051.79	6.84	8.78	588.69
Minimum	50.00	1.00	1.00	-1981.00
Maximum	9972.00	43.00	54.00	1995.00

Table 22

Summary of Linear Mixed Effects Model for FTO with number of words of TCU as predictor (Study 2)

Fixed effects	Mean	SE	95% CI
Intercept	-0.33	21.02	-41.12 – 41.36
Number of words	4.26	1.35	1.54 – 6.87

Note. Mean represents the posterior mean unstandardized beta coefficient, 95% CI represents the credible interval around the mean**.**

Table 23

Summary of Linear Mixed Effects Model for FTO with number of syllables of TCU as predictor

Fixed effects	Mean	SE	95% CI
Intercept	-1.51	20.52	-42.12 – 39.40
Number of syllables	3.51	1.06	1.48 – 5.63

Note. Mean represents the posterior mean unstandardized beta coefficient, 95% CI represents the credible interval around the mean**.**

Table 24

Summary of Linear Mixed Effects Model for FTO with TCU duration as predictor

Fixed effects	Mean	SE	95% CI
Intercept	8.58	20.61	-31.77 – 48.98
Number of syllables	0.01	0.004	0.003 – 0.02

Note. Mean represents the posterior mean unstandardized beta coefficient, 95% CI represents the credible interval around the mean**.**

Figures

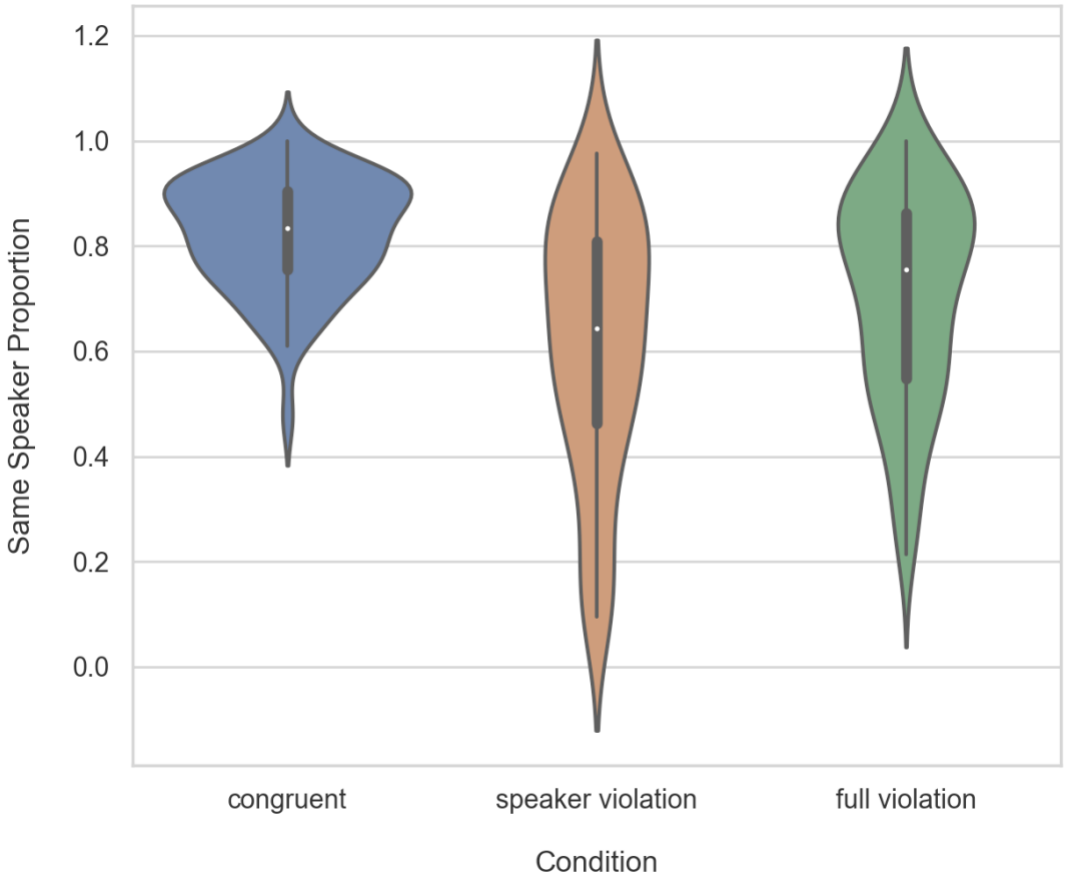

Figure 1. Violin plot showing the proportion of stimuli perceived as spoken by the same speaker in the *congruent*, *speaker violation* and *full violation* conditions in Experiment 1. The white dot indicates the median, the thick gray line represents the interquartile range, and the thin gray line represents the rest of the distribution barring outliers. The overall shape indicates the kernel density estimation of the underlying distribution.

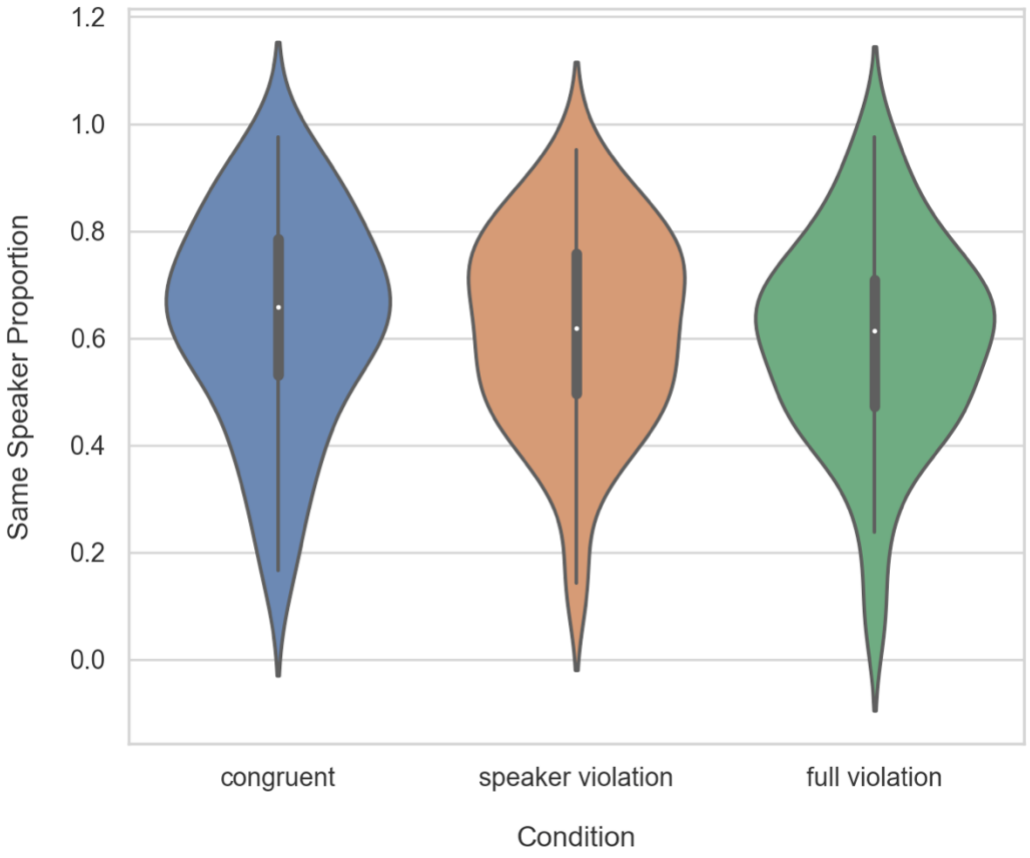

Figure 2. Violin plot showing the proportion of stimuli perceived as spoken by the same speaker in the *congruent*, *speaker violation* and *full violation* conditions in Experiment 2. The white dot indicates the median, the thick gray line represents the interquartile range, and the thin gray line represents the rest of the distribution barring outliers. The overall shape indicates the kernel density estimation of the underlying distribution.

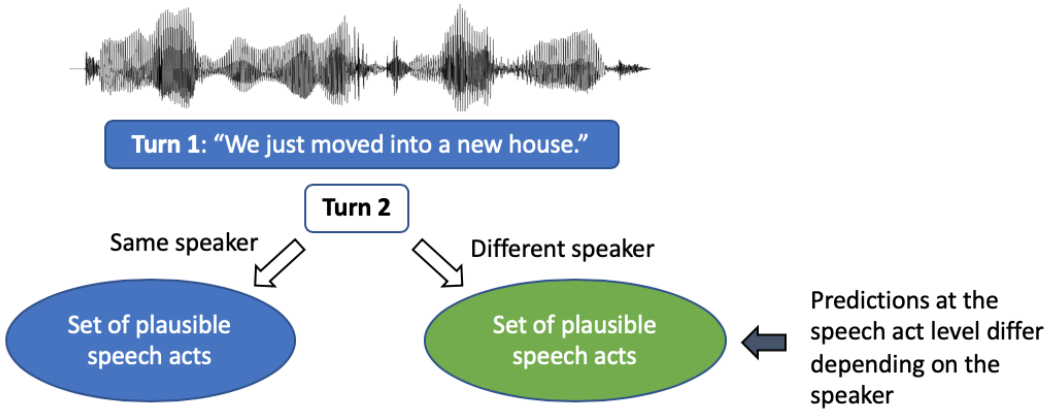

Figure 3. Listener's communication model.

Note. This figure is a schematic illustration of the state of the communication model of the listener after hearing turn 1 but before the onset of turn 2. The listener presumably has separate probabilistic predictions for turn 2 if the same speaker continues and turn 2 if a different speaker continues.

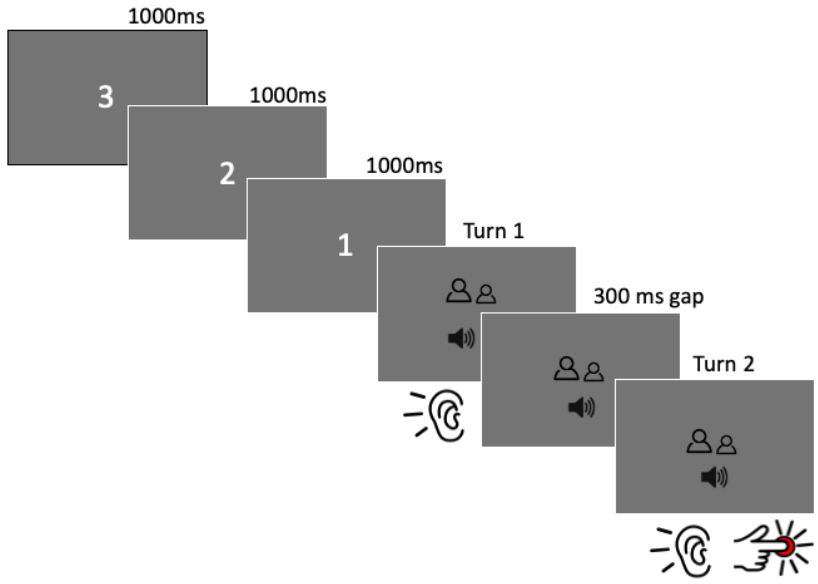

Figure 4. Presentation of each stimulus in a trial of the experiment.

Note. Each trial began with a countdown timer where the numbers "3", "2" and "1" were displayed sequentially for 1000 ms each (centered in white font). Participants then listened to turn 1 while the screen displayed an audio symbol as well as an image of one or two people (depending on whether they would hear one or two speakers). Participants then heard a 300ms silence followed by turn 2. The visual display of the screen remained constant until participants pressed the keyboard button. As soon as participants pressed the button, or 4000ms after the offset of turn 2, the audio file cut out and the next trial began automatically.

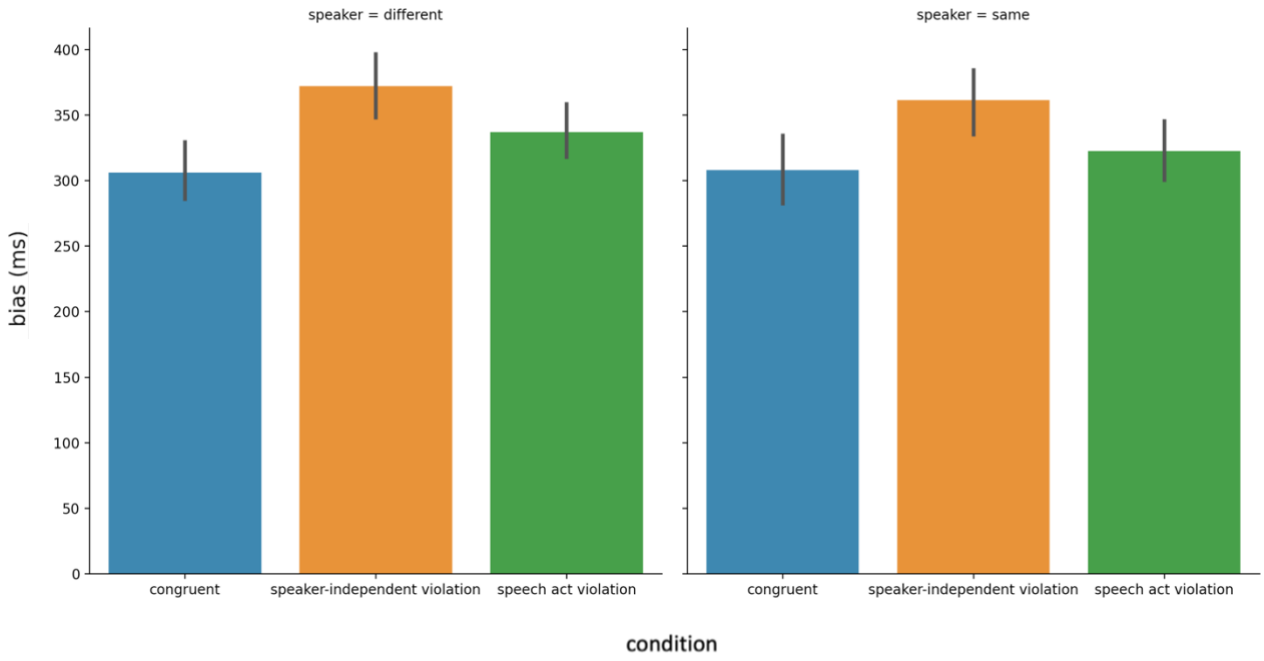

Figure 5. Mean bias scores and standard errors separated by congruency and speaker.

Note. This figure presents mean bias values separated by each condition of the experiment. Bias values for different speakers are plotted on the left; bias values for the same speaker are plotted on the left. The data show the same pattern across different and same speaker: congruent trials were responded to most quickly, followed by speech act violation trials. Speaker-independent violation trials show the largest mean bias values.

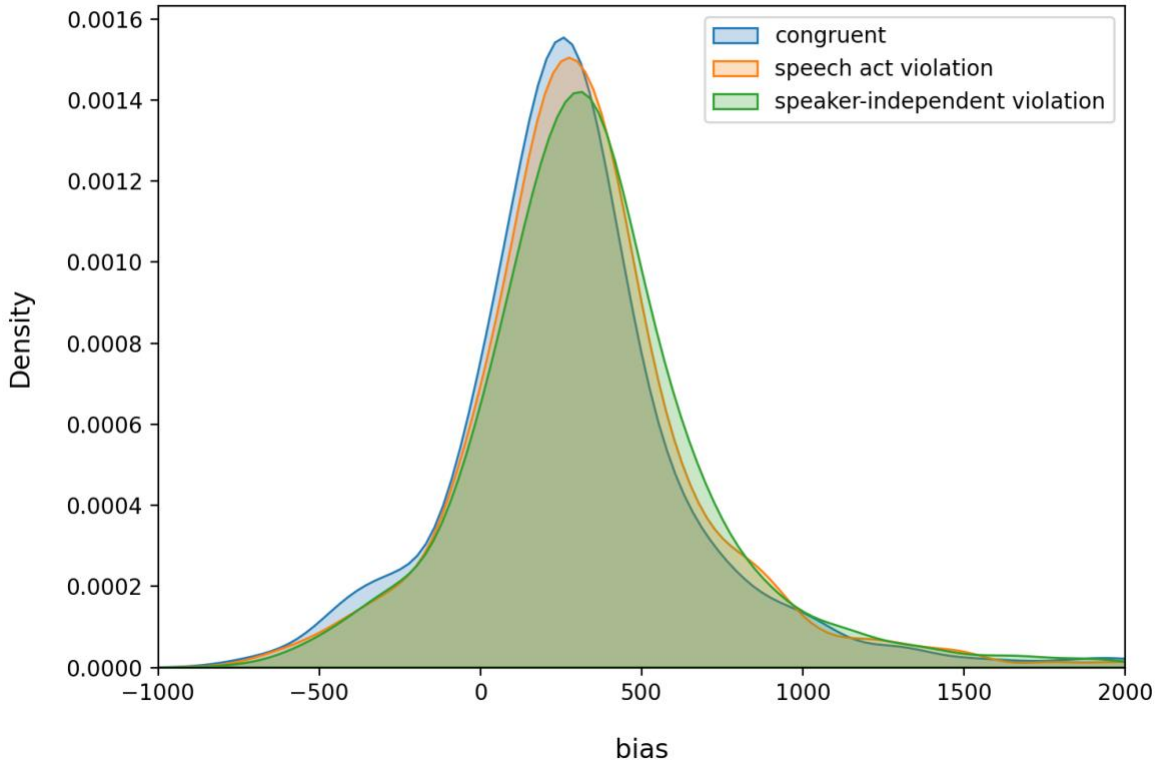

Figure 6. Kernel density estimate plot of bias separated by congruency.

Note. This figure presents a kernel density estimate plot of the three distributions of bias data separated by the three experimental conditions in the experiment.

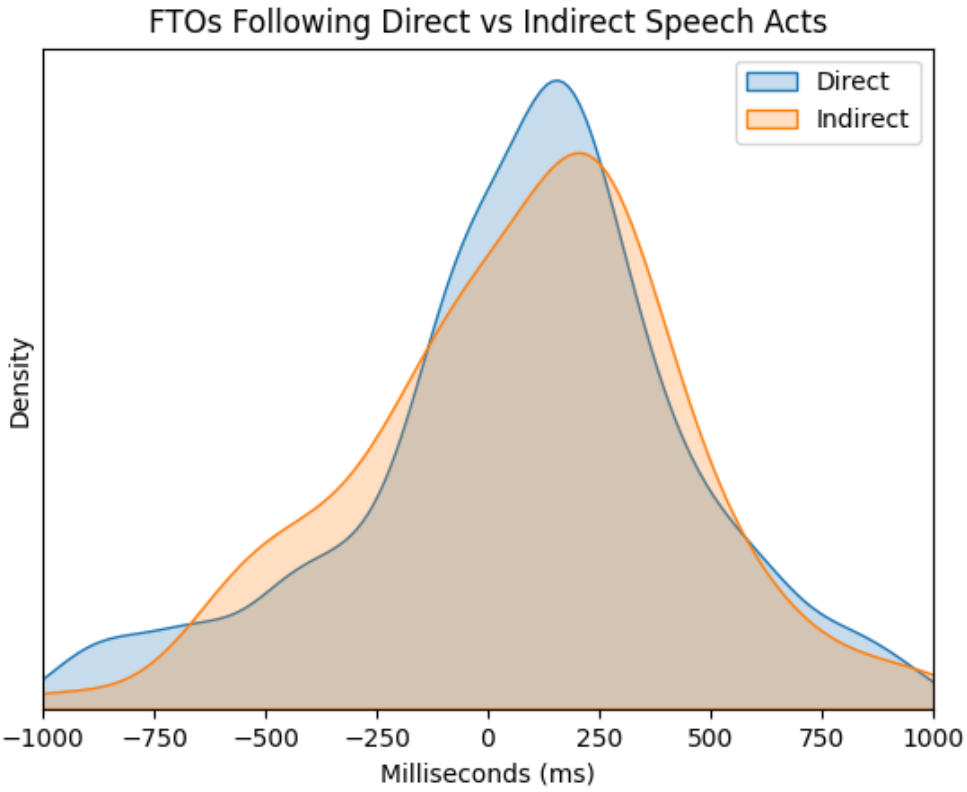

Figure 7. Kernel density estimate plot of FTOs following direct vs. indirect speech acts.

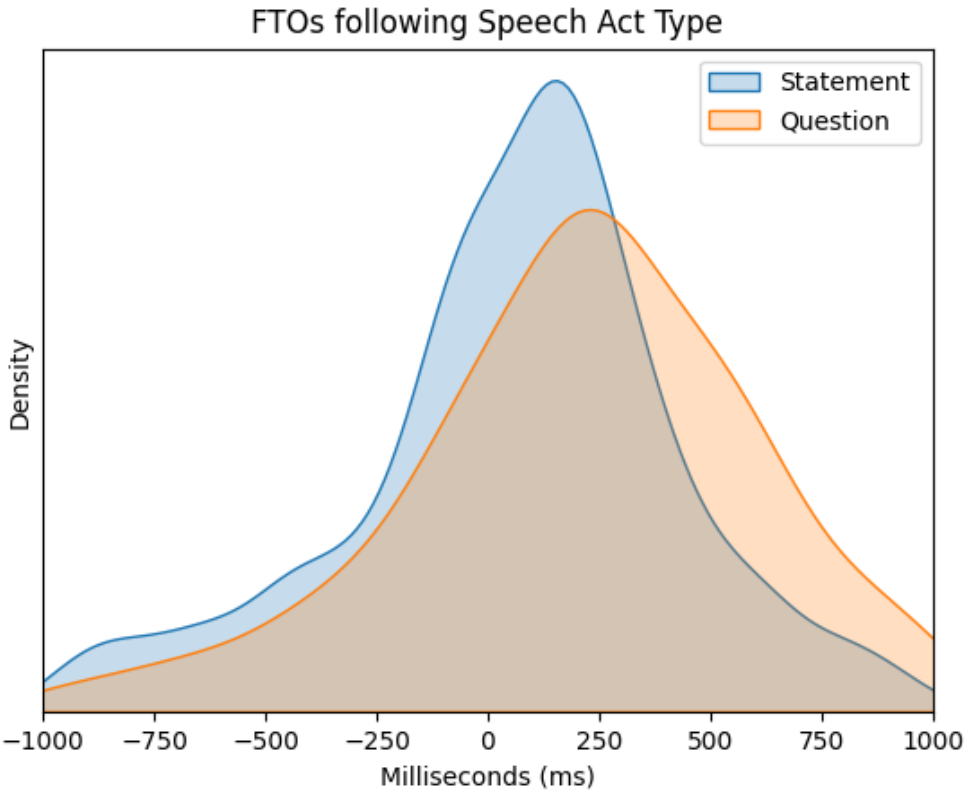

Figure 8. Kernel density estimate plot of FTOs following question vs. statement speech acts.

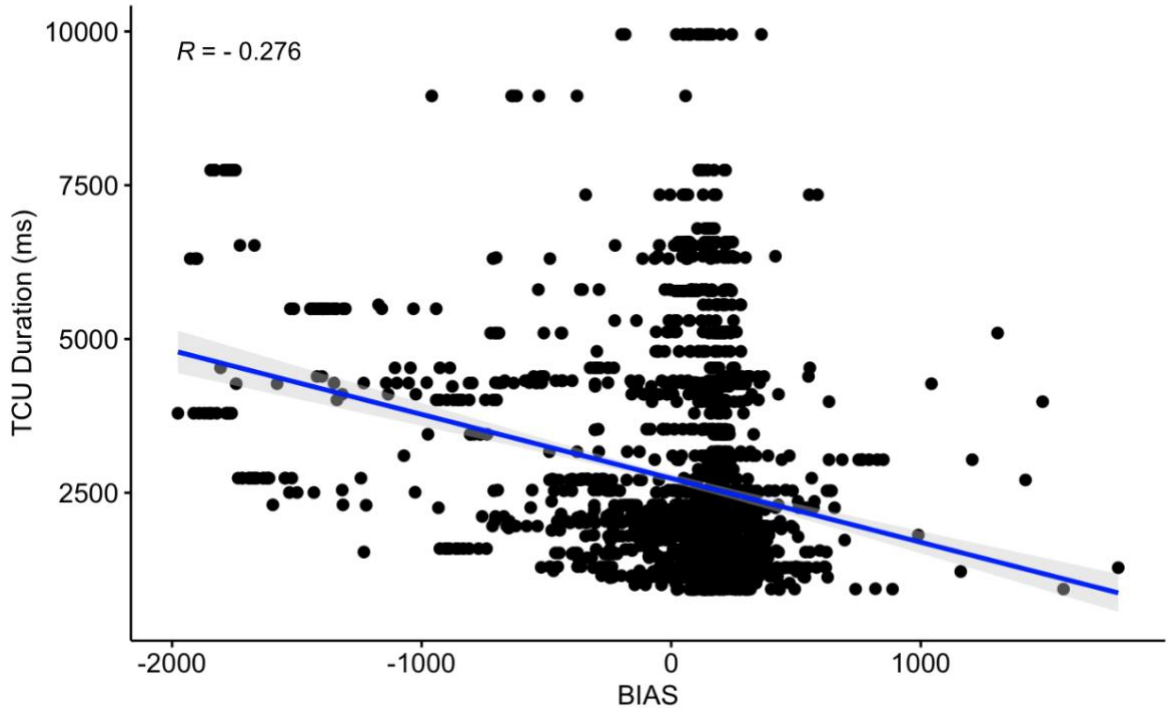

Figure 9. Scatterplot showing the inverse relationship between TCU duration and *bias* in de Ruiter et al.'s 2006 turn-end estimation experiment.

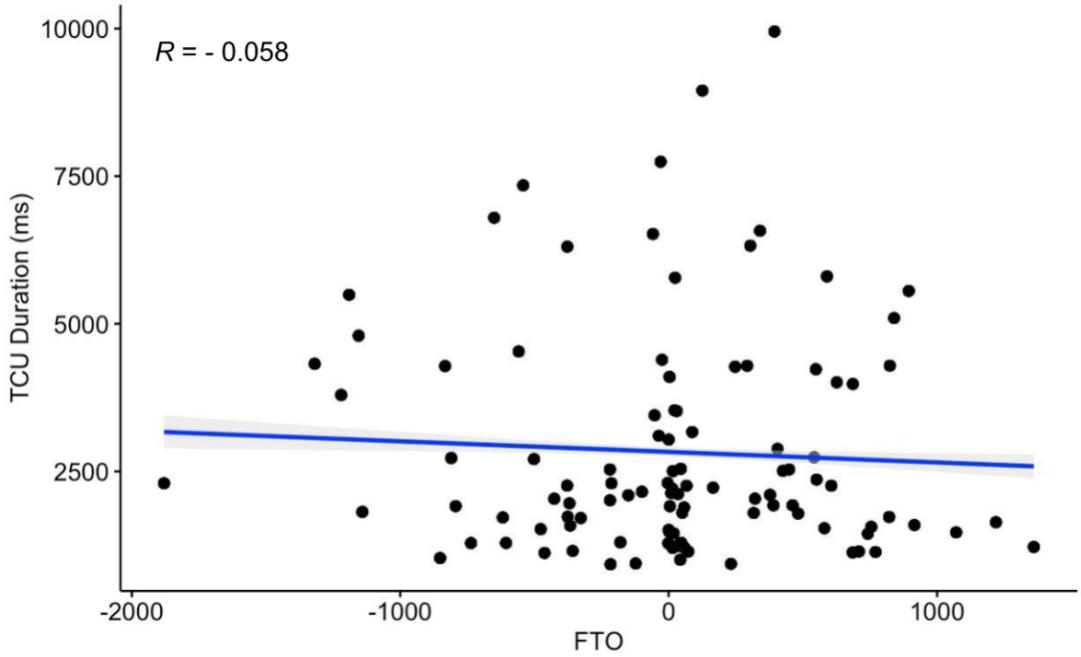

Figure 10. Scatterplot showing the relationship between TCU duration and FTO from de Ruiter et al.'s 2006 turn-end estimation experiment. FTOs correspond to turns that were recorded in natural dialogue and then used as stimuli in the button press experiment.

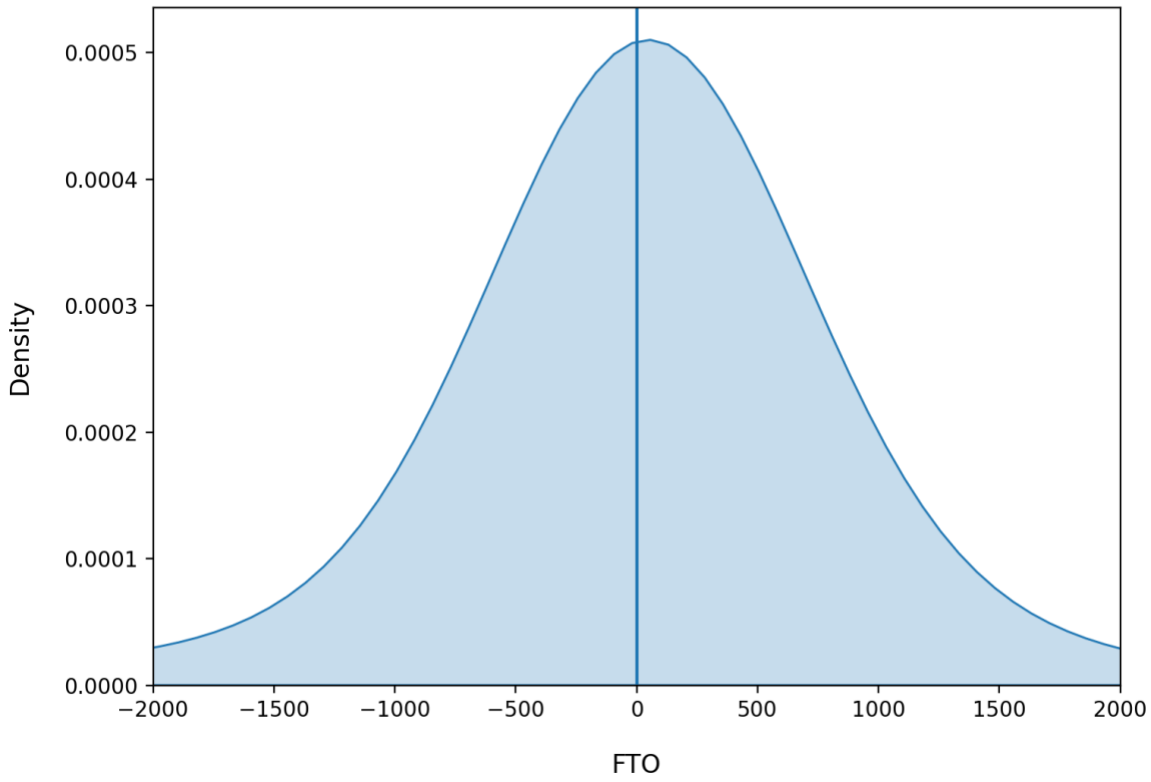

Figure 11. The distribution of Floor Transfer Offset (FTO) for the Switchboard corpus of natural conversations.

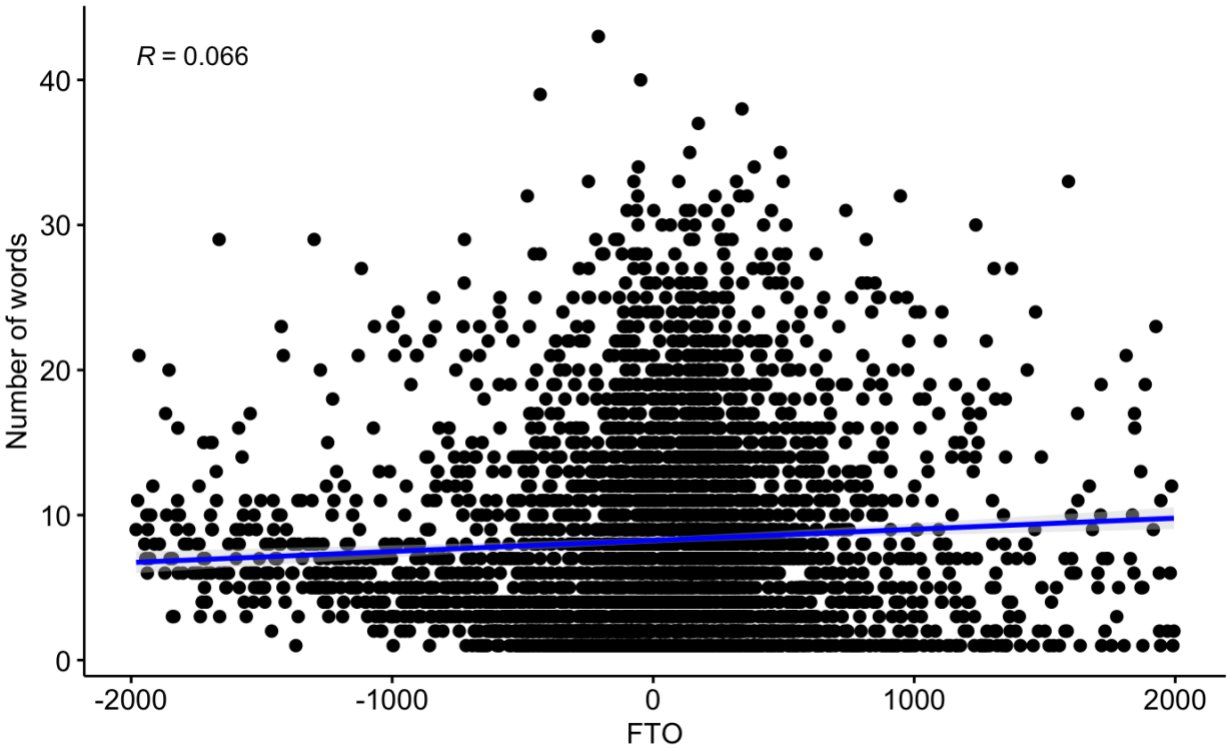

Figure 12. Scatterplot showing the relationship between the number of words per TCU and the subsequent FTO in the Switchboard Corpus.

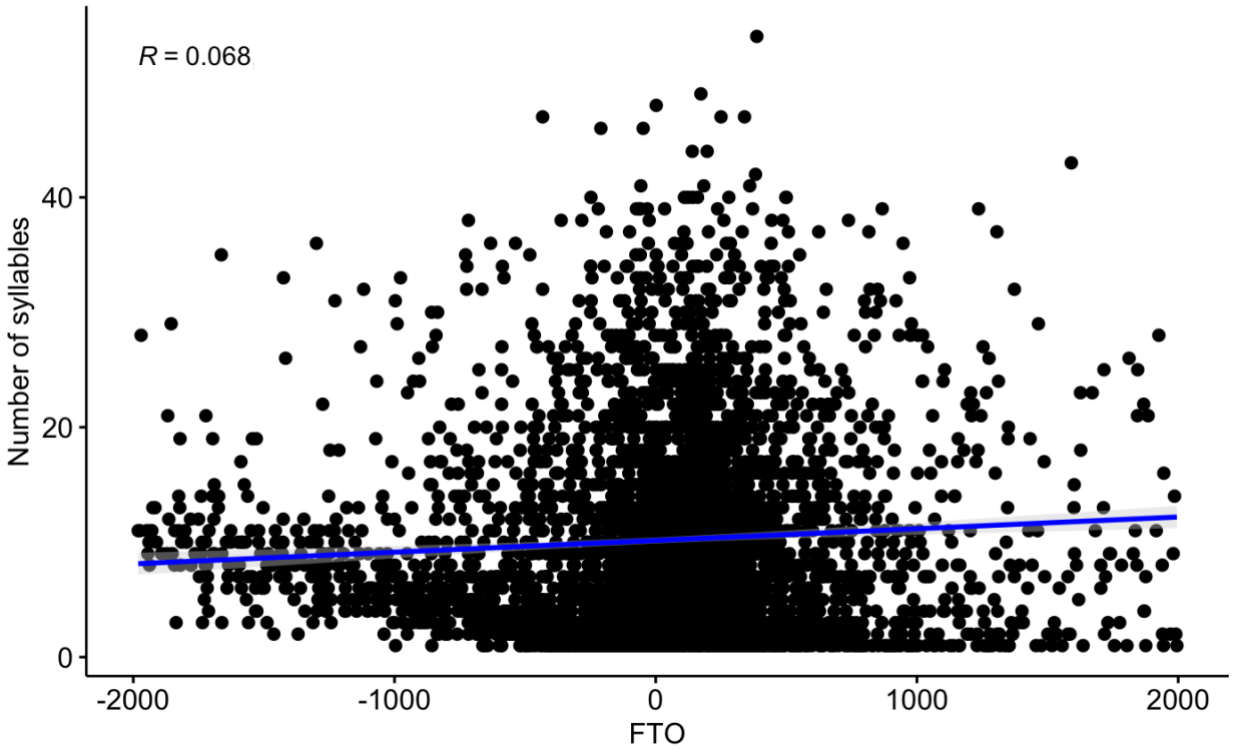

Figure 13. Scatterplot showing the relationship between the number of syllables per TCU and the subsequent FTO in the Switchboard Corpus.

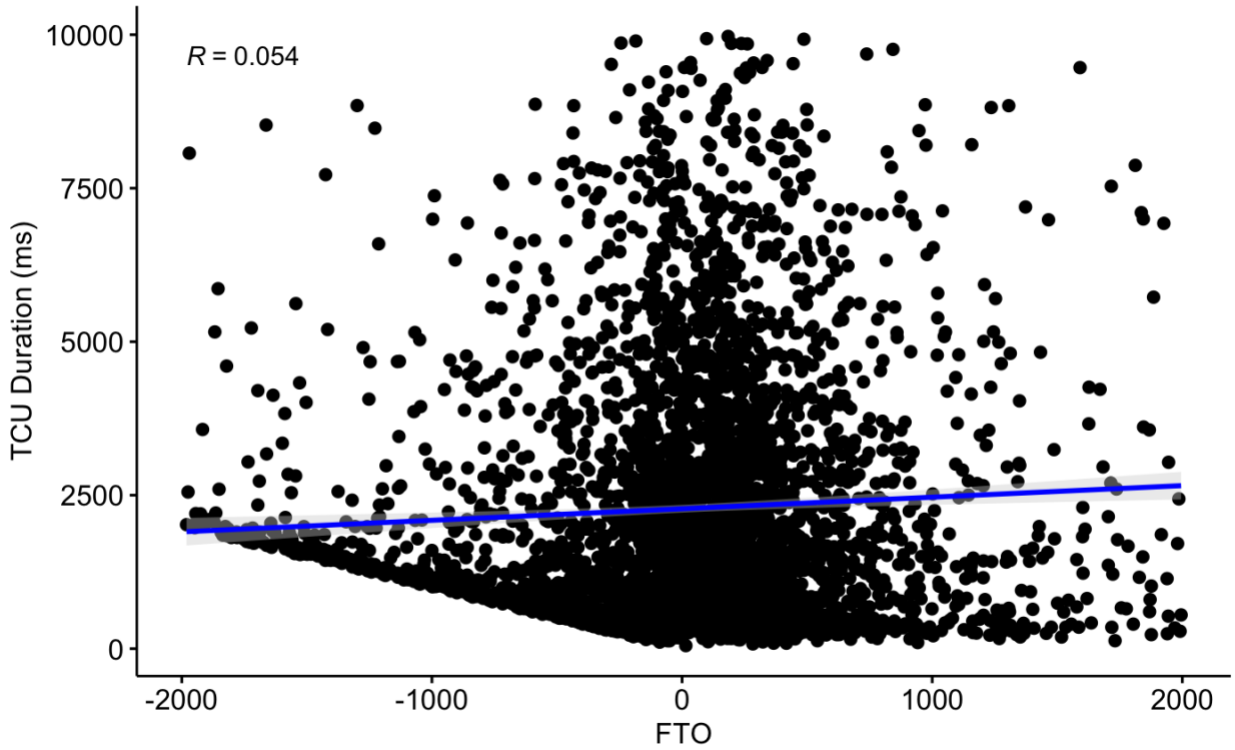

Figure 14. Scatterplot showing the relationship between TCU duration and the subsequent FTO in the Switchboard Corpus

Appendix

Appendix A: Mapping of Jurafsky et al.'s (1997) speech act tags of the Switchboard Corpus onto statement, question, command and other. For the full annotation manual see https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

DAMSL Tag	SWBD-DAMSL Tag	Description	Mapping
sd	sd	Statement-non-opinion	Statement
b	b	Acknowledge (Backchannel)	Other
sv	sv	Statement-opinion	Statement
%	%	Abandoned or Turn-Exit	Other
aa	aa	Agree/Accept	Other
ba	ba	Appreciation	Statement
qy	qy	Yes-No-Question	Question
x	x	Non-verbal	Other
ny	ny	Yes Answers	Other
fc	fc	Conventional-Closing	Other
b^r	b	Acknowledge Self-Repeat	Other
sd^e	sd	Statement Expansions of y/n Answers	Statement
qw	qw	Wh-Question	Question
sd(^q)	sd	Statement w/ Quotation	Statement
bk	bk	Response Acknowledgement	Other
nn	nn	No Answers	Other
qy^d	qy^d	Declarative Yes-No-Question	Question
h	h	Hedge	Statement
bh	bh	Backchannel in Question Form	Statement
^q	^q	Quotation	Statement
bf	bf	Summarize/reformulate	Statement
sd^t	sd	Statement about Task	Statement
aa^r	aa	Agree/Accept Repeat	Other
+@		Continuation w/ error	Other
o	bc	Other	Other
na	na	Affirmative Non-Yes Answer	Statement

^2	^2	Collaborative Completion	Other
b^m	b^m	Repeat Phrase	Other
ad	ad	Action Directive	Command
qo	qo	Open-ended Question	Question
qh	qh	Rhetorical Question	Statement
^h	^h	Hold before answer/agreement	Other
qy^g	qy	Tag question	Statement
o@	bc	Overloaded TCU	Other
ar	ar	Reject	Statement
sv(^q)	sv	Statement-Opinion w/ Quotation	Statement
ng	ng	Negative Non-No Answers	Statement
no	no	Other Answers	Statement
sd^r	sd	Statement, Self-repeat	Statement
br	br	Signal Non-Understanding	Other
sd@	sd	Statement w/ error	Statement
qr	qy	Or Question	Question
fp	fp	Conventional Opening	Other
qrr	qrr	Or-question tacked onto yes-no question	Question
ny^r	ny	Yes w/ repeat	Other
nd	nd	Dispreferred Answer	Statement
sv^t	sv	Opinion about task	Statement
nn^r	nn	No Answer w/ repeat	Other
fe	ba	Exclamation	Statement
fc^m	fc	Conventional Closing w/ mimic	Other
%@	%	Abandoned w/ error	Other
sv^e	sv	Opinion, expansion of y/n answer	Statement
t3	t3	3rd Party Talk	Other
qy^t	qy	Yes/No Question about Task	Question
t1	t1	Self-talk	Other
ba^r	ba	Appreciation w/ repeat	Other
bd	bd	Downplayer	Other

^g	^g	Tag Question	Statement
sv^r	sv	Opinion w/ repeat	Statement
sv@	sv	Opinion w/ error	Statement
qw^d	qw^d	Declarative wh-question	Question
b@	b	Backchannel w/ error	Other
ft	ft	Thanking	Other
fa	fa	Apology	Statement
aa^m	aa	Accept w/ repeat	Other
sd^m	sd	Statement w/ mimic	Statement
ad^t	ad	Action-directive about task	Command
br^m	br	Signal Non-understanding w/ mimic	Other
aap	am	Accept Part	Other
sd^c	sd	Statement about communication	Statement
qw^t	qw	Wh-question about task	Question
co	cc	Offer	Question
x@	x	Non-speech w/ error	Other
sd*	sd	Statement w/ transcription error	Statement
am	am	Maybe	Statement
ar^r	ar	Reject w/ repeat	Statement
na^r	na	Affirmative non-yes w/ repeat	Statement
na^m	na	Affirmative non-yes w/ mimic	Statement
cc	cc	Commit	Statement
"	bc	Other	Other
ba@	ba	Appreciation w/ error	Statement
bk^r	bk	Acknowledge w/ repeat	Other
qy^r	qy	Yes-no Question w/ repeat	Question
fc^t	fc	Conventional Closing about task	Other
sv^m	sv	Opinion w/ mimic	Statement
+		Segment	Other
sv*	sv	Opinion w/ transcription Error	Statement
arp	nd	Dispreferred Answer	Statement

sd(^q)^t	sd	Statement w/ quotation about task	Statement
qy^h	qy	Yes-no Question hold before answer	Question
qy@	qy	Yes-no Question w/ error	Question
bk^m	bk	Acknowledge Answer w/ mimic	Statement
aa@	aa	Accept w/ error	Statement
qy^g^t	qy	Yes-no Question Tag Question about Task	Question
by	bc	Sympathy	Statement
fc^r	fc	Conventional Closing, repeat	Other
sd,o@	sd	Statement, other	Statement
qy^m	qy	Yes-no question w/ mimic	Question
qy^c	qy	Yes-no Question about communication	Question
fp^m	fp	Conventional Opening Mimic	Other
qy^d^t	qy^d	Declarative Yes-no question about task	Question
qw^r	qw	Wh-question repeat	Question
qr^d	qy	Declarative Or-question	Question
co^t	cc	Offer about task	Question
qw^h	qw	Wh-Question, hold before answer	Question
bc	bc	Correct Misspeaking	Other
+*		Continuation w/ transcription error	Other
sd^e^t	sd	Statement expanding y/n answer about task	Statement
na^t	na	Affirmative Non-yes answer about task	Statement
qw@	qw	Wh-question w/ error	Question
fx	sv	Explicit Performative	Statement
sv,o@	sv	Opinion, other w/ error	Statement
sd^e@	sd	Statement expanding y/n answer w/ error	Statement
qy^2	qy	Yes-no question w/ collaborative complete	Question
bf@	bf	Summarize/reformulate w/ error	Statement

ny^m	ny	Yes Answer w/ Mimic	Statement
bd^r	bd	Downplaying w/ repeat	Other
b*	b	Continuer w/ transcription error	Other
^2@	^2	Collaborative Completion w/ error	Other
qy^d^r	qy^d	Declarative Yes-no Question w/ repeat	Question
qy^d@	qy^d	Declarative Yes-no Question w/ error	Question
qrr^t	qrr	Or-question about task	Question
qo^t	qo	Open Question about task	Question
ny@	ny	Yes Answer w/ transcription error	Statement
nn^m	nn	No answer w/ mimic	Statement
bh^m	bh	Rhetorical Question w/ mimic	Statement
bf^r	bf	Reformulate w/ repeat	Statement
ad(^q)	ad	Action direction w/ quotation	Command
^q^t	^q	Quotation about task	Statement
sv(^q)@	sv	Opinion w/ quotation and error	Statement
sd^e^r	sd	Statement reply to y/n questions w/ repeat	Statement
sd^e^m	sd	Statement reply to y/n questions w/ mimic	Statement
sd^2	sd	Statement collaborative completion	Statement
qrr^d	qrr	Declarative Or-question	Question
o*	bc	Other w/ transcription error	Other
nn^e	ng	No answer to y/n question	Other
fo	bc	Other forward-function	Other
^2^g	^2	Collaborative completion, tag question	Question
sd(^q)@	sd	Statement w/ quotation and error	Statement
sd(^q)*	sd	Statement w/ quotation and trans. error	Statement

qy^g@	qy	Yes-no tag question w/ error	Question
qy^g*	qy	Yes-no tag question w/ transcription error	Question
qy^d^m	qy^d	Declarative Yes-no question w/ mimic	Question
qy(^q)	qy	Yes-no Question w/ quotation	Question
qo^d	qo	Declarative Open Question	Question
qh^m	qh	Rhetorical Question w/ mimic	Statement
oo	cc	Offer	Question
o^r	bc	Other w/ repeat	Other
no^t	no	Other answers about task	Statement
ng^r	ng	Negative Non-no answer w/ repeat	Statement
h^r	h	Hedge w/ repeat	Other
ad^r	ad	Action directive w/ repeat	Command
ad^c	ad	Action directive about communication	Command
ad@	ad	Action directive w/ error	Command
aa*	aa	Accept w/ transcription error	Statement
sv^c	sv	Opinion about communication	Statement
sv^2	sv	Opinion w/ collaborative completion	Statement
sv,sd,o@	sv	Opinion, Statement, Other w/ error	Statement
qy*	qy	Yes-no Question w/ transcription error	Question
qw^g	qw	Wh-Question tag question	Question
qw^d^t	qw^d	Declarative Qh-Question about task	Question
qr^t	qy	Or-question about task	Question
qh@	qh	Rhetorical Question w/ error	Statement
o^c	bc	Other about communication	Other
nd^t	nd	Dispreferred answer about task	Statement

na@	na	Affirmative Yes-no Answer w/ error	Statement
fw	bc	You're Welcome	Statement
fp^r	fp	Conventional Opening w/ repeat	Other
co^c	cc	Offer about communication	Question
bh^r	bh	Backchannel in Question Form w/ repeat	Statement
bh@	bh	Backchannel in Question Form w/ error	Statement
bf^m	bf	Summarize w/ mimic	Statement
ba^m	ba	Appreciation w/ mimic	Statement
b^m^t	b^m	Repeat phrase about task	Other
aa^t	aa	Accept about task	Statement
aa^2	aa	Accept w/ collaborative completion	Statement
^q@	^q	Quotation w/ error	Statement
^q*	^q	Quotation w/ transcription error	Statement
%*	%	Abandoned w/ transcription error	Other
x*	x	Non-verbal w/ transcription error	Other
sd^q	sd	Statement w/ quotation	Statement
qy^g^r	qy	Yes-no Tag Question w/ repeat	Question
qy^g^c	qy	Yes-no Tag Question about communication	Question
qy^d^h	qy^d	Declarative Yes-No Question hold	Question
qy^c^r	qy	Yes-no about communication w/ repeat	Question
qw^m	qw	Wh-question w/ mimic	Question
qw^c	qw	Wh-question about communication	Question
qw*	qw	Wh-question w/ transcription error	Question
qh^r	qh	Rhetorical Question w/ repeat	Statement
qh^h	qh	Rhetorical Question w/ hold	Statement
oo^t	cc	Open Offer about task	Question

o^t	bc	Other about task	Other
ny^e	na	Yes Answer Plus Expansion	Statement
ny^c	ny	Yes Answer about communication	Statement
no^r	no	Other Answer w/ repeat	Other
nn*	nn	No Answer w/ transcription error	Statement
ng^m	ng	Negative Non-no Answer w/ mimic	Statement
h^t	h	Hedge about task	Other
fc@	fc	Conventional closing w/ error	Other
fa^c	fa	Apology about communication	Statement
cc^r	cc	Commit w/ repeat	Statement
br^r	br	Signal Non-Understanding w/ repeat	Other
bk@	bk	Acknowledge w/ error	Other
bf^t	bf	Reformulation about task	Statement
bf^g	bf	Reformulation Tag Question	Statement
bf*	bf	Reformulation w/ transcription error	Statement
bf(^q)	bf	Reformulation w/ quotation	Statement
bc^r	bc	Correct Misspeaking w/ repeat	Other
b^m^r	b^m	Continuer w/ mimic and repeat	Other
b^m^g	b^m	Tag Question Continuer w/ mimic	Statement
b^m@	b^m	Continuer w/ mimic and error	Statement
am^r	am	Maybe w/ repeat	Statement
ad*	ad	Action directive w/ error	Command
aa,o@	aa	Accept, Other w/ error	Statement
^q^r	^q	Quotation Repeat	Statement
^h^r	^h	Hold w/ repeat	Statement
^2*	^2	Collaborative Completion w/ transcript error	Other
t1^t	t1	Self-talk about task	Other

sv^e^r	sv	Opinion Answer w/ repeat	Statement
sv;sd	sv	Opinion, statement	Statement
sv,qy^g@	sv	Opinion, Yes-no Tag Question	Statement
sv(^q)*	sv	Opinion w/ Quotation and Error	Statement
sd^t*	sd	Statement about task w/ transcription error	Statement
sd^r@	sd	Statement w/ repeat and error	Statement
sd^m@	sd	Statement w/ mimic and error	Statement
sd^m*	sd	Statement w/ mimic and transcription error	Statement
sd^e(^q)^r	sd	Statement exp of y/n quest. w/ quote/repeat	Statement
sd;sv	sv	Statement, opinion	Statement
sd;qy^d	sd	Statement, Declarative Yes-no Question	Statement
sd;no	sd	Statement, Other Answer	Statement
sd,sv	sv	Statement, Opinion	Statement
sd,qy^g	sd	Statement, Yes-no Tag Question	Statement
sd(^q)^r	sd	Statement w/ Quotation and repeat	Statement
qy^h@	qy	Yes-no Question, hold w/ error	Question
qy^d^c	qy^d	Declarative Yes-no Question about comm	Question
qy^d*	qy^d	Declarative Yes-no Question w/ transc. error	Question
qy^d(^q)	qy^d	Declarative Yes-no Question w/ quotation	Question
qy,am,o@	qy	Yes-no Question, Maybe, Other w/ error	Question
qw^t@	qw	Wh-question about task w/ error	Question
qw^r^t	qw	Wh-Question w/ repeat about task	Question
qw^d^m	qw^d	Declarative Wh-Question w/ mimic	Question
qw^d^c	qw^d	Declarative Wh-Question about communication	Question

qw^d@	qw^d	Declarative Wh-Question w/ error	Question
qw(^q)	qw	Wh-Question w/ quotation	Question
qrr@	qrr	Or-question w/ error	Question
qr^d*	qy	Declarative Or question w/ trans. error	Question
qr(^q)	qy	Or-question w/ quotation	Question
qo^r	qo	Open-ended question w/ repeat	Question
qo^d^c	qo	Decl Open-ended question about comm	Question
qo@	qo	Open Ended Question w/ error	Question
qh^g	qh	Rhetorical Tag Question	Statement
qh^c	qh	Rhetorical Question about communication	Statement
qh*	qh	Rhetorical Question with transcription error	Statement
qh(^q)	qh	Rhetorical Question with quotation	Statement
oo(^q)	cc	Offer with Quotation	Question
ny^t	ny	Yes Answer about task	Statement
ny^c^r	ny	Yes Answer about communication w/ repeat	Statement
ny*	ny	Yes Answer w/ transcription error	Statement
no@	no	Other answer w/ error	Statement
nn^t	nn	No Answers about task	Statement
nn^r^t	nn	No Answers w/ repeat about task	Statement
nn^r@	nn	No Answers w/ repeat and error	Statement
ng^t	ng	Negative Non-no Answers about task	Statement
ng^r,o@	ng	Neg Non-No Answers w/ repeat, other w/ error	Statement
na^m^t	na	Aff Non-Yes Answers w/ mimic about task	Statement
na,sd,o@	na	Affirmative Non-Yes Answer, Statement, Other	Statement
h^m	h	Hold w/ mimic	Statement

h@	h	Hold w/ error	Statement
h,sd	h	Hold, statement	Statement
fw*	bc	You're Welcome w/ error	Statement
ft^t	ft	Thanking about task	Other
ft^m	ft	Thanking w/ mimic	Other
fo^c	bc	Other forward function about communication	Other
fc,o@	fc	Conventional Closing, Other w/ error	Other
fa^t	fa	Apology about task	Statement
fa^r	fa	Apology w/ repeat	Statement
cc^t	cc	Commit about task	Statement
br,o@	br	Signal Non-Understanding, Other w/ error	Other
bk^t	bk	Response Acknowledgement about task	Statement
bk,sd,o@	bk	Acknowledgement, Statement, Other w/ error	Statement
bh,sd,o@	bh	Backchannel, Statement, Other w/ error	Statement
bf^2	bf	Reformulation w/ collaborative completion	Statement
bf,nn,o@	bf	Reformulation, No Answer, Other w/ error	Statement
bd@	bd	Downplayer w/ error	Statement
ba^m@	ba	Appreciation w/ mimic and error	Statement
ba,fe	ba	Appreciation, Exclamation	Statement
b^t	b	Continuer about task	Other
b^r@	b	Continuer w/ repeat and error	Other
b^m,sd,o@	b^m	Cont. w/ mimic, statement, other w/ error	Statement
b^2	b	Continuer w/ collaborative completion	Other
ar^m	ar	Reject w/ mimic	Statement
ad,qy@	ad	Action directive, Yes-no Question w/ error	Command
ad,o@	ad	Action directive, Other w/ error	Command
aap^r	am	Accept Part w/ repeat	Statement

aap^m	am	Accept Part w/ mimic	Statement
aa^r,o@	aa	Accept w/ repeat, other w/ error	Statement
aa^h	aa	Accept & hold	Statement
aa,ar	aa	Accept, Reject	Statement
^h^t	^h	Hold about task	Statement
^h@	^h	Hold w/ error	Statement
^g@	^g	Tag Question w/ error	Question
^2^t	^2	Collaborative completion about task	Other
^2^r	^2	Collaborative completion w/ repeat	Other
%@*	%	Abandoned w/ errors	Other
%,o@	%	Abandoned, other w/ error	Other