NORTHWESTERN UNIVERSITY

Predicting Social Dynamics in Interactions Using Keystroke Patterns

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Media, Technology, and Society

Department of Communication Studies

By

Adam Goodkind

EVANSTON, ILLINOIS

March 2023

This thesis is dedicated to all of the doctors, nurses and therapists who kept me on my feet and enabled me to complete my doctorate.

Acknowledgements

The road to completing my thesis has been a wild ride, filled with many detours. I couldn't have done it alone, though, and would like to acknowledge the following people for helping me to get here.

- My advisor, Darren Gergle, who has been incredibly supportive through the years and has been a constant source of knowledge and motivation. Most importantly, he has always been patient with me, no matter how naive my questions are, even when I'm finishing up a dissertation.
- Anne Marie Piper, for serving on my committee, and always reminding me of the importance of our research. I only hope we get to collaborate more in the future.
- David-Guy Brizan, for teaching me C++, then being a great research partner, and then serving as a valuable sounding board on my committee.
- Andrew Rosenberg, for taking a chance on an unproven programmer and introducing me to the world of keystroke dynamics, which has fascinated me for years. Your clear explanations of concepts in statistics and NLP helped to set me up on this path.
- Klinton Bicknell, for bringing me into the world of word surprisal, neural networks, and GAMs. I learned so much from our collaborations.
- Janet Fodor, for properly introducing me to psycholinguistics and introducing me to the Implicit Prosody Hypothesis.
- My colleagues and co-conspirators in the CollabLab and the Language & Computation Lab

- My research assistant, Elana Laski, for being incredibly reliable and diligent with annotations, and acting as a great sounding board for ideas.
- My dissertation coach, Ilana Emmett, for helping me organize and plan out this unwieldy project.
- My mom, dad, stepdad, and sister, for constantly supporting throughout this entire process, on every level.
- Wallie, for always being the bestest boy and forcing me to get out of the house multiple times a day.
- And finally, my wife Shaina, who has been a constant source of companionship, fun, motivation and support throughout this entire process. She kept me sane during the pandemic, as we slowly lost our minds at the same pace, and has been by my side through thick and thin. I would not be where I am today without her.

Abstract

Every day, we communicate through computers on projects ranging from a group lunch order to booking a flight to learning critical medical information. And every day, we also *mis*communicate through computers: We don't pick up on an intentionally humorous response, or we miss the criticality of a request. This is made more frustrating because if these responses or requests were made in a face-to-face setting, these underlying intentions would be easier to pick up through tone of voice or the rate of speech, i.e. spoken prosody (Pierrehumbert and Hirschberg, 1990).

The COVID-19 pandemic, and its effects on remote working, have added a tragic emphasis to the need for a better understanding of computer-mediated communication, as text-based CMC has come to occupy an even more central role in our lives (Microsoft, 2021; Teevan et al., 2022).

My thesis aims to use timing patterns in typing, called *keystroke dynamics*, to detect underlying motivations and intentions, and make the information normally available only in face-to-face interactions also accessible in a text-based interaction, where prosodic information is assumed to be lost (Plank, 2016). If this information can be recovered and utilized, it will make text-based conversations more expressive and increase the bandwidth of information that it is possible to exchange via text.

For my thesis, I recruited 196 participants who took part in a 16-minute conversation where they exchanged movie and TV recommendations. Following the conversation, participants completed a survey that asked them to rate their opinion on aspects of their partner, as well as the conversation itself. My thesis uses all of these sources of information to investigate the underlying dynamics of a conversation.

The first study in my thesis used keystroke timing to infer characteristics of *dialogue acts* in a conversation, or the illocutionary function of an utterance (Stolcke et al., 1998). I use typing patterns to answer whether these different dialogue acts have different typing patterns associated with them. This is important because a dialogue act such as a question would necessitate a very different type of reply compared to a statement. If a computer agent or a human interlocutor could gather more information about the type of dialogue act being produced, then they could also generate a more appropriate response.

Study 2 looks at adjacent pairs of utterances and infers underlying sentiment changes as well as the effect of a participant's opinion of their partner. A unique aspect of typing in dialogue, as opposed to in isolation, e.g. answering essay prompts or typing a thesis, is that a participant's utterances are dependent on, among other factors, previous context as well as the participant's overall impression of their partners. I find that typing patterns provide additional information about underlying sentiment and opinions, beyond only lexical information. This is the primary concern of the burgeoning field of *affective computing* (Buker and Vinciarelli, 2021; Picard, 2000): Not only is it important to understand the sentiment of a single utterance, but also how shifts in sentiment are manifested, as well as the overall emotions of a user.

Finally, my third study looks at the complex sentiment of rapport (Tickle-Degnen and Rosenthal, 1990), and how well a neural network can predict low rapport between partners. Because rapport is multidimensional, keystroke patterns provide an ideal production modality given that their patterns are sensitive to a number of influences, both social and cognitive. In a series of experiments, I test the predictive power of the full set of keystrokes, as well as subsets based on the participant's role in the conversation and subsets based on temporal slices. While the temporal subsets provide roughly the same amount of predictive accuracy, the subset of keystrokes collected when a participant is receiving recommendations is especially accurate. Predicting low rapport of a receiver is important in many settings, such as patient/provider or IT professional/user, where it is essential to maintain high rapport so that provided recommendations are well-received. Keystroke patterns allow for a continuous, non-obtrusive method for monitoring these feelings of rapport.

My findings have implications for bandwidth-mediated theories of computer-mediated communication, as well as channel expansion theories (Gergle, 2017; Walther, 2011, *inter alia*). In addition, my studies are based on the cognitive concept of Implicit Prosody (Fodor, 2002b), and so my findings could provide support for this theory. Finally, my findings also bring up many ethical issues surrounding keystroke monitoring, and these are discussed as well.

Overall, my studies show how keystroke patterns are sensitive to a number of social dynamics and can detect signals that lexical information alone is less sensitive to. Uncovering these signals and sharing them with interlocutors, whether humans or computer agents, can improve the expressiveness of a conversation.

As a final note, my data as well as the code used to collect the data are publicly available at https://github.com/angoodkind/KiDcorpus. This corpus of data should be valuable to researchers in many different areas, and can be used to expand upon the foundations established in my thesis.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Every day, we communicate through computers on projects ranging from a group lunch order to office presentations to critical medical decisions. And every day, we also *mis*communicate through computers: We don't pick up on an intentionally humorous response, or we miss the criticality of a request. This is made more frustrating because if these responses or requests were made in a face-to-face setting, these underlying intentions would be easier to pick up through tone of voice or the rate of speech, i.e. spoken prosody. My thesis aims to use timing patterns in typing, called *keystroke dynamics*, to detect these underlying motivations, and make the information normally available only in face-to-face interactions also accessible in a text-based interaction, where prosodic information is assumed to be lost.

This thesis explores text-based computer-mediated collaboration through multiple levels of granularity. These levels of granularity are important because conversations and user intent are best understood at multiple levels of context. For example, an utterance might be best understood through the context of the preceding utterance or an utterance might be best understood through the context of the entire preceding conversation. As a realization of this uncertainty, Faggioli et al. (2021) built a hierarchical conversational model that maps all dependencies from local semantic dependence within an utterance to dependencies between an utterance and the entire preceding conversation; this mapping allowed the researchers to better trace a single conversational thread through the entire conversation.

The notion of dynamic context in a conversation is fundamental to the task of conversational analysis, and thus is essential for understanding latent variables that are underlying an entire conversation (Park et al., 2018). Clark (1996) calls this "layering," in which common ground, which emerges throughout the course of a conversation, requires understanding an utterance at multiple levels of meaning. These multiple levels are created in a "context space," which ties together an utterance with the context necessary to understand it (Reichman, 1978).

It is also important for my thesis to investigate a conversation at multiple levels of granularity because causation and psychological intention often do not exist at the same level of granularity. For example, Peng et al. (2022) shows that while the cause of emotional distress can exist at the level of an entire conversation, the expression of distress or psychological intent may only exist at the level of a single utterance. Therefore, being able to study a conversation at multiple levels of granularity is essential for understanding the mindset that underlies the words of a conversation.

In addition to studying a conversation at multiple levels, this thesis also investigates keystrokes as an both independent variable for predicting mindsets, as well as a dependent variable that is predicted by a known mindset. The reasoning behind this is that these studies are designed to study correlations rather than causation. Because an instrument such as an experimental intervention was not used (cf. Liebman and Gergle (2016b)), it is difficult to make assertions about whether typing behavior causes changes in mindset or vice versa.

The work in this thesis uses keystroke patterns because the typing modality of language production combines the spontaneous and dynamic elements of spoken language production with the static elements of finalized written text (Barrett et al., 2018a; Chen et al., 2021; Pinet et al., 2016). Keystroke patterns include everything from typing speed (Bajaj and Kaur, 2013), to the location and duration of pauses (Medimorec and Risko, 2017), to the mistakes and revisions a typist makes (Brizan et al., 2015; Pinet and Nozari, 2021). These production patterns can illuminate cognitive processes and social dynamics that are not overtly evident from surface-level or visible word choice; however, these processes exist in the latent patterns involved in moment-by-moment production, and can provide insight into why a user is taking a certain action. While the ultimate aim of my thesis research is to make these patterns more immediately beneficial to human-to-human interactions through a computer, they should also be expanded and refined for human-to-computer interactions, e.g. talking to a chatbot. Making keystroke patterns evident to a human interlocutor might include converting typing patterns into a visualization that a message recipient can see; in the case of chatbots, this could mean sending a computer agent both word-based features based on visible message text, and also augmenting these with keystroke-based dynamic production features.

However, keystroke data, especially when collected remotely (as I do in my thesis) can be especially noisy and unpredictable. As a simple example, if a user pauses in their typing, there could be myriad reasons for the pause (Conijn et al., 2019): a typist could be thinking and planning or they could be distracted by a fly in the room.

Further, typing patterns are only revealing about a user when they can be compared to a baseline for that user. For example, if User A types faster than User B, this does not provide any information about the cognitive state of each user. Moreover it wouldn't even answer who the more adept typist is: User B could be a fast typer but is having a bad day. Although the studies in my thesis control for individual participants, the studies do not establish a baseline against which to compare typing patterns in the experimental dialogue.

Before proceeding though, it is important to acknowledge the ethical issues surrounding keystroke collection. This will be expanded upon in the Overall Discussion (Chapter 8) but it is critical to address at the beginning. Keystrokes are a biometric and can be used for identification, similar to a fingerprint or face recognition (Monrose and Rubin, 1997a). In addition to personal identification, keystroke patterns can also be used to identify demographics such as gender or education level (Brizan et al., 2015, inter alia). However most major internet browsers make it easy for developers to collect keystrokes (Acien et al., 2021). Thus, it is important to consider ethical implications when collecting keystroke information, as this information is both private and highly informative. Efforts are being made, though, to anonymize keystroke information so as to keep this information private (Monaco and Tappert, 2017).

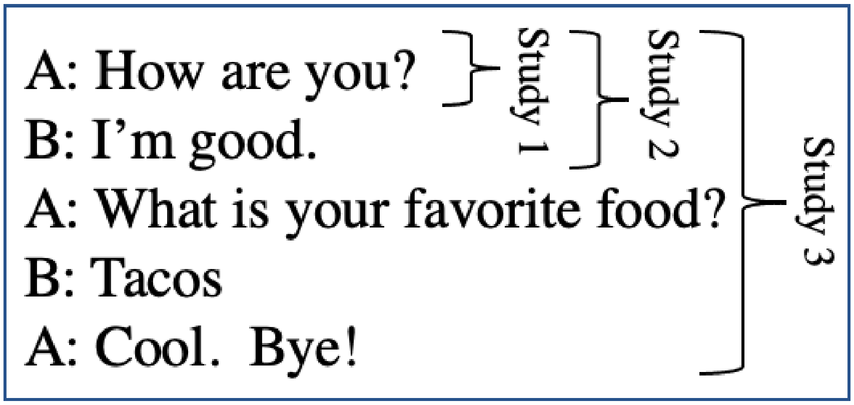
To further complicate the establishment of a baseline typing pattern, it is also well-known that typing patterns differ depending on the task at hand, as well as the application within which they are typed (Barghouthi, 2009). While the experiments in my thesis were run in a single interface with a single experimental prompt, findings such as those in Barghouthi (2009) and subsequent studies point to the notion that keystroke patterns detected in one setting might not translate to another setting and therefore may be less generalizable. Despite these limitations of studying keystrokes, my thesis represents an important starting point for combining human-computer interaction with keystroke dynamics, while basing its hypotheses on principles from cognitive science.

While the application of my findings will be to benefit computer-mediated communication, my studies are also motivated by theories from the cognitive sciences. In particular, the Implicit Prosody Hypothesis (e.g. Fodor, 2002b, and expanded on in Section 2.1) can provide insights into why cognitive processes loom large over typing production. Implicit prosody posits that even when people are not speaking out loud they still use prosody when reading silently or typing. The evidence for this is that “hearing” a voice in one’s head helps to add structure to language, and is seen in eye tracking or comprehension tests (Breen, 2014a). As such, one aim of my thesis is to elucidate how typing patterns parallel spoken prosody so that the information contained in spoken prosody, e.g. altered meaning from a different tone of voice, can also be used in a text-based environment.

As an example, when we speak to a person we dislike, we tend to use a different tone-of-voice, a different speaking rate, and make different word choices (Fujie et al., 2004). In response to a terrible idea, we may say *That’s a great idea!* but the phrase is laden with irony and its facetiousness is readily evident. In text-based communication, while we can sometimes use fonts (Heath, 2021), emoticons (Yuasa et al., 2006), or punctuation (Gregory et al., 2004) to convey prosody, often the correlation between (written) orthography and (spoken) prosody is insufficient or misunderstood (Heath, 2021). As a result, our understanding of our relationship with an interlocutor may be inaccurate and uncertain.

My studies aim to understand the underlying mindset or sentiment of a user, in order to make text-based communication more multi-dimensional and better represent the true intentions of speakers.[1] However, see the Overall Discussion, specifically Ethical Considerations, for a discussion of circumstances in which a user may not *want* their true intentions or mindset to be known to their interlocutor.

1.1 Studies Overview



A: How are you?
B: I'm good.
A: What is your favorite food?
B: Tacos
A: Cool. Bye!

Figure 1.1
A comparison of the scope of each of my studies. Study 1 looks at the dialogue act functions of an individual utterance; Study 2 looks pairs or *dyads* of utterances; Study 3 looks at entire conversations.

For each conversation, I used crowdsourcing to randomly pair two participants, who engaged in a 16-minute conversation. One participant began by discussing their movie and television preferences, and then the other participant provided recommendations. I collected timing information on every keystroke, as well as the final transmitted message.[2] I then used various machine learning techniques, ranging from linear models to neural networks, to study the connections between keystroke timing and features of conversations.

The three studies were structured as follows:

[1] As a note on terminology, I will often use the term *speaker* to refer to the person producing language, whether they are speaking out loud or producing messages in a text-based environment.

[2] Details of data collection are provided in Chapter 4.

Study 1 - Utterance-level dialogue acts: An utterance's functional purpose in a dialogue will change the way the words are typed. When engaged in a conversation, some utterances are intended to make reference to previous remarks, e.g. a clarifying question, while other utterances are intended to progress the conversation forward, such as a statement introducing a new topic. The function of an utterance is referred to as a dialogue act (Stolcke et al., 2000). Sometimes the same or similar words can function in both of these capacities. For example, in the dialogue *A: It's nice outside. B: Okay*, B's response could convey different messages: *Okay* can be said as a statement, to acknowledge A's assertion, or *Okay* could be said inquisitively, as if to question why A made that assertion. This can be confusing in a text-based communication context, since an interlocutor only receives the static or final word-form, rather than any inflection or timing patterns, where these features can be informative as to whether the utterance is, e.g., referring to previous utterances or is being used to introduce a new topic (Lai, 2012).

Study 1 aims to differentiate between typing patterns of dialogue acts that require very different responses. In the example above, A's response to *Okay* should look very different depending on whether B's utterance was an acknowledgment or a question. In addition, Study 1 investigates whether 8 primary types of dialogue acts have unique typing patterns associated with them. These findings are useful for elucidating the cognitive properties that underlie the production of different dialogue acts, and whether certain dialogue acts require more or less cognitive effort to produce.

Additionally, findings such as those above would parallel findings in studies such as Dhillon (2008), which found similar pause differences in spoken language production. By illustrating these parallels, it demonstrates the correspondence between spoken prosodic difference and type-written timing differences, allowing future researchers to design text-based algorithms, such as chatbots, using principles derived from spoken prosody.

Study 2 - Turn-level sentiment analysis: Study 2 analyzes sentiment at the conversational level. The first half of the study uses a collection of keystroke predictors to predict sentiment categories, such as positive, negative, extreme and neutral. Since these sentiment categories have been shown to be detectable in isolation, e.g. Brizan et al. (2015), this part of Study 2 also acts to demonstrate the sentiment findings from typing in isolation also extend to typing in dialogue.

However, dialogue is unique in that it is a joint action, and should not be considered in isolation (Clark, 1996). As such, Study 2 also investigates elements of sentiment that are unique to dialogues. Rather than studying only the sentiment of an isolated utterance, Study 2 also looks at how *change* in sentiment between utterances affects typing production patterns. In addition, a user will naturally form an opinion of their conversational partner, whether in spoken conversation or text-based conversation (Koudenburg et al., 2017; Murray et al., 2010). As such, Study 2 used regression modeling to determine how typing patterns are affected not only by the sentiment of an individual utterance but also by the user's overall opinion of their conversational partner.

Study 3 - Conversation-level rapport: Study 3 investigates whether the level of rapport experienced by conversational partners can be predicted using typing patterns. This is important for communication because in many settings – conversing with a friend, a customer, or a medical patient – establishing rapport is important in order to achieve successful collaborative outcomes, such as better peer tutoring (Olsen and Finkelstein, 2017) or patient satisfaction in a therapeutic setting (Leach, 2005).

This study uses participants' self-reported impressions after the experimental conversation, which asked them to measure aspects such as rapport, enjoyment, and self-awareness. It then builds machine learning models using keystroke-based features to accurately predict a typist's sense of established rapport.

Study 3 also looks at how different subsets of the conversation can predict rapport, where subsetting either looks at timing slices of a conversation or focuses only on certain roles that a participant is playing. This is important because rapport prediction is helpful *during* a conversation, where making adjustments to perceived rapport can improve the outcome of an interaction (Gidron et al., 2020). Moreover, in a setting such as a customer service interaction, the person of interest would be the customer, who is only receiving, not providing, service.

Each study investigates dialogue at different levels of granularity: single utterances, pairs of utterances, and entire conversations. Since my goal is to ultimately make the underlying motivations of latent typing patterns salient to a partner, it is important to only extract the most relevant data that is actually connected to those motivations, rather than extracting every pattern in typing production.

My thesis is structured as follows: In the next two sections of the introduction, I review the motivations for my thesis research and then enumerate the research questions that each study will answer. Chapter 2 then reviews prior literature that is relevant to the entire thesis. Chapter 4 explains how the data was collected for my thesis and exactly what data was collected. Chapters 5, 6 and 7 each contain the individual studies of my thesis. Each of these chapters reviews specific prior literature, explains results, and then includes a discussion of that specific study. Finally, Chapter 8 ties together each of the individual studies in this thesis and provides an overall discussion of how each study elucidates the central theme of my thesis.

1.2 Motivation

1.2.1 Importance of better understanding text-based CMC

The COVID-19 pandemic, and its effects on remote working, have added a tragic emphasis to the need for a better understanding of computer-mediated communication, as text-based CMC has come to occupy an even more central role in our lives. In a Pew Future of Work study from December 2020, researchers reported that 57% of respondents often or sometimes use chat-based platforms such as Slack or Google Chat (Pew Research Center, 2020).

Interestingly, when compared to video-based platforms such as Zoom, text-based platform usage was consistent across educational and income demographics. On the other hand, for video-based communication platforms there was a clear divide with better-educated and higher-income individuals using these platforms more often.

This shifting platform usage trend has a strong effect on the work environment, especially as "Generation Z" enters the workforce with very different technology expectations than prior generations, specifically in expecting more remote work and remote communication tools (Janssen and Carradini, 2021). The 2021 Work Index Trends report from Microsoft notes that a manager's role is increasingly centered around keeping a team connected and monitoring employees' well-being and mindset (Microsoft, 2021). However, Muir et al. (2017, p. 526) points out that when power relationships are asymmetric, e.g. a (higher-powered) manager and a (lower-powered) employee, "managers can find maintaining positive working relationships and good levels of rapport with virtual team members particularly challenging when relying on instant messaging to communicate." While it is impossible to remotely interact with every employee multiple times a day, as a manager might do in a physical office, keystrokes provide a unique way to do just that.

Importantly, though, in comparison to straightforward email monitoring using keywords, keystroke patterns also offer anonymity, in that keystroke patterns can be analyzed independently of the actual lexical content of what is being typed (see Section 2.2.6). This can be thought about with an analogy to spoken language: If a close friend is especially stressed, but their words are hard to discern, we can usually still tell that it's our friend and that they're stressed because of the uniqueness of their voice and spoken stress patterns. Similarly, keystroke patterns can offer this same level of unique anonymity, where timing patterns are unique, regardless of the actual words being typed. At the same time, insights at this level of cognition raise serious ethical and privacy issues. These are addressed regularly in my thesis, as well as in the IRB submitted and approved for data collection (Section 4.4.1).

As of this writing (March 2023), while many workers are returning to their physical offices, 36% of workers are still working remotely (Flynn, 2022). While Zoom usage will likely be replaced by face-to-face meetings in physical offices, text-based communication will also likely still retain a more central role given its wider usage in work settings. (It will be very helpful to revisit this in the future, but given the popularity of platforms such as Slack and Microsoft Teams before the pandemic, it seems unlikely that their usage will suddenly shrink.) Regardless of how these trends continue to change in the immediate future, text-based CMC plays an important role in our day-to-day routines. This points to the importance of my thesis, as I aim to better understand the latent emotions and thoughts that lie behind the salient text that interlocutors see in text-based communication.

1.2.2 Ubiquity of Computer-Mediated Communication

To understand society's changing work and lifestyle settings even more profoundly, it is best to reflect on the title of Gergle (2017), "Discourse processing in *technology*-mediated environments" (emphasis added). Although "computer"- and "technology"-mediated environments are synonymous in many respects, subtle and less-subtle differences exist. The most apparent difference is that computer-mediated environments seem to conjure pictures of a user sitting down and using a laptop or desktop computer to communicate via a chat-based client such as Instant Messaging. Technology-mediated environments, though, seem to expand this picture to scenes such as controlling a smart TV with your voice, or engaging in a video chat on a smartphone.[3]

The difficulty, though, arises from the sources of information missing in these (non face-to-face) contexts. For example when chatting online with a customer service agent we do not have physical knowledge about the agent, such as facial expressions, and we cannot even be sure of their name or age. We use all of these to help make more informed decisions about how to communicate more efficiently. We also use information from prosody, such as rising or falling tones, as well as hesitations or rewordings (e.g. Fodor, 2002b; Shriberg et al., 2000; Snow, 1994; Trott et al., 2019). Given this, many early CMC researchers, using traditional communication principles, concluded that CMC is an impoverished environment, with narrow bandwidth and limited social presence (Walther and Parks, 2002). However, research such as Reid et al. (1997) then posited that social presence simply takes a longer time to develop via CMC, and so using the same timeframe to

[3] Although new technologies such as smart TVs may use mobile-based, touchscreen communication, the studies in my thesis were all completed on desktop or laptop computers, with full QWERTY keyboards. Nonetheless, one motivation for these studies is that they are first steps in understanding more broad phenomena such as mobile communication. Once fixed-keyboard communication is more thoroughly understood, it can be extended to (often noisier) mobile-based communication.

compare CMC to face-to-face communication does not provide an accurate picture of CMC. In my own work, this is one of the reasons for a longer experiment with a starting prompt, so that participants have more time to develop a relationship.

One of the main goals of this thesis, then, is to show that the information traditionally considered to be limited to face-to-face interactions or even audio/visual online conversations, is actually available in latent information available in keystroke patterns. This, then could add support to the Social Information Processing (SIP) theory (Walther, 1992). If reported rapport between participants is still high, and participants still exhibit and take advantage of differences in timing patterns, it could provide evidence for a central tenet of SIP: "...communicators are just as motivated to reduce interpersonal uncertainty, form impressions, and develop affinity in on-line settings as they are in other settings." (Walther and Parks, 2002, 535).

However, given that much of this information might only be present in latent keystroke timing patterns, designers could use the results of my thesis research to make this latent cognitive information more overt. This is underscored by Tidwell and Walther (2002)'s argument that unlike face-to-face communication, CMC offers individuals only limited opportunities to observe others unobtrusively or to gain information about them indirectly. But by using keystroke patterns to make information about mindsets more overt, we can enrich the CMC environment to allow for more successful interactions and collaborations.

1.2.3 Uniqueness of keystroke dynamics

While my thesis is focus on typing and keystroke patterns, research into spoken language research is exploding, with good reason (Sisman et al., 2021) More and more companies are turning to voice assistants for customer service, troubleshooting, and even solving complex issues (Rozumowski et al., 2020). In combination with these developments, more and more products are adding the capability to control devices with voice commands, from cars to televisions to toaster ovens. As such, there are many good reasons to be pushing ahead with research into spoken-language conversational research.

Procedure	Keystroke Research	Speech Research	Text Analytics
Data collection	Trivial using a keylogger and non-obtrusive	Can be non-obtrusive using just a microphone, but Svec and Granqvist (2010) shows, because different microphones often have different sampling rates and sensitivity levels, inter-microphone comparison is often difficult if not impossible.	Trivial using e.g. a Google Form or web-scraper
Measuring production data	Trivial, including disentangling overlapping keystrokes from different users	Extracting timing data from speech can be extremely complex and labor-intensive. The average time for measuring word onset is 4 hours for 1 hour of speech (Bazillon et al., 2008; Novotney and Callison-Burch, 2010). The task multiplies in complexity when voices are overlapping.	Timing data is not available for most written data. For research such as Kalman et al. (2013a), the timestamps when a message was sent are sufficient.
Available features	Timing, edits, final version	Timing, edits, final version	Final version

Table 1.1
A comparison of research stages in keystroke analysis, speech analysis, and text analytics. Most of these steps are more straightforward and simple in keystroke research.

That being said, we still use, and specifically type on, computers *a lot*. Statistics show that we spend upwards of 10 hours a day using technology, and it is not uncommon for technology use to be described as "nearly constant" (Twenge and Farley, 2021). Especially with a move towards more remote work arrangements, what was formerly "water cooler" chat can be replaced by an SMS text message or a direct message on Slack. To play on an old workplace cliche, many meetings *are* becoming emails.

Focusing specifically on improving CMC, keystroke research presents many advantages over research using other modalities of language production. To gain a better sense of this, Table 1.1 compares typing research to other modalities of language production, which should illuminate the relative advantages of this type of research. In the tables below, "text analytics" consists of fully written text in a final static form. For example, this could be an essay on a test or an email sent to a recipient. This excludes research into the process of writing, where content is being dynamically produced, possibly by hand, e.g. pen and paper. "Speech research" consists of studying vocal language production, whether a monologue or an interaction involving two or more participants.

1.2.4 Parallels to the Implicit Prosody Hypothesis

Studies such as Goodkind and Rosenberg (2015), Plank (2016), and Barrett et al. (2018b) show that timing phenomena present in speech data are also present in typing data. Fodor (2002a) describes this as “silent prosody” or the Implicit Prosody Hypothesis (see Section 2.1 for a full explanation of the theory and terminology). This is the phenomena experienced when a person is producing or comprehending language silently while still being affected by non-explicit prosodic contours. For example, when reading silently, people often still hear a voice in their mind, where this voice is akin to an out-loud voice with pauses and tone. This voice adds prosodic contours to static text, and is reflected in the speed at which people read different parts of the text.

The same phenomenon applies to typing, where even though people are *producing* language rather than comprehending it, they still hear a voice in their mind as they type, whether the person is interacting with another agent or producing a thesis in solitude; importantly, this silent “voice” alters the speed at which people type. Although my thesis does not explicitly investigate the precise parallels between spoken prosody and silent prosody in typing, many of the typing features I use have an analog in spoken language.

This is important because (explicit) prosody, i.e. prosody that is audible in a spoken language or visible in a signed language, signals everything from importance of the words being produced (Swerts and Geluykens, 1994), to how much the words being produced are part of shared knowledge or common ground (Mushin et al., 2003), to when a speaker is planning to yield the floor to their interlocutor(s) (Gravano and Hirschberg, 2009). By looking at similar phenomena in Implicit Prosody, it becomes possible to understand which timing-related choices exist only in the mind, compared to those timing adjustments that are explicitly produced and only intended to aid spoken interactions. These timing features in typing can then be used to infer underlying cognition or motivations behind the words being typed.

Just as prior research on typing, such as Priva Cohen (2010), Logan and Crump (2011), and Plank (2016), has shown that findings in typing research can apply to the larger domain of cognitive science, the studies in my thesis can have the same type of benefit.

1.2.5 New keystroke dataset

The data collected for my thesis is also now publicly available. The corpus, called the Keystrokes in Dialogue (KiD) Corpus, is located at https://github.com/angoodkind/KiDcorpus. As of this writing (March 2023), the public repository is not complete; however, the following data will be made available:

- The full keystroke logs for each participant, complete with timing data
- The full message text and timing for each participant
- The survey responses from each participant, covering their impressions of their partner and the conversation
- Basic demographic information for each participant, including age, gender, and education level

1.2.6 Theoretical contributions

My thesis also makes contributions to three areas of inquiry: Human-Computer Interaction, Language Prosody, and Keystroke Research.

1.2.6.1 Human-Computer Interaction

My thesis focuses on human-to-human computer-mediated interaction, and studies the information that is present in the latent patterns of keystrokes. In the future, this information can be used to make social information more visible to conversation partners, for example by displaying a visualization of typing patterns that represents their mindset, so that one speaker can understand that their partner is excited or confused.

Adding explicit typing data to text-based CMC would add valuable empirical data to the social influence approach to media richness, as well as the channel expansion theory (Fulk et al., 1995; Walther, 2011). If adding typing data to CMC changes a user's perception of that modality and they elect to use it for more purposes, this bolsters support for a theory such as channel expansion, which posits that properties ascribed to various mediums are not fixed (Carlson and Zmud, 1999; Walther, 2011). This would also point to the limitations of media richness theory, which defines media richness based on *a priori* properties of media.

On a more practical level, this would also invite designers to use text-based CMC for a wider array of purposes. Channel expansion theory's core argument is that as individuals gain more experience with a particular communication medium, the medium becomes richer for them (Carlson and Zmud, 1994). Put another way, as users gain experience with text-based CMC that is augmented with visualized typing data, they will learn how to encode and decode more expressive messages, and use it for more purposes.

These findings can also be extended to settings such as moderating an online discussion forum. While a moderator may not be able to review every single post, my findings will allow moderators to detect questionable postings. For example, if the typing patterns of the poster seem indicative of a personal opinion when it should be a fact, or extreme anger, this could raise a flag. The moderator could then review that post, or that user, more closely.

My thesis is concerned with human-to-human CMC, as opposed to human-to-computer CMC. While the findings from my thesis can eventually be applied to technologies such as automated assistants and chatbots, both of these introduce new variables, and thus fall outside the scope of my thesis.

1.2.6.2 Keystroke Research

Keystroke production is still a relatively understudied domain of language production, when compared to speech analysis and text analysis. Moreover, very few studies exist that investigate keystrokes in dialogue: most study keystrokes in monologue, or isolated settings such as writing essays or entering passwords (see Section 2.2.5 for a review of keystroke research in dialogues). Thus, these studies will help to expand the domain of keystroke research.

In addition, the web interface developed and used in my data collection (described in Section 4.4.3) is valuable for future researchers running task-based or interactive studies while simultaneously collecting detailed keystroke information. The code base for the interface is available in the same repository as the dataset corpus (https://github.com/angoodkind/KiDcorpus). Since the data is publicly available, it can be used to address numerous other questions about keystrokes and keystrokes in dialogue, e.g. related to demographics, power dynamics, or remote keystroke collection. As mentioned previously, not only is the amount of extant keystroke datasets small, but keystroke data from interactions is almost non-existent.[4] Thus, the research conducted for my thesis could also benefit future researchers who are investigating related questions.

[4] See https://vmonaco.com/datasets/ for a thorough list of available keystroke datasets.

Chapter 2

Related Work

This thesis brings together three areas of research that have thus far not been fully integrated: dialogue analysis, keystroke pattern analysis, and speech prosody. I begin by setting up prosody, especially *implicit* or *silent* prosody. I then introduce keystroke dynamics, providing a background as well as the diversity and depth of keystroke research. Following this, I present relevant aspects of sentiment analysis and rapport with a partner. Throughout this section, I highlight relevant prior studies that have preceded the studies in my thesis, and show how these studies can be expanded by the studies proposed in my thesis.

2.1 Explicit and Implicit Prosody

Prosody is the study of all the elements of language that contribute toward acoustic and rhythmic effects, but primarily those that occur above the individual sound or phoneme level, and instead are focused on longer segments of language. For example, prosody studies the different tone of voice we use when making a statement versus asking a question. It also studies the rate of speech, such as why we enunciate a complex word more slowly and precisely. Finally, prosody looks at why certain words are said more energetically than others, and how a speaker decides at what volume to produce a word or sentence (e.g. Pierrehumbert and Hirschberg, 1990; Selkirk, 1995). Taken together, these aspects of language production are called “prosodic contours of speech,” as they shape the language being produced.

2.1.1 Implicit Prosody Hypothesis

Almost all studies of prosody investigate *explicit* prosody, i.e. sounds that are audibly perceptible. This thesis, though, builds upon the concept of *silent* or *implicit* prosody (Fodor, 2002b; Lovric, 2003). The Implicit Prosody Hypothesis (IPH) investigates the prosodic contours projected silently onto stimulus, e.g. during silent reading or when typing on a computer. The IPH says that the prosodic contours and boundaries which are audibly apparent, such as speaking rate, also impact the speed at which we comprehend and read language, e.g. we read at the same rate that the words would usually be spoken.

To measure silent reading speeds, researchers use high-precision procedures such as eye-tracking, which can measure exactly how long an eye is focused on a single word or sequence of words. Prior research has observed changes in reading time that take place at specific points in a sentence where spoken prosodic variations usually occur. For example, Ashby and Clifton Jr. (2005) found that words with two stressed syllables (e.g. *ULtiMAtum*) are read more slowly than words with one stressed syllable (e.g. *inSANity*). The researchers take this as evidence that readers routinely assign stress patterns to silently read words. For an overview of other empirical observations see Breen (2014b).

2.1.2 Implicit Prosody and Keystrokes

Keystroke analysis is well-suited to picking up on a user’s implicit voice. As pointed out in Galbraith and Baaijen (2019):

> [S]peaking prevents the monitoring of inner speech. By contrast, writing, partly because of its slower output, but mainly because it is produced manually rather than vocally, allows-—indeed encourages—-monitoring of inner speech.

Previous studies of dialogue usually collect one of two types of data: spoken conversations, e.g. the Switchboard Corpus (Godfrey et al., 1992), where timing metrics are available but laborious to measure, and text-based messaging, e.g. an online chat such as (Liebman and Gergle, 2016a), where entire messages are analyzed, sometimes along with the time when the message was transmitted.

By studying keystrokes in dialogue, I can gain insight into the use of silent prosody during interactions rather than in isolated activities such as silently reading. This also provides a significant opportunity for HCI and text-based CMC: It was traditionally assumed that when spoken speech is absent from a conversation, as in a text-based dialogue, that some source of information is lost, or at least significantly altered (Daft and Lengel, 1986). By investigating *implicit* prosody in text-based chats, I am able to capture aspects of sentiment and thought that were previously thought to be limited to *explicit* prosody. Finding evidence of prosody in text-based chat bolsters constraint-based theories of CMC, since it points to the notion that the same processes are involved in text-based CMC, but represented differently (Clark, 1996; Gergle, 2017).

As an example, hesitancies and revisions (that a speaker produces when they are stressed or confused) are evident in explicit prosody, but are difficult to detect in a finalized textual message. The keystroke patterns that go into the creation of that message, though, may provide evidence of the unstable thought patterns underlying the message.

2.1.3 Spoken prosody and the current studies

Spoken prosody is relevant to all of the studies in my thesis, and so insights from speech could be very instructive for my own investigations of typing. All of my studies will be explained in more detail in their respective chapters, and so the list below is a brief introduction to the relationship between prosody and the studies in my thesis.

A more comprehensive list of parallels between spoken prosody and prosodic patterns in typing is produced in Appendix C. As my thesis does not look for direct analogs between speech and typing timing patterns, but rather is motivated by speech prosody and the IPH, these direct parallels are not immediately germane.

- Study 1 looks at *dialogue acts*, and the unique characteristics of producing different types of dialogue acts. Stolcke et al. (1998) and many studies since have shown that prosodic properties of speech improve the accuracy of predicting the type of dialogue act.

- Study 2 looks at how sentiment and a user's opinion of their conversational partner affects their typing patterns. Studies such as Gravano et al. (2011) show how prosodic characteristics of speech affect social perceptions, while Li et al. (2017a) demonstrates that sentiment analysis can be significantly improved using prosodic information.

- Study 3 looks at how well typing patterns can predict the rapport that the typist feels towards their partner. Lubold and Pon-Barry (2014) found that elements of spoken prosody can be used to facilitate detection of rapport levels.

The evidence above points to how important prosodic information can be in collaboration. In fact, it is also well-established that prosodic contours are important for successful communicative outcomes (e.g. Pierrehumbert and Hirschberg, 1990). For that reason, providing additional temporal-based information to a text-based conversation can aid all participants involved in a conversation.

As an example of a connection between spoken prosody and typing patterns, Kalman and Gergle (2009) finds that the same types of sounds elongated in spoken prosody are also produced as repeated keystroke characters in typed text. Moreover, the authors find that much like the same speech prosodic contour can be used for multiple dynamic effects, typists employ repeated letters for different effects, as well.

Ballier et al. (2019) provides an interesting preliminary look at the relationship between speech prosody and keystroke patterns, as well. Their findings are limited, and are primarily at the syllable-level, as opposed to higher-level analysis, such as sentence- or paragraph-level. Nonetheless, their results are still interesting and relevant to the present study, and provide a rationale for the distinction in my studies between *inter*-word pause timing and *intra*-word pause timing, because syllable-level pauses would primarily be evident only within a word.

Finally, it is well-established that many syntactic unit boundaries are well-marked by changes in spoken prosody (Vicsi and Szaszák, 2010), i.e. a noun phrase is demarcated by longer pauses or larger pitch changes at its edges. Plank (2016) looks at the granularity of syntactic information available in keystroke data. The authors ask whether pause times can be used to locate the boundaries of every prepositional phrase, or only the boundaries of every sentence. By adding pause time data to a feature-set in a bidirectional Long Short-Term Memory (bi-LSTM) model, the researchers are able to improve over baseline accuracy in part-of-speech tagging and chunk labeling. For this reason, the studies in my thesis also look at pause times at phrasal boundaries, which are very similar to syntactic unit boundaries. If these timing differences are important in dialogue typing, as well, then it points to the use of silent prosody in typing.

2.2 Keystrokes

Keystroke studies enjoy a long history, going back to at least the 1920s Coover (1923). In World War II, Allied forces analyzed the unique production timing patterns of telegraph operators, called the "Fist of the Sender". Since each operator had a unique temporal signature, and individual telegraph operators travelled with specific troop battalions, this analysis allowed Allied forces to track the movements of different Axis troop units Banerjee and Woodard (2012).

But while keystroke dynamics has its origin in identifying individuals by the timing of the dots and dashes they produced, modern studies of keystrokes have exploded in both the breadth of human behavioral traits that are studied, as well as the fine-grained level of detail at which these areas are studied. This section will begin with an overview of how keystroke timing is measured, and how these measures are built into higher-level features. Following this, because my thesis will study complex social processes, this section will then show the diversity of areas that keystroke analysis touches upon, to set up why keystroke analysis is an ideal method to measure multidimensional behavioral patterns.

It is also important to explicitly mention that producing language on a traditional keyboard is still a highly relevant phenomenon that requires more detailed study. While speech-based computer-mediated interactions continue to grow in popularity, keyboards are still used on a near-daily basis in many facets of everyday life. As stated by Conijn (2020), quoting Brandt (2014):

> Writing is omnipresent in our society and plays, more than ever, an important role in our daily communication, work, and learning (Brandt, 2014). As Deborah Brandt puts it, millions of people (including myself) spend more than half of their working day "with their hands on keyboards and their minds on audiences" (Brandt, 2014, cover).

Keystroke analysis has also moved beyond QWERTY keyboards, and can be applied to tablets and smartphones that use touchscreens and swiping across multiple "keys" at once (Saevanee et al., 2012; Villani et al., 2006). This is important for the applications of my research, because many computer interactions are not limited to users sitting down at desktops: they can take place in conference rooms, on the go, and with each participant using a different modality. As an example of this diversity, a recent report from the Pew Research center found that 60% of Americans prefer to get their news from mobile devices, while the other 40% prefer desktop computers or television (Pew Research Center, 2019).

Further, keystroke analysis is not intrusive in the way that attaching sensors for galvanic skin response or an iris scan require significant interruption in activities (Fairclough, 2009). Rather, keystroke analysis can repeatedly and continuously measure a typist's behavior without any incursion into the daily keyboarding habits of the user (Locklear et al., 2014; Vizer and Sears, 2017).

2.2.1 Advantages of keystroke-based analysis

The primary advantages of keystroke research are that it is relatively inexpensive and unobtrusive to collect data, and relatively easy to analyze. As an example, one recent study analyzed 136,000,000 keystrokes from 480,000 participants (Dhakal et al., 2018). Prior studies, on the other hand, have estimated that transcribing an hour of speech data requires a trained researcher or professional anywhere from 4-10 hours (Bazillon et al., 2008; Novotney and Callison-Burch, 2010).

On the other hand, accurately determining the final text of a typing session is trivial, and timing measures such as pauses or keypress duration are not difficult to accurately determine in typing data (Dahlmann and Adolphs, 2007). Even when different computers with different keyboard layouts are used within a single experiment the measured timing differences are often negligible (Bridges et al., 2020; Pinet et al., 2017).

Another unique advantage of keystroke analysis is that the keylogs keep revision data completely intact. For examples, a user's final text might be "A bee," but the keystroke log might look like the figure below:

```
[SHIFT] [A] [SPACE] [B] [U] [G] [DELETE] [DELETE] [DELETE] [B] [E] [E]
```

In this case, we can easily recover the revised text. On the other hand, in spoken language production, if a revised word or phoneme was not fully articulated, it would be difficult-to-impossible to retrieve.

Similarly, natural dialogue contains a significant amount of overlap, where one speaker begins talking before another speaker has stopped talking (Heldner and Edlund, 2010a). Overlapping speech is much more difficult to transcribe than single-speaker speech, and just as with incomplete or corrected speech, it is difficult to extract meaningful timing data, such as pause timing.

This is important information to be able to recover because revisions contain valuable information about a typist. For example, Lindgren et al. (2019) points out that a revision that is immediate versus a revision after a pause indicates different underlying cognitive processes.[1]

Similarly, small revisions such as typos versus larger revisions such as correcting an entire idea also implies meaningfully different cognitive processes. Further, and pertinent to my thesis, Lindgren et al. (2019) points out that revisions can occur because of how a writer initially perceives

[1] While I do not delineate revision types in my analysis, but rather group together all revisions, the data is structured in such a way that this is trivial to extract, and will be a topic of future studies.

the reader, or changes their perception of the reader, analogous to audience design principles (Clark and Murphy, 1982; Horton and Gerrig, 2016).

2.2.2 Keystroke features

Before proceeding in any keystroke study, it is important to isolate the features being studied, and why they are being studied. To give an idea of the wealth of features available in keystroke analysis, Figure 2.1 reproduces a figure from Conijn (2020), which provides a succinct illustration of fundamental available features, which can be combined and expanded upon. There exist two primary dimensions that run through all of these features: the time interval *between* keystrokes, and the time duration for which a key was pressed. These features have analogs in speech data: the time taken to type a word is similar to the time taken to speak a word; the duration for which a typist holds down a key is similar to the intensity or loudness of speech.[2]

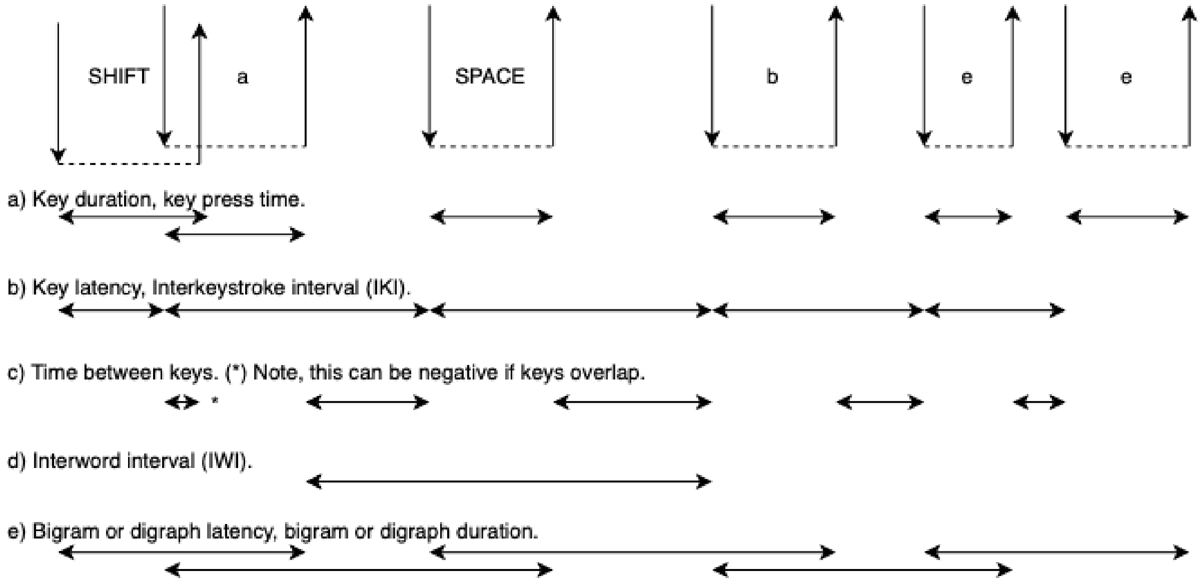

SHIFT
a
SPACE
b
e
e
a) Key duration, key press time.
b) Key latency, Interkeystroke interval (IKI).
c) Time between keys. (*) Note, this can be negative if keys overlap.
*
d) Interword interval (IWI).
e) Bigram or digraph latency, bigram or digraph duration.

Figure 2.1
A set of features available from extracted keystroke timing. Reproduced from Conijn (2020)

[2] A more intuitive analog in speech to keypress duration would be phoneme or letter duration. However studies such as Lee et al. (2015a) show that boredom and strong emotions affect keypress duration, because of the intensity of typing. Similarly, a highly emotional spoken phrase would be characterized by altered voice energy or volume.

Using the two dimensions of keystroke features (latency and duration), features of prior work can be organized into methodological categories. Table 2.1 is similar to a list of categories in Conijn et al. (2019); in addition I have also added a column with expected analogues in spoken language production.

Category	Examples in keystrokes	Analogues in speech
Pause timings or latencies	Interkeystroke intervals (IKI) between or within words, (e.g. Medimorec and Risko, 2017), or initial pause time, (e.g. Allen et al., 2016).	Direct parallels in speech, where pauses between words and word duration are used. Pauses can be unfilled (silence) or filled (e.g. *um* and *uh*) (Clark and Fox Tree, 2002).
Keystroke duration	How long a key is held down for. The duration of a keypress is often associated with excitement and emotional response (e.g. Epp et al., 2011)	Energy in speech (manifested as loudness or intensity) is indicative of emotion and cognitive load (e.g. Mijic et al., 2017). It is important to note that keystroke duration does not parallel speech duration, e.g. elongated syllables. In typing, something like repeated letter, e.g. *hiiii* better parallels elongated syllables (Kalman and Gergle, 2009).
Revision behavior	The number of backspaces (see Deane (2013)), or time spent in revision (Goodkind et al., 2017).	Utterances are often repaired and restarted in the middle of a phrase. How often and where a repair exists is useful for inferring cognitive properties of a speaker (Blacfkmer and Mitton, 1991).
Fluency or written language bursts	Sequences of text production without interruptions, such as the number of words per burst after a pause or revision (e.g. Baaijen et al., 2012; Van Waes and Leijten, 2015)	Language learners, whether children or second language learners, often only speak a small sequence of words fluently, with a pause, and then a resumption of speech (Housen and Kuiken, 2009).
Verbosity	The number of words (see Allen et al. (2016)), or the number of unique words or lemmas (Goodkind et al., 2017)	The number of unique words used, and different *types* of words (e.g. nouns, verbs, etc.) are often measured in speech as a metric of cognitive development (Yu, 2010).

Table 2.1
Categories of keystroke features, along with possible parallels in speech production.

As mentioned above, one advantage of features such as those outlined in Table 2.1 is that they are also infinitely expandable. For example, rather than grouping all inter-keystroke intervals, a researcher can subdivide this feature into linguistically-delineated features, e.g., intervals-in-verbs, intervals-in-nouns, etc. Prior research has found this approach to be more accurate than approaches without subdivision (Brizan et al., 2015; Goodkind et al., 2017; Locklear et al., 2014). Further, a feature can be subdivided by its statistical properties, e.g. the mean of all measurements, the standard deviation, the minimum value, the top quantile, etc. (Abadi and Hazan, 2020; Kołakowska, 2015, 2018).

2.2.3 Emotion and keystrokes

Prior research has shown that short pieces of text, such as blog entries, can be used to identify the emotions of a writer (Gill et al., 2008). However, just as emotion identification in spoken language is aided by a combination of text analysis and speech analysis, many studies have also shown that a combination of keystroke patterns alongside text analysis improves results (Kołakowska, 2013; Lee et al., 2015a; López-Carral et al., 2019).

Emotion can be classified either discretely or continuously, and keystrokes seem sensitive to both (Epp et al., 2011). In a discrete classification, emotions are categorical, e.g. happy, sad, neutral (Cowen et al., 2019). In a continuous classification, emotions are evaluated on dimensional spectrums, e.g. *valence*, or the degree of negativity or positivity, and *arousal*, the intensity of the evoked emotion (e.g. Lee et al., 2014). Lee et al. (2014) and Lee et al. (2015a) used the same general experimental framework but presented emotional stimuli visually and aurally, respectfully. The studies found that emotional valence affected keystroke duration, where a more negative emotion led to longer keypresses, interpreted as less energetic responses. On the other hand, arousal affected typing speed, in that the more intense the emotion, the quicker a subject would type.

For a more realistic and open-ended response, López-Carral et al. (2019) presented subjects with emotional images, varying in emotional valence and arousal, and then had the subjects type captions for the images. Interestingly, the researchers found effects similar to Lee et al. (2014)

and Lee et al. (2015a), where valence was negatively correlated to keystroke duration and arousal was negatively correlated to typing speed. However, López-Carral et al. (2019) found much more significant influences.

The results of all of these studies seem to point to two takeaways. First, emotion, evoked in different ways, can affect keystroke patterns. Secondly, the more naturalistic a typing experience is, the more strongly emotion affects typing. This points to the utility of my experiments, in that a dialogue will present a more naturalistic setting than responding to individual stimuli or typing a sequence of numbers.

My studies use takeaways from both research lines: sentiment and rapport are tested as both categorical variables as well as continuous spectrums that can be evaluated using a regression-type analysis. The studies cited herein seem to point to the value of both approaches.

Another important takeaway from these studies is that typing patterns are useful both as dependent and independent variables. My studies also extend this line of research. For example, Study 1 uses a binary classification task to measure how well a collection of keystroke variables can predict the correct dialogue act. On the other hand, within Study 2 I investigate how well the sentiment of an utterance along with the participant's overall opinion of their partner can predict the timing of specific keystroke patterns.

2.2.4 Cognition and keystrokes

Some foundational models of cognition, such as Rumelhart and Norman (1982) actually used typing to create holistic models of the interaction between language production and motor control. These models have been refined over the years, and more recent models of cognition via typing are able to detect two distinct, hierarchical cognitive processes during typing production: an "outer" loop that controls cognition at the level of word retrieval, and an "inner" loop that controls intraword, letter-by-letter word execution (Logan and Crump, 2011; Yamaguchi et al., 2013).

Vizer and Sears (2017) created a *continuous* classification system to measure cognitive demand. Unlike many classification studies that output a single discrete classification at the end of a training instance, a continuous classification system is constantly updating and changing its predictions. In a conversational situation such as a game or a troubleshooting call, cognitive demand will change, and so making predictions after all data has been collected is not necessarily useful or an accurate picture of changing demands. This is why Study 3 also examines the effectiveness of using subsets of typing data to predict a typist's mindset. For example, Study 3 predicts overall rapport given only the keystrokes from the first half of the conversation. While this does not constitute a *continuous* measure of rapport, it does provide a foundation for further subsetting.

Effects of cognitive changes are relevant to my thesis because providing recommendations requires a different amount of cognition from receiving recommendations. Moreover, as seen in studies such as Branigan et al. (2011), different perceptions of an interlocutor require different amounts of effort to formulate an appropriate response. If these differences carry over to dialogues, then noticeable differences should exist when a dialogue act changes in Study 1 or a participant's role changes in Study 3.

Importantly for the studies in my thesis, keystroke analysis has also been shown to be sensitive to the same temporal and intensity patterns seen in spoken language. As noted in Ballier et al. (2019, p. 363), "It may not be the case that the variation of typing speed mirrors the variation of speech rhythm, but comparable grammars of chunking can be carried out for speech and keylog data."

This observation is important because it demonstrates that typing production also taps into the same cognitive processes manifested in speech or language comprehension. As an example, in psycholinguistics it has been repeatedly observed that more uncommon words, or words with lower frequency, are more difficult and take longer to comprehend and produce. Along that line, Nottbusch et al. (2007) found that keystroke pause duration is correlated with both word frequency and word length.

In other studies, Plank (2016) found that pauses in typing correspond to boundaries in syntactic units (e.g. a noun phrase or verb phrase), and therefore can be used as a shallow syntactic parser. Similarly, Goodkind and Rosenberg (2015) found that typing patterns are sensitive to whether a word is part of a multiword expression or is a singleton. For example, the pauses around the phrase "muddying the water" would be more pronounced than the pauses around "sipping the water."

These studies provide motivation for the subdivisions of timing features used in my studies. For example, rather than solely considering the average overall dwell time for a user, my studies also look at features such as the average dwell time before or within content words.

2.2.5 Keystrokes in chats

While the vast majority of keystroke studies test a typist in isolation, keystroke analysis of chats has proven useful for a handful of other goals.

Buker and Vinciarelli (2021) used an experimental setup very similar to my own, and investigated similar questions. However, whereas Studies 2 and 3 in my thesis investigate a participant's opinion of their partner, Buker and Vinciarelli (2021) investigates how well keystroke dynamics can predict different personality traits. Roffo et al. (2014) found they could infer personality and identity in chats using keystrokes (although most of their features were based on lexical and stylographic textual features).

Borj and Bours (2019) used keystroke analysis to identify liars in a chat. The central finding was that being deceitful required more deliberate effort and less natural thoughts, and this different mode of thinking was evident in different typing patterns. This is relevant to Study 1 in my thesis, where a statement-dialogue-act might only report facts, whereas an opinion-dialogue-act might require personal imagination, though not necessarily deceit.

What ties all of these studies, as well as my studies, together is the notion that typing patterns reflect innate features of a typist. While this type of investigation is very relevant to HCI, my thesis instead focuses on the interactive aspect, and looks at how keystrokes can elucidate dimensions of an interaction, rather than a person in an interaction.

Each of the studies above *could have*, in theory, been conducted in isolation, since they only study each typist on their own. As a significant distinction, my thesis advances the stance taken by Clark (1996) that language is a game "designed for two" and can best be understood when looking at a dyad, or call-and-response, between speakers. The studies in my thesis use typing patterns to better understand the nature of a relationship, and how these patterns reflect the way a speaker feels towards their partner.

2.2.6 Ethical issues with keystroke collection

This section should conclude with a discussion of the ethical issues surrounding keystroke analysis. Keystrokes are a "biometric" or personal identifier, like a fingerprint or iris scan (Banerjee and Woodard, 2012; Epp et al., 2011; Locklear et al., 2014; Monrose and Rubin, 1997b). As such, collecting keystrokes without a participant's knowledge would be ethically murky at best, but more likely strictly unethical. Moreover, it is relatively easy to collect keystrokes, as all major browsers allow extensions to keep keylogs without a user even giving explicit permission (Morales et al., 2020). Because keystroke patterns can reveal information such as gender, age, education level, and native language (e.g. Goodkind et al., 2017; Tsimperidis and Arampatzis, 2020), the information contained in our keystroke patterns should be protected.

All of the experiments in my thesis obtained IRB approval, and all participant identities are anonymized during and after the experiment (see Section 4.4.1). I did wait to notify participants only *after* the experiment that their keystrokes were logged, so that they were not self-conscious about their typing. This notification, though, included the option to not share keystroke data if they object.

The importance of anonymization is especially relevant today, as technology firms devour enormous amounts of data and create massive open data sets. Because keystroke patterns can identify an individual, simply removing a proper name or email would be insufficient. This is specifically mentioned in Forsyth (2007), which was concerned with military-grade privacy masking. However, they acknowledge that names and usernames are often misspelled or abbreviated.

Nonetheless, recent advances in keystroke analysis have found success with "anonymizing" keystrokes, where the specific keys are unknown but only the typing rhythms overall are measured (Monaco and Tappert, 2017). As another attempt at further anonymization, Leinonen et al. (2017)

instead seeks to automatically remove all traces of keystroke patterns, such as revisions and timestamps, in order to truly deidentify text.

The success of studies such as Monaco and Tappert (2017) also points to how powerful keystroke pattern analysis can be. Given that the verification of an individual can still be made from keystroke patterns alone, without the context of the actual keys or letters produced, this demonstrates the extent to which typing patterns and practices are an innate and reliable signal, similar to the vocal quality of each individual, where the timbre of a voice is consistent regardless of exactly which letter they are pronouncing.

2.3 Dialogue

The analysis of dialogue between multiple entities differs substantially from traditional linguistic analysis. By "dialogue," I mean spontaneous or quasi-spontaneous interactions between two or more entities, where the utterances of one entity bear some relationship to utterances of the other entities. This can of course cause problems, where it becomes difficult to disentangle the direction of influence. Niederhoffer and Pennebaker (2002, p. 347) describes the problem succinctly: "What Person A says at Time 1 influences what Person B says at Time 1. But what Person B says at Time 1 also directly influences what Person A says (in response) at Time 2." While my thesis does not directly address this problem, it does take it into account in the controls and limitations of the studies.

In contrast to interactions, linguistics has traditionally concerned itself with planned, static written language that is independently motivated, with little-to-no interaction with other sentences. For example, a sentence such as "The cat the dog the man hit chased meowed." is of interest to those studying linguistic structure (namely center embedding), but would be very unlikely to occur in a spontaneous conversation, at least without significant pauses and pitch changes.

Another interesting way to view this distinction between dialogue and monologue is through the concatenation of two propositions. Kasher (1972) defines a sentence as "...a series of sounds that have a meaning.” As a continuation, Krauss and Fussell (1996) shows that in a *dialogic* view of conversation, *meaning* emerges through the conversational process, rather than from a single sentence or single utterance per se. Put another way, the utterances of interlocutors are tightly connected, and their meaning is shaped not only by the utterance itself, but also by utterances that were previously produced and may be produced in the future, including utterances from the speaker themself and from other speakers (Clark, 1996; Garrod, 1999). In my thesis this is reflected by the fact that features are engineered not only from a message in isolation, but also how a message relates to preceding and following messages.

As mentioned in Section 2.2, the vast majority of keystroke studies are conducted with a single entity typing text in an isolated environment, whether engaged in free typing or fixed-text typing. While a handful of studies such as Borj and Bours (2019); Bukeer et al. (2019); Roffo et al. (2014) use keystroke patterns within conversations to identify deception or gender, my thesis will make a novel scholarly contribution in using keystroke patterns to analyze the dialogue itself, and the interactive process that emerges in a dialogue, rather than each partner in isolation.

2.3.1 Conversation Analysis

The formal study of dialogue is known as Conversation Analysis (CA), and evaluates the unique dyadic nature of conversation. Conversation has been traditionally studied in naturalistic settings, such as a recording of an interaction between a telephone operator and an inquiring party (see Horton (2017) for an overview), rather than as a controlled experiment. My thesis uses a semi-naturalistic setup, where the dialogue is spontaneous, but the prompts and assigned roles are constant between experiments.

Rather than studying individual utterances, conversations are studied at the pair-level, which contains an utterance from one participant and an adjacent response utterance from another participant, an *adjacency pair*. Many types of adjacency pairs exist, such as *question-answer*, *greeting-greeting*, and *inform-acknowledge* (Stivers, 2012). Because of this, conversation has been studied at the group level (more than one person), as opposed to studying individuals and their mental processes in conversation. Again, keystroke-level analysis of conversational text will allow us to bridge the gap between individuals and dyadic constructs: Language production is a reflection of internal cognitive processes, while the text of a conversation also reflects group-level interaction. Both of these are trivial to capture and measure using the data collection methodology in my thesis.

A unique feature of conversational analysis that is not available in monologic speed is "turn-taking." This is the notion that one participant speaks, and then another participant speaks. However, recent studies have shed light on the degree to which orderly turn-taking is an idealization, rather than a reflection of everyday conversation. Heldner and Edlund (2010b) and Levinson and Torreira (2015) have shown that as much as 30-40% of corpora contain overlap and prolonged pauses.

Because overlap is pervasive in conversation, typing analysis again provides a unique advantage. Whereas in speech research the process of disentangling overlapping speech is difficult, in typing data it is trivial to connect keystrokes to each interlocutor, and also measure the length of time that multiple speakers were simultaneously typing, or overlapping.

Studies 2 and 3 further utilize the turn-taking nature of a dyadic construct, and specifically a spontaneous dialogue. Edelsky (1981) makes an important distinction in turn-taking between an *exclusive floor* and a *cooperative floor*. An example of an exclusive construct is a professor delivering a lecture, where they hold the floor until they entertain a question. On the other hand, cooperative floors exist only by the cooperative nature of a conversation, where one interlocutor waits for the other, but is free to interrupt at any time. For this reason, Study 2 only looks at turns that are not interrupted, so as to avoid any confounds introduced when the cooperative nature of a dialogue is violated. Study 3 looks at the ratio of interrupted and uninterrupted turns, to measure whether experimental partners are exhibiting the same level of cooperativeness.

My thesis will add more data to the language production processes that underlie turn-taking. Since CMC does not contain explicit non-verbal cues, and yet users engaged in CMC still report positive conversational experiences that are marked by few prolonged pauses, and few instances of one participant trying to type simultaneously to another participant typing a message, then signals aside from intonation must also exist in conversation.[3] If cues such as intonation or gesture are not available, then my thesis may shed light on what cues are available, which help promote a positive conversational experience.

2.3.2 Sentiment analysis in dialogue

Most sentiment identification in dialogue has been performed on audio data (Shon et al., 2021; Yeh et al., 2019). The prosodic features used in these studies, though, do have analogs in typing patterns. Yeh et al. (2019) used loudness, pitch and duration, while Shon et al. (2021) found that classifying emotion using "semi-labeled" input from both speech and text, separately, improved the accuracy of their system. This is helpful for my own studies, since I utilize both timing information and textual information.

A key difference between sentiment analysis in monologic text and sentiment analysis in dialogue is that monologues lacks context, in that there is little to no moment-to-moment coordination between the producer and their audience. Dialogue studies, though, highlight the "joint action" of language use (Clark, 1996).

2.3.3 Computer-mediated communication in dialogue

As mentioned in the Introduction, it is important to constrain our discussion of dialogue to a specific subset: text-based computer-mediated communication (CMC). Although I will continue to use the term CMC, the term sometimes is too restricting. A recent review of literature titled its chapter on CMC "Discourse processing in *technology*-mediated environments" (Gergle, 2017). Although "computer"- and "technology"-mediated environments are synonymous in many respects, subtle and less-subtle differences exist. The most apparent difference is that computer-mediated environments seem to conjure pictures of a user sitting down and using a laptop or desktop computer to communicate via a chat-based client such as AOL Instant Messaging. Technology-mediated

[3] However, see Riordan (2011) for a discussion of non-verbal cues used in CMC, such as emoticons and font changes.

environments, though, seem to expand this picture to scenes such as controlling a smart TV with your voice, or engaging in a video chat on a smartphone.

Below I provide a brief overview of relevant theories of CMC. While my thesis does not aim to provide a novel theoretical contribution to CMC, the findings of my studies bolster current evidence for certain theories. Further, the importance of theoretical inquiry should be underscored, because in a rapidly-changing world of technology: “We can’t keep up with new innovations, so we need theory and models that can.” (Scott, 2009, p. 754). While all of the theories below do not hold face-to-face communication as a “gold standard” (Gergle, 2017), they do all help to explain what strategies we use to replace the verbal and nonverbal cues used in face-to-face conversation, which are lost in text-based communication. (Riordan, 2011).

2.3.3.1 Theories of CMC

Social Presence Theory posits that the cues present in face-to-face communication are filtered out, and that a lack of cues makes communication feel less intimate and involved (Short et al., 1976). As Walther (2011) points out, though, this theory can be challenged by the fact that, on various occasions, people intentionally choose to use alternative forms of communication, even when face-to-face communication is available.

Other approaches such as Media Richness Theory (Daft and Lengel, 1986) and Social Information Processing (SIP Walther, 1992, 2018) posit that different mediums make different channels of information available to its users and that users adapt to the affordances of different channels. Further, users are willing to adjust the time-course of information transmission to accommodate the medium, since relationship development takes longer in a computer-mediated environment versus a face-to-face encounter (Walther and Parks, 2002). As opposed to a bandwidth-based theory, where a necessary hierarchy is erected with certain mediums lacking channels available in other mediums, the channels available in a given medium are adaptable.

The affordance-based theory of Clark and Brennan (1991) looks at eight different constraints, each unique to the specific modality or technology being used to communicate. Further, Clark and

Brennan (1991) also frames grounding within a cost-benefit analysis, in that each technology or modality imposes different constraints or costs on communication, which in turn impose different costs on grounding. As an example, if a conversation is extremely important or higher-stakes, such as a space shuttle launch, an interlocutor will be willing to pay a higher cost or expend more effort in order to establish common ground. This may take the form of a repeated phrase, or a more verbose statement that provides every detail and additional word.

To relate affordances to text-based chats (and keystroke analysis), we can look at the affordances of Reviewability and Reviseability. The former affordance speaks to the ability to review one's own utterances, or a partner's utterances, and is manifested by the ability to look at the history of a chat in most platforms. The latter affordance concerns the ability to change or edit an utterance, both before and after it is transmitted. As (Chafe and Tannen, 1987) points out, the amount of effort a speaker needs to put into being understood is drastically different when a message can be planned, reviewed and revised. Since spoken language is not effortlessly planned or revised, the result of these affordances is that speakers are held less accountable than writers, since an addressee expects written language to be more accurate and articulate (Horton, 2017).

One of the main goals of this thesis is to show that the information traditionally considered to be limited to face-to-face interactions or even audio/visual online conversations, is actually available in latent information available in keystroke patterns. This, then could add support to a social information-based theory. For example, if in Study 3 participants consistently report high rapport, it could point to the idea that participants take advantage of differences in timing patterns to adapt their communication to a text-based medium, ad thereby provide evidence for a central tenet of SIP: "...communicators are just as motivated to reduce interpersonal uncertainty, form impressions, and develop affinity in on-line settings as they are in other settings." (Walther and Parks, 2002, 535).

Chapter 3

Research Questions

Building upon the results of prior research, my thesis aims to expand upon these results by addressing the following question.

Study 1 examines dialogues at the level of the individual utterance, and specifically asks how dialogic function affects the way an utterance is produced.

RQ 1a) Can typing patterns predict differences in pairs of dialogue acts, where each member of the pair would require a very different response?

RQ 1b) Does each dialogue act have a consistent set of typing patterns associated with it?

Study 2 then looks at conversational sentiment analysis by modeling the sentiment of complete turns (composed of one or more successive utterances). In addition, it looks at the sentiment change between two participants' successive turns.

RQ 2a) Does keystroke information provide additional information about user sentiment, above standard lexically-determined sentiment values?

RQ 2b) Does keystroke information provide additional information about changes in user sentiment, when sentiment changes from one turn to the next?

RQ 2c) Are typing patterns sensitive to a user's opinion of their partner, when considered independently from the sentiment of a user's utterances?

Finally, **Study 3** uses typing patterns and stylographic features to predict whether a subject is enjoying a conversation. The study asks not only whether typing data from the entire interaction can predict rapport levels, but also how well specific subsets of data can predict rapport.

RQ 3a) Can typing patterns over an entire conversation be used to predict low levels of rapport between partners in an interaction?

RQ 3b) Can a random subset of keystroke data predict conversational rapport as well as a complete set of keystrokes?

RQ 3c) Does a subset of keystrokes from the first half of a conversation predict low rapport as well as a subset of keystrokes from the second half of a conversation?

RQ 3d) Does a subset of keystrokes from when a subject is providing recommendations predict low rapport as well as a subset of keystrokes from when a subject is receiving recommendations?

Chapter 4

Data Collection Methodology

In Chapters 5, 6, and 7 I will go into specific methodologies for each study. Before delving into the specifics of each study, though, this chapter first describes the characteristics of participants as well as the data collected, and then describes the overall methodology used for data collection.

The overall goal of my experimental setup was to emulate an online text chat environment, akin to what a person would encounter in chatting with online customer service, or engaging in a conversation on Slack. At the same time, behind the scenes, I aimed to collect not only the text appearing in the conversation, but the timing of every individual keystroke, from when a key was pressed to when it was released. This included not only visible keys such as letters and numbers, but also non-printing keys such as CONTROL or SHIFT. A screenshot of the experimental apparatus is provided in Figure 4.7. The individual components of the experimental setup will be explained in Section 4.4.3.

The experiments were run on the Prolific crowdsourcing platform, which is similar to Amazon's Mechanical Turk. However, Prolific has better quality control and is more respectful of participant privacy (Palan and Schitter, 2018). Participants were randomly paired, and pseudonyms were used to maintain anonymity. Further details are provided below in Section 4.3.

The experimental apparatus described below was used for data collection in all of the studies of my thesis. The entire experiment can be broken down into three phases:

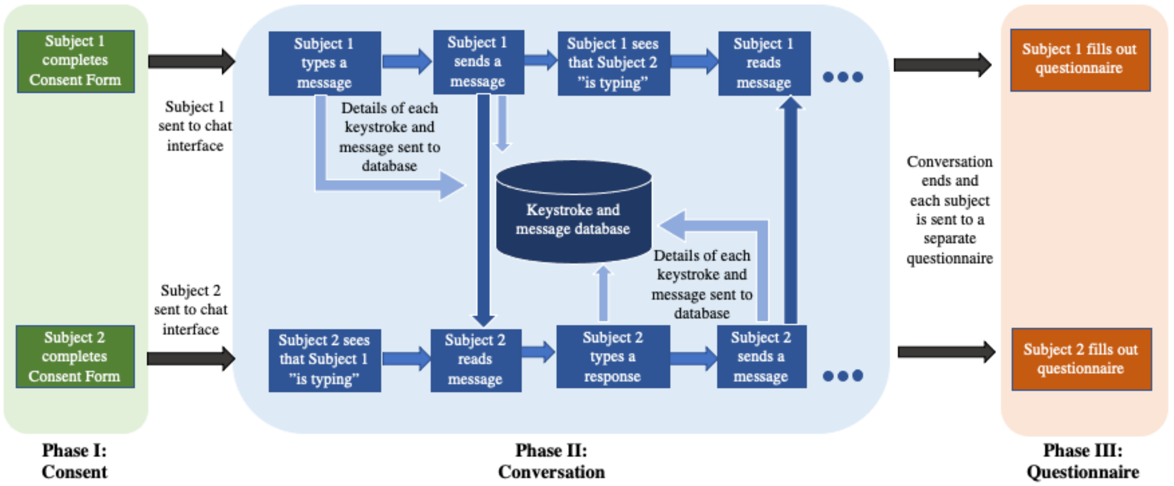

Subject 1 completes Consent Form
Subject 1 sent to chat interface
Subject 2 sent to chat interface
Subject 2 completes Consent Form
Subject 1 types a message
Subject 1 sends a message
Subject 1 sees that Subject 2 "is typing"
Subject 1 reads message
Details of each keystroke and message sent to database
Keystroke and message database
Details of each keystroke and message sent to database
Subject 2 sees that Subject 1 "is typing"
Subject 2 reads message
Subject 2 types a response
Subject 2 sends a message
Conversation ends and each subject is sent to a separate questionnaire
Subject 1 fills out questionnaire
Subject 2 fills out questionnaire
Phase I: Consent
Phase II: Conversation
Phase III: Questionnaire

Figure 4.1
A map of all 3 phases of my data collection experiment. The horizontal lines show how the experiment proceeds for each participant. The lines pointing to the central database are intended to convey that all keystroke and message information is being sent simultaneously, in the background, while the conversation is taking place.

- Phase I: The initial advertisement and instructions for the experiment were sent to potential participants. Importantly, and as will be discussed in Section 4.4.1, the advertisement and initial consent did not mention keystrokes, but rather only that I was interested in collecting opinions about movies and television shows. The participants were then directed to a consent form, which followed IRB guidelines. Upon consenting, the participant was routed to the conversational apparatus, or the actual experiment.

- Phase II: This is the bulk of the experiment. The participant took part in a 16-minute conversation about movies and TV shows. In the first half, one participant recommended entertainment to the other participant; in the second half the participants switched roles. At the end of the 16 minutes, each participant was routed to an individual post-conversation questionnaire.

  Importantly, the term "message" in Figure 4.1 is not just the sum of individual keystrokes. For example, if a participant types "`A SPACE B A T DELETE DELETE E D ENTER`," then 10 individual keystrokes will be sent to the database. However, only the final message, "`A BED`," will be recorded as the transmitted message in the database (and seen by the other participant). This highlights an advantage of using keystroke data: both the production process and final product are preserved.

- Phase III: The participant rated multiple aspects of the conversation itself, as well as aspects of their partner (see Section 4.4.5 for details about the questionnaire). Upon completion of the questionnaire, the participant was routed to a final consent form that informed them as to the true nature of the experiment (i.e. studying keystroke patterns themselves), and the participant was asked to consent again, now with knowledge of the entire experiment. The participant was also provided with the option to contact the researchers if they were uncomfortable with this information being collected. None of the participants had further questions, nor did any withdraw consent after full disclosure.

4.1 Demographics

In total I collected 102 usable conversations, comprised of 204 participants. No participants participated multiple times in the experiment. Partners were assigned randomly: The order in which participants joined the experiment determined which room they were assigned to and they were paired with the participant who joined right before or right afterwards. All participants in my study were required to currently reside in the United States and be native English speakers.

In my final data, 119 participants identified as female, 84 identified as male, and 1 preferred not to say. Almost half of the conversations (49) involved partners who identified as different genders. 34 conversations took place between participants who both identified as female, while 17 took place between two participants who identified as male.

The average age was 34 years old, while the median age was 32. The youngest participant was 18, while the oldest was 72. As can be seen in Figure 4.2, even though age constraints were removed, the vast majority of participants still fell between 20-40.

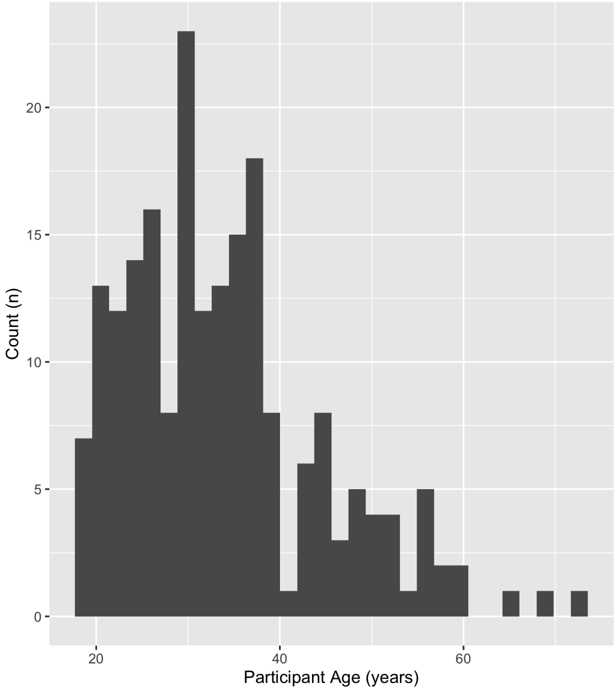


Figure 4.2
The overall age distribution of participants

The average age difference between partners in conversations was 12 years apart, with a median difference of 10 years apart. Figure 4.3 illustrates why removing age constraints did not materially affect the nature of the conversations.[1] The flat trendline strongly supports the notion that an increase in age difference did not impact the enjoyment ratings assigned by partners. In other words, a conversation between two 25-year-old partners was, on average, roughly as enjoyable as a conversation between a 25-year-old participant and their 40-year-old partner.

The majority of participants, 117 or 57%, have obtained an undergraduate degree or are in the process of obtaining an undergraduate degree. 58 participants have (only) a high school education, and 29 participants have or are completing graduate school (including 2 PhDs). Of those participants with current data as to their student status, 130 (64%) are no longer students while 41 are currently students.

[1] A smaller version of this analysis was run after the pilot study, which led to the decision to remove age constraints.

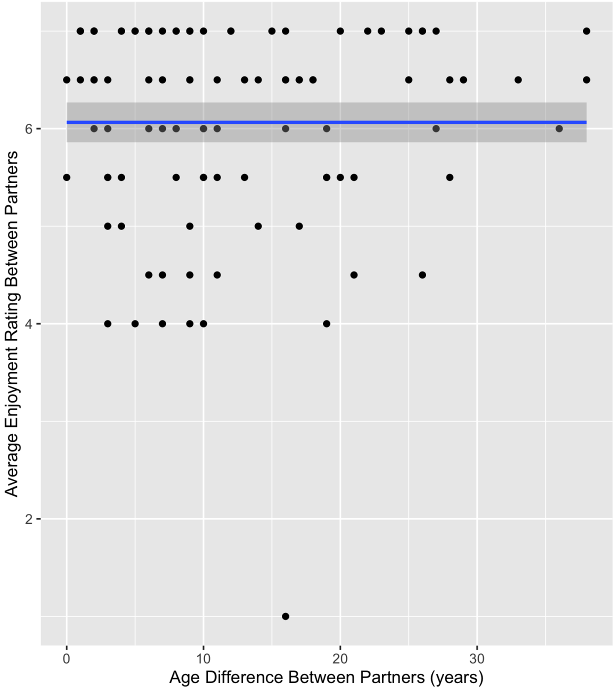


Figure 4.3
The average enjoyment rating between partners, as a function of the age difference between them.

Table 4.1 provides a detailed breakdown of employment status. Since the financial motivations underlying crowdworkers are important, it also seems important to highlight that the motivations for the crowdwork in my experiments were likely heterogeneous and of varying neediness.

Finally, the participants had familiarity with Prolific, and had also successfully completed a large amount of previous experiments. In this way, I ensured that typing patterns were not due to lack of familiarity with the platform, or due to a participant simply trying to rush through the experiment.

Employment Status	*n*
Full-Time	78
Part-Time	35
Unemployed (and job seeking)	22
Not in paid work (e.g. homemaker, retired or disabled)	19
Due to start a new job within the next month	4
Expired data or Other	46

Table 4.1
Employment status of participants

the average platform approval rating for participants was 99%. The lowest score was 96%, and 168 (82%) participants had a 100% approval rating. The average participant had completed 617 studies, with a median of 511 studies. Only 14 participants had completed less than 100 studies, while 27 had completed more than 1,000 studies.

As a final note, the demographics listed above are the only meaningful statistics released by Prolific. Prolific takes privacy very seriously, and does not release demographic information that could allow for a participant to be identified. Moreover, none of my experiments use or control for any demographics, as this was outside the scope of my studies. However, prior studies have shown that keystrokes can predict demographics such as gender or education level (Cascone et al., 2022; Tsimperidis and Arampatzis, 2020); therefore it is important in future studies to control these other factors in order to independently understand keystroke patterns.

4.2 Summary of Collected Conversation Data

The final experimental data was reviewed to ensure that anonymity was maintained and that participants remained engaged throughout the conversation. After all of the collected data was sanitized in this manner, it comprised 102 conversations, 4,895 messages, and 355,408 individual keystrokes. The average conversation contained 48 messages, with a median conversation length of 42 messages. A distribution of the conversation lengths can be seen in Figure 4.4.

The average message was made up of 10.7 words, with a median message length of 8 words. The distribution of different message lengths can be seen in Figure 4.5.

Interestingly, the message count between each partner did not differ significantly; however the *word* count between partners was strongly dependent on the role of the participant, as it related to whether they were providing or receiving recommendations. To verify this, a Welsh Two Sample t-test showed that the difference in the number of messages was almost non-existent ($p = 0.94$); on the other hand, the average word count of a recommendation provider's message was 11.1 words while a recommendation receiver's average message was only 10.3 words ($p < 0.01$). The difference

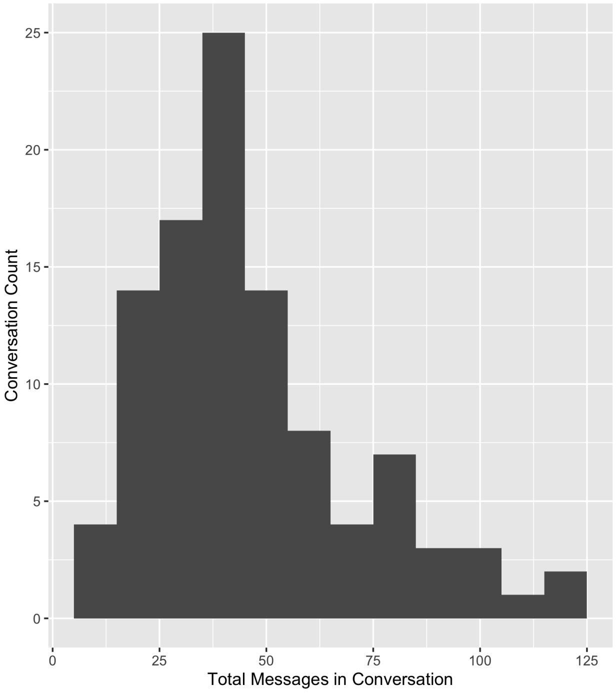


Figure 4.4
Distribution of conversation lengths, as measured by the total number of messages sent by both participants.

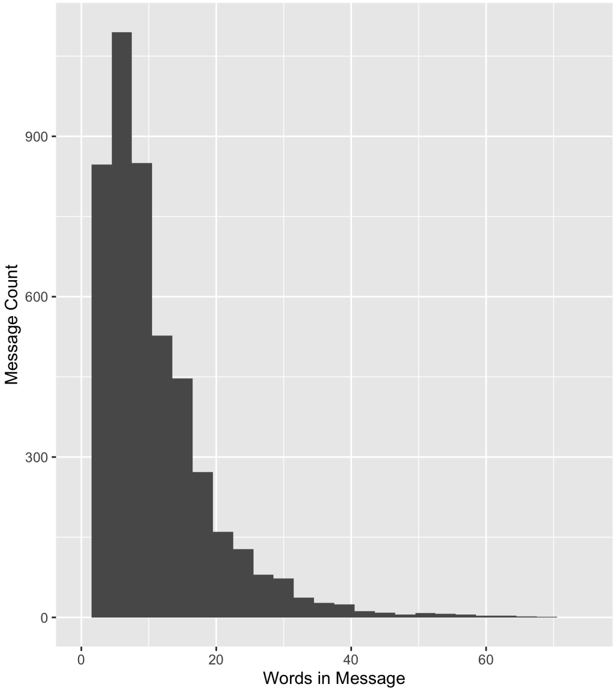


Figure 4.5
Distribution of message lengths, as measured by the number of (whitespace-delimited) words.

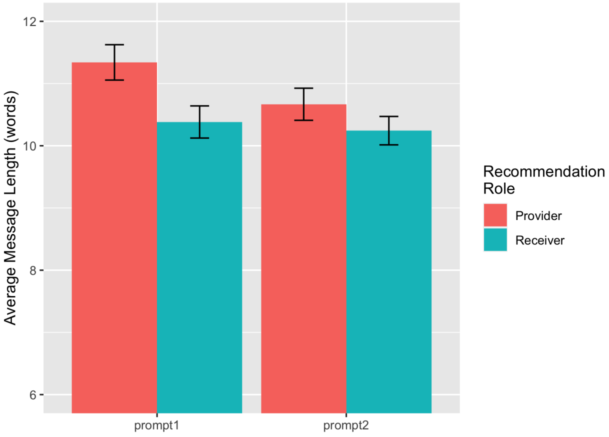


Figure 4.6
The average word count for each role type, as role relates to providing or receiving recommendations.

is illustrated in Figure 4.6, and broken up by prompt to illustrate that the differences existed in both halves of the conversation.

This will be expanded upon in Chapter 7, but it is interesting to note how this conforms with Herbert Clark's notion that language is a form of "joint action" (Clark, 1996), where every request or question from a recommendation recipient is responded to with a reaction or answer. On the other hand, as shown in Study 3, while the quantity of messages is equivalent, the contents of those messages as well as production patterns differ significantly.

4.3 Prolific Crowdsourcing

Because of the COVID-19 pandemic, collecting data in a traditional manner, such as having university students visit a lab for in-person experiments, was unfeasible and unsafe. Therefore, it was decided to use an online crowdsourcing platform instead. For my data collection I decided to use Prolific (Palan and Schitter, 2018) to run my experiments, rather than more popular crowdsourcing platforms such as Amazon Mechanical Turk (MTurk, Paolacci et al., 2010).

The benefits of Prolific over Amazon's Mechanical Turk are numerous. See Appendix B for an enumeration of Prolific's benefits. Studies such as Peer et al. (2022) demonstrate that the average data quality on Prolific is higher than on similar platforms, and users of Prolific are less likely to be using it as their primary source of income, which could make them less desperate to complete a task quickly.

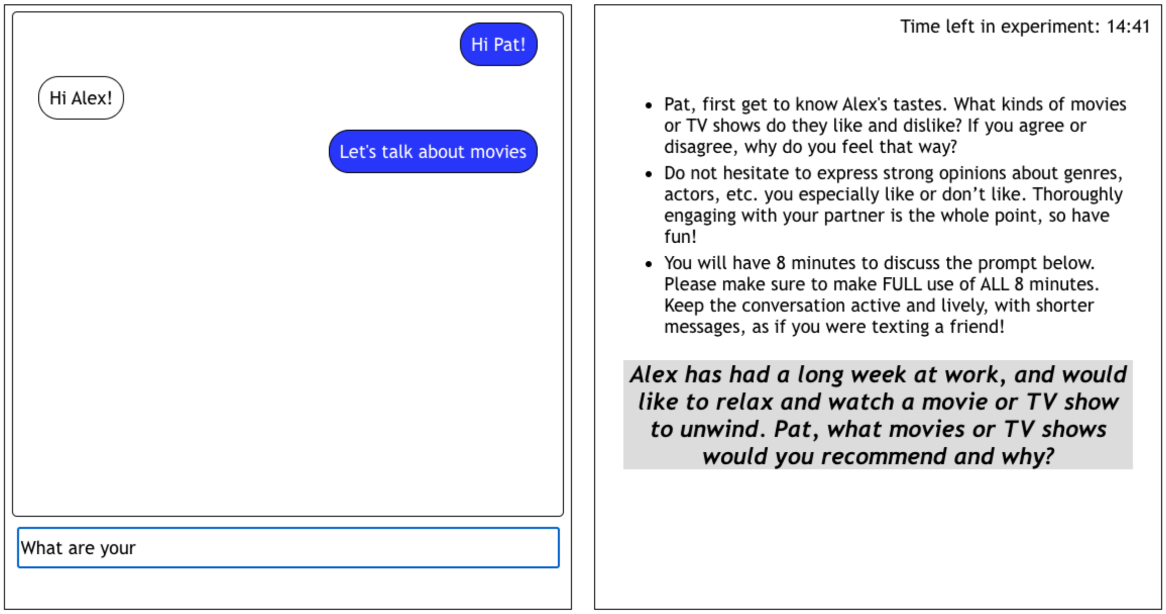
4.4 Experimental design details

4.4.1 IRB Approval

All experiments were approved by the Northwestern Institutional Review Board (IRB) before any data was collected for analysis. The IRB ruled the study exempt from further review, based on two qualifications: the experiments only involve “tests, surveys, interviews, or observation,” and are only “benign behavioral interventions.” Both of these are considered low-risk by the IRB. Further, all participants were required to be located in the United States, so that there was no conflict with international privacy laws such as the EU’s General Data Protection Regulation (GDPR).

A critical element of my experiment was that keystrokes were collected without participants knowing that I was also recording them at this level of granularity. This was important because I wanted to capture “naturalistic” conversations and language production, without participants feeling self-conscious of the way they were typing. Because of this, my experiment was considered to be deceptive, or at least providing "incomplete information" at the beginning. The advertisement for the experiment is reproduced in Figure A.1. Participants were initially led to believe that the purpose of the study was to understand why people prefer or disprefer certain movies, genres, etc., and what their rationale was. No mention was made before the experiment that the study was also investigating keystrokes and typing patterns.

However, keystroke data is private information; after the experiment we disclosed the full objectives of the study, and provided the participants the option to withdraw from the study but still receive full compensation. In actuality, none of the participants chose to withdraw consent after full disclosure.



Hi Pat!

Hi Alex!

Let's talk about movies

What are your

Time left in experiment: 14:41

- Pat, first get to know Alex's tastes. What kinds of movies or TV shows do they like and dislike? If you agree or disagree, why do you feel that way?
- Do not hesitate to express strong opinions about genres, actors, etc. you especially like or don't like. Thoroughly engaging with your partner is the whole point, so have fun!
- You will have 8 minutes to discuss the prompt below. Please make sure to make FULL use of ALL 8 minutes. Keep the conversation active and lively, with shorter messages, as if you were texting a friend!

Alex has had a long week at work, and would like to relax and watch a movie or TV show to unwind. Pat, what movies or TV shows would you recommend and why?

Figure 4.7
The experiment chat interface. This apparatus or web interface is what participants were viewing in Phase II of my experiment (see Figure 4.1).

This method of consent was also approved by the IRB. For further details of the IRB approval, see Appendix A.

4.4.2 Experiment Design Iterations

The experimental apparatus described in Section 4.4.3 is the finalized version that was used for data collection in my thesis. However, prior to this in both peer-testing and a pilot study, the experimental setup was improved through an iterative design process. Using feedback from both colleagues as well as participants in the online pilot study, adjustments were made to the timer displayed during the experiment, age constraints of the participants, and the wording of the experimental prompts. Appendix D provides the rationale for these changes and how they improved the data being collected.

4.4.3 Experiment Apparatus

The experiments were hosted on an Amazon Elastic Compute Cloud (Amazon EC2) instance. Because of the small footprint of the experimental apparatus, I was able to use a free-tier `t2.micro` instance, with only 1 GB of memory, running on Amazon's own Linux distribution.[2]

The actual experimental apparatus was written in JavaScript, HTML, and CSS. It was mostly built around a React (Meta Platforms, 2022) and `socket.io` framework (Rauch, 2013), in order to display, transmit and record every keystroke as soon as it was activated. The experiment worked on almost every major browser. The only exception, perhaps because of a CSS issue, was that the experiment did not work properly on Firefox. Participants were informed of this before starting the experiment.

The backend of the experiment automatically created new chat rooms for every two participants. In other words, when the first participant joined they were automatically routed to Room 1. The second participant was also routed to Room 1. When the third participant joined they were automatically routed to Room 2.

As seen in Figure 4.7, the experiment had two main components. On the left-hand side was the chat interface. This operated similar to a messaging interface on a phone, where the participant could see their own text in real-time as they entered it. In my interface, autocorrect, autocomplete and spellcheck were all turned off. Turning off autocomplete prevented participants from "entering" a long text string just by clicking their mouse. Turning off autocorrect meant that only the exact text entered would be transmitted, rather than a corrected version of the keystrokes entered. Finally, since spellchecks can differ from browser to browser and person to person, eliminating this option removed this variable from the experiment.

Once a participant pressed ENTER, the text was transmitted to their partner and displayed in the window on the top left. Similar to many text-based messaging interfaces, an "is typing..." indicator was added, so that participants would know when their partner was active. Importantly, the "is

[2] All of the code is publicly available. The front-end code is available at https://github.com/angoodkind/dialogue-keystrokes; the back-end code is available at https://github.com/angoodkind/keystrokes-collablab-backend.

typing“ indicator had a lag of two seconds, so that if a participant stopped typing for a moment, the indicator would not continuously disappear and reappear. The exception to this, though, was that the indicator would disappear as soon as a message was transmitted. This was important so that reaction times to messages would not be affected by a participant thinking their partner was still typing, and therefore waiting to respond.

The right-hand side displayed the conversation prompts as well as a countdown timer. The details of the prompt texts is discussed in 4.4.4. The experiment included two prompts so that in the first prompt one participant could provide recommendations, and in the second prompt that participant would be the receiver of recommendations. The timer on the top kept track of how much time was remaining in the experiment, so that the participants knew when they needed to wrap up.

Upon joining the room, participants were assigned the name either Pat or Alex. The purpose of assigning names was to help protect anonymity while still allowing each participant to refer to the other by their “name.” The names Pat and Alex were chosen because they are relatively gender-nonspecific, at least within the United States.

After each prompt was displayed for eight minutes, the experiment ended and the participants were automatically redirected to a post-experiment questionnaire, as well as a full disclosure about the experiment (see Section 4.4.1).

All keystrokes and transmitted messages were sent to a Firebase Real Time Database (Moroney, 2017). For every keystroke the database logged details of the key, e.g. not only that SHIFT was used, but whether it was the SHIFT key on the right-hand or left-hand side of the keyboard. The database also logged the exact time, using millisecond-level UNIX timestamps, when the key was pressed and released. For each transmitted message, the exact message as it was send to a partner was recorded, along with the time it was transmitted (approximately equal to when the ENTER key was pressed).

4.4.4 Conversational prompts

As mentioned previously, the entire experiment consisted of two prompts. The reasoning behind this was that in the first prompt, Subject 1 would recommend movies and TV shows to Subject 2, and in the second prompt Subject 2 would recommend movies and TV shows to Subject 1. This balancing between recommendation provider or recipient was to ensure that each participant had a chance to lead the conversation and respond to the conversation, assuming different "roles" in each section. If this balance had not been taken, then a possible confound in the experiments would be that a certain typing pattern is specific to one role (recommendation provider or recipient) rather than a behavior related to the overall social dynamics that I am interested in studying.

The topics of movies and TV shows were chosen because they are almost universal within the United States. While every person does not watch the same amount of TV and movies, it seems safe to assume that by age 18 almost every person has watched a fair amount of entertainment. Although I did not pre-screen for this or ask participants about this, this will be added in future iterations of my experiment so that viewership frequency can be factored in.

Some dialogues conveyed that one participant knew a lot about movies or TV and was possibly an expert in the subject. For example, one participant said of their partner, "my partner was well versed and knowledgeable about [movies and TV shows]," whereas another participant said "[my partner] did not have the necessary information to provide accurate recommendations." This variation is to be expected and is an accurate reflection of everyday encounters where people interact with others of varying subject matter expertise.

The two conversational prompts were:

Prompt 1

Alex has had a long week at work, and would like to relax and watch a movie or TV show to unwind. Pat, what movies or TV shows would you recommend and why?

Pat, first get to know Alex's tastes. What kinds of movies or TV shows do they like and dislike. If you agree or disagree, why do you feel that way?

Prompt 2

Pat is bored, and would like to watch a really thought-provoking or stimulating movie or TV show. Alex, what movies or TV shows would you recommend and why?

Alex, now get to know Pat's tastes first. What kinds of movies or TV shows do they like and dislike? If you agree or disagree, why do you feel that way?

Both prompts

You will have 8 minutes to discuss the prompt below. Please make sure to make FULL use of ALL 8 minutes. Keep the conversation active and lively, with shorter messages, as if you were texting a friend! Do not hesitate to express strong opinions about genres, actors, etc. you especially like or don't like. Thoroughly engaging with your partner is the whole point, so have fun!

The types of movies to be recommended were also changed between prompts in order to help the conversation stay dynamic. If the movie types were the same in both prompts, then it is possible that the second half would simply become a reflection of the first half, where Subject 2 simply recommended the same entertainment that Subject 1 had recommended in the first half.

Moreover, as stated in the experiment's instructions, discord was also encouraged. Although most conversations stayed positive, I also wanted to see how a person's typing behavior would change when reacting to a statement with which they strongly disagreed.

Finally, in order to maintain the conversational nature of the experiment, participants were encouraged to keep messages short. In my pilot study, participants would sometimes type paragraph-length messages where they stated all of their likes and dislikes, or all of the movies they would recommend. Since the goal of my overall thesis, though, is to look for typing analogs to conversational speech, I wanted the typed dialogues to more closely mimic the short, spontaneous utterances produced during a face-to-face conversation.

4.4.5 Post-experiment Questionnaire

Upon completion of the conversational phase of the experiment, participants were redirected to a questionnaire about the conversation in which they participated.

These questions were based on multiple related prior studies. Liebman and Gergle (2016a) and Liebman and Gergle (2016b) also studied human-human text-based dialogue, in which two participants had to resolve a moral dilemma. Similar to my own studies, Liebman and colleagues were also investigating the nature of computer-mediated communication. My questions were also based on those from Pecune et al. (2019), which investigated human-AI interactions for movie recommendations. Since Pecune et al. (2019) was also interested in what made a recommendation appealing, many of its questions were relevant to my own studies.

For the questions below, participants had to provide ratings from 1-7, based on a Likert scale (Joshi et al., 2015). The final two questions were open-text responses.

1. To what degree did you enjoy the conversation? [1=Not at all, 7=I enjoyed it a lot]
2. To what degree do you think your partner enjoyed chatting with you? [1=Not at all, 7=They enjoyed it a lot]
3. To what degree did the conversation go smoothly? [1=Not smooth at all, 7=Very smoothly]
4. Hypothetically, how much do you think you'd enjoy watching a movie with your partner? [1=Not at all, 7=I would definitely enjoy it]
5. How would you rate the level of rapport established between you and your partner? [1=No rapport, 7=Lots of rapport]
6. How likely do you think it is that you'll end up watching one of the movies your partner recommended? [1=Not likely at all, 7=Very likely]
7. In a few sentences, how would you describe your partner and the overall conversation?
8. Do you have any additional comments or questions for the study authors? [optional]

The responses to these questions were skewed very positively. Figure 4.8 shows this distribution. The studies in my thesis will discuss these distributions in more detail, but it seems that most subjects had high opinions of their partner as well as the overall conversation.

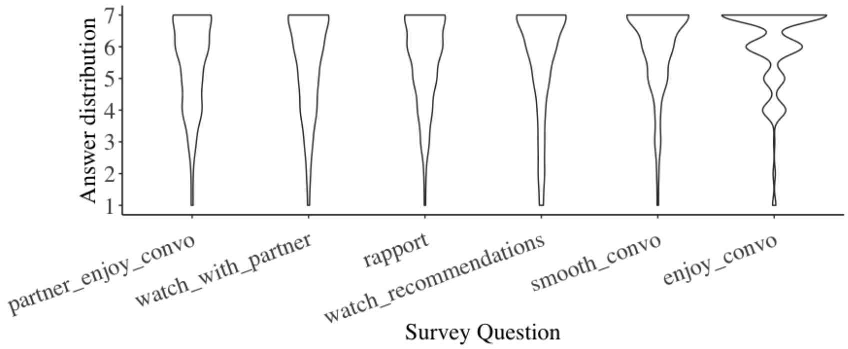


Figure 4.8
The distribution of responses to survey questions. All responses were heavily skewed towards positive ratings.

4.5 Keystroke Collection

From a methodological standpoint, it is also necessary to mention the nature of the keystroke timestamps that I collected. The majority of prior keystroke experiments were conducted in a lab setting, often with a desktop computer (e.g. Brizan et al., 2015; Killourhy and Maxion, 2012; Kuzminykh et al., 2020). This is done both to reduce latency between a keystroke and when that keystroke appears, as well as to establish consistency between each run of an experiment.

In my case, however, there existed variation in computer type (laptop, desktop, tablet with a physical keyboard), browser choice (Chrome, Safari, etc.), and internet connection (slow, fast, wired, and wireless). Future studies will pinpoint the differences between browsers and connections. In addition, the dataset released alongside my thesis includes details about the computers each participant used, so that participants can be partitioned in future studies.

4.5.1 Collected Features

The raw features sent to each of the database are enumerated in Tables 4.2 and 4.3. For the keystroke database, all information was collected for every individual keystroke. Although a separate entry was recorded for the key-press and key-release, these were immediately combined in the database

Keystroke Database

Feature	Description	Example
Experiment ID	An identifier for the session with two participants	E001
Subject ID	An identifier for the individual participant. This was used as the key to link keystrokes to personal demographics and questionnaire answers.	S001
Prompt Number	Which prompt the keystroke occurred within, so that I can reference whether the participant is the recommendation provider or receiver	Prompt2
Utterance ID	An identifier for the utterance within which the keystroke occurs. This was used as the key to link keystrokes to messages.	E001-S001-1
Raw Keystroke Char	A raw representation of a keystroke. For example, even if `SHIFT` was being held down, the E-key would be recorded as `e` rather than `E`. In addition, non-printing keys such as `SHIFT` would be recorded as such.	b
Visible Keystroke Char	The visible representation of a keystroke. A capital letter that is the result of a held-down modifier key would be recorded as such. Non-printing keys receive a null entry.	B
Key-press Time	A UNIX timestamp recorded when a key was pressed down	1642087297175
Key-release Time	A UNIX timestamp recorded when a key was released up	1642087297278
Existing Text In Message	The text that currently exists in the participant's textbox, that has not been sent yet.	"A "

Table 4.2
In the schema for the keystroke database, imagine that the participant is typing the `B` in the message `"A Bee"`

into a single entry. For the message database, all information was collected for every individual transmitted message.

4.6 Engineered Features

The studies in this thesis all require engineered features beyond the raw features that were collected during data collection. A complete list is provided below. The tables are broken down by engineered features of each keystroke, an individual word, a complete message, a participant overall, and an entire conversation. Each study/chapter will also make mention of the specific features used for that particular study.

Message Database

Feature	Description	Example
Experiment ID	An identifier for the session with two participants	E001
Subject ID	An identifier for the individual participant. This was used as the key to link keystrokes to personal demographics and questionnaire answers.	S001
Prompt Number	Which prompt the keystroke occurred within, so that I can reference whether the participant is the recommendation provider or receiver	Prompt2
Utterance ID	An identifier for the utterance within which the keystroke occurs. This was used as the key to link keystrokes to messages.	E001-S001-1
Sent Text	The full message transmitted	A Bee
Time Sent	A UNIX timestamp recorded when the message was transmitted	1642087297278

Table 4.3
In the schema for the message database, imagine that the participant has sent the message `"A Bee"`

Each of the features in Table 4.4 was measured for every individual keystroke. This includes both printing and non-printing keystrokes, as well as printed characters that were deleted and not in the final transmitted text.

Since my thesis also considers linguistic context, I also calculated features based on the complete word within which each keystroke took place. These engineered features are enumerated in Table 4.5.

The features in Table 4.6 were applied at the utterance level. More specifically, the features only apply to utterances that were transmitted and therefore viewable by both partners.

The features in Table 4.7 were calculated for each participant. These features therefore reflect overall traits of a participant, rather than features that are specific to a single utterance or even subset of utterances.

Finally, the features in Table 4.8 are applied to the entire conversation. Some features reflect the differences between participants, but because these are contrastive features they only apply to the pair of partners rather than an individual.

Feature	Description
Time Since Conversation Began (minutes)	The conversation is considered to have "began" when the first user transmits their first message
Interval Between Keystrokes	The time gap between when the previous key was released and this key was pressed. Cam be negative in the case of SHIFT being held while a letter is entered.
Keystroke Duration	The time elapsed from when this key was pressed to when it was released
Flight Time	The time between when the previous key was pressed to when this key was pressed. This will always be positive and can be more reliable.
Time Since Utterance Start	The time between when the first key in the utterance was pressed to when this key was pressed
Position in Word	The keystroke's position relative to the entire utterance. Position ∈ Beginning of Utterance, End of Word, Space, Beginning of Word, End of Utterance
Within Word	The word within which this keystroke occurs. Words are delimited by spaces, and may include nonsense words
Within Keystroke Sequence	The space-delimited sequence within which this keystroke occurs. This sequence will includes keystrokes such as DELETE

Table 4.4
Individual keystroke-level features

Feature	Description	Example
Word Length (characters)	The number of printed characters in a word	`Bee` = 3
Lemma	The lemmatized version of a word, or the "canonical" form of a word. This allows for comparison of the same root word but in different forms, e.g. past tense or pluralized. Lemmatization was performed using the `hunspell` package in R (Ooms, 2022).	`drove` = `drive`
English Word	Checks whether the word is a valid English word, using the `hunspell` English dictionary (Ooms, 2022)	`drive` = TRUE `droive` = FALSE
Semantic Category	Divides words into Function and Content words, where function words are words with a grammatical function, such as determiners and conjunctions and content words have specific meanings, e.g. nouns and verbs (Pennebaker et al., 2003)	`dog` = CONTENT `the` = FUNCTION

Table 4.5
Features added to each word, which is a space-delimited character segment.

Feature	Description
Continuation	Denotes whether this message is a continuation of a turn, where the same participant sent the previous message as well
Utterance Length (words)	The number of space-delimited words in a transmitted utterance
Time Delay After Previous Message Received	The interval between when when a message was transmitted and composition of the next message begins. For a message that is a continuation of a turn by the same participant, this number will always be positive, i.e. a person cannot begin typing a new message before the current message is transmitted. For different participants, this interval can be negative if a participant begins composing a message before their partner has transmitted a message.
Sentiment Score	For each transmitted utterance, sentiment scores were calculated using the `VADER` package in R (Hutto and Gilbert, 2014). A score above 0 denotes positive sentiment, while a score less than 0 denotes negative sentiment. Neutral sentiment (score = 0) is also possible
Sentiment Difference	The difference between the sentiment score of this utterance and the previous utterance

Table 4.6
Features added to each full transmitted message

Feature	Description
Overall typing rate	The elapsed time of all utterances the participant typed, as opposed to the elapsed time of the entire conversation, divided by the total number of keystrokes
Intra-word typing rate	The average interval between each keystroke within a word, rather than before or after the word. This metric tends to be a more accurate measurement of motor-based typing ability as opposed to language skills (Logan and Crump, 2011).
Inter-word typing rate	The average pause time before a word is typed, or the interval between a SPACE and the first letter of a word. Unlike intra-word typing rate, this rate is a more accurate measurement of lexical recall, i.e. retrieving a word from memory (Logan and Crump, 2011).
Average Length of Utterances	The average length of each utterance, with separate features for the number of words and the number of keystrokes
Edit Rate	The average rate of BACKSPACE or DELETE keypresses per visible keystrokes
Average Pause Before Responding	For non-overlapping messages, when a participant responds to the other participant's message, this is the average gap between receiving a message and initiating a reply
Questionnaire Answers	These are summarized in Section 4.4.5

Table 4.7
Features calculated for a participant overall.

Feature	Description
Conversation Length	Three separate features: the number of utterances, words, and characters in a conversation
Average Utterance Length	Two separate features: the number of words and characters in each utterance
Ratio of turns	A measure of how equally the participants produced contributions: the ratios of utterances, words, and characters
Average Questionnaire Scores	The average rating that each partner assigned to the other
Age Difference	The differences in ages between partners

Table 4.8
Features calculated for an entire conversation.

Chapter 5

Study 1: Keystroke patterns in dialogue acts

The study looks at how dialogic function, or “dialogue acts” (DAs), correlates with timing changes in typing production patterns. Dialogue acts are critical for understanding both computer-mediated human-human dialogues as well as human-computer dialogues. As Core and Allen (1997) explains, “the system must keep track of how each utterance changes the commonly agreed upon knowledge common ground” (p. 1). Dialogue acts are essential for this because they provide evidence as to whether knowledge is agreed upon and so a conversation can build upon it, or whether previously shared knowledge is still being questioned.

As outlined in Section 3, this study investigates the following two research questions:

RQ 1a) Can typing patterns predict differences in pairs of dialogue acts, where each member of the pair would require a very different response?

RQ 1b) Does each dialogue act have a consistent set of typing patterns associated with it?

By answering these questions, a system can better identify how each utterance functions in each conversation. As will be explained further below, different dialogue acts require different amounts of cognitive effort to produce (Gnjatović and Delić, 2013). Since keystroke production is sensitive to cognitive effort (see Section 2.2), using keystrokes to further understand dialogue acts is a fruitful direction for identifying different dialogue acts.

To answer these questions, this study compares dialogue acts from two different perspectives:

Study 1a takes certain keystroke timing metrics of an utterance and measures the differences in these metrics *between* a set of two different dialogue acts, where a proper response to each member of that pair would look very different. For example, I compare Opinion vs Non-opinion utterances, and find that these utterance types can be differentiated based on the typing patterns that are characteristic of each dialogue act type. The distinction between whether an interlocutor is saying an opinion versus non-opinion is important, as a proper response to an opinion, such as "I agree," would not be a proper response to a non-opinion.

Study 1b measures how timing metrics are produced *within* each and every dialogue act. Put another way, Study 1a asks if a set of certain timing metrics are unique for each dialogue act in a pair and can be used to distinguish one dialogue act from the other; its models take the form `Dialogue act` $\sim$ `Timing metric`. Study 1b asks if each dialogue act has clear-cut timing signature for each keystroke metric, where the question is whether certain production patterns are consistent for that dialogue act, *though the pattern need not be unique between that dialogue act and the other dialogue acts*, but rather just well-defined within that dialogue act. The models for Study 1b take the form `Timing metric` $\sim$ `Dialogue act`. I find that while some dialogue acts do have some reliably distinct keystroke timing metrics associated with them, the overall results are a bit murky. No dialogue act has consistent patterns for each keystroke metric, although most DAs have multiple unique typing patterns. Nonetheless, the findings are promising and point to avenues for future research.

As a final clarification, Study 1 is *not* a dialogue act classification task. While the findings of this study should be extended to improve the accuracy of classification, that task is outside the scope of this study. However, it is especially important to improve dialogue act classification by finding new sources of information, e.g. keystroke patterns. The reasoning behind this is that newer neural network models, especially context-aware attention-based transformers, have extracted very complex textual patterns (e.g. Malhotra et al., 2022). Thus, it could be fruitful to also look “beneath” the text to augment surface-level lexical information with latent keystroke production data.

5.1 Methodology

In order to perform dialogue act analysis of the collected dialogues, labeling of DAs was performed using both automatic classification as well as manual human verification of the labels. A total of 4,874 utterances were used for this study. Only 21 utterances from the original data (Section 4.2) could not be used because of technical issues such as no printed characters.

One luxury of typing data as compared to speech data is that utterance segmentation in typing data is trivial. For example, as Edlund et al. (2005) points out, when segmenting speech data annotators usually look for certain amounts of silence. This process is complicated on its own, but also made further complicated by the fact that silences also exist *within* sentences. In comparison, an utterance or sentence in typed messages is trivial to distinguish. Sentences within an utterance are offset by punctuation; an utterance is offset by the transmission of a message.

Although each message was considered an utterance, this also presented complications for assigning a single label to each sent message. As an example, if Subject 1 asked Subject 2 how they were doing, and Subject 2 replied (in a single message), “I’m well. How are you?” then the reply constitutes two different acts: 1) “I’m well” is a backward-facing response to a previous question, while 2) ”How are you?” is a forward-facing question.

To circumvent this issue, only the first sentence of multi-sentence utterances was evaluated. This affected 530 utterances of the 4,874 utterances. A qualitative review by my research assistant and me showed that very few utterances contained radically different sentence types. More often than not, sentences that were significantly different were transmitted as distinct messages. As an example of a multi-sentence message where both sentences are similar, one participant (Subject 16) said, "Kevin Hart is always funny. I also like The Rock."

Nonetheless, future studies should also look at every sentence of multi-sentence utterances. One danger in discarding everything but the first sentence is that Study 1 could have thrown out either valuable information, significantly different information, or the majority of information in a message if the first sentence was very short while the proceeding sentences were long.

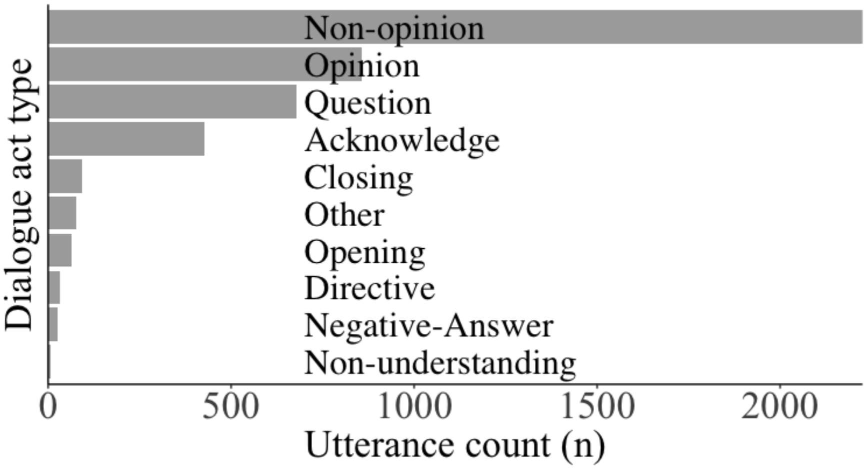


Figure 5.1
The distribution of high-level dialogue act categories within my collected dataset. These categories were mapped from the original 27 dialogue acts (see Table 5.1 for details).

As another approach to the issues of multi-sentential messages, Ivanovic (2005) which performed dialogue act classification on text messages, considered every individual sentence to be an utterance. This approach would not work for my own studies, though, because it would not allow me to look at metrics such as the pause times before an utterance for each DA. This timing issue was never problematic in the original Switchboard Corpus (Godfrey et al., 1992; Marcus et al., 1993), which was used for early dialogue act studies such as Jurafsky et al. (1997). Spoken conversations do not have an analog to a partial unsent message, i.e. messages are “sent” in real-time as they are spoken. An extended pause, e.g. between sentences, constitutes an utterance boundary.

As a first step in utterance labeling, all utterances were automatically classified using the `DialogTag` library.[1] Although the library’s author does not provide many details about how their model was trained, it is based on a Transformer model from Hugging Face, and uses the BERT uncased language model (Devlin et al., 2019). The classifier’s tuning was performed using a cross-entropy loss function.

Because my dataset is relatively small, the 27 different tags used by this classifier were grouped into 10 higher-level tag categories, to avoid overly sparse categories. These categories are Open-

[1] https://github.com/bhavitvyamalik/DialogTag

ing, Closing, Non-opinion (statement), Opinion (statement), Question, Acknowledge, Directive, Negative-answer, Non-understanding, and Other. The grouping is outlined in Table 5.1, while the distribution of grouped categories is illustrated in Figure 5.1.

Some of the original categories looked especially intriguing, especially for how they relate to the social elements overall thesis, and could require unique cognitive processes and thus exhibit unique timing patterns. For example, *collaborative completion*, *repeat phrase*, and *hedge* would all be informative concerning the social dynamics of a dialogue. Regarding collaborative completions, Poesio and Rieses (2010, p. 1) states, "Collaborative completions are among the strongest evidence that dialogue requires coordination even at the sub-sentential level."

Unfortunately, these categories occurred very infrequently in the data (< 5 utterances) and the classifier was unable to accurately detect them. In future studies, these types of DAs should be intentionally evoked, so that their timing patterns can be studied.

Final classification of dialogue acts was performed manually by a research assistant and me. We used the `DialogTag` classifications as a baseline, but made personal judgment calls if we felt a classification was incorrect. Approximately 15% of dialogue act labels were changed. This probably speaks to the limitations of the classifier; because of technical limitations it was the only one available to us. Because each utterance was considered in isolation, the classifier did a poor job on utterances that were Acknowledgments (of previous utterances), instead classifying them as Statements. Future work will use more sophisticated and context-sensitive classifiers.

Further, it seemed that the classification of Non-Opinion (statements) was too liberally applied. It seems that the algorithm adhered too rigidly to the coder's heuristic in the original SWBD-DAMSL coding scheme, "When in doubt, it is probably [a statement]" (Jurafsky et al., 1997, p. 22).

In order to rectify this, a keyword search was performed on utterances labeled as a Statement that searched for common terminology used when expressing an opinion. The list was composed of keywords on common lists used for teaching English to students. The words/phrase were: *I think, my favorite, love, hate, best, worst*. Utterances were manually reviewed to remove false-positive, e.g.

Original Dialogue Act	Mapped Dialogue Act Category	Forward/Backwards DA
Acknowledge (Backchannel)	Acknowledge	Backward
Agree/Accept	Acknowledge	Backward
Appreciation	Acknowledge	Backward
Backchannel in Question Form	Acknowledge	Backward
Conventional-closing	Closing	Forward
Action-directive	Directive	Forward
Negative Non-no Answers	Negative-Answer	Backward
No Answers	Negative-Answer	Backward
Statement-non-opinion	Non-opinion	Forward
Signal-non-understanding	Non-understanding	Backward
Conventional-opening	Opening	Forward
Statement-opinion	Opinion	Forward
Apology	Other	NA
Collaborative Completion	Other	NA
Hedge	Other	NA
Hold Before Answer/Agreement	Other	NA
Offers, Options Commits	Other	NA
Or-Clause	Other	NA
Other	Other	NA
Quotation	Other	NA
Repeat-phrase	Other	NA
Declarative Yes-No-Question	Question	Forward
Open-Question	Question	Forward
Rhetorical-Question	Question	Backward
Self-talk	Question	Backward
Wh-Question	Question	Forward
Yes-No-Question	Question	Forward

Table 5.1

The rows are sorted by the mapped column, which is the mapping used in this study. The first column contains all of the dialogue acts present in my data. The final column is whether the dialogue act has a forward or backward function, in line with Jurafsky et al. (1997).

Mapped Dialogue Act	Utterance Count	Example Utterances
Non-opinion	2246	*It is on Netflix.* *I watched Gone Girl.*
Opinion	955	*Best personality in the world, I think.* *The whole premise is so good!*
Question	728	*What type of movies do you like?* *Did you see the latest spiderman movie?*
Acknowledge	562	*Oh definitely.* *Yes!*
Closing	107	*It was nice chatting!* *Have a great day :)*
Opening	99	*Hi Alex!* *Oh, how rude of me, hello Pat.*
Other	98	*Training Day* *Less now!*
Directive	32	*Check out the trailer.* *You should give it a watch.*
Negative-Answer	28	*No, not really.* *Not yet!*
Non-understanding	18	*Who?* *4 hours?*

Table 5.2
The mapped dialogue act categories, along with a count of utterances as well as examples of each category.

the television show *Love Island*. Of 2,450 utterances initially classified as non-opinion statements, this filtering changed 350 utterances from Statement to Opinion.

5.2 Results

Given the size of my collected dataset, it would not be feasible to run a model that tries to predict every single dialogue act against every other, e.g. `All dialogue acts ∼ Timing metric`, using various keystroke-timing metrics as predictors. As can be seen in Figure 5.1, the dataset is simply too unevenly distributed, and too small to obtain robust statistics for each dialogue act.

Rather, I chose certain dialogue acts pairs of interest, where the contrast between the two would be important to distinguish because an appropriate response, either from a human or computer agent, would look very different (e.g. Matsumoto and Araki, 2016). Further, many of these DAs have distinct patterns in speech prosody, and so testing the distinction in typing is an important indicator of whether typing patterns bear parallels to spoken prosody (Benus et al., 2006; Hirschberg et al., 2005).

Moreover, it seems that the cognitive processes that go into the production of each could be very different. Gnjatović and Delić (2013) points to the variation in cognitive complexity of different dialogue acts, where different DAs require different cognitive efforts for retrieval and integration. These cognitive costs could manifest themselves as pauses or mistakes in typing. Thus, if a CMC system could also pull information from typing patterns, then it could make better inferences about the dialogue act being produced.

The pairs I chose to investigate are: Non-opinion vs Opinion, Statement vs Question, and Forward-facing vs Backward-facing dialogue acts. As an example of an in/appropriate response, if a partner expressed an opinion, an appropriate response would be to say *I agree*. However, if a partner made a statement of fact, then responding with *I agree* would not be appropriate.

To improve the modeling in both experiments, all predictors were standardized. However, it is important to clarify the scopes of Study 1a and 1b, respectively, so as to explain the different standardization procedures. When standardizing my data I had the option to perform by-subject standardization either using the entire dataset, or only the subset of two dialogue acts under consideration.

The fundamental question in Study 1a is whether two different dialogue act categories, e.g. Statements and Opinions, are produced distinctly from one another. Study 1a does not ask whether Statements and Opinions are produced distinctly within the entire set of all dialogue acts. Given the scope of Study 1a, standardization was performed only on the subsetted data, since I am interested in the distinctions within subsets, rather than overall distinctions.

In Study 1b, though, I am interested in whether each dialogue act has a robust timing signature. As such, in Study 1b standardization is performed on the entire dataset of all dialogue acts. See the discussion of this study (Section 5.3) for further details.

5.2.1 Experiment 1a: Timing patterns predicting dialogue acts

In order to test how well a set of keystroke metrics predicted differences in a dialogue act binary, I used a model with the predictors below. An exhaustive iterative process was used to test this final model; the process is outlined in Appendix E. Each model was built in `R` using the `lme4` package to run logistic regression models.

1. The pause between the previous message being sent and the beginning of the current utterance (pre-utterance gap)
2. Time gap between words 1 and 2
3. The interaction of the pre-utterance gap and the gap between words 1 and 2
4. Typing speed of word 1 (keystrokes/word duration)
5. Typing speed of word 2 (keystrokes/word duration)
6. The interaction of word 1 speed and word 2 speed
7. Utterance typing speed (keystrokes/utterance duration)
8. Utterance average inter-keystroke interval (IKI)
9. The interaction of speed and IKI
10. The edit count (pressing BACKSPACE or DELETE)
11. Typing speed variability (SD of IKI)

The full set of predictors was then applied to the each dialogue act binary seen in Table 5.3. Because standardization was performed for each binary, the model coefficients do not directly represent any definite unit of measurement. Rather, they are similar to a z-score and also represent the direction of the effect. The significance of each predictor was also measured, since the coefficients between models are not directly comparable.

After running the models, the effect of each predictor could be calculated. A more detailed discussion follows in the Discussion section, but below I also provide a high-level overview of the results:

The pre-utterance gap was significantly different for all 3 models. The gap was shorter for non-opinion utterances as compared to opinions, longer for questions versus statements, and longer for backward-facing DAs compared to forward-facing DAs.

While average IKI was only significant for opinions versus non-opinions, the typing speed of all 3 models was significant. An important distinction between typing speed and IKI is that the typing speed takes into account the duration of the entire utterance, rather than IKI which only considers the intervals between keys. Thus, typing speed would also be affected by pauses at the beginning and end of an utterance. The utterance typing speed was slower for non-opinion utterances, faster for questions versus questions, and slower for backward-facing DAs.

The amount of editing was significantly different for two of the models, as well. Non-opinion utterances had more edits than opinions, and statements had more edits than questions.

In terms of the typing speeds of words 1 and 2, only a few models showed significant differences. Word 1 was typed faster for statements versus opinions; conversely, word 2 was typed slower for statements. Word 2, however, was typed faster for forward-facing DAs.

	Dependent variable: Dialogue Act Binary		
Covariate	Non-opinion / Opinion	Question / Statement	Backward / Forward
Pre-utterance gap	0.05^{**}	-0.08^{***}	-0.09^{***}
Inter-key interval (IKI)	0.06	-0.08^{*}	0.05
Utterance speed	0.08^{**}	-0.16^{***}	0.10^{**}
IKI:speed	0.02	0.03	0.03
Edit count	-0.01^{**}	0.02^{***}	-0.001
Speed variability (sd)	0.02	0.05	-0.05
Word 1-2 gap	0.001	0.01	-0.02
Pre-utt-gap:word 1-2 gap	-0.05^{*}	0.02	0.01
Word 1 speed	-0.03	0.07^{***}	0.05
Word 2 speed	-0.004	-0.06^{**}	0.07^{**}
Word 1:2 speed	-0.04	-0.02	-0.03
Observations	2,965	3,592	4,111
Log Likelihood	-1,739.17	-1,610.93	-1,191.23
AIC	3,504.33	3,247.87	2,408.47
BIC	3,582.26	3,328.29	2,490.65

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.3
Results of predicting dialogue act binaries using different timing metric covariates

5.2.2 Experiment 1b: Dialogue acts predicting timing patterns

Experiment 1a established important timing metrics for dialogue act distinctions at multiple levels of granularity. Using the subset of metrics established in the first half, experiment 1b then flips the dependent and independent variables, looking at how robust each timing metric is for each type of dialogue act. As mentioned before, though, robustness is not the same as unique. Different dialogue acts could have robust timing signatures, but those signatures need not be unique from all others.

As a toy example, Study 1a demonstrated that keystroke metric 1 and keystroke metric 2 were able to distinguish between statements and opinions. But it is possible that those metrics were only useful for distinguishing those two dialogue acts. By then asking how well all dialogue acts predict keystroke metric 1, I can gain insight into whether that metric is a useful metric for dialogue act segmentation overall. If that metric was only distinct for two dialogue acts, but not useful for differentiating a plethora of other dialogue acts, then it is probably not a useful feature to use in future dialogue act classification tasks.

Because experiment 1b considers all dialogue acts, by-subject standardization of values was performed across all dialogue acts, rather than just across the subsetted dialogue acts in experiment 1a. For this reason, the (numeric value of) coefficients reported below are not directly comparable to the coefficients in experiment 1a.

One thorny element of these models was how to code contrasts. No dialogue act is inherently a "reference" dialogue act against which all other dialogue acts should be compared does not exists. For example, a Statement might make sense as a neutral reference point, but this is not an inherent property of a Statement utterance, and it would deprive the model of the ability to compare Statements to an overall average utterance, to determine if a different exists. In other words, the models should measure the extent to which *each* dialogue act differs from the (grand) mean of all utterances. Neither dummy coding, contrast coding, or any sum-to-zero contrast coding would return the results of all levels of a factor, since one level would need to be used for references.

In order to avoid this issue, I used grand mean contrast coding. I first calculated the grand mean of the response variable, which is the mean of all observations. For each level of dialogue act, I then calculated its deviation from the grand mean. These deviations were then used as the contrast weights. This allowed every level of the categorical variable to be reported and discussed, rather than leaving one factor level (a dialogue act) as a reference level.

Another issue to consider is that different dialogue acts have different average word counts (see Figure 5.2). As such, not adding word count as a covariate to other models leaves open the possibility that word count, not dialogue act alone, is predicting the dependent variable. Therefore, word count influence was verified and then added as a covariate.

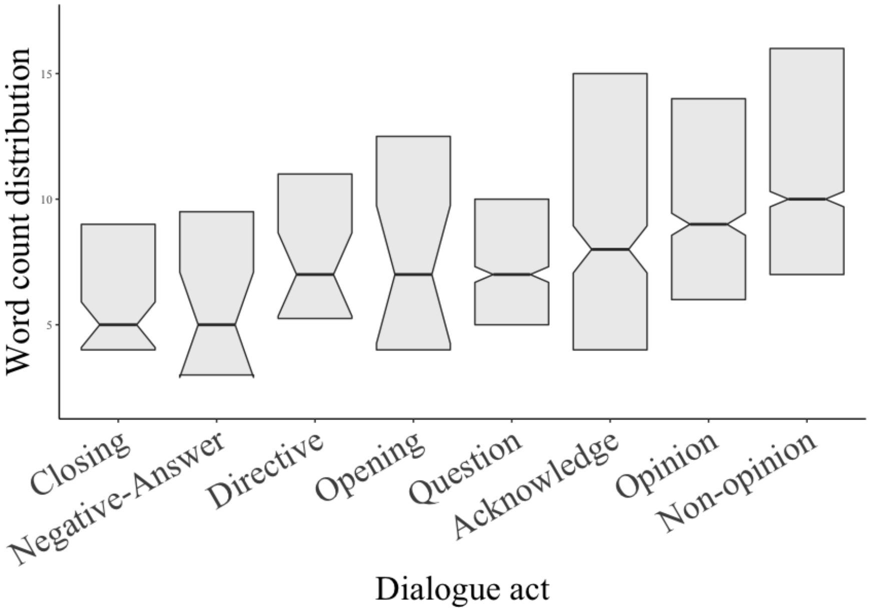


Figure 5.2
The distribution of word counts within each dialogue act, ordered from shortest word count to longest word count. Closing DAs have the shortest average word count will non-opinion statements have the longest average word count. This result makes sense since closings are more formalized and regular, while non-opinion statements vary significantly in content. However, further work is required in the future since my data also had significantly more instances of Opinions, Non-opinions, and Acknowledgments.

Because of the strong relationship between different types of dialogue acts and utterance length, I first tested that direct effect of dialogue act category as a predictor of word count.

As can be seen in Table 5.4, several dialogue acts do have distinct word counts associated with them. Specifically, Non-opinion and opinion statements are significantly longer than an average utterance. Acknowledgment utterances are also significantly longer. On the other hand, Questions and closing utterance are significantly shorter than an average utterance. Directives, negative answers, and opening utterances are not significantly longer or shorter than the average utterance. Overall, an ANOVA run on the model found that the dialogue act was a significant predictor of word count ($F(6,4100) = 19.58, p < 0.0001$).

	Dependent variable:
Factor Level	Word count
Grand mean (Intercept)	9.98^{***} (0.41)
Acknowledge	1.18^{*} (0.58)
Closing	-2.42^{*} (0.97)
Directive	-1.35 (1.45)
Negative-Answer	-2.36 (1.61)
Non-opinion	2.75^{***} (0.44)
Opening	2.18 (1.61)
Opinion	1.48^{**} (0.49)
Question	-1.46^{**} (0.51)
Observations	4,108
R^2	0.03
Adjusted R^2	0.03
Residual Std. Error	8.83 (df = 4100)
F Statistic	19.58^{***} (df = 7; 4100)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 5.4

Results of whether dialogue acts alone can predict word count. A positive coefficient signifies that that particular dialogue act has a higher word count, and vice versa. Standard errors are report in parentheses. The linear model that produced these results used deviation coding, with the grand mean as the reference level. Since no specific dialogue act was used as a reference level, all dialogue acts are reported as deviations from the grand mean.

Dependent Variable	Dialogue act	Word count	Overall model	Reference table
Word count	19.57****		19.58****	5.4
Utterance speed	9.55****	1.55	8.55****	E.2
Edit count	6.29****	1272.92****	164.6****	E.3
Speed variability	5.09****	63.29****	6.93****	E.4
Pre-utterance gap	3.89****	18.8****	6.02****	E.5
Word 1 - word 2 gap	1.87+	5.75*	2.36*	E.6
Word 1 speed	2.53*	1.78	2.43*	E.6
Word 2 speed	4.45****	0.05	3.90***	E.6

*+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001*

Table 5.5

F scores measuring the influence of changes in dialogue acts and changes in word counts on keystroke timing response variables. The overall model fit, which included all dialogue acts + word count, is also reported. Links are provided to the reference tables of each model, which specify the effect of each individual dialogue act. For models with interactions, overall effects are not reported since the interactions had minimal effects and the models looked very similar to models without interactions.

Since the initial model found that dialogue acts and word counts were significantly related, all of the subsequent models controlled for word count by adding it as a covariate with dialogue acts.

The remainder of the results are organized as follows: Table 5.5 provides an overview of the influence of dialogue act and word count on keystroke timing metrics. The table also provides links to each individual model, so that the readers can delve into the details of the direction and size of the effects of each dialogue act in how it influenced each timing metric.

As a brief preview, in every model run dialogue acts overall were significant factors at the .05 α level. The lone exception is the model measuring the gap between the first two words of an utterance, although these were approaching significance with α less than .10.

The first model looked at whether dialogue acts had distinct speeds at which they were typed. A fixed effect model using dialogue act and word count as predictors found a number of distinctive typing speeds for specific dialogue acts. They are illustrated in Table E.2. Directives, Opinions, and

Questions are typed more quickly. Acknowledgments, Negative-answers, and Opening utterances are typed at a slower pace.

In an ANOVA run on the model, typing speed was significantly affected by the dialogue act of the utterance ($F(8,4099) = 9.6, p < 0.001$). Interestingly, word count alone was not a significant predictor of typing speed ($F(1,4092) = 1.6, p = 0.2$). Acknowledgments, Negative answers and Openings were significantly slower, while Directives, Opinions and Questions were typed significantly faster than average.

The next metric I tested was edit count (using BACKSPACE or DELETE) and whether different dialogue acts have distinctive amounts of editing. In these models it was especially important to control for word count, since every additional keystroke typed is an additional opportunity to make an edit.

As expected, word count was highly influential on edit count ($F(1,4092) = 1273, p < 0.001$). The reasoning behind this is that every additional keystroke presents an additional opportunity to make an edit. Therefore, if more words are produced, it is more likely for more edits to also be produced. Interestingly, while only opening utterances consistently had more edits than other types of utterances ($p < 0.05$), the overall effect of dialogue act was still important.

In addition, the intercept in this model was also marginally significant ($t = 2.0, p < 0.05$). Since the intercept in a deviation coded model is the unweighted grand mean of all dialogue act categories, this might speak to the underweighted influence of non-opinion statements, which comprise the majority of utterances, but would have had equal weight in calculating the grand mean.

The next model investigated if the difficulty of word retrieval varied more or less within certain types of dialogue acts. This difficulty was operationalized by using the standard deviation of the pauses before each word, since this pause is likely affected by lexical retrieval rather than motor control (Logan and Crump, 2011).

Interestingly, the ANOVA on the model showed that both dialogue act ($F(7,4092) = 5.1, p < 0.001$) and word count ($F(1,4092) = 63.3, p < 0.001$) significantly impacted word retrieval variability. It should also be noted that while dialogue act was influential on variation, the size of the effect of word count was much larger. This will be further discussed in the Discussion.

Next, I looked at the pause time between the previous utterance being sent and the current utterance being initiated. For this analysis I eliminated conversational openings as a dialogue act, since this gap did not represent a part of the conversation, but rather just a gap after the timer began.

While overall the preceding pause was dependent on dialogue acts, only Questions had a significantly different (longer) gap before they were initiated ($p < 0.05$). The word count of an utterance also significantly affected the gap before it was initiated, where a negative coefficient indicated that a shorter utterance was preceded by a shorter gap. This seems logical if an entire utterance needs to be retrieved before production begins.

Finally, I investigated whether the dialogue act influenced the speed at which the first word was typed, the speed of the second word, as well as the gap between the first word and the second word. Whereas the first word might represent an instantaneous reaction to the previous utterance, the second word might represent more of a decision within the current utterance. The gap between the first two words might be representative of how linked the two words are within the same phrase.

When typing the first word, the speed was different for different dialogue acts, although the effect size was small. Interestingly, word count had very little effect on the speed at which the first word was typed, which might point to the first word being produced before full utterance planning is performed. The only individual dialogue act that was consistently different was Non-opinion utterances, which exhibited significantly faster production ($p < 0.01$).

For the second word in the utterance, the speed was significantly affected by dialogue act. Regarding specific dialogue acts, Opening utterances had a slower second word ($p < 0.05$), while Closing utterances had a slightly faster second word ($p < 0.10$). This might speak to more planning for a conversational opening, resulting in slower execution speed, whereas a closing is more ritualized and requires less planning.

When inferring the length of the gap between words 1 and 2, Opening utterances are not different from the mean, whereas Closing utterances have a *shorter* gap (as opposed to their more quickly typed second word in the previous model). Acknowledgments have a slightly longer gap ($p < 0.05$), whereas in the previous model they had a more slowly typed second word. In modeling the gap between words, when the word count of the entire utterance increases, the length of the gap between the first two words also increases ($p < 0.05$).

5.3 Discussion

For the sake of discussion, the tables from Sections 5.2.1 and 5.2.2 have been distilled into Tables 5.6 and 5.7, respectively. The features mentioned are those with significant F scores. The research questions I set out to answer are:

RQ 1a) Can typing patterns predict differences in pairs of dialogue acts, where each member of the pair would require a very different response?

RQ 1b) Does each dialogue act have a consistent set of typing patterns associated with it?

Dialogue act binary	Features
Non-opinions	shorter pre-utterance pause, typed slower, more edits
Opinion	longer pre-utterance pause, typed faster, fewer edits
Question	longer pre-utterance pause, typed faster, fewer edits, word 1 typed slower, word 2 typed faster
Statement	shorter pre-utterance pause, typed slower, more edits, word 1 typed faster, word 2 typed slower
Backward	longer pre-utterance pause, typed faster, word 2 typed slower
Forward	shorter pre-utterance pause, typed slower, word 2 typed faster

Table 5.6
Features of each dialogue act binary that had significant F-score for distinguishing the pair. The features are compiled from the individual tables in Section 5.2.1. Each feature is relative to the other level of the pair, not an overall comparison.

It is important to reiterate the differences between the two individual experiments comprising Study 1. In the first half I used various timing metrics as predictors of a binary dialogue act category distinction. In this case the predictors had to have unique values for each dialogue act in order

Dialogue act	**Features**
Non-opinion	higher word count, longer pre-utterance pause, word 1 typed faster
Opinion	higher word count, typed faster, word 1 typed faster
Question	lower word count, typed faster, less variability in typing speed, longer pre-utterance pause, word 2 typed faster
Acknowledgement	higher word count, typed slower, more variability in typing speed, word 2 typed slower, longer gap between words 1 and 2
Closing	lower word count, word 2 typed faster, gap between words 1 and 2 shorter
Opening	typed slower, more edits, word 2 typed slower
Directive	typed faster
Negative-answer	typed slower

Table 5.7
Features of each dialogue act with significant F-scores for defining that dialogue act. These features are compiled from the individual tables in Section 5.2.2. The dialogue acts are organized by descending frequency of occurrence. Each feature is relative to the grand mean of all utterances.

to be considered important. A predictor with the same values for each dialogue act level would not be considered a significant predictor. On the other hand, the second half of Study 1 used all dialogue act categories as predictors of specific timing metrics. For these models, each dialogue act could have the same coefficient value for the timing metric, but as long as that value accurately and consistently predicted the timing metrics (less variance), it sufficiently demonstrated a reliable timing metric for that dialogue act, even if the timing signature may not be discernible from that of other DAs.

Regarding **RQ 1a**, it seems that typing patterns are able to distinguish certain pairs of dialogue acts. Evidence for this claim is illustrated in Table 5.3, where a number of keystroke features are significant predictors of different dialogue acts. This is especially notable because it points to the notion that dialogue acts can be detected not only from their word choice, but also from the lexically-independent temporal production patterns. This could be due to a number of factors such

	Metric							
Dialogue Act	Word count	Pre-utterance gap	Typing speed	Speed variability	Edit count	Word 1 speed	Word 2 speed	Gap b/w words 1-2
Non-opinion	↑	↑				↑		
Opinion	↑		↑			↑		
Question	↓	↑	↑	↓			↑	
Acknowledgement	↑		↓	↑			↓	↑
Closing	↓						↑	↓
Opening		↓			↑		↓	
Directive			↑					
Negative-answer			↓					

Table 5.8
Dialogue acts with significant keystroke timing differences as compared to the grand mean. The colored arrows represent significant F-scores.

as the cognitive complexity of various DAs, including the amount of recall and planning that needs to go into production.

RQ 1b is best answered using Table 5.7, which shows the typing features that are significantly associated with individual dialogue acts. While some dialogue acts like Questions or Acknowledgments have a large set of unique predictors, other dialogue acts such as Directives and Negative-answers have very few unique typing patterns associated with them. As such, the answer to the second research question seems uncertain. *Some* dialogue acts appear to be distinguishable by typing features, while others do not. Similarly, some typing features seem to have unique timing patterns in a number of dialogue acts, while others appear to be very similar across all dialogue acts.

As mentioned at the beginning of Section 5.2, the first set of experiments was limited to distinguishing differences in three opposing pairs of dialogue act categories. Since I am using n dialogue act categories, there exist at least n^2 category comparisons because supersets of categories can be created. However, I focused on pairs that would be the most important dialogue acts to distinguish, because when a conversation partner or computer agent wants to generate an appropriate, relevant and natural-sounding response, this distinction would be very important

As an example, Opinions and Non-opinions may look similar on a lexical level, but an appropriate response to each would look very different: If a user responds to a Non-opinion statement such as "Today is Tuesday" by saying "I agree," then this response is not appropriate; on the other hand, if a user responds "I agree" to the Opinion statement "Today feels like Tuesday," then this is an appropriate response.

An interesting distinction among all binary pairs surrounds the pauses before utterances. The pause before an utterance is produced can be thought of as the period of time when cognitive planning takes place (Baaijen et al., 2012). Tasks of varying complexity require different amounts of planning before typing them (Conijn et al., 2019). Evidence from my study shows the different dialogue acts can be thought of as requiring different amounts of cognitive effort to produce.

For example, Questions have longer pre-utterance pauses (Tables 5.3 and E.5). This trend is also seen within the larger subset of backward-facing dialogue acts (Table 5.3). In contrast to forward-facing dialogue acts, backward-facing dialogue acts require the participant to process and incorporate the previous context in a conversation. This would be manifested in longer pauses before producing a question or backward-facing dialogue act.

Conversely, Questions are then typed faster and have fewer edits. However, this is supported by Baaijen et al. (2012), which found that longer pauses result in more “well-formed bursts [of typing]” (p. 246). The notion of a well-formed burst is also supported by the findings in Table E.4, which show that Questions are typed at a more steady pace with less variability in speed.

Further support for the notion that Question utterances are preplanned before they’re produced comes from Table E.4. Because Questions have significantly less variability than other dialogue acts, it stands to reason that more pre-planning goes into Questions, so that they are produced at a more consistent rate.

The findings regarding pre-utterance pauses point to the potential utility of typing patterns to increase the richness of online conversations. For example, if a computer agent knew with high probability that an utterance was going to be a question or a backward-facing utterance, then the agent could make a more educated guess as to which words deserve more attention when parsing the utterance. In addition, a computer agent could constrain the possible referents of a backward-facing DA if they know that it pertains only to the previous conversation.

An important distinction that a conversational partner must make, whether human or computer, is in the factual nature of an utterance. This was the motivation for studying the distinctions between Non-opinions and Opinions in Study 1a. Non-opinion utterances have a shorter pause beforehand, but are then typed slower with more edits. On the other hand, Opinion utterances have a longer pause before they are produced, but are then typed more quickly with fewer edits. This has an interesting parallel with speech prosody, where truthful speech is found to have more pauses, disfluencies, and corrections (Benus et al., 2006; Hirschberg et al., 2005). This could again point to the typing process being partially guided by silent prosody (Fodor, 2002a).

While I am not going so far as to say that an opinion is a lie, an opinion is likely based less on empirical fact than a non-opinion utterance. Because an opinion is not rooted in empirical fact, it might take longer to retrieve, as seen in the longer pre-utterance pause, but then produced more fluidly, as seen in the faster typing and fewer edits, because the speaker does not need to ensure that their wording is aligned with an objective reality.

In either case, the multiple distinctions between Opinion and Non-opinion utterances again points to the usefulness of utilizing typing features for human and computer agent conversational partners. These typing metrics could provide information about how objective or subjective an utterance is, so that a partner can possibly respond with another fact or another opinion.

When specifically distinguishing between Forward vs Backward dialogue acts, it was encouraging to observe that both pre-utterance gaps and overall typing speed were reliably different. The advantage to making this distinction could be important in a human-human dialogue, where e.g. a computer system highlights detects a backward-facing dialogue act and highlights certain points of the prior dialogue. This distinction could also be helpful for a computer agent, as well. As "attention" has become more important in neural network models, dialogue act identification could also be used to guide where the focus of conversational agents should be, so that the agent can generate an appropriate response. This is similar to Su et al. (2019), which uses an attention-based response generation system where attention is directed by semantic and contextual information in utterances.

From a methodological standpoint, the three models run in Exp. 1a are interesting because they also shed some light on which dialogue act binaries are most distinct and which are perhaps too granular to be distinguished. To clarify, the first model compared two single categories from the DAMSL annotation scheme (Jurafsky et al., 1997); the second model compared all statement categories against all question categories; the third model roughly used the entire set of dialogue acts and split it into one subset for backward-facing dialogue acts and a complimentary subset of forward-facing dialogue acts. The most distinct binary was Statements vs Questions, which could point to the unique processes involved in the creation of each of these utterance types.

For Exp. 1b, the results are more difficult to interpret succinctly. Looking at Table 5.5, it seems clear from these results that dialogue acts do have a significant effect on almost every timing metric measured. However, while the results do not definitively point to a set of metrics that can consistently used for dialogue act classification, the overall ANOVAs do point to promising future research. Almost all of the typing metrics, as a whole, are significantly affected by the dialogue act within which they are produced. While it is hard to pin down exact typing features or exact dialogue acts from my experimental results, the overall results seem to point to the unique typing features of a number of dialogue acts.

The patterns I've observed in Study 1b are also important in informing a computational system as to where *not* to look for dialogue function information. For example, the typing speed of the first word and the pause after demonstrate relatively weak predictive power of dialogue acts, as seen in Table 5.5. In other words, these findings point to the uniformity of this word speed and pause location across all dialogue acts.

A possible explanation for this comes from the theory of chunking during recall and language production. Chunking is a fundamental idea in cognitive psychology (Miller, 1956a,b) where individual items are grouped together into a single unit so that retrieving them from memory only requires a single action. This theory has been widely applied to language comprehension and production, where words are not retrieved individually but rather as a group or phrase, e.g. "pain in the neck" recalled as a single item rather than 4 discrete words (e.g. McCauley et al., 2017).

Moreover, chunking has been recognized to exist in typing, where longer pauses occur between phrases and in periods of cognitive overload (Leijten and Van Waes, 2013; Schilperoord, 2002). The findings above could imply that across all dialogue acts, the first words and selection of the second word are retrieved as a chunk, because the gap between the words and the respective typing speeds are similar across all instances.

In conjunction with this notion, it seems that another place *not* to look for meaningful typing patterns (and possibly even linguistic patterns) is in the opening utterance. As seen in the summary table above, Opening utterances have more edits and are typed slower. Both of these features point to a higher cognitive load (Brizan et al., 2015). Perhaps this points to the impact of social dynamics on typing in conversations, where lack of familiarity is more influential than the cognitive demands of a task. Regardless, these findings point to the idiosyncratic nature of the Opening utterance, and why data from these utterance types should be considered less important.

It is also critical to address the extremely small R^2 values of some models, despite the significance of the overall model. The overall significance is most likely due to statistical power, but the low fit points to the exploratory or inferential nature of this study. Specifically, the results of this study should be considered theoretically interesting, but in need of significant refinement before being put into production.

As a final methodological note, because there may exist a nearly limitless number of combinations of outcomes, the binomial outcomes I am trying to predict are not exhaustive, i.e. they do not cover the entire complete spectrum of response options such as heads versus tails in a coin flip. However, Popescu-Belis (2005) raises two interesting points regarding the classification of dialogue acts. They review many different tagsets and point out that the number of tags is a compromise between theoretical grounding and human annotation ability. Further, though, they conclude that DAs should be considered multi-dimensional, and so a single tag should not be considered adequate. As such, there likely exists latent outcome variables that I am not considering. This raises the possibility that specific dialogue act classifications may be too broad or too specific.

To tie all of these observations back to one of the themes of this thesis, this study also investigated keystroke features that have correlates in spoken prosody. In other words, pauses in typing are similar to pauses in spoken dialogue and edits in typing are similar to disfluencies in speech. But as Wei et al. (2022) points out, while prosodic features are important for dialogue act classification, an end-to-end classifier also must be able to prioritize prosodic features used for classification. Study 1 perhaps, in parallel, points to which keystroke features should be prioritized.

The fact that spoken prosodic features are useful for dialogue act classification and that these keystroke features are useful for the same task in written discourse is not proof that silent prosody exists, but it does provide evidence that similar cognitive processes are taking place. Moreover, this study also showed that not every keystroke feature with a prosodic correlate is useful in dialogue act classification, but rather that some are more important and should be prioritized when interpreting pragmatic intentions in type-written text.

In addition, Study 1 points to the promise of keystroke patterns in helping to identify underlying illocutionary force in utterances within an online dialogue. Whether this is a manifestation of silent prosody or a manifestation of another latent process, it seems clear that a connection exists between keystrokes and motivations or intentions. These connections can be utilized by a computer agent facilitating a human-human dialogue or a computer agent generating appropriate responses in a conversation.

5.4 Future work

Future work will need to establish a robust baseline to prove that, e.g. keystroke patterns provide more accurate distinctions than lexical information alone. Nonetheless, the results of Study 1a demonstrate that a number of keystroke features are helpful in distinguishing pairs of dialogue acts.

Future work should collect more data overall and collect more data for dialogue acts with few examples, in order to provide a more robust answer to this question. It will be discussed further below; nonetheless the results of study 1b seem inconclusive.

5.4.1 Limitations of future work

In the future, this type of experiment should be repeated on a different type of task. Because all of the conversations had a controlled time-frame as well as a specific task, i.e. recommending movies and TV shows, it is difficult to draw a stronger conclusion from my study saying that the findings apply to dialogue acts in general, or just my particular task. The low R^2 values, for instance, point to the fact that extending these models to other tasks might be difficult.

Future studies should also use a more sophisticated modeling technique. Logistic regression, while well-established and accepted within statistical social sciences, also lacks flexibility (Tolles and Meurer, 2016). I chose this technique because the primary concern of Study 1 was to differentiate between binary dependent variables (Exp. 1a) or multi-level single independent variables (Exp. 1b). However, typing patterns, especially among many typists, are not consistent across typists or even within a single typist. As such, future studies of typing behavior should use a more sophisticated modeling technique, including taking advantage of modern advances in modeling keystrokes using deep neural networks, e.g. Chang et al. (2021b).

Another limitation of this study is that it lacks an objective measure of success or improvement over current methods. In order to evaluate this, a classifier would need to be trained on a dataset with keystroke information available (like the dataset for this thesis) using a current state-of-the-art text-based classifier trained only on overt features such as word choice. Theses results would then need to be compared with a classifier that uses the same text-based features but where each observation would be augmented by keystroke-derived features. The two models' predictions could then be compared to see if keystrokes provide a meaningful improvement in accuracy or appropriateness of responses.

This is not unreasonable, and would need to be done to show that keystroke-tracking would be worthwhile for an improved conversational experience, beyond current capabilities.

In future work within this thesis, I will also use an expanded feature-set. For the initial study, I tried to use the features that were the most cognitively informative, and that could inform the binary distinctions I felt were most important for a person to have a satisfying conversation with a computer agent or human partner. As an example, Study 1 looked at utterance speed, which would account for pauses within the utterance. However, as studies such as Conijn et al. (2019) and Baaijen et al. (2012) have shown, not all pauses are the same, e.g. pausing between phrases implies less of a disruption than pausing in the middle of a phrase.

Finally, the experimental prompts I used did not evoke the large variety of dialogue acts I was aiming for and which would have made the models more informative. The diversity of dialogue acts may also be informative about different properties of specific dialogue acts. In the future, a different experimental setup should be used, perhaps using a problem-solving game or a similar scenario.

In addition, future studies should control for any specific roles that a user is playing as well as a chronological marker of when an utterance occurred in a conversation. Regarding roles, in the case of my experiments each user either was providing or receiving recommendations, depending on which portion of the conversation they were in. The reasoning for these additional investigations is based on two factors: 1) The sociological theory of "personae" posits that people may exhibit different traits depending on their social role (D'Onofrio, 2020). Because participants perform different roles in each half of the experiment, this theory seems germane. 2) During the course of conversations, conversational participants become more familiar with each other and tend mimic one another in linguistic style, or "converge" (Danescu-Niculescu-Mizil and Lee, 2011). It stands to reason then that linguistic-based typing patterns should also change as a conversation proceeds. This was partially verified by running an ANOVA predicting word counts based on conversation position, which was trending towards significance, $\chi^2 = 3.03, p = 0.08$.

Regardless of limitations, though, it does seem that this study shows the potential for keystrokes to inform the classification of dialogue acts, as well as the processes that go into producing different dialogue acts. Resolving this would result in better collaboration between humans and computer agents. For example, Pecune et al. (2019) showed that when an agent provides a more social explanation of movie recommendations, it improves the perceived quality of the interaction. In the same way, better understanding the motivations and illocutionary force underlying the text of utterances could lead to higher quality responses and overall interactions.

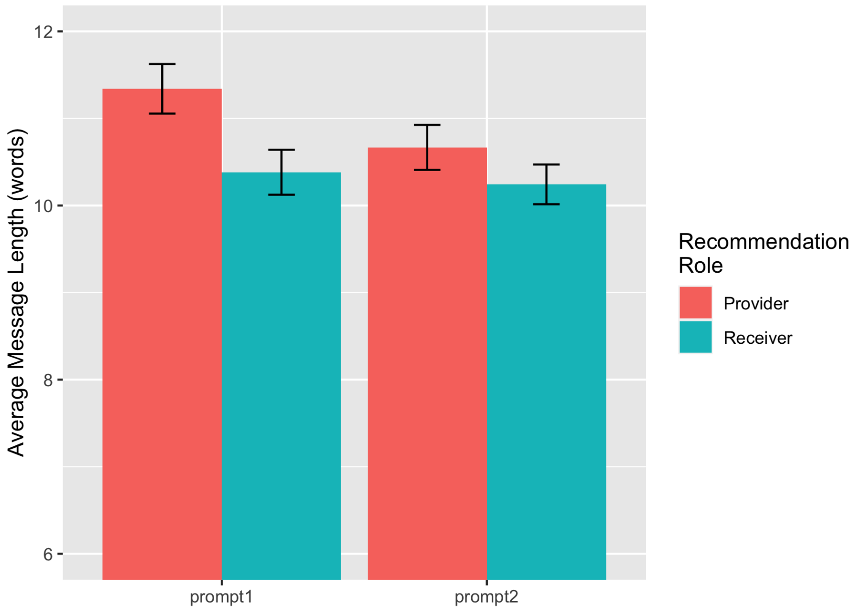


Figure 5.3
Word counts differ significantly depending on the recommendation role of the participant. Participants played different roles in each half of the conversation, referred to as Prompt 1 and Prompt 2.

Chapter 6

Study 2: The relationship between keystrokes and sentiment

This study examines how sentiment, sentiment change, and opinions in a conversational dyadic relationship affect the way a user types. This is important for a number of reasons: Most importantly, not all underlying user sentiment is apparent from word choices, and often sentiment based on word choices can only be measured after a complete message or conversation is complete. On the other hand, if sentiment also affects keystroke patterns, then this allows for real-time or continuous sentiment measurement, where user sentiment is being measured *as* a conversation progresses. Since studies such as Lee et al. (2014) and Lee et al. (2015a) have shown that emotion affects the way a user types in isolation, this study's findings about sentiment in dialogue will be fruitful for future developments such as a more empathetic chatbot.

Moreover, a user's underlying opinions are almost never overtly obvious from word choice. Aside from a user uttering a message such as, "I am enjoying this conversation," these underlying opinions are almost never realized in text. However, if keystroke patterns reflect opinions, then it would be possible to use information from typing patterns to deduce underlying opinions.

This study also demonstrates the usefulness of incorporating keystroke data into sentiment analysis for dialogue. While using keystrokes to infer sentiment is not new (e.g. Epp et al., 2011;

Lee et al., 2015a; López-Carral et al., 2019; Vizer, 2009; Yang and Qin, 2021), it has not been extended to conversational data. This seems to be a natural extension of sentiment measurement using keystrokes: If typing patterns are viewed as implicit prosody, and it has long been known that prosody is used to a greater extent in dialogues versus isolated speech (Blaauw, 1994; Bruce and Touati, 1990; Hieronymus and Williams, 1991), then typing pattern differences should be more apparent in dialogue than monologue. As such, I will also demonstrate that adding keystroke data to a lexical text-based classifier can add further accuracy to sentiment prediction in dialogues. The reasoning behind this is that while both word choice and keystroke timing are sensitive to cognitive patterns, each of these might be sensitive to different cognitive (e.g. Logan and Crump, 2011). Therefore, by adding an additional source of cognitive information, a researcher can learn more about underlying sentiment.

Further, conversational settings introduce the notion of sentiment *change*, where we can measure how much the sentiment changed between turns, rather than simply looking at the sentiment of a turn on its own. This is important in an interaction setting, where it is important to understand if conversational partners are on the same emotional level, or different levels as seen in sentiment change between turns. Moreover, sentiment *change* is unique from sentiment *per se* in that a change may be less likely to be evident on a lexical level. For example, a user's sentiment might shift very negatively, although they restrain themselves and use similar or neutral lexical choices. This restraint might affect keystroke production without affecting word choice. This is partially supported by findings such as Lee et al. (2014, 2015a), which show that emotion affects typing patterns. However, the "gold standard" manually-annotated sentiment that I use as an outcome variable is based on lexical evidence, and would therefore not detect these latent signals. Future studies will need to create an outcome variable based on subjective self-reported sentiment change, and then test how well a lexical model and a keystroke model can predict this change, in order to understand the improvement from keystrokes.

Finally, in a social or conversational environment, a user will also develop opinions about their conversational partner, which can affect the sentiment with which the user produces language.

Because of this unique feature of conversational language, I will also show how opinions can interact with sentiment to affect typing patterns, so that in the future typing patterns can be used to infer opinions.

As outlined in Section 3, Study 2 sought to answer the following research questions:

RQ 2a) Does keystroke information provide additional information about message sentiment as well as sentiment change between turns, above standard lexically-determined sentiment values?

RQ 2b) Are typing patterns sensitive to a user's opinion of their partner, when considered independently from the sentiment of a user's utterances?

The study is made up of two experiments:

Experiment 2a aims to replicate previous studies that used keystrokes to predict sentiment classification: Very Negative vs. Very Positive and Extreme (positive or negative) vs. Neutral sentiment. Because exact sentiment prediction can be difficult to judge, these distinctions are also often used in training sentiment models, e.g. Epp et al. (2011). Experiments 2a also examines a unique aspect of conversation interaction: sentiment change. Rather than only considering sentiment in an isolated message, I examine how changes in sentiment from message-to-message also are reflected in typing patterns. I found that adding keystroke information to estimates made by an algorithmic lexical-based model does significantly improve the amount of deviance explained by the baseline lexical model.

Finally, one aspect unique to interactions is that a user forms an opinion of their partner. This opinion may not only affect how a user types in general, but also affect how they type messages of certain sentiments. As an example, perhaps a user already does not like their partner; as a result, no special effort is required to type a negative message, whereas typing a positive message requires great effort. Experiment 2b looks at the distinct influences of message sentiment and a user's opinion of their partner, to see if opinions also have an effect on keystroke timing. While I do find that certain overall opinions have an independent effect on typing patterns, it does not appear that

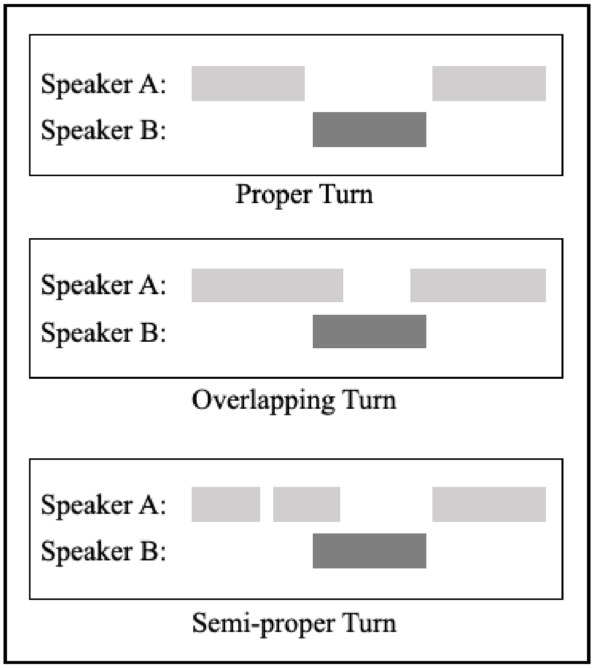


Figure 6.1
Different turn types within a dialogue. In a *proper turn*, which is the only turn type used in Study 2, a turn begins after the preceding turn ends, and then it ends before the onset of the following turn. In an *overlapping turn*, the turn begins or ends while the other interlocutor is typing. In a *semi-proper turn*, a turn begins after the first the first message of the preceding turn is sent, but before the conclusion of the entire turn.

this affect consistently exerts an independent influence. Nonetheless, a number of models show promising independent effects which should be studies further in future studies.

The two experiments use models that flip the dependent and independent variables. Experiment 2a predicts sentiment from a combination of keystroke predictors. Experiment 2b uses sentiment and opinion predictors to predict keystroke timing patterns. The reasoning behind these setups will be expanded upon in the discussion section. In brief, the two experiments are answering different questions: Exp 2a asks whether a set of keystroke features can make distinctions about sentiment; Exp 2b asks if sentiment and parter-opinion have independently robust effects on different keystroke metrics.

This study will focus exclusively on what I term “proper” turns (see Figure 6.1). This is similar to the original conception of turn-taking in the origins of conversational analysis (Sacks et al., 1974;

Current Turn Sentiment	Following Turn Sentiment		
	Negative	Neutral	Positive
Negative	8%	1%	3%
Neutral	2%	14%	4%
Positive	7%	5%	55%

Table 6.1
The sentiment of the current turn, broken down by the sentiment of the following turn. As can be seen, for each sentiment level, the highest proportion of following turns is the same sentiment level. Furthermore, the majority of adjacency pairs (55%) are made up of positive turns followed by positive turns.

Wilson et al., 1984). Overlapping turns, while ubiquitous in naturalistic conversation, also introduce a large amount of variation into language production, before, during, and after the interruption. For example, if a user starts typing while they know their partner is typing, these typing patterns are likely to be different as compared to when they know their partner isn't typing, because the user knows that their partner is also producing language. Similarly, if another message pops up in the middle of a user typing their message, this new message could likely distract the user. In fact, the way that a speaker responds to an interruption is not ubiquitous, but rather highly dependent on cultural norms and gender norms (Tannen, 1984). This additional variability will be expanded in the discussion section of this study (Section 6.4).

While understanding overlapping turns is essential for the future development of using keystroke patterns during online conversations, this is outside the scope of this study.

This distinction is also important because comparing full turns is more similar to comparing asynchronous dialogues such as back-and-forth tweets. A tweet is not made up of a single sentence, but rather multiple ideas, which make them more analogous to full turns. This highlights the importance of studying full turns in Study 2, rather than individual utterances, because full turns provide of view of an entire idea creation rather than a single component message in that idea. In fact, conversational analysis sometimes uses the turn, rather than utterance, as its base unit (Crookes, 1990).

6.1 Related Work: Sentiment and dialogue

As mentioned in Section 2.2.3, keystroke patterns have previously been utilized for sentiment detection primarily when typing monologues in isolation. In this study, though, I highlight two specific extensions of this work: sentiment analysis in dialogue and sentiment analysis using additive models.

6.1.1 Sentiment analysis in dialogue (versus monologue)

As devices such as Alexa and Siri evolve from simple question-answering voice assistants to pseudo-social companions (Pradhan et al., 2019), it is important to detect sentiment in human-computer interactions rather than only detecting sentiment in isolated language production. Bertero et al. (2016) trained a model to perform sentiment analysis alongside speech recognition in a real-time dialogue system. As they point out, when emotion detection can be done quickly, it allows for speech recognition accuracy to also be increased. In other words, rather than simply decoding a speech signal, this decoding can be aided when the emotional dimension of the speech is also provided. My study, in a similar vein, examines the relationship between sentiment and keystroke patterns, so that future researchers can create better computer-agents, such as chatbots, that generate a more relevant and emotionally-appropriate response to a user.

The question remains, though, as to how well research from isolated sentiment analysis can transfer to sentiment analysis in dialogue. As Zhang et al. (2019) points out, sentiment in dialogue is sensitive to both the speaker themselves as well as previous context. Gergle (2017) similarly points out that sentiment in dialogue is simultaneously sensitive to individual-, group-, and even network-level properties. In an isolated setting, however, previous context and other group-level properties does not exist or at least does not exert the same influence on the present language choices.

Ghosal et al. (2020) created an utterance-level model of sentiment analysis in conversations. However, the model was significantly improved by incorporating elements of commonsense knowledge, in order to better understand relationships and sentiment shifts that were not apparent on a purely lexical level. This study has a similar aim, by demonstrating that keystroke knowledge can also complement lexical knowledge to improve sentiment understanding.

Welch et al. (2019) provides an interesting parallel to my own study. The researchers used longitudinal asynchronous dialogue data to predict not only the content of the next message, but also the timing of when the next message would be sent. As a caveat, because their data was asynchronous it was on a very different time scale than my keystroke data. Nonetheless, similar to my own study it shows the use of timing-related data for understanding a dialogue and the temporal or linguistic relationships between messages.

Finally, Ganesan et al. (2022) presents an interesting parallel to my own study of sentiment change. The researchers used large language models such as RoBERTa (Liu et al., 2019) to predict "moments of change" in sentiment, such as a sentiment switch or an escalation of the same sentiment between successive online posts. They compared these results to a transformer model that also incorporated "psychological features" of the user. However, the purely lexical model outperformed this augmented model. Given this finding, it will be informative to see if cognitive-based features are more informative in synchronous conversations as opposed to the asynchronous posts used in Ganesan et al. (2022).

As a last note, it is important to study sentiment in dialogue because dialogues can elicit different emotional reactions than monologues consisting of the same content. Stranc and Muldner (2019) studied student sentiment after watching a teacher deliver a lecture as a monologue versus delivering the same material in dialogue. They found that, while controlling for retention of material, dialogues provided more positive sentiment in student comments after watching the lecture. A study such as this is important because it highlights the need to study sentiment specifically in dialogue, rather than viewing dialogue similarly to monologue.

6.1.2 Sentiment analysis using generalized additive models (GAMs)

Additive models such as those employed in this study are well-suited to capture a nonlinear relationship between typing patterns and sentiment level. Moreover, an additive model is made up multiple smoothing functions, and so it can be used to model phenomena where a latent or unidentified factor is also influencing an outcome. As Hastie and Tibshirani (1987) points out in one of the first studies of GAMs, these models have “the advantage of being completely automatic, i.e. no ‘detective work’ is needed on the part of the statistician.”

As an example of the flexibility of GAMs in sentiment analysis, Qi and Li (2014) used these models to predict sentiment from nouns and noun phrases, rather than the traditional approach of using verbs and adjectives to locate emotional word choices. Because nouns, in isolation, look largely the same in negative and positive language, it was important to use a GAM that could also detect latent factors.

In another interesting application of GAMs, Wang et al. (2022) used these models to determine consumer satisfaction, assuming it was related to, but not identical to, sentiment in reviews. The additive models seemed to be capable of teasing apart these variables.

In my own study I am using dialogue data, where the language a user produces and the way they produce it is inherently complex. Language production in dialogues can be influenced by many factors such as a user’s state of mind, by a previous turn, or by the beginning of the conversation. Using GAMs will allow my models to account for multiple variables which may or may not be an explicit part of a model.

6.2 Methodology

Study 2 uses the same data that was collected for Study 1 (see Chapter 4 for details of the data collection procedure). However, rather than studying individual messages (as in Study 1), this study concatenates adjacent messages from the same user into a “turn”, and then looks at adjacent pairs of turns, called an *adjacency pair* (Schegloff and Sacks, 1973). A "turn" is equivalent to a sent

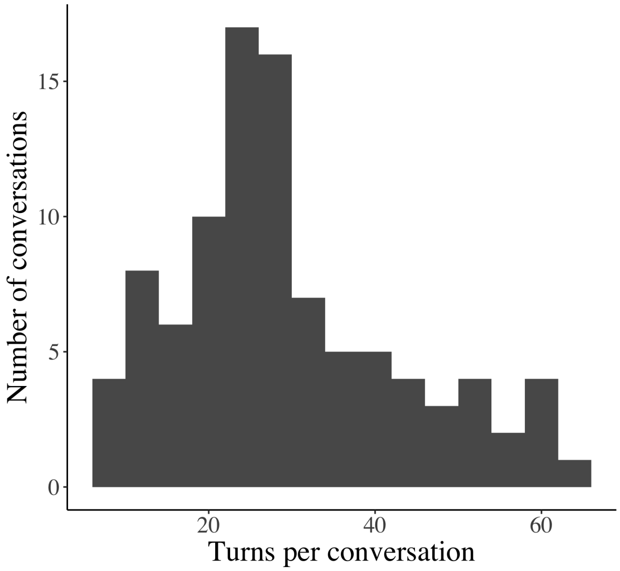


Figure 6.2
The distribution of the number of turns per conversation.

Turn Type	Occurrences	Length (characters)	Word Count
Proper	676	79	16
Overlapping	905	91	18
Semi-proper	1154	88	17

Table 6.2
Occurrence count and features of different turn types. These turn types are illustrated in Figure 6.1.

message. In other words, a turn is not delineated by punctuation, but rather only when the ENTER key was pressed to transmit a message.

The final dataset included 2,890 turns, with a mean word count of 30 words per turn. However, there exists significant variability in the number of turns per conversation: the shortest conversation comprised 8 turns while the longest conversation was 66 turns. The standard deviation for turns in a conversation was 14 turns. The variation in turns per conversation is illustrated in Figure 6.2.

Although there were 2,890 turns in the entire dataset, as Table 6.4 illustrates, each model in Study 2 used only between 262-450 proper turns. The size of this dataset is relatively small compared to datasets used for comparable tasks. For example, the IEMOCAP dataset, considered one of the gold standards in conversational sentiment analysis, was trained on 5,810 utterances (Busso et al., 2008).

As mentioned elsewhere in this thesis, keystroke data has high variability: this is not only because language data is inherently messy, but specifically because timing such as pauses in typing data can occur for many reasons, some of which are impossible to detect. As an example, a typist can pause for cognitive reasons, e.g. thinking of what to say next, physical reasons, e.g. they are tired, or completely unrelated reasons, e.g. distracted by a fly in the room (Dahlmann and Adolphs, 2007; Leijten and Van Waes, 2013). For this reason, the initial filter on all data in Study 2 was to only keep the first 95% quantile of keystroke data, to prevent especially long pauses from skewing the findings, and then center and scale the remaining data (Epp et al., 2011).

To illustrate why this was done, it is unlikely that a 20 second pause had a different underlying motivation than a 10 second pause, and so the distinction is not meaningful. In future work, it may be useful to take an approach similar to Baaijen et al. (2012), in which pauses are binned together, e.g. a bin for pauses less than 1 second, or a bin for pauses between 5 and 10 seconds. By taking this approach, no pauses are eliminated but some precision is lost.

6.2.1 Feature selection

The feature-set for Study 2 was slightly modified from Study 1. Many features were reused when creating the optimal model. However, since this study uses whole turns rather than individual utterances, this introduces certain features that would not have been available when looking at single messages, e.g. the pause been messages or multiple phrasal boundaries.

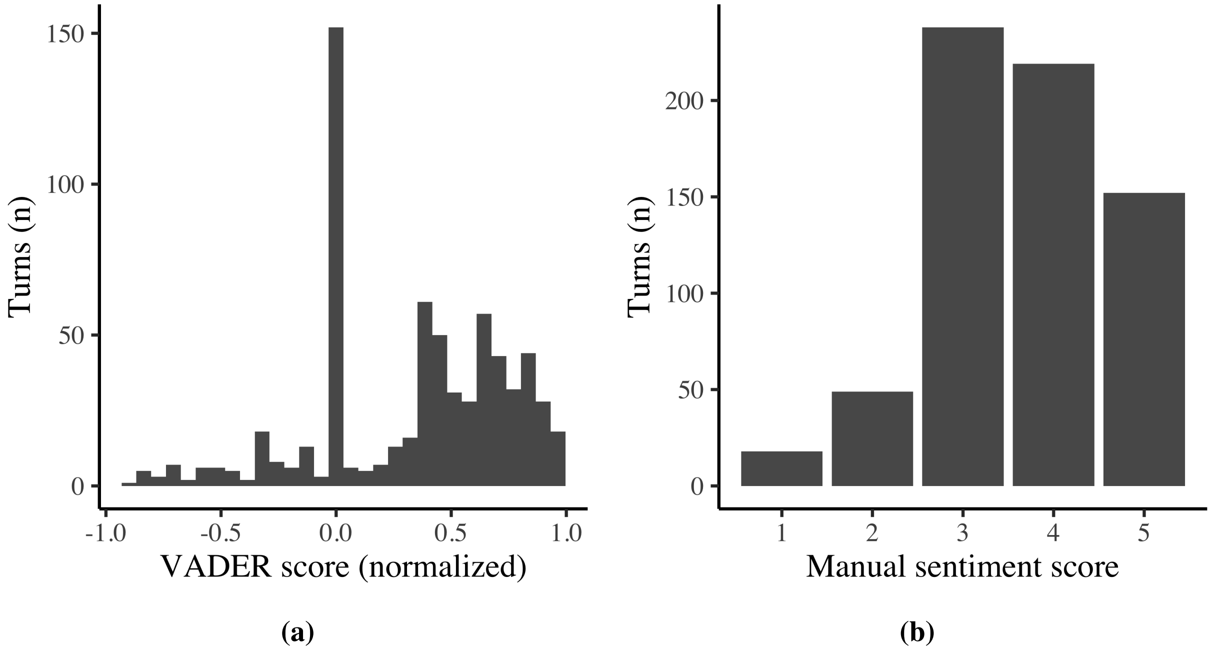
Keystroke timing features were selected primarily for their cognitive relevance, rather than for strictly practical reasons. For example in keystroke dynamics research it is common to measure the timing of every keystroke bigram, trigram, or *n*-gram. Because the current dataset is relatively small, though, this feature would have high variability.[1] The turn features that were tested are below.

- Pauses (inter-keystroke intervals or IKIs) before, within and after a word (Conijn et al., 2019)
- Pauses at phrase, sentence and message boundaries (Galbraith and Baaijen, 2019)
- Dwell time, or how long a key is depressed (Lee et al., 2014, 2015a)

[1] However, Study 3 will use *n*-grams.

- Edits or deletions during a message (Olive et al., 2009)

Galbraith and Baaijen (2019) noted that the importance of different types of pauses depends on where they occur in a text, i.e. in the middle of a sentence, at the end of a sentence, or at a phrase boundary (delimited by a comma or semicolon). In other words, not all pauses should simply be aggregated and averaged, but rather different pause locations imply different reasons for pausing, and these different reasons could be more or less influenced by sentiment change.



(a) **(b)**

Figure 6.3
Figure 6.3a illustrates the distribution of algorithmically-determined sentiment values, using VADER. Figure 6.3b illustrates the manually-assigned sentiment scores, using human annotators. As can be seen, in both instances most turns are neutral, and more turns are positive than negative.

The sentiment measures used for this study were collected in two ways: algorithmically and manually. Algorithmic sentiment analysis was done using the `VADER` Python package (Hutto and Gilbert, 2014).[2]

[2] I also ran a pilot study using the `sentimentr` R package for sentient analysis. However, too many turns were labeled neutral sentiment, but only because the algorithm could not make any decisions, whereas VADER was trained on social media and seems to better understand language that appears neutral on the surface but in informal settings is used to convey sentiment. Despite choosing one algorithm over the other, it should be noted that the results of both were

Manual sentiment analysis was performed by a research assistant and me. Annotation guidelines were drawn up that followed Mohammad (2016) and specified what signals to look for in a turn in order to assign a sentiment score. The inter-annotator agreement, measured by Cohen's κ, was 0.94, which is considered "near perfect agreement" (McHugh, 2012). While I only had one additional annotator available, it should be noted that manual text annotation is essential for sentiment analysis learning (Bobicev and Sokolova, 2017).

6.2.2 Generalized Additive Models (GAMs)

Finally, I chose to use additive models built in `mgcv` in order to capture nonlinearities in the data (Wood, 2022, 2006). In addition, each predictor also output an *effective* degree of freedom (edf), which represents how many knots or change-points actually exist for the predictor. As an example, an edf of 1 would mean that the predictor is essentially linear, whereas an edf of 3 would imply two points at which the slope of the line changed. By having a sense of how many change-points actually exist, it is possible to make further inferences about how many groups actually exist in the data, since a different group would necessitate a change-point.

In addition, GAMs are considered to be an especially interpretable method of machine learning (Chang et al., 2021a). As an example, Hegselmann et al. (2020) presented the visual output of GAMs to healthcare professionals who needed to assess clinical levels of risk. They found that the doctors were able to "mentally simulate" the output of the additive models, and felt comfortable making an informed decision.

All predictors of interest used the default thin plate regression spline, while a smoothing spline was also added to use the individual subject as a random effect. An additional parameter was added to increase penalization so that less informative smooths could be reduced to an edf of zero. Without this additional penalization, a non-informative spline could only be reduced to a linear function (Marra and Wood, 2011). This penalization is similar to LASSO regression for linear models; since

strongly correlated. A Pearson correlation coefficient was calculated at $r(2888) = .58, p < 0.0001$, and a Wilcoxon paired t-test found no significant difference ($p = 0.65$).

the dataset was small, preventing overfitting was important. Aside from these changes, no other hyperparameters were set. Because the dataset was so small, hyperparameter tuning would have most likely had a minimal effect. To assess model fit as well as model comparison, recent work has suggested using AIC comparison as well as ANOVAs (Wood et al., 2016). Both of these tests have been specifically adapted for generalized additive models in the aforementioned `mgcv` package.

6.2.3 Selecting an optimal additive model

To test the various hypotheses, an optimal keystroke model was selected. In order to select this model, the response variable was the full continuous sentiment values in VADER, from -1 to +1; for predictors, different subsets of keystroke patterns were tested in combinations. Ultimately, the most accurate model used four smoothed predictors: overall IKI mean, average dwell time within content words, average IKI before function words, and average IKI at all phrasal boundaries. The model also included a random effect spline for each individual subject. The implications of these predictors providing the most accurate fits will be included in the discussion section.

Predictor	Effective df	Referential df	F score	p-value
Inter-keystroke Interval (IKI)	4.0	9	1.454	0.007
Dwell within content words	2.7	9	1.35	0.003
IKI at beginning of function words	0.8	9	0.195	0.111
IKI at phrasal boundaries	≈ 0.0	9	0	0.892
By-subject	12.7	159	0.089	0.220

Table 6.3

An ANOVA highlighting the importance of each (smoothed) predictor in the optimal keystroke model. While not all predictors attained statistical significance, the combination of these four predictors explained 12.9% of the deviance of the sentiment values. This was the best fitting model of all of those tested. Effective df represents how many change-points were *actually* needed for the predictor's best fit, whereas referential df is the expected degrees of freedom.

This optimal model was compared to a baseline model that used all IKIs and all dwell times as predictors. Compared to this model, a one-way ANOVA comparing the baseline model to the optimal model provided a marginally significant improvement ($p = 0.07$) using a χ^2 test. The optimal model also had an AIC of 482, whereas the baseline model had an AIC of 487.2, which provides evidence that the additional predictors did not only add complexity while providing a better fit.

Further, the optimal model also used full penalization of less informative splines. The usefulness of penalizing less informative splines is illustrated by fact that the AIC of the optimal model improved from 487 to 482, while the effective degrees of freedom only increased by 4.2. This increase in effective df should be compared to the difference in *referential* df between the models which was 18. The referential degrees of freedom represent the maximum possible increase in degrees of freedom necessary to accommodate the complex model, whereas the difference in effective degrees of freedom shows how many additional degrees were actually required for this extended GAM.

Figure 6.4 below shows the partial effects of each predictor. As can be seen, some predictors have a strong linear or at least monotonic effect, while others are nonlinear and not easily definable by a polynomial. One advantage of additive models is that they provide the ability to capture this.

A *dis*advantage of additive models, however, is that using splines rather than linear predictors makes the effect size and direction of effect difficult to interpret (Wood, 2013). Since non-linear effects, such as IKI time in Figure 6.4, can have varying slopes and directions, an interpretable β coefficient cannot be derived to make clear a predictor's precise effect. As described by Wood (2013), though, the p-values use a Wald t-test that uses a null hypothesis that $s(x) = 0$. A low p-value, then, indicates a low likelihood that the splines that make up the function are jointly zero.

6.3 Results

Experiment 2a examines the degree to which keystroke information can augment lexically-determined sentiment information. Experiment 2b then looks at the effects of users' opinions on different keystroke patterns, when user opinion is considered independently of sentiment information.

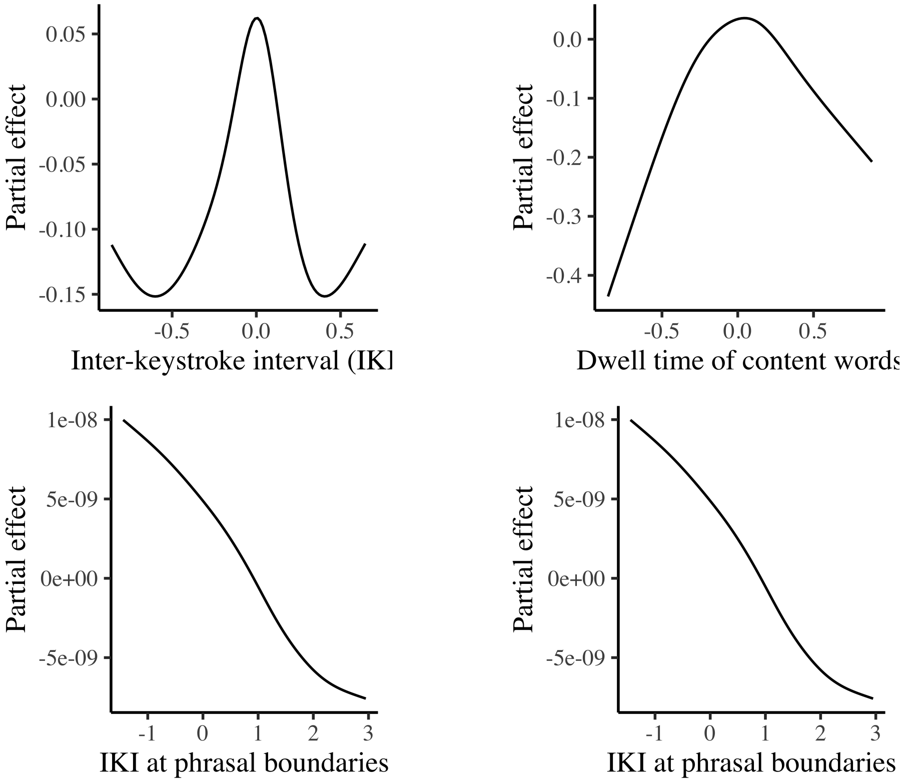


Figure 6.4
The nonlinear functions that define each predictor in the optimal keystroke model

6.3.1 Experiment 2a

The first experiment sought to determine the extent to which adding keystroke information to lexical sentiment information could improve the model's predictive power. The lexical baseline model used traditional algorithmic sentiment analysis based solely on printed text (VADER, Hutto and Gilbert, 2014). Table 6.4 shows the results of adding this information to baseline lexical models for four different sentiment analysis tasks.

Table 6.4 is broken down into four different prediction tasks. The percentage of deviance explained by the model is reported for each model, rather than reporting R^2 or adjusted R^2; the reasoning for this is that R^2-derived measures are only accurate when sums of squares are used, whereas the GAM fitting process makes this an inaccurate measure (Wood, 2006). The base model uses an off-the-shelf sentiment analysis library, VADER, which only considers (surface-level) lexical

Prediction Task	n (turns)	Base VADER Model Deviance Explained (edf, AIC)	Keystroke Model Deviance Explained (edf, AIC)	Combined Model Deviance Explained (edf, AIC)	ΔAIC	p-value (χ^2 test)
Exact Sentiment $(1-5)$	450	26.9% (25.2, 1184.6)	15.3% (32.4, 1264.9)	27% (24.7, 1184.5)	-0.1	$p = 0.28$
Positive (4,5) v. Negative (1,2)	303	21.6% (14.3, 230.9)	7.1% (8.5, 256.9)	33% (28.1, 229.1)	-1.8	$p = 0.04^{*}$
Extreme (1,5) v Neutral (3)	262	34.4% (26.1, 287.9)	25.5% (37.4, 342.5)	46.5% (44.6, 281.2)	-6.7	$p = 0.005^{**}$
Sentiment Change (-4−4)	386	12.3% (3.0, 1191.9)	2.8% (4.4, 1234.5)	15.4% (5.7, 1183.3)	-8.6	$p = 0.008^{**}$

Signif. codes: $*** - p < 0.001, ** - p < 0.01, * - p < 0.05, \dagger - p < 0.1$

Table 6.4

The model results for four prediction tasks. The n represents the number of turns, not the participant count. The model results report the percentage of deviance explained by the model, with the effective degrees of freedom and the AIC in parentheses. The p-values were determined using an ANOVA comparison of the base model versus the combined model, with a χ^2 test. The influence of the individual predictors used for the combined models are unpacked in Table 6.5. For most models, the addition of keystroke information resulted in a significant improvement in model fit.

content, but takes into account intensifiers and dependencies. The keystroke model uses only the predictors mentioned in the Methodology section that make for an optimal model. Finally, the ΔAIC metric and p-value are derived from an ANOVA comparison between the baseline model and combined model, to assess the value of adding keystroke data to a lexical baseline.

The first task had a goal of predicting the exact sentiment of a turn, i.e. 1, 2, 3, 4 or 5. Although the keystroke predictors reduced AIC slightly, by 0.1, this was not statistically significant[3].

The second task used a binomial model to predict either positive sentiment (a value of 4 or 5) or a negative sentiment (a value of 1 or 2). Adding keystroke information resulted in a statistically significant improvement in AIC, by 1.8; deviance explained improved by 11.4% ($p < .05$) from 21.6% to 33%.

The third task also predicted a binomial outcome, i.e. whether the sentiment was extremely negative/positive (a value of 1 or 5) or whether the sentiment was neutral (a value of 3). An ANOVA

[3] This is ironic since the optimal model was fit using this prediction task. Nonetheless, it seems that this task might not be ideally suited to keystroke information: Perhaps the exact sentiment values are too fine-grained to be detected in keystrokes, or perhaps the lexical model is exceptionally good at predicting this.

also showed a statistically significant improvement in model fit, and improved AIC by 6.7; deviance explained improved by 12.1% ($p < .01$), from 34.4% to 46.5%.

The final task predicted the sentiment change between the preceding turn (from another participant) and the current turn, using a continuous measure rather than a categorical measure such as *same* or *different*. This prediction task is unique to dialogues, because language produced in isolation does not have a previous turn produced by a different conversant. Adding keystroke data to this lexical model also provided a statistically significant improvement, and improved AIC by 8.6; deviance explained improved by 3.1% ($p < .01$) from 12.3% to 15.4%.

I will delve further into the final task as well as the prior three in the following discussion, but it is perhaps telling that cognitive signals such as keystroke timing patterns are highly informative as to changing mindsets. On the other hand, a purely lexical analysis may be less effective at detecting changes.

Table 6.5 breaks down the Combined Models in Table 6.4 into individual predictors. Each column in Table 6.5 is a different prediction task, which corresponds to a row in Table 6.4. As can be seen, when combining lexical information (VADER) with keystroke information, the lexical information appears to be more influential than the keystroke information. This can be expected as the VADER model was built with a very large number of parameters and is likely more nuanced. As a note, though, p-values are difficult to interpret when using additive models (Wood, 2013), and so even though these have been specifically built for GAMs, these values should not always be taken at face value. Rather, the most reliable measurements are the ANOVA model comparisons in Table 6.4.

Nonetheless, the fact that keystroke information can still be valuable in addition to lexical information demonstrates that keystroke information is not merely redundant to lexical information, but can provide complimentary information.

The p-values in Table 6.5 are derived from a type III ANOVA of each model that used a χ^2 test. For predicting the exact sentiment score ($p = 0.08$) as well as predicting extreme versus neutral sentiment ($p = 0.01$), the by-subject variation provided significant additional information. In other

Predictor	Exact (F, edf))	Positive v Negative (χ^2, edf)	Extreme v Neutral (χ^2, edf)	Change (F, edf)
VADER	$p \approx 0.0$ **(12.3, 1.2)**	$p < 0.0001$ **(31.2, 2.1)**	$p \approx 0.0$ **(55.4, 3.4)**	$p \approx 0.0$ **(6.2, 0.9)**
IKI	$p = 0.85$ (0.0, 0.0)	$p = 0.15$ (7.4, 4.2)	$p = 0.20$ (16.0, 9)	$p = 0.9$ (0.0, 0.0)
Dwell Time (Content Words)	$p = 0.23$ (0.1, 0.3)	$p = 0.32$ (0.01, 0.0)	$p = 0.20$ (1.1, 0.6)	$p = 0.007$ **(0.6, 1.2)**
Pre-word Pause (Function Words)	$p = 0.89$ (0.0, 0.0)	$p = 0.16$ (9.9, 5.9)	$p = 0.50$ (0.0, 0.0)	$p = 0.24$ (0.1, 0.5)
Boundary pause	$p = 0.80$ (0.0, 0.0)	$p = 0.79$ (0.0, 0.0)	$p = 0.45$ (0.0, 0.0)	$p = 0.02$ **(0.4, 1.0)**
Subject RE	$p = 0.08$ **(0.2, 21.1)**	$p = 0.13$ (17.1, 1.5)	$p = 0.01$ **(41.0, 3.1)**	$p = 0.51$ (0.0, 0.0)

Table 6.5
The table above shows the influence of each individual predictor on the outcome of the combined models (lexical + keystrokes) in Table 6.4. F-scores of 0.0 indicate no variance, while edf's of 0.0 indicate linear predictors.

words, each subject was unique in how their production patterns differed when delineating sentiment in these two tests, as opposed to the delineators in the other two experiments.

The model predicting sentiment change was also notable in that two keystroke-based predictors provided significant additional information: the dwell time within content words ($p = 0.007$) and the pauses at phrasal boundaries ($p = 0.02$).

6.3.2 Experiment 2b

Experiment 2b looks at sentiment and keystroke timing from a different perspective, where sentiment and opinion scores predict keystroke patterns (the inverse of Experiment 2a). The goal of understanding how a participant's opinions about a conversation affect their keystroke timing patterns. Whereas in Experiment 2a, the sentiment value was the response variable and the keystroke patterns are the predictors, in 2b the keystroke patterns are the response variable and the sentiment value plus participants' opinions are the predictors. This change was made so that the independent effects of many different user opinions could be tested on each keystroke feature, where keystroke metrics are held constant while opinion and sentiment change.

In Experiment 2b, a baseline model measured how well manually annotated sentiment scores could predict keystroke patterns. This baseline was then compared to an expanded model that added a participant's opinion rating as a predictor. By comparing the two models, I am able to better isolate the impact of a user's opinions on the way that they type.

Table 6.6 summarizes the results of Experiment 2b.

	Opinion Question						
Keystroke Feature	Watch with partner	Smooth convo	Enjoy convo	Watch recommendations	Rapport	Mean	Self-opinion
Pre-turn pause	$p = 0.38$	$p = 0.78$	$p = 0.12$	$p = 1.0$	$p = 0.85$	$p < 0.0001^{***}$	$p = 0.62$
IKI	$p = 0.20$	$p < 0.0001^{***}$	$p < 0.0001^{***}$	$p = 0.11$	$p < 0.0001^{***}$	$p < 0.0001^{***}$	$p = 0.16$
Dwell	$p = 0.09^{\dagger}$	$p < 0.0001^{***}$	$p < 0.0001^{***}$	$p = 0.18$	$p < 0.0001^{***}$	$p = 0.08^{\dagger}$	$p = 0.17$
Edit ct	$p = 0.01^{*}$	$p = 0.15$	$p = 0.38$	$p = 0.08^{\dagger}$	$p = 0.09^{\dagger}$	$p = 0.09^{\dagger}$	$p = 0.07^{\dagger}$
Pre-word pause	$p = 0.19$	$p = 0.10$	$p = 1.0$	$p < 0.0001^{***}$	$p = 0.32$	$p = 0.28$	$p = 0.29$
Boundary pause	$p = 0.22$	$p = 0.08^{\dagger}$	$p = 0.43$	$p < 0.0001^{***}$	$p = 1.0$	$p = 0.14$	$p = 0.13$
Before send pause	$p = 0.98$	$p = 1.0$	$p = 1.0$	$p < 0.0001^{***}$	$p = 0.11$	$p = 0.10$	$p < 0.0001^{***}$

Signif. codes: $*** - p < 0.001, ** - p < 0.01, * - p < 0.05, \dagger - p < 0.1$

Table 6.6

The table above shows the effects of each opinion rating on the timing of the respective keystroke features. Each full opinion question is listed in Section 4.4.5. The *p*-values are derived from an ANOVA comparing a baseline model using just the manual sentiment rating to an expanded model that used the sentiment rating as well as the opinion ratings. The mean score is the average of the first five opinion questions. The final question was not averaged in because it partially depended on self-reflection, rather than the participant's opinion of their parter. As can be seen, a number of different partner opinions affected the inter-keystroke interval and dwell times. In addition, the final question asking what the partner thought of the participant (self-awareness) affected the interval before a participant sent a message.

The responses of interest were:

- Mean pre-turn pause - the mean interval between when the previous message was sent and the current message was begun
- Mean inter-keystroke interval - the mean interval between each keystroke, analogous to typing speed
- Mean dwell time - the mean time of how long a key is pressed, i.e. key-down to key-up
- Edit count - the total number of times a participants uses the BACKSPACE or DELETE keys
- Mean pre-word pause - the mean interval that occurs before a word is typed, i.e. between SPACE and the first letter of the word
- Mean boundary pause - the mean interval surrounding a phrasal boundary, delimited by a comma, period, question mark, etc.
- Mean pre-send pause - the mean interval between when the last character of a message is typed and the ENTER key is pressed to send the message

The full text of the opinion questions that make up the predictors of interest is outlined in Section 4.4.5. An example question was: *Hypothetically, how much do you think you'd enjoy watching a movie with your partner? [1=Not at all, 7=I would definitely enjoy it]*

Table 6.6 shows the significance of different opinions on typing patterns. The values were derived from an ANOVA comparing a baseline model to a test model that added opinion ratings as a predictor. Thus, the *p*-values reflect the influence of the additional opinion predictor in predicting keystroke patterns, or to what extent different opinions affect the way a participant types while controlling for the sentiment of a turn.

As a point of clarification, the column labeled "Mean" is the average of all proceeding opinion values. These questions asked a participant what they thought of the conversation or partner *per se*. On the other hand, the final question also required introspection and so I did not want to confound the other opinion values with this latter value.

Looking at each keystroke pattern (i.e. each row), the overall mean inter-keystroke interval and mean dwell time are strongly influenced by a number of different opinions. Moreover, the two opinion values that do not have a statistically significant or marginally significant effect on keystroke patterns still have a p-value below 0.20. This result will need to be further investigated in future studies.

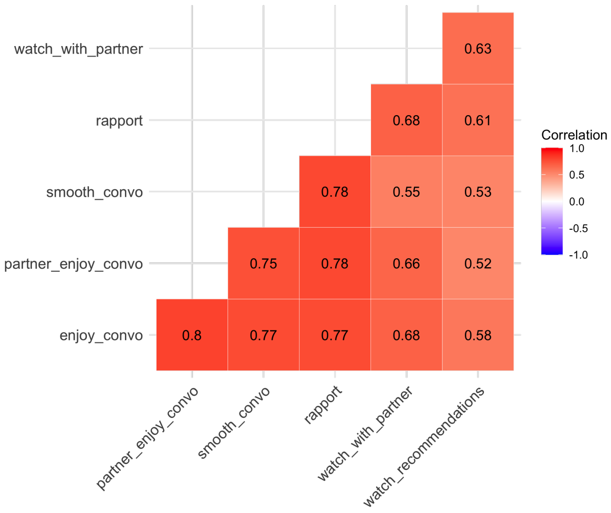


Figure 6.5
The correlation of scores between each opinion question. As can be seen, all of the correlations are very strong in a positive direction. In a post-hoc test, all correlations were statistically significant to an extremely low α.

Before looking at the influence of specific opinion scores, it is also important to note the non-independence of each survey question. As seen in Figure 6.5, all of the answers to the individual questions were strongly positively correlated to one another. The similarity can be seen in Figure 4.8, where the distribution of responses to most questions was very similar.

However, the influence of different individual opinions (i.e. each column in Table 6.6) shows interesting results. Specifically, a participant's opinion on whether they will watch a partner's recommendations is closely related to a number of keystroke patterns, although the direction of causality between opinions and keystrokes is unclear and will require follow-up experimentals. Moreover, the mean opinion score also significantly affects multiple keystroke patterns. This latter result might be the result of the high levels of correlation seen between individual opinion values, as seen in Figure 6.5. Because the mean score essentially accounts for many opinions, it may be that this multi-faceted score is the most accurate predictor for physical change as manifested in keystroke timing.

A final result is the significant influence of self-awareness on the pause time before a message is sent. One way to look at this is that the less confident a participant is that their partner is enjoying the conversation, the more that participant might hesitate before sending a message.

6.4 Discussion

As a reminder, Study 2 sought to answer the following research questions:

RQ 2a) Does keystroke information provide additional information about message sentiment as well as sentiment change between turns, above standard lexically-determined sentiment values?

RQ 2b) Are typing patterns sensitive to a user's opinion of their partner, when considered independently from the sentiment of a user's messages?

As briefly noted in the introduction to Study 2, I limited my dataset only to *proper* turns, or turns without interruption. This was done in order to avoid introducing further variation into the timing surrounding turns, where a user seeing an "is typing" indicator might disrupt their flow of language production. This is further complicated by the notion that different cultures respond differently to conversational interruption. Tannen (1984) classifies these conversational styles into two groups: "high involvement" speakers do not mind interruptions or overlapping speech and will even intentionally use simultaneous speech to show agreement or enthusiasm; "high considerateness" speakers, on the other hand, are more concerned with being considerate of others and prefer not to impose on the conversation as a whole or on specific comments of another conversant.

These styles of interruption also have an intriguing tie-in with Study 1, on dialogue acts. Choe (2018) and Tannen (1984) point out how an interruption can also be perceived differently depending on whether it references previous context or whether it is completely unrelated and intends to advance the conversation in a new direction. Choe (2018) uses this idea to evaluate "listenership" in a multi-party instant message conversation, and finds that people naturally find ways of adapting strategies from spoken conversation into text-based conversation in order to demonstrate their level of involvement and engagement. Given these findings from previous research, it will be especially fruitful to evaluate typing patterns in overlapping turns, not only because these turn types constitute the majority of a conversation, but also because different behaviors in response to interruptions could be telling about the cultural expectations of a user.

Before delving into the individual results, the differences between the VADER sentiment model, my own manual sentiment annotation, and other algorithmic models should be further clarified. The primary difference between the VADER ratings and the manual ratings is that the VADER ratings were determined algorithmically while the annotation ratings were done by (a small number of) human annotators. In my experiments the annotation scores were held as a "gold standard" because the human annotators were influenced by not only the explicit lexical content but also implicit connotations of a message, which could be missed by an algorithmic sentiment determination (see Appendix F for the full guidelines, based on Mohammad (2016)). Although it is possible or probable that VADER missed some implicit meanings, the advantage that VADER has over similar software is that VADER was tuned for microblog-like language (Twitter) and its results generalize better over multiple social media domains (Hutto and Gilbert, 2014, p. 216). Since the spontaneous conversations in my experiments use an informal language style that is more similar to social media posts (as compared to the periodicals that similar sentiment libraries were trained on), VADER seemed to be the most germane library for my purposes.

Creating the optimal keystroke model was interesting, specifically which keystroke predictors were the most informative. For example, the optimal model used the dwell time of only the content words. However, in building the optimal model, an earlier model used all dwell times. While overall dwell time was a significantly informative predictor, it also made the by-user random effect less informative. However, when looking at only the dwell time within content words, the by-user random effect was much more significant and the overall model fit improved. This seems in line with similar findings from Lee et al. (2014) and Lee et al. (2015b) which found that dwell is significantly correlated to emotional arousal at a unique user-by-user level. This makes sense in that emotional arousal is more related to content words (Niederhoffer and Pennebaker, 2002), and so limiting dwell times, connected to emotion, to only content words, should be more informative and more tailored to individuals. In the future if researchers are building an agent that is informed about user emotions from keystrokes, the predictors of the optimal model point to the notion that not all keystroke patterns will be helpful. For example, if dwell times in function words are not unique to individuals then incorporating these dwell times into a user template might make the template less tailored to that particular user.

It should also be noted that the optimal keystroke model created for these experiments was created using only keystroke features. It is possible that when keystroke metrics are used as predictors along with lexical sentiment scores, then a different set of keystroke metrics might be optimal. In other words, there might be significant overlap between keystroke features in the optimal model and lexical sentiment scores. As such, adding keystrokes to sentiment would be largely redundant. This issue should be addressed in future work that builds an optimal model that is also based on a keystroke metric's low correlation to lexical sentiment scores.

Experiment 2a provided an answer to RQ **2a**, showing that keystroke information can provide additional or complimentary information in addition to the information provided by a lexically-determined sentiment baseline value. These results are seen in Table 6.4, where a combined model that includes both lexical information and keystroke information outperforms a model that only uses a standard sentiment analysis library.

Table 6.5 echoes results in previous literature that found that sentiment analysis tools designed to assess sentiment in isolated text could also be useful for conversational dyadic text (Ojamaa et al., 2015; Zhang et al., 2021). Since there exist only a handful of studies that perform sentiment analysis on dialogue, as compared to monologue, this additional validation is also valuable to the general research community.

It is difficult to compare the results of my models to those used for e.g. training VADER (Hutto and Gilbert, 2014), because the datasets were different sizes and the tasks were different.

Nonetheless, it appears that the VADER sentiment analysis library that was trained on isolated data will still provide information on sentiment in conversations.

However, beyond demonstrating the transferability of lexical models, my experiment also illustrated the utility of keystroke models in determining conversation-based sentiment. This can be seen as an extension of the plethora of literature reviewed in Yang and Qin (2021) which has demonstrated that keystroke information can be useful for detecting sentiment in isolated settings. Taken as a whole, Experiment 2a demonstrated that lexical and keystroke sentiment models designed for isolated text can be extended to lexical and keystroke information produced in a conversational setting.

Experiment 2a also provided a significant contribution to existing research by showing that aside from keystroke information being useful for determining the sentiment of a turn *per se*, keystroke information is also useful for predicting how sentiment values change from one user's to the other user's turn. In fact, keystroke information provides the most statistically significant amount of additional information for predicting change, as compared to the other prediction tasks in Experiment 2a. This is similar to the neural network model built in Hazarika et al. (2018), which predicted emotions of utterances based on lexical content as well as language production patterns. Whereas Hazarika et al. (2018) used audio and visual features of a user involved in a spoken conversation, I was able to derive similar results using keystroke patterns. As mentioned previously, monitoring keystrokes is significantly less intrusive than a video camera and microphone, which points to an advantage of the approach taken in this study.

Sentiment change is especially tricky to keep track of and predict within conversations: Text written in isolation proceeds in a somewhat linear and logical fashion; conversely, sentiment in a conversation can jump around, where sentiment is constantly dependent on changing context as well as more or less recent context (see Zhang et al., 2021). When designing a chatbot or an augmented text chat platform, if a human or computer agent could take advantage of keystroke information from an interlocutor to detect when sentiment has shifted, then this could act as a trigger. The human or computer agent would have evidence that the interlocutor is not on the same page as themselves, and so a shift in tone or content is necessary.

To investigate an additional source of variability in dialogues, Experiment 2b sought to answer the final research question, RQ **2b**. Whereas Experiment 2a demonstrated that keystrokes patterns are informative about sentiment, Experiment 2b answers whether user's opinions of their partner and the conversation itself also affect typing patterns in addition to the sentiment of an utterance.

Table 6.6 shows that the user's opinions are strongly correlated with the way that they type; specifically, opinion scores are related to typing patterns in a way that is independent of the sentiment of the text they are typing. As a concrete example, a user could type something very positive or very negative, and that positive sentiment would be reflected by typing information. However, the sentiment of the text would not be the only factor that contributes to the way they type: the user's opinions of their partner would also independently be associated with the typing metric. This seems to answer RQ **2b**, in that typing patterns are independently sensitive to both the specific utterance sentiment and overall user opinions.

This finding makes sense in light of findings such as those in Gidron et al. (2020) and Barnett et al. (2018), which showed that the rapport between a participant and an experimenter can effect executive function and experimental test results. In my own experiments, then, it would also stand to reason that opinions would effect typing patterns. The reason for this is that typing execution is governed by multiple cognitive processes, including motor execution, lexical recall and executive function (Dagum, 2018; Logan and Crump, 2011; Rumelhart and Norman, 1982). As such, if rapport can interfere with executive function, then changes in rapport levels can likely also change typing patterns. Note, however, that this conclusion does not go so far as to hypothesize about the direction of effect of a rapport-typing connection.

This finding is important if researchers intend to use typing patterns to derive underlying or unobservable motivations. In a spoken conversation, speakers can use auditory cues to infer some of this information, as in Ondobaka et al. (2017). But this information is not readily available in text-based dialogue. Nonetheless, Experiment 2b seems to suggest that this information is present in the typing signal, and is therefore recoverable and can be made manifest.

Based on the results, it seems that the user's opinions most strongly affect the broad typing patterns, i.e. overall inter-keystroke intervals (IKIs) and overall dwell times. It is possible, though, that my data should have been subset in a different fashion, and so opinions actually affect a different and more specific subset of typing patterns. This should be investigated in future work.

Another interesting result was that multiple user opinions affect the pause time before a user sends a message, i.e. their hesitancy before pressing ENTER. This pause time is the gap between when a user finishes composing a message and they then transmit the message. Since pauses in typing can be a sign of uncertainty (Schilperoord, 2002), a delay in this location could represent uncertainty or a lack of confidence in whether a user wants to send the message they have just composed.

It would be interesting in future work to also further investigate the relationship between a user's opinion of themselves (self-opinions in Table 6.6) and the hesitancy in sending a message. As shown in work such as Jokinen (2010), in spoken language a connection exists between hesitation and uncertainty; it should be investigated as to whether this same relationship holds in text-based communication as well, where a lower self opinion is correlated with greater hesitation in transmitting a message.

Pauses before sending a message are also interesting because my findings leave open the question of whether this hesitation is a "verbal cue" or not. Kalman (2007, 42) points out that, "To be considered a nonverbal cue in text-based CMC, an element must be expressed differentially by different writers or in different contexts, and this variance must be communicated to the reader." Although my experiments used a by-subject random effect, Table 6.5 shows how this random effect was not always significantly influential on the predictive task. Therefore, a more uniform experiment should be performed where more subjects produce a more uniform quantity of samples, which could provide evidence as to how consistent a user is about using this pause, but also how consistently different each user is. This is important because the goal of my research is not just to elucidate cognitive processes and how they differ under different circumstances, but also which processes are intended to *communicate* underlying emotions, rather than just being a reflection of those motivational differences.

The question still remains though, *Why does all of this matter?* When building computer agents that can form trustworthy and natural-feeling relationships with humans, it is important to not just match wording or sentiment, but also use timing indications such as response times to gauge a user's motivations. Li et al. (2017b), for example, shows the importance of combining these factors to make a human-like robot. Zhao et al. (2016) also shows how rapport in virtual agents is partially based on timing patterns.

The importance of this is highlighted in Bothe et al. (2017), which brings up a critical point relating to emotion detection in dialogue: as humans and computers interact in more high-risk environments, such as a car assembly lines or repairing a satellite in orbit, the detection and conveyance of emotion in language production becomes crucial. As an example of applying my own research, one applicable domain is in telehealth. For a remote healthcare provider, it is important to understand not only what a patient is saying, but also the emotion valence of how they are saying it. Keystroke analysis of dialogues could be an important conveyance of this emotional content.

As a similar example, an important underlying trait that can be detected in typing patterns is cognitive load (Brizan et al., 2015, inter alia). Khawaji et al. (2014) goes even further, and shows how keystroke and mouse movements reflect both trust and cognitive load. These two human factors could be important in a setting where e.g. a repair-person is delivering complex directions via CMC or a doctor is involved in a difficult discussion: it is important to be able to judge both the opinion of an interlocutor as well as the degree to which they can comprehend the conversation, where an understandable conversation could be represented by a lighter cognitive load.

These findings are also important in light of findings such as those in Bos et al. (2002), which examined trust development in four different communication settings: face-to-face, video, audio, and text chat. They found that the emergence of trust was the worst and slowest in text chat. However, findings such as those in my Experiment 2b could be used to make user trust more salient and thereby make trust development or lack of trust more obvious to a user or their conversational partner. This could allow the user to consciously or intentionally improve their level of trust.

Synthesizing the findings of both experiments, though, it seems that keystroke patterns provide information about underlying social motivations. This information includes both a more nuanced sentiment, as compared to what can be detected lexically, as well as overarching opinions of a conversational partner which may never be overtly realized in a conversation.

6.4.1 Unexpected findings

The abundance of neutral messages was surprising, in that I had anticipated a much more "lively" discussion about TV and movie preferences since they can be very personal and strongly held beliefs. However, this distribution was similar to that in Ojamaa et al. (2015), which also found that the majority of the turns in their data had neutral sentiment. For the sake of Study 2, the abundance of neutral sentiment was less than ideal, since it made the examples of strong sentiment more rare. In retrospect this makes sense, in that the goal of most turns is simply to advance the conversation, e.g. forward functioning dialogue acts in Study 1 (Chapter 5).

Similarly, as illustrated in Figure 6.3a (algorithmic sentiment analysis) and Figure 6.3b (manual sentiment analysis), a large proportion of turns have positive sentiment, and a majority of survey opinions are positive. As noted by Svennevig (2014, 302), the goal of a first-encounter conversation between strangers (like those in my experiment) is to establish common ground and establish a relationship of cooperativeness. Thus this large positive skew should have been anticipated, and a more deliberate methodology should have been employed to elicit negative utterances and negative partner opinions.

6.5 Future work

One of the most important next steps in this research is to establish not just a correlation between underlying motivations and typing patterns, but to also establish a direction of causality. Similar to a study such as Liebman and Gergle (2016b), an intervention should be added to these conversations so that a conclusion can demonstrate whether changing sentiment causes changing typing patterns, or whether a (preconceived) opinion causes changes in typing patterns. In my own study, it seems safe to conclude that both of these factors are correlated to changes in typing patterns, although the experiments lack the ability to say whether sentiment or opinion causes changes in typing.

As workplaces become more distributed (Pew Research Center, 2020; Teevan et al., 2022), it will also be important to study how sentiment and opinion influence typing in larger groups, such as a meeting with more than two participants. In addition, it is important to study typing patterns in situations where power dynamics are more prominent. For example, Willemyns et al. (2003) studied boss-employee trust relationships through the lens of communication accommodation theory. It points to the necessity of more conscientious discourse management in order to establish and build this type of relationship, especially in a distributed environment. In general, this type of study should move past first-encounter dialogues (Ojamaa et al., 2015), and be extended to longitudinal studies that investigate how relationships and typing change over time, as familiarity between users develops.

Finally, the research in Study 2 must be extended to overlapping turns. As mentioned in numerous studies such as Gravano and Hirschberg (2011), the majority of turns in a dialogue overlap one another. Therefore, limiting a study or (eventual) tool to only “proper,” non-overlapping turns artificially eliminates a large proportion of a dialogue.

Part of my rationale for not studying overlapping turns in this initial study is that if a user is typing, and then sees a typing indicator pop up for their interlocutor, this may be a distraction or cause the user to stop typing and wait for their interlocutor. However, a future study should investigate user behavior when a typing indicator appears. Tegos et al. (2020) showed that most users rely on a typing indicator popping up as proof that their interlocutor is engaged in the conversation. One could imagine in the case of investigating typing and underlying relationships that if rapport level is high, then a user will not feel to stop typing and wait for their interlocutor to send a message, whereas if a user is less confident in their relationship, they might stop and wait to see what their interlocutor says before proceeding.

Chapter 7

Study 3: Predicting rapport levels based on keystroke patterns

Study 3 aims to classify users, based on their reported level of rapport during a conversation: one group reported a very high level of rapport during a conversation, while the other group reported medium to low levels of rapport. Since rapport has been shown to be detectable in speech prosody, and prior studies have shown that typing patterns can parallel aspects of speech prosody, the aim of this study is to discern if feelings of rapport are also detectable in typing patterns. Moreover, this study investigates how keystrokes from different subsets of a conversation predict rapport levels. By better understanding this, future researchers will be able to pinpoint areas of a conversation in which levels of rapport could be more accurately detected. Since rapport detection in language production is dependent on when the language is produced within a dialogue, this is especially important (Tickle-Degnen and Rosenthal, 1990).

Study 3 aims to detect as many cases of the medium-to-low rapport group as possible, since in a situation where we are trying to improve an interaction, this would be the group most in need of detection and intervention.

However, rapport is a very complex social dynamic (e.g. Tickle-Degnen and Rosenthal, 1990). This adds to the importance of finding additional sources of information, e.g. keystroke timing, that are sensitive to rapport levels. Many definitions of rapport exist (see Section 7.1 and Table 7.1 for examples); one of the most apt comes from early work on rapport, though: "…an individual's experience of harmonious interaction with another person, often described as 'clicking' or 'having chemistry"' (Tickle-Degnen and Rosenthal, 1990, p. 286) Despite the complexity of rapport, typing patterns provide a promising way to understand and measure this underlying social dynamic, since typing is sensitive to a number of cognitive and social processes (Schilperoord, 2002). Study 3 collected survey data from each subject after a conversation, where each question in the survey asked the subject to rate the conversation and their partner on different dimensions of rapport. I then took the mean of all ratings, and used this as a measure of rapport during the conversation. I then extracted features from typing patterns during the conversation, and used machine learning to predict rapport ratings based on these typing features.

Since establishing high rapport can be an important element of a successful interaction or productive collaborations , whether doctor/patient, salesperson/customer, manager/employee, etc., it could be very helpful to be able to continuously measure and predict rapport as an interaction unfolds. This study pursues that line of inquiry by looking at not only how well a full typing session can predict rapport, but also how well different subsets can predict final rapport. If this is successful, it could allow rapport level to be measured before an interaction concludes, so that one party can make adjustments to their interaction style in order to improve rapport.

As an example, if a customer is interacting with a salesperson online, the salesperson would want to make sure that the customer is enjoying their interaction. Furthermore, the salesperson would want to know as early as possible whether the customer is enjoying themselves, so that the salesperson can make adjustments to their approach.

However, study 3 investigates whether each portion of a conversation is equally valuable in predicting overall rapport, in order to establish if the influence of rapport is stronger in temporal slices of a conversation, as well as in slices that are segmented by the role that the typist is playing in the overall experimental task. Finally, an additional subset randomly samples a proportion of keystrokes, to see if the aforementioned slices *per se* are responsible for changes in predicting rapport, or if the type of subsetting is important. (See Figure 7.1 for a visual representation of these subsets.)

As a reminder, Study 3 aims to answer these research questions:

RQ 3a) Can typing patterns over an entire conversation be used to predict low levels of rapport between partners in an interaction?

RQ 3b) Can a random subset of keystroke data predict conversational rapport as well as a complete set of keystrokes?

RQ 3c) Does a subset of keystrokes from the first half of a conversation predict low rapport as well as a subset of keystrokes from the second half of a conversation?

RQ 3d) Does a subset of keystrokes from when a subject is providing recommendations predict low rapport as well as a subset of keystrokes from when a subject is receiving recommendations?

The motivation for these questions is to find the most valuable subsets of an interaction for predicting rapport. Tickle-Degnen and Rosenthal (1990) hypothesized that different components of rapport are more important or less important at different stages of an interaction; by slicing a conversation temporally, we can show which of those components is most important for overall, final rapport. In addition, study 3 is motivated by the ultimate application of my research: recommender "systems." A system could be a chatbot providing automated recommendations to a customer or a doctor providing well-being recommendations to a patient. In both cases, though, the user of interest is *receiving* recommendations, and so it is important to see if a user in this role is expressing signs of rapport.

To answer these questions, Study 3 used a uniform set of features for each model, but different subsets of keystroke data went into calculating each feature. A machine learning model was then trained on each subset, and the models' predictive accuracy was compared. As discussed in Section 7.2, though, the most important class to predict was when a subject reporting lower levels of rapport, rather than accurately predicting when a subject reports high rapport. Therefore, the comparison metrics reflect how well the model detected this class, rather than, e.g., overall accuracy.[1] The

[1] It should be noted that for all of the models, the ZeroR, simply classifying every instance as the majority class, would be 136/192, or 71%. But by definition, this would misclassify every instance of the minority class; therefore

features utilized in this study include both keystroke timing signatures as well as stylometric features. However, in each experiment, feature importance is also assessed because the same feature derived from a different subset of data could make that feature more or less valuable for predicting overall rapport.

Overall, I found that keystrokes are useful for detecting low rapport, especially when keystrokes from an entire conversation, rather than a random subset, are utilized. In addition, I found that keystroke patterns from when a user is *receiving* recommendations are more useful at predicting low rapport than keystrokes patterns from when a user is *providing* recommendations.

7.1 Related Work: Rapport

Rapport is essential to establishing a good relationship or good interaction, but is notoriously difficult to define and measure.

7.1.1 Defining rapport

Tickle-Degnen and Rosenthal (1990), a seminal study on the conceptualization of rapport, defines it as "…an individual's experience of harmonious interaction with another person, often described as 'clicking' or 'having chemistry."' While this definition is comprehensive, and has been adopted by many subsequent researchers, the definition still resorts to phrases such as "harmonious interaction" and "having chemistry," which are themselves difficult to define.

Lubold and Pon-Barry (2014) defines rapport in terms of a feeling of closeness, and many rapport-related questionnaires ask about a feeling of connectedness. Along those lines of connectedness, other studies on rapport ask subjects to what degree the partner "paid attention" to the subject, or the conversation was "engrossing" and "worthwhile" (LaBahn, 1996). Regardless of definitions, it seems that researchers must repeatedly resort to abstract concepts in order to explain rapport.

it is not a germane comparison. All of the classifiers have an overall classification accuracy near 71%, but this is not reported because Study 3 is concerned with correctly detecting the minority class.

These definitions point to the possible utility of connecting rapport to keystroke analysis. Table 7.1, below, highlights how keystroke timing could add a concrete measurement to some of the more abstract definitions of rapport.

The motivation behind why keystroke timing analysis could be a more accurate method to measure rapport comes from Chung and Pennebaker (2014, p. 5):

> A consistent finding is that many of the word categories used to reliably classify psychological states can be considered to be a part of language style as opposed to language content. That is, **how** people say things is often more revealing that **what** they are saying. [Emphasis added]

In other words, this study (as well as previous studies in this thesis) look at not only lexical choices (*what* is said) but also timing characteristics of keystroke production (*how* it's said).

7.1.2 Measuring rapport

Early attempts to quantify rapport mostly focused on the relationship between psychotherapists and their patients, since rapport is critical to building a productive therapeutic relationship. Anderson and Anderson (1962) quantified rapport by looking at the proportion of matching definitions between a therapist and client, where higher rapport was correlated with a greater proportion of matching definitions, since both parties saw certain concepts in the same light.

Relying on word-matching, though, is a crude estimate; by using both linguistic choices and production patterns, I hope to be able to capture more robust similarities. Müller et al. (2018) ran a study with many similarities to my own study. They aimed to predict low rapport based on non-verbal cues. These included facial expressions and speech prosody. They found these non-verbal cues to be extremely helpful in establishing high rapport and in their own experimental predictions. For example, facial expression and body posture can signal "attention," which is an important element of rapport. Further, Müller et al. (2018) hypothesizes that these non-verbal cues are more often imitated by the other participant, and this reflection is important for establishing rapport.

Definition (source)	Findings from keystrokes and prosody
"…an individual's experience of harmonious interaction or 'clicking' …" (Tickle-Degnen and Rosenthal, 1990)	Pauses during typing, as well as increased mistakes, are associated with increased cognitive load or strain (Schilperoord, 2002). In a "harmonious" interaction, it seems that one would expect fewer prolonged pauses and fewer mistakes, because the subject is more comfortable with expressing their thoughts.
"…the perception that a relationship has the right 'chemistry' and is enjoyable." (LaBahn, 1996)	Multiple studies have found that typing patterns are sensitive to a typist's emotions (Epp et al., 2011; Lee et al., 2015a). Notions such as enjoyment, which are a component of rapport, seem to fit this category, and are therefore likely to be perceptible in typing patterns.
"engrossing…involving… worthwhile…" (Grahe and Bernieri, 2002)	Studies that looked at the connection between keystrokes and emotions also study the intensity of emotions, not just the positivity/negativity of the emotions themselves (Maalej and Kallel, 2020). It seems that a conversation that is engrossing or involving will evoke more intense and less apathetic contributions. This should be reflected in the energy with which keystrokes are produced, as realized in keystroke dwell time.

Table 7.1
Definitions of rapport and possible quantification by keystroke patterns

These findings are especially germane since my thesis approaches typing patterns as a manifestation of "silent prosody." However, whereas prosody in Müller et al. (2018) is visible/audible to other participants, my study aims to show that prosody manifested in typing patterns (that are likely not visible to a partner) is also valuable.

Herzig et al. (2016) performed a study that bears a number of parallels to my own study, where they aimed to predict customer satisfaction of an interaction based on previously compiled personality profiles, as well as *affective*, as opposed to purely lexical features, of prior conversations. Their goal was to be able to make predictions, and perhaps route the customer to the appropriate representative, before the current interaction even began. In my own study, this is similar to Experiment 3b, where I aim to predict rapport based only on the first half of a conversation. I also use features beyond purely lexically-based features, and look at production patterns, including semantically-contextualized production patterns.

In terms of the actual classifications of rapport, two primary methods exist: self-reports from participants, and external observers who assign a rapport rating to an interaction they observe. This study relies on participants' answers to a questionnaire after the experiment, which is a form of self-reporting. Future studies could also use external annotators, as both self-reports and external annotation have been shown to yield similar results, where a strong correlation exists between self-reporting ratings and an external annotator's ratings, even when the external judge only views a small slice of an interaction (Carney et al., 2007).

7.1.3 The complexity of rapport

Seo et al. (2018) highlights the complex interactions of individual verbal behaviors that contribute to a sense of high rapport. For example, asking an off-topic question, such as personal information, can increase rapport in certain settings, whereas in other settings or at the wrong time it can seem rude. As such, measuring straightforward semantic similarity between turns would not be a sufficient surrogate for rapport estimation. Similarly, some statements or questions are properly responded to with short responses, while others have more appropriate long responses. Therefore, in this case, measuring simple turn length, or turn length similarity, would also be insufficient.

The model devised by Tickle-Degnen and Rosenthal (1990) provides an attractive apparatus for experimentation, as the study breaks down rapport into three dimensions: *attentiveness*, *positivity*, and *coordination*. But the study also found that each dimension does not exude equal influence on rapport throughout the course of a conversation. For example, early in an interaction positivity and attentiveness are most important for establishing good rapport; in the later stages of an interaction coordination and attentiveness are more influential on good rapport. For this reason, my study breaks down conversations into a first half and a second half, in order to assess both the predictive capability of each half as well as which features are most important. Combining my findings, i.e. which temporal slice is most important, with Tickle-Degnen and Rosenthal (1990)'s most important components during that slice, will provide insight into which components of rapport are the most accurate predictors of overall, final rapport.

Finally, Raj Prabhu et al. (2020) provides a robust confirmation of my own research methods, specifically combining a subject's answers to multiple questions into a single metric, which they call *Conversation Quality*. Because the perception of quality is so multi-faceted, they affirm that it can only be derived from the answers to multiple questions. Similar to the ultimate goals of my thesis research, they "intend to quantify spontaneous interactions by accounting for several socio-dimensional aspects of individual experiences" (p. 196). By quantifying an experience, researchers can compare different experiences and also quantify how much a certain change improved an experience.

7.1.4 Rapport and spoken language production

As mentioned in the introduction, Study 3 utilizes both keystroke timing features and stylographic features. These stylographic features, however, go beyond metrics such as the average length of an utterance, and instead are focused on elements such as the ratio of utterance length between one subject and their partner. This is related to the notion of "coordination," which Lubold and Pon-

Barry (2014) found to correlate strongly to rapport. This ratio, rather than outright measurement, highlights the differences with studying underlying motivations in interaction rather than in isolation. In other words, understanding a typist in dialogue requires more than understanding the typist's own metrics, but also understanding the relationship of those metrics to their partner's metrics.

Michalsky and Schoormann (2017) shows that when a subject finds their partner to be more likable and socially attractive, the subject puts more cognitive effort into matching their partner's style, where style-matching is another form of coordination. In this study, greater cognitive effort is manifested by typing rate and edit patterns. As such, it will be interesting to see if these variables are predictive of rapport levels.

Finally, it seems that rapport is ideal to study through keystroke analysis, given the sensitivity of this type of analysis to changes in cognitive load (Brizan et al., 2015, *inter alia*). Barnett et al. (2018) and Barnett et al. (2020) found that when an examiner intentionally established either high or low rapport with a subject, even though the experimenters did have meaningful interactions with the subjects, the level of rapport affected performance on cognitive tasks such as the Stroop test and word association tests. In these investigations, it was found that high rapport improves results on cognitive assessments.

Findings such as those by Barnett and colleagues seem to underscore the importance of studies in my thesis. If keystroke analysis can provide accurate predictions of perceived rapport, and rapport helps improve cognitive functioning, then increasing rapport will create a more productive interaction, e.g. more accurate movie recommendations because the partners in an interaction are able to better articulate their thoughts.

7.2 Methodology

7.2.1 Dataset partitioning

In order to test the predictive accuracy of different subsets of keystroke data, the entire dataset was partitioned in different ways, as illustrated in Figure 7.1. The random subset used 50% of the keystrokes of each subject in each conversation, so as to include approximately the same number of keystrokes as the other partitioning methods. In addition, if one subject had 100 messages and the other subject had 50 messages, then taking an equal proportion of each would keep the comparative ratio between the two subjects the same.

As can be seen in Figure 7.1, the role partitions included keystroke data from both halves of the conversation; vice versa, the half partitions included data from each role.

7.2.2 Determining number of classes

This study used each subject's post-conversation questionnaire responses to construct a 6-dimensional vector of conversational impressions for each subject. These 6 questions are reproduced in full in Section 4.4.5. They asked participants about enjoyability, conversational smoothness, partner connections, etc. In other words, each question attempted to get at a different dimension of overall rapport (LaBahn, 1996; Lubold and Pon-Barry, 2014; Tickle-Degnen and Rosenthal, 1990).

Before proceeding with any prediction preprocessing, the subjects' survey answers had to be divided into the appropriate number of classes, so that classes were as discrete as possible without being too small. As can be seen in Figure 7.2, survey responses were heavily positively skewed (a Likert scale from 1-7 was used for each question, with 1 being low enjoyability, low connection, etc. and 7 being high ratings of those qualities). Looking at the average response value to all questions, 40 subjects had a 7.0 average, i.e. every question received a 7 rating. Only 15 subjects had a mean answer of less than 3.5.

In order to determine the optimal number of clusters, I used `NbClust`, an R package that tests 30 different distance algorithms to measure the appropriate number of clusters (Charrad et al., 2014). Each algorithm uses different instantiations of cluster analysis such as within cluster sums of squares, average silhouette and gap statistics. The majority of distance metrics recommended 2 clusters, while a handful recommended 3 clusters. This study will use 2 clusters, although studies can also investigate classification for 3 clusters. Figure 7.3 uses PCA to visualize 2 versus 3 clusters. As can be seen in Figure 7.3b, there is not clear partitioning between the highest 2 clusters, which

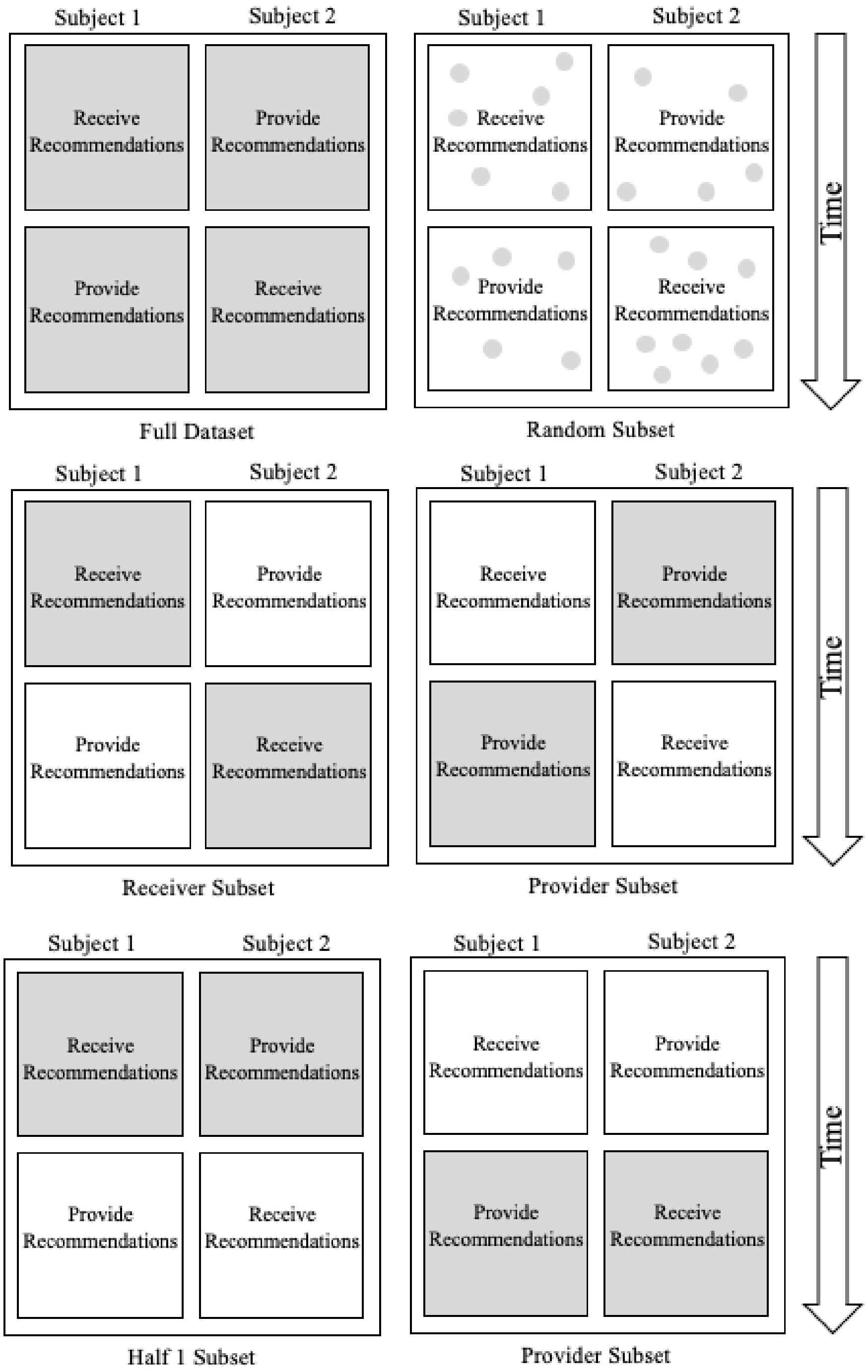


Figure 7.1
Each subset was created by dividing the full dataset in different ways, either by receiver/provider roles, first half/second half, or random sampling. The grayed out areas indicate the selected data.

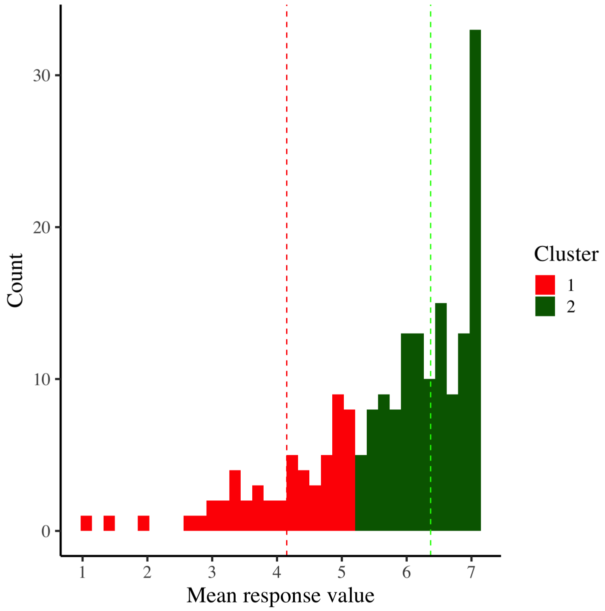


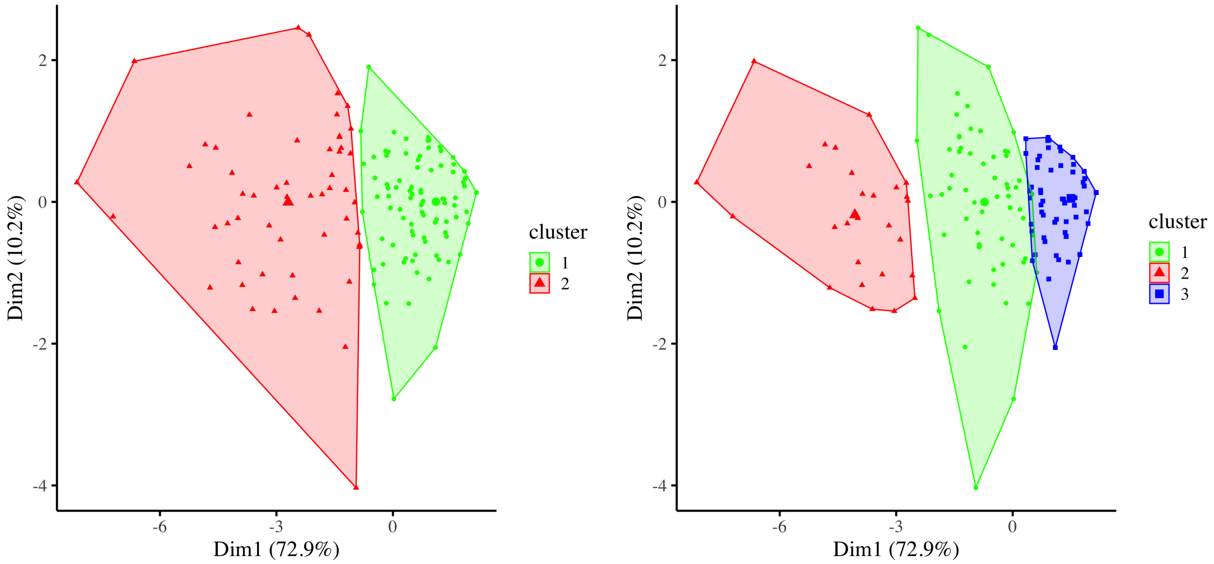
Figure 7.2
A histogram of the mean response value from each subject. The bottom lines represent the mean response value for each cluster. As can be seen, the overall response value was heavily positively skewed.

would hint that these clusters are not necessarily well-separated. Once it was determined that 2 clusters was optimal, k-means clustering was used to classify each subject into a cluster.

Nonetheless, class size was still skewed. Cluster 1, made up of subjects who had lower rapport ratings, included only 56 subjects, with a mean question response of 4.15 (out of 7.0). Cluster 2 included 136 subjects with a mean response rating of 6.38. Because of this imbalance, synthetic minority over-sampling (SMOTE) was used to balance training data (Chawla et al., 2002).

7.2.3 Metrics selection

As mentioned previously, the minority class, consisting of subjects who reported lower rapport levels, was the class of interest. For this reason, overall accuracy was not used, since the majority class consists of those subjects who feel higher rapport; in a real-life setting these partners would not necessarily require extra attention. In order to use more meaningful metrics, this study focuses on the following 4 metrics:



(a) Creating 2 clusters, or classes, from the 6-dimensional answer vector for each subject. PCA was used to visualize the clusters.

(b) Creating 3 clusters, or classes, from the 6-dimensional answer vector for each subject. PCA was used to visualize the clusters.

Figure 7.3
The difference in cluster partitioning between two clusters and three clusters. In partitioning the data into3 clusters, more overlap exists between the top two clusters.

- **Area under the ROC curve (AUC)** - Studies such as Wardhani et al. (2019), among others, suggest that AUC is more robust than F1-score when evaluating imbalanced data. Further, AUC uses prediction probabilities, whereas F1 looks at correct/incorrect predictions. Because rapport is not a discrete binary, probabilities also seem more appropriate for the problem at hand.

- **Matthews correlation coefficient (MCC)** - The MCC is a convenient metric in that it is a single numeric representation of all cells in a confusion matrix (Chicco and Jurman, 2020). However, its performance can sometimes degrade on imbalanced data (Zhu, 2020). As a result, it is still reported in the results as all elements of the confusion matrix are important for this task; however, AUC will still take precedence.

- **Positive predictive value (PPV)** - The PPV measures proportion of positive cases (those who *actually* report low rapport) against the number of predicted class members. Unlike a straightforward true-positive or true-negative rate, the PPV also takes into account the *prevalence* of a class, which is the proportion of the class of interest within the entire dataset (Altman and Bland, 1994). For this reason, it is germane to this study, where low rapport prevalence could vary in different situations; therefore, it is reported but also secondary to AUC.

- **F1 score** - Although F1 scores may not be the most appropriate for imbalanced data, it does balances precision and recall, and takes the harmonic mean of both, thus giving them equal weight. While empirical evidence exists to show that F1 core might not be ideal for my scenario, the scores are still reported here for the sake of completeness, and because they are widely used and understood in the machine learning community.

7.2.4 Feature selection

The full list of features, with brief explanation, is enumerated out in Table 7.2. Most features are repeated from previous studies. However, since Study 3 uses data from both sides of an entire conversation, it also includes dyadic features such as ratios between subject language production and conversational partner language production.

7.2.5 Classifier choice

All of the models in this study were built using the `tidymodels` ecosystem in R (Kuhn and Silge, 2022). This allowed for all model types to be built from the same syntax and code blocks, which aids reproducibility within this study as well as future reproducibility.

A tuning subset of data, a holdout set consisting of 5% of the total data, was run using H2O's automatic machine learning (AutoML) (LeDell and Poirier, 2020). The algorithm tested 83 different

Feature	Comments
Inter-keystroke interval (IKI)	Similar to Epp et al. (2011), times were log-transformed and pruned to the top 95%
IKI of content words	Content words (Pennebaker, 2011) have intrinsic meaning and include nouns, verbs, etc.
IKI of function words	Function words include determiners, pronouns, etc.
IKI at word beginning	This measures the hesitation before selecting a word, rather than mid-word when the lexical item has already been retrieved (Logan and Crump, 2011).
IKI in mid-word	This measures motor execution, rather than lexical retrieval (Logan and Crump, 2011).
IKI at phrasal boundaries	As described in Galbraith and Baaijen (2019), these pauses reflect different cognitive processes that focus on phrase planning rather than word retrieval.
IKI before sending message	This reflects hesitation before transmitted a message.
Dwell time	Dwell time is strongly connected to emotion (Lee et al., 2014).
Dwell time of content and function words	If dwell time is more connected to emotions, and content words have more intrinsic meaning, then the semantic words types should be measured separately.
Edit count	The frequency of edits should reflect uncertainty in the language being produced (Olive et al., 2009).
IKI of lexical density	Lexical density measures the ratio of different types of words, such as nouns to total words (e.g. Khawaja et al., 2014). Study 3 measures the IKI timing ratio of content or function words to all words.
Turn count ratio	The ratio of subject turn count to partner's turn count, in order to test the importance of coordination between partners (Gravano and Hirschberg, 2009).
Turn-type ratio	Similar to the above, this tests if partners overlap the other with the same frequency, as opposed to how often they let the other partner complete a turn.
Word count ratio	This looks at the ratio of how many words each partner is producing. If rapport is high, word count should be approximately equal (Erkens et al., 2005).

Table 7.2

A list of features used in Study 3

models. The top scoring models, as measured by the range of AUCs, were random forests, boosted trees, and neural networks.

In future studies, more sophisticated modeling techniques should also be considered, including deeper neural networks, LSTMs with recurrence, and Transformer models. These more sophisticated neural networks are especially useful for modeling keystroke patterns on mobile devices (Chang et al., 2021b; Stragapede et al., 2022). However, Study 3 only used a relatively simple multilayer perceptron because it allowed for straightforward comparisons between models and was thus useful for an initial investigation.

For each subset, the tuning set was then used to calculate optimal hyperparameters for each model. In other words, models were retrained for each subset of data. The rationale for this is that the objective of Study 3 was not to measure how well subsets perform on a model tuned from a full dataset, but rather how well each subset performed as a complete dataset, i.e. training and testing.

Because my dataset was small and imbalanced, using traditional partitioning of the data would result in very small sample sizes. For example, if I had used a 75%/25% training/testing split, then the minority class in the testing set would only have 14 instances, which seems insufficient for obtaining reliable prediction results.

In order to circumvent this size issue, I used cross-validation on the entire dataset (minus the tuning set). Specifically, I use Monte Carlo cross-validation, also known as *repeated random sub-sampling validation* (Burman, 1989; Shao, 1993). Unlike *k*-fold cross validation, in this setup the folds are not mutually exclusive, and are randomly resampled with replacement. As a result, the test set would not need to be broken up into *k* groups; rather, overlap is allowed between folds. The disadvantage to this approach, though, is that it leaves open the possibility that a datapoint would never be selected. That being said, I ran 25 repetitions for each model, which should minimize the risk of datapoints being left out.

Model	Dataset	Correct Predictions	Mean Certainty
Neural Net	Receiver	36	0.59
Neural Net	Half 2	32	0.53
XG Boost	Random	31	0.52
Neural Net	Half 1	30	0.52
Neural Net	Full	28	0.52
Neural Net	Random	30	0.51
XG Boost	Half 1	27	0.49
Random Forest	Half 1	27	0.49
Neural Net	Provider	24	0.48
Random Forest	Receiver	26	0.48
Random Forest	Random	26	0.47
XG Boost	Receiver	27	0.47
Random Forest	Half 2	19	0.44
Random Forest	Full	18	0.44
XG Boost	Half 2	22	0.44
Random Forest	Provider	20	0.44
XG Boost	Full	20	0.43
XG Boost	Provider	22	0.42

Table 7.3
This table is made up of each model/dataset combination. Results are arranged by the mean certainty for each combination in predicting the minority class. As can be seen, the neural network trained on the receiver subset had both the highest number of correct minority class predictions, as well as the highest average confidence in predicting the minority class.

7.3 Results

As mentioned earlier, all models were tested on all relevant subsets. However, for those experiments reported here, only the neural network model results are reported. This decision was made because the neural network reported the strongest AUC results on the full dataset. This partially represents the model's high confidence in its predictions. This can be visualized in Figure 7.4 below, where the neural network's predictions have a higher density in the high probability end of the x-axis. A density plot was appropriate because the number of instances (of the minority class) was constant, which makes the densities appropriate for comparison. In addition, Table 7.3 is arranged by the mean confidence in predictions of the minority class. Most of the top spots are occupied by the neural network.

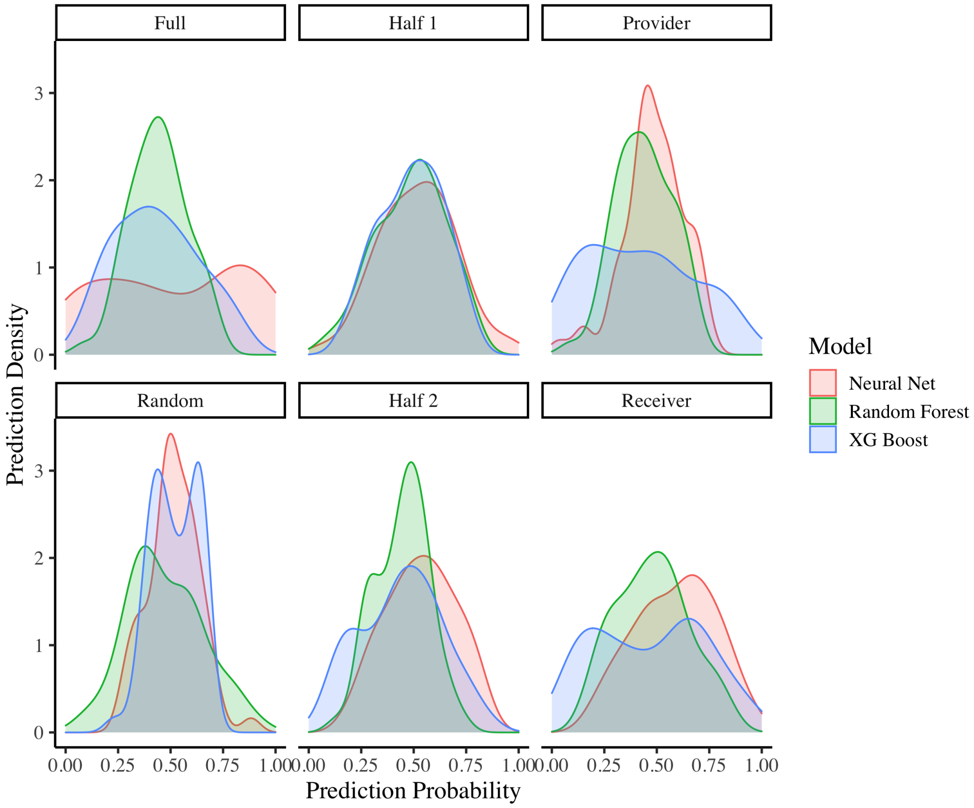


Figure 7.4

A density plot of mean predictive confidence in predicting the minority class. In many instances, the neural network has the greatest density of high-confidence predictions. This is not exclusive to the neural network, though, and some classifiers are more confident on certain subsets.

Dataset	Mean AUC (SD)	Mean MCC (SD)	Mean PPV (SD)	Mean F1 (SD)
Full dataset	0.71 (0.08)	0.27 (0.14)	0.47 (0.11)	0.48 (0.10)
Random subset	0.60 (0.11)	0.07 (0.19)	0.34 (0.11)	0.39 (0.13)
p-value	$< .0001$	$< .0001$	$< .00001$	$< .001$
Effect size (d)	1.18	1.15	1.23	0.81
df	44	44	48	46

Table 7.4
For each dataset, the AUC, MCC, PPV, and F1 score of the neural network are reported. In addition, for the comparison of the two datasets, the p-value, effect size, and degrees of freedom are reported. In this comparison, the differences in each metric score are significant.

For each experiment, the AUC, MCC, PPV and F1 score of the neural network model trained on that subset are reported, and the distributions are visualized. In addition, for each metric a two-sample *t*-test was run on each pair, and those results also reported.

Finally, the most important features for each classifier are visualized, although these are from the boosted tree models. The reasoning behind this was two-fold: (1) single feature importance is difficult to assess in neural networks, and (2) the boosted tree still had very good performance on many datasets.

7.3.1 Experiment 3a: Full dataset vs random subset

The first experiment tested whether rapport ratings could be equally well predicted from the entire dataset as compared to a random subset. This test was performed first so that any significant differences detected in Experiments 3b and 3c could be more appropriately attributed to that specific subset, rather than subsetting in general. The random subset used 50% of each subject's total keystroke output.

As can be seen in Figure 7.5 and Table 7.4, in all cases each metric was significantly better for the full dataset. All effect sizes are considered large (Cohen, 1988).

As can be seen in Table 7.4, the variance is lower for all full datasets across metrics. A possible explanation for these results, then, is that adding more keystrokes reduced uncertainty and led to more confidence in model predictions.

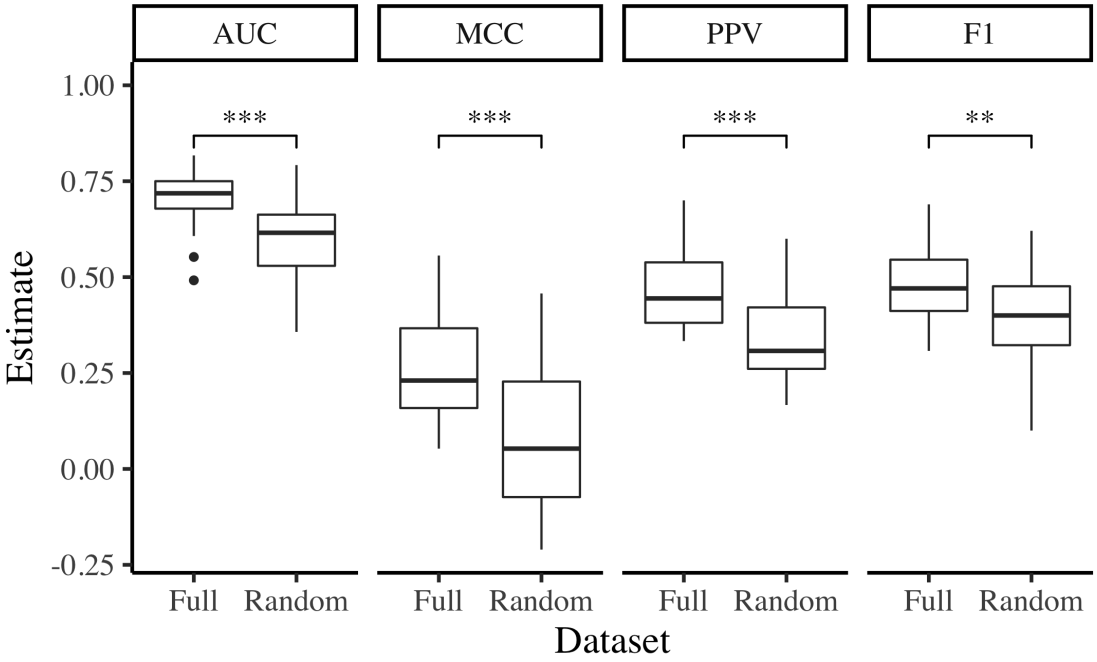
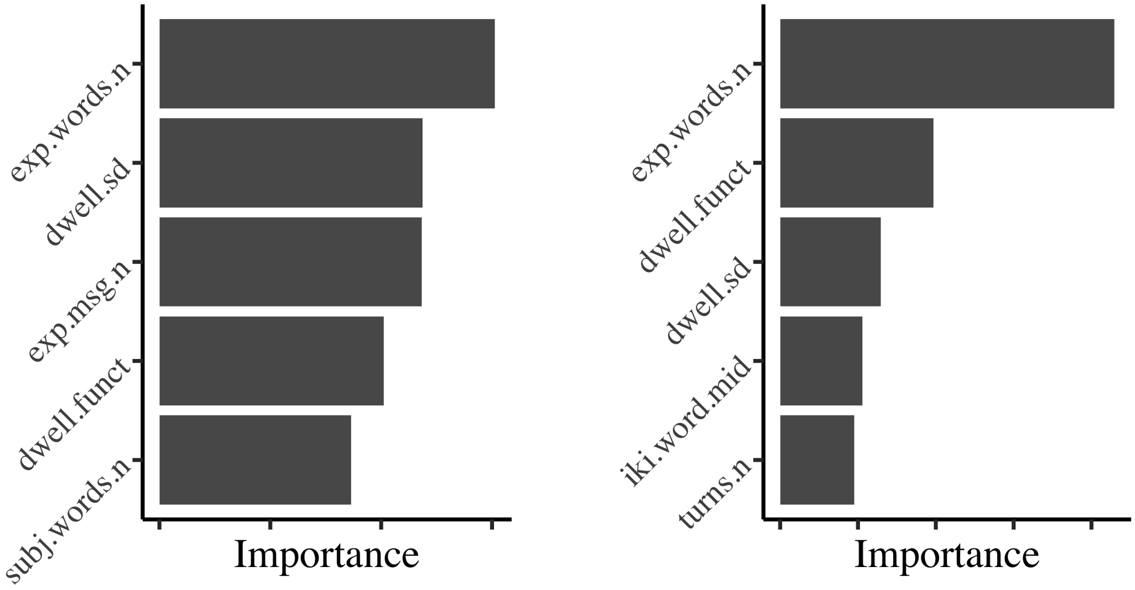


Figure 7.5
The distributions of metric scores for the full dataset versus a random subset. All metrics were significantly higher for the full dataset as compared to the randomly selected subset. It is important to bear in mind that different metrics are calculated on different scales, and so comparing, e.g. AUC to MCC is not meaningful. The only meaningful comparisons are within each metric.

Because this study is inferential, I also wanted to extract *which* features were the most important. The `vip` package in R calculates importance using the Shapley values of each feature, where Shhapley values are similar to coefficient values in linear regression (Shapley, 1953). Crucially, though, because of the opaque nature of neural networks, these charts were made using the optimal boosted tree models. Although the results of the boosted trees were not consistently as accurate as the neural networks, they still scored highly, and so the important features of the boosted trees should not be radically different than the important features in the neural network models.

The five most important features in the boosted trees for each dataset are visualized in Figure 7.6. In both cases, the absolute number of words in the experiment were the most important predictor. However, key dwell times and typing rate (IKI) were also important.



(a) Full dataset

(b) Random subset

Figure 7.6
These figures illustrate variable importance for the full dataset and random subset. Feature importance is based on the boosted tree models rather than the neural network models, due to the opaque nature of neural networks.

Dataset	Mean AUC (SD)	Mean MCC (SD)	Mean PPV (SD)	Mean F1 (SD)
Half 1 subset	0.66 (0.09)	0.21 (0.15)	0.42 (0.11)	0.45 (0.11)
Half 2 subset	0.61 (0.01)	0.12 (0.17)	0.35 (0.09)	0.41 (0.12)
p-value	.09	.055	.01	.23
Effect size (d)	0.48	0.56	0.74	0.34
df	47	47	46	48

Table 7.5
For each dataset, the AUC, MCC, PPV, and F1 score are reported for the neural network models. In addition, for the comparison of the two datasets, the p-value, effect size, and degrees of freedom are reported.

7.3.2 Experiment 3b: First half subset vs second half subset

The next experiment subsetted keystroke data by whether the keystroke was typed in the first 8 minutes of the experiment or the last 8 minutes. The tests in this experiment were aimed at answering whether data from the initial half or latter half of the conversation was a better predictor of whether rapport was classified as high or medium-to-low.

The results of subsetting by conversation half were not as clear as Exp 3a. As can be seen in Table 7.6, most of the differences in metric scores were marginally significant, and all less than 0.10. However, only one metric, PPV, was significant below the 0.05 alpha level. In addition, the effect size of the AUC and F1 comparisons is considered small while the effect size for the MCC and PPV is considered moderate.

Finally, the five most important features for each dataset are visualized in Figure 7.8. In the first half, both keystroke and stylometric features were important, where stylometric features are measures of writing style such as word count. However, in the second half, only stylometric features were important. Although it fell outside the scope of study 3, future work will delve into why typing patterns were less important in the second half, as it may point to the notion that certain features are reflective of rapport only when those features occur within certain temporal slices of an interaction.

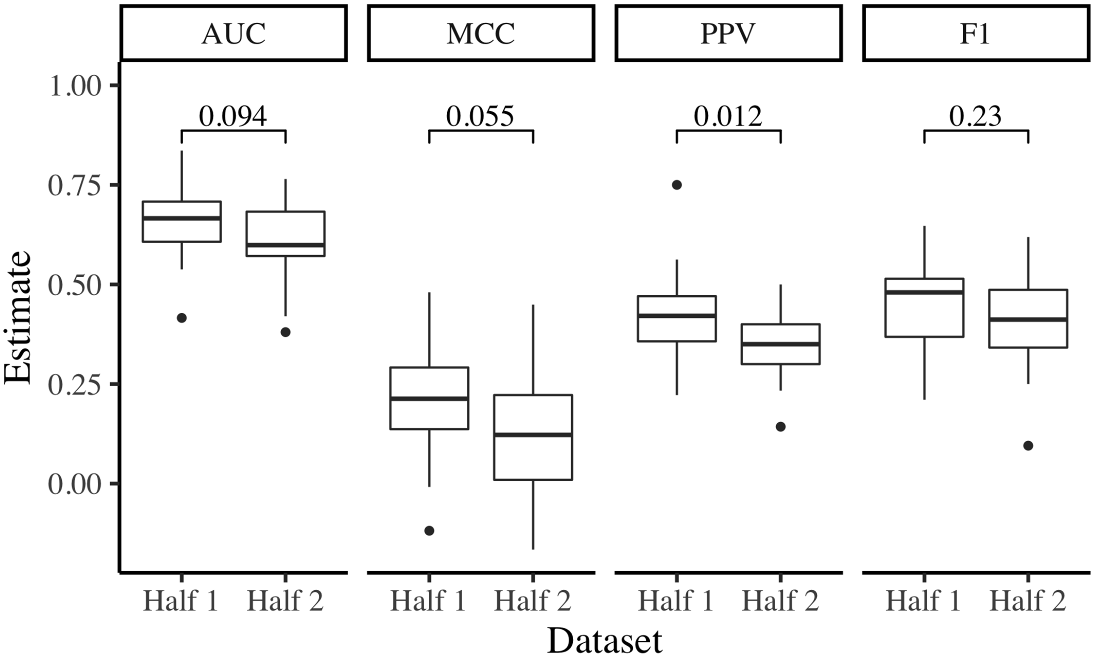
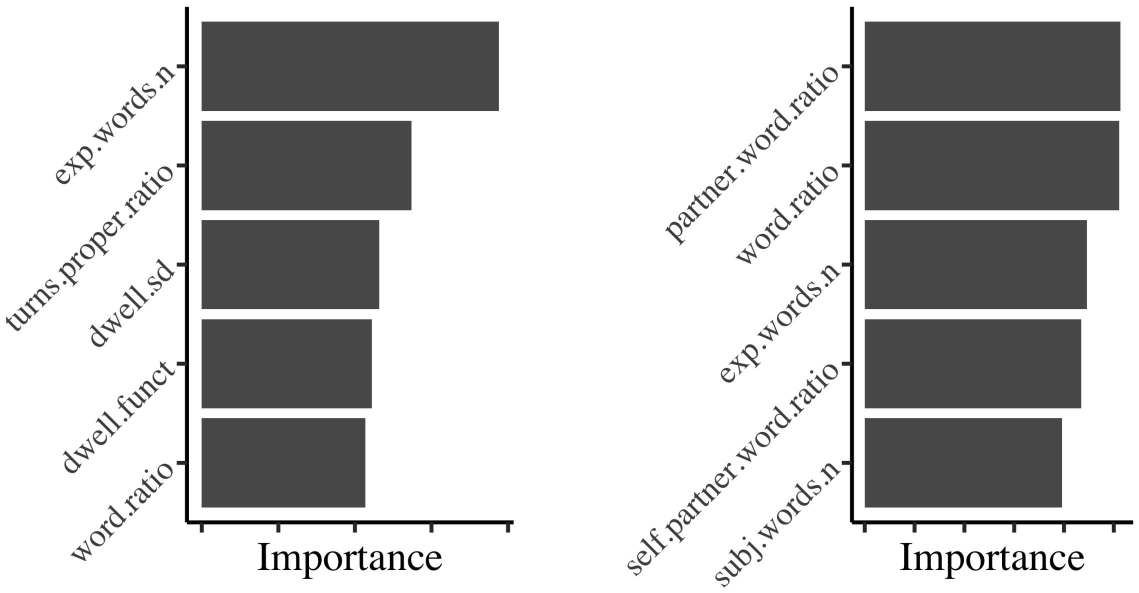


Figure 7.7
The distributions of metric scores for the first half subset versus the second half subset. The metrics are better for the first half subset as compared the second half, but only marginally so. It is important to bear in mind that different metrics are calculated on different scales, and so comparing, e.g. AUC to MCC is not meaningful. The only meaningful comparisons are within each metric.



(a) Half 1 subset

(b) Half 2 subset

Figure 7.8
These figures illustrate variable importance for the first half subset and second half subset. Feature importance is based on the boosted tree models rather than the neural network models, due to the opaque nature of neural networks.

7.3.3 Experiment 3c: Recommendation provider vs recommendation receiver

The final experiment investigated how well typing data from subjects in different "roles" predicted rapport ratings. By roles, I mean when a subject was *providing* movie recommendations compared to when a subject was *receiving* recommendations. This experiment tests whether the task at hand influences typing patterns, specifically how well typing patterns when occupying different roles predict feelings concerning rapport.

The results seem to paint a clear picture that typing patterns when a subject is receiving recommendations is a better predictor of rapport than typing patterns when a subject is providing recommendations. Moreover, the effect size of all of the differences reported above are considered large.

Dataset	Mean AUC (SD)	Mean MCC (SD)	Mean PPV (SD)	Mean F1 (SD)
Provider subset	0.59 (0.07)	0.10 (0.12)	0.35 (0.10)	0.39 (0.11)
Receiver subset	0.73 (0.10)	0.32 (0.16)	0.47 (0.10)	0.53 (0.11)
p-value	<.000001	<.00001	<.0001	<.0001
Effect size (d)	-1.65	-1.53	-1.21	-1.36
df	42	45	48	48

Table 7.6
For each dataset, the AUC, MCC, PPV, and F1 score are reported. In addition, for the comparison of the two datasets, the p-value, effect size, and degrees of freedom are reported.

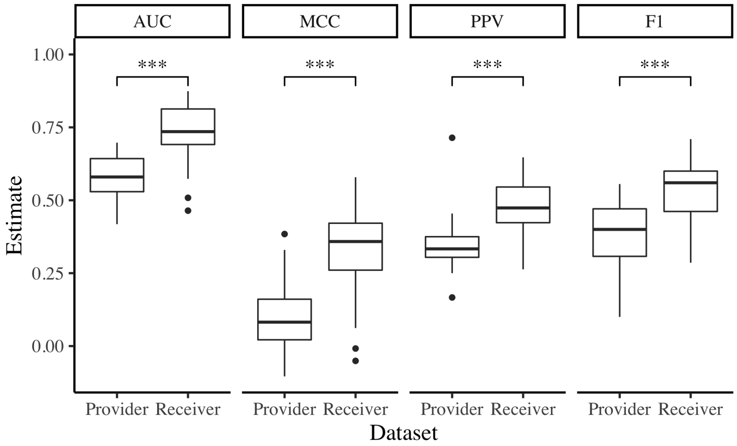
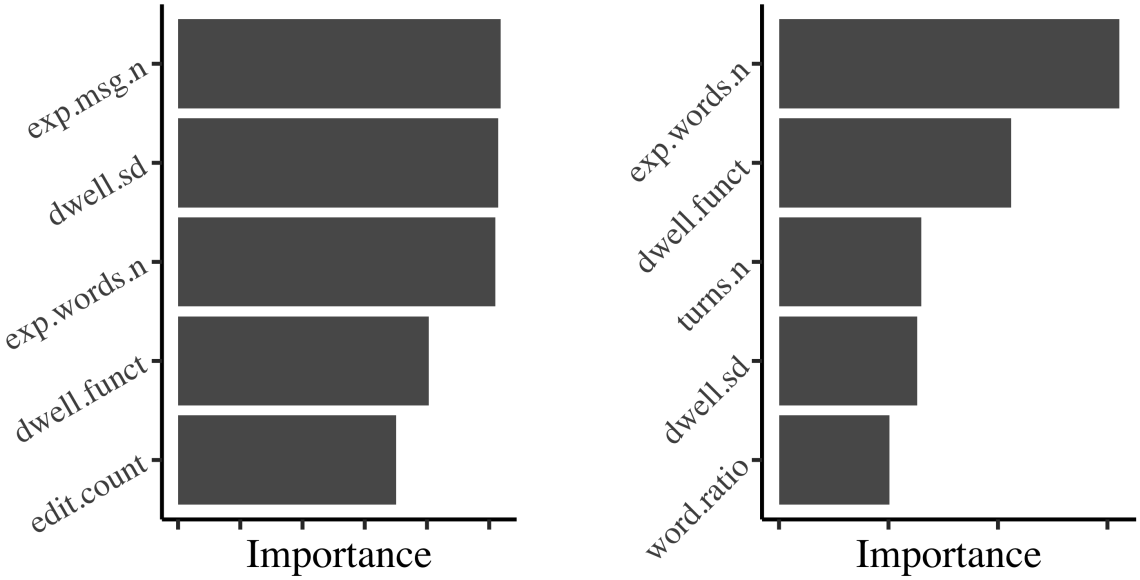


Figure 7.9
The distributions of metric scores for the Provider subset and Receiver subset for the neural network model. The metric scores from the receiver subset were significantly higher than the scores derived from data from the provider subset. It is important to bear in mind that different metrics are calculated on different scales, and so comparing, e.g. AUC to MCC is not meaningful. The only meaningful comparisons are within each metric.

Finally, the five most important features for each dataset are visualized in Figure 7.10. In both cases, the absolute number of words or messages in the experiment were the most important predictor. However, key dwell times were also important in both cases.



(a) Provider subset

(b) Receiver subset

Figure 7.10
These figures illustrate variable importance for the Provider subset and Receiver subset. Feature importance is based on the boosted tree models rather than the neural network models, due to the opaque nature of neural networks.

7.4 Discussion

The comparisons performed in Experiments 3a, 3b and 3c provide an intriguing picture of how accurately certain subsets of a conversation predict overall feelings of rapport during that conversation. Moreover, the different types of important predictors help to fill in this picture.

Looking at the overall findings, it is interesting how well the neural network performed on many of the datasets. One of the strengths of a neural network is its ability to perform feature engineering. However, this strength is sometimes limited only to deeper networks (Seide et al., 2011). Since this study used a multi-layer perceptron, it is possible that this deep feature engineering was a key to the model's success. This seems worth mentioning because it points to notion that perhaps the features constructed for this study (as well as the entire thesis) were too fine-grained or misguided, which is why the network's own constructed features were superior. Although this is a possible advantage of neural networks, one disadvantage is the "black box" nature of neural network predictions. To be more precise, it is very difficult to understand why the model made the predictions it did (Linzen et al., 2018). Nonetheless, methods do exist for understanding predictions, and these should be explored in the future to better understand exactly which dimensions of language production best predict underlying mindsets.

For the boosted tree model, which was fairly strong in its predictive accuracy, it is possible to extract variable importance, as seen in Figures 7.6, 7.8, and 7.10. Of note are two features types: subject/partner word ratios and key dwell times.

Ratios of words, messages, etc. between a subject and partner are interesting because they point to the importance of coordination between partners (Scissors et al., 2009), which is an *extrinsic* marker. On the other hand, a partner cannot see a subject's individual keystrokes, and so these features are considered intrinsic and not communicated.

If more coordination is a sign of higher rapport between partners, then it stands to reason that ratios between a subject's language production and their partner's production should be closer to equal, or 1:1. Indeed, a quick analysis of word ratios and message ratios both found that for cluster 2, the group with higher rapport ratings, these ratios were much closer to 1.0. For word-level ratios, this difference was found to be significant, while it was not statistically significant for message ratios.

A possible reason that word ratios were more important than message ratios was the nature of the experimental task: subjects were encouraged to engage in a "text message"-like conversation; similar to Ling and Baron (2007), this could result in strong reciprocity between a message and a response, regardless of the content of those messages. However, in high rapport conversation, the content or at least length of the messages would be more equal, which would result in the statistically significant differences in the ratio of word counts but not message counts.

The other interesting feature that plays an important predictive role in every dataset, and is the most important keystroke-based feature, is dwell time, or how long a key is held down for. As observed in studies such as Lee et al. (2014) and Lee et al. (2015b), dwell time is more strongly associated with emotional responses. It would then stand to reason that rapport would be more closely connected to dwell time than, e.g., typing speed. Moreover, the most important sub-feature of dwell times is often not the overall timing, but rather the timing associated with a semantic subset of words, such as function words or content words (Chung and Pennebaker, 2014; Pennebaker, 2011). Since the ultimate goal of my thesis research is to use keystroke timing to infer the way a human feels when talking to another human or computer agent, it is essential that a keystroke-based system also knows where to look for keystroke information.

The fact that these two types of features were both important also highlights an important theoretical dispute in conversational coordination: Is the phenomenon of coordination *mediated* by the environment or *unmediated* and the result of the user's own mental state? Another way to view this is through the lens of audience design, and whether coordination is intended for the audience, or is a natural response.

Like many previous studies, Branigan et al. (e.g. 2011); Garrod and Pickering (e.g. 2009), the fact that both intrinsic typing patterns and extrinsic word transmission are important points to a synthesis of these two theories. This has wide-ranging implications for the future design of HCI systems such as chatbots. On the one hand, a chatbot will need to use a measure of reflection in order to encourage a feeling of connection with a human user. On the other hand, a system should also monitor the cognitive effort, e.g. keystroke patterns, behind the coordinated word production of the user, to better understand the user's mindset and whether a feeling of connectedness exists.

Experiment 3a sought to answer **RQ 3a** and **RQ 3b**. **RQ 3a** was concerned with classifying rapport level over an entire conversation, and it appears that the distinction between high and lower rapport can be predicted. We can see this in Table 7.4, where the full dataset had a mean AUC of .71, which is above chance, and considered "acceptable discrimination" (Hosmer Jr et al., 2013, p.

177). On the other had, regarding **RQ 3b**, it appears that a random subset cannot predict rapport as well as a full dataset, with the AUC being considered "poor discrimination."

However, it is possible that this was due to the method of random sampling. For example, when subsetting a contiguous chunk such as the first half, in a 5-letter word, there would be 3 mid-word keystrokes, 1 word-initial keystroke, and 1 word-final keystroke. In the random sampling, hypothetically every keystroke could have been a mid-word keystroke. This speaks to the importance, in future work, of using a repeated random subsampling, so as to balance out these types of disparities, or only sampling full words and sentences.

The fact that the random subset was significantly worse was surprising because studies such as Pecune et al. (2018) or Olsen and Finkelstein (2017) used external annotators to make rapport judgments from very brief slices of a conversation, and found that these judgments were highly accurate. However, the fact that the judgments in other studies were accurate while the random subset in this study was significantly inferior, possibly points to the need to use a contiguous subset, rather than a random subset. As Zhao et al. (2014) points out, rapport is a dyadic phenomenon, co-constructed over time by both members of the dyad. It is possible, then, that tracking this continuous development is also important for rapport judgment.

That being said, the medium and methodology of my thesis compared to prior studies is vastly different. Moreover, there exists serious methodological issues with my random subset; specifically, the random subset was just *a* random subset. Because randomization was used, the random subset should have been resampled a number of times so that the full dataset could be compared to the randomization process in general. I will discuss this further in the Future Work section.

Experiment 3b provided answers regarding **RQ 3c**, the comparison of data from the first half of the conversation to the second half, which is also intriguing. Although not all of the differences in metrics were significant, they were all at least approaching significance ($p < .10$). In answer to **RQ 3c**, then, it appears that keystroke data from the first half of an experiments does predict final rapport as well as keystroke data from the latter half of a conversation.

However, the comparison of PPV, which specifically measured how well a classifier predicts low rapport (the "positive" case), was significant ($p = .01$). This finding aligns well with previous research on trust development in HCI. Tolmeijer et al. (2021) performed a longitudinal study of rapport and found that while first impressions are strongly influential on overall impressions, sometimes a negative first impression can be slightly improved with subsequent interaction. It seems that the results of Experiment 3b corroborate a very similar conclusion: data from the first half (first impressions) is significantly more accurate at predicting a poor final relationship. However, because not all metrics were significant, it could be interpreted as the second half data (subsequent dialogue) slightly improving a final impression.

In designing an intelligent computer agent, evidence from Experiment 3b would imply that the initial parts of a dialogue should be more heavily weighted and closely monitored, but that subsequent dialogue should not be disregarded.

Finally, Experiment 3c answers **RQ 3d** by shedding light on the importance of different tasks or roles in relationship building. The findings show that typing data from when a subject who is receiving recommendations is significantly better at predicting rapport than typing data from when a subject is providing recommendations. In fact,the AUC of 0.73 is higher than the AUC of the full dataset.

The fact that the receiver subset is more accurate than the provider subset seems to intuitively make sense, as a recipient would judge the quality of recommendations they are receiving, and use these judgments to form an impression of their partner (the provider). On the other hand, a provider is primarily only outputting recommendations, and so the provider has very little feedback information on which to judge their partner and form an impression.

It is possible that these findings follow from the theory put forth by Clark and Schaefer (1989) that the establishment of common ground requires both a *presentation* phase and *acceptance* phase. Gergle (2017) enumerates the varied ways that acceptance can be signaled, beyond verbal feedback. As such, perhaps typing patterns are also an internal signal of acceptance, and are therefore more strongly connected to a sense of rapport in the conversation.

For the larger aim of my thesis research, it is also beneficial that recipient information is more informative regarding rapport. In recommender systems or customer service chatbots, the user is the recipient of the information. For example, in a digital mental health app, a designer would want a computer agent to provide counseling to a user of the app. Since recipient typing data seems more informative about rapport, a system could better rely on the typing patterns being produced by the user (receiver) in order to assess how the recommendations are being received, and the level of rapport between the user and computer agent.

7.5 Future work

In future iterations of these experiments, a few changes should be made to both data collection methods and data processing methods.

The first change that should be made in future work is a new experimental setup that generates a greater proportion of low rapport levels. As mentioned repeatedly throughout this study, a limitation on classifier performance was the imbalanced dataset (although methods such as SMOTE and metrics such as PPV were used to partially circumvent this). Nonetheless, having more examples of low rapport interactions will improve classifier performance. This improvement is critical if future dialogue systems aim to detect low rapport in its many guises.

Regarding overall prediction methodology, this study used classification, where subjects were divided into high rapport and low rapport classes. However, rapport is not binary; rather, it exists on a continuous scale. For this reason, a regression analysis would also be very appropriate. Moreover, when I was analyzing the 6-dimensional vector for each subject, based on their answers to the 6 survey questions, I also ran a Kaiser analysis to determine the “true” number of factors/dimensions in the 6-dimensional vectors (Auerswald and Moshagen, 2019). Similar to Liebman and Gergle (2016b), I found that because the answers to each question were highly correlated, there is really only a single factor in the answers. The upshot of this is that it would be appropriate to reduce the 6 answers to a single mean, and use that number as the response variable in a regression analysis.

This will be done in a future analysis; for my dissertation I chose to begin the process using less granular clustering.

Regarding specific analyses, it would be helpful to also test two changes. The first change regards the random subset. In these experiments, I only used a single random subset, due to methodological constraints. As a result, my conclusions about the random subset can really only be extended to that specific random subset, rather than random subsetting in general. In future iterations, I would resample a random subset, and test all of these samples against the full dataset.

To extend random subsetting, it would also be helpful to determine what proportion of the full dataset can be randomly selected so as to achieve comparable results. Study 3 used 50% of the data, but it is possible that subsetting only 25%, for example, would achieve similar results while lowering the amount of data that needs to be extracted. Collecting less data would also help to improve the anonymity of the keystroke data, providing greater privacy while still achieving comparable performance (see Manandhar et al. (2019) for an example of a continuous but anonymous authentication system using keystrokes).

Regarding input features, this study only looked at overall ratios, such as word ratios between what a subject produced and their partner produced. However, studies such as Lubold and Pon-Barry (2014) also looked at turn-by-turn ratios, to measure coordination between interlocutors, where coordination is usually a sign of connectedness. This is important if the goal of a future system is to continuously monitor rapport, since a full dataset would not be available in the middle of a conversation.

Moreover, as mentioned multiple times, rapport is a dyadic feature that emerges from an interaction. Therefore, it would be helpful, especially for human-to-human dialogue management, to also take into account the mindset of the partner, based both on final questionnaire data and the partner's typing patterns.

Chapter 8

Overall Discussion and Future Work

In the three individual studies in my thesis, I demonstrated that keystroke patterns are associated with different underlying intentions in a conversation, where those intentions may not be evident from visible word choice alone.

Study 1, by better identifying dialogue acts, lets us better understand what part of a conversation a user considers to be agreed upon as common ground, and which direction the user wants the dialogue to proceed in: a user might have a question about previous material. which means that they do accept the previous context as part of the common ground, or they may want to advance the conversation forward because both participants agree on what's been said (e.g. Convertino et al., 2008, but see their further clarification).

Study 2 uses keystrokes to better detect sentiment in utterances. While sophisticated methods exist that measure sentiment based on word choice alone, I show that adding keystroke information can improve these results. Because I am studying an interaction rather than a monologue, sentiment is also sensitive to changes from turn to turn. I show that keystrokes are also sensitive to these changes in sentiment, in addition to the sentiment of the utterance on its own. Another unique element of a dialogue is that a user forms an overall opinion of their partner. I show that this overall opinion also influences typing patterns, independently of the sentiment of that specific utterance.

Finally, Study 3 looks at how well keystrokes can predict the level of rapport established between a user and their partner. Further, study 3 asks which subsets of conversational data best predict rapport? I divided conversations chronologically into a first half and second half, divided a conversation by the task being accomplished, and randomly subset the conversational data. Using the findings from this study, rapport can be measured *during* a conversation, not just at the end.

The overall findings have not only theoretical implications for computer-mediated communication, but also practical implications for designing future CMC-related interfaces. If underlying intentions can be better identified, then a CMC system can utilize this information and either communicate it visually to a human partner or augment existing lexical information that a computer agent processes when conversing with a user. While the studies within my thesis were not designed to test this application directly, the research presented here is intended to act as a foundation for implementing these improvements to a CMC system.

CMC Theory

A main contribution of my thesis is the support it provides for a channel expansion theory (Carlson and Zmud, 1999), since the high levels of rapport achieved between anonymous users indicates that they were engaged in relatively rich conversation. Traditional theories of CMC such as media richness or bandwidth-limited theories posited that the capabilities of a medium are single-dimensional and determined *a priori*. A channel expansion theory, though, argues that as individuals gain more experience with a particular communication medium, the medium becomes richer for them (Carlson and Zmud, 1994). In other words, a medium's richness is perceptually determined by its users rather than being an intrinsic property of the medium itself. Walther (2011) situates this by saying that with experience, users learn how to encode and decode affective messages using a particular channel. In other words, the nature of media and their potentials are socially constructed (Fulk et al., 1987).

Kalman et al. (2013b) also points out that media richness theory does not explore the influence of the chronemic variables that are transmitted in lean media. In other words, media richness theory usually focuses on only the text of a text message and disregards elements such as the time between messages. But studies like my thesis show that a large amount of information is embedded in *how* language is produced, not just *what* language is produced. When this is not taken into consideration in analyzing a medium of communication, that analysis is necessarily incomplete and not an apt comparison between mediums.

In order to see how my experiments demonstrate this, it is instructive that Walther (2011) calls CMC users "cognitive and behavioral misers," who prefer to do a task using less effort than using more effort. As compared to face-to-face (FtF) communication, CMC is considered more effortful. If CMC is more effortful, though, then the overwhelmingly high rapport reported in study 3 shows that users are willing to incur greater effort, perhaps by making more use of the affordances of the experimental CMC platform (Clark and Brennan, 1991), in order to achieve the goal of the experiment, i.e. communicating recommendations and their rationales.

The high rapport ratings also support the filtered cues theory put forth in Walther and Parks (2002). In their critique of bandwidth-mediated theories, the researchers point out a confound in many previous studies that found FtF communication to be more expressive than CMC. While it has been well-established that CMC-based relationships take longer to develop than relationships built from in-person interactions, many previous studies used the same time limits for both mediums. For this reason I gave my participants 16 minutes to discuss a relatively specific topic (as opposed to completely spontaneous conversation or the more free-wheeling conversations in the Switchboard corpus (Godfrey et al., 1992)). It seems that this was an adequate amount of time to establish a high level of rapport, which helps to quantify the long*er* timeframes for relationship-building in CMC.

One question that my thesis brings up is whether certain theories of CMC are now outdated, or at least need heavy revision. At least in an environment such as the Prolific experimental platform, many participants have a very different relationship with CMC. Peer et al. (2022) reports that on some online behavioral research platforms, a majority of users use the platform as their primary source of income and spend the majority of their day on the platform. These users have a very different relationship with technology in general and CMC specifically. The computer-mediated environment, for these users, is not one option among many, but is the sole option. In the Electronic Propiniquity Theory, the number of options matters, and users with only one option felt closer to the person they were communicating with (Walther, 2011). This could also explain the abundance of high rapport ratings among users in my experiments.

Nevertheless, as Walther (2011) reasons, “We can’t keep up with new innovations, so we need theory and models that can (p. 754 in Scott 2009)." My experiments seem to add empirical evidence for existing theories of CMC that show that users are willing and able to expand the CMC medium to work for the purposes of their communication needs. However, it is possible that the shifting and more ubiquitous role of CMC will necessitate revisions to existing theories, as well.

Affective computing

My thesis adds to the rapidly growing field of *affective computing* (AC). This area of research recognizes that a computer agent needs to understand not only the (linguistic) content that a human is producing, but also recognize the emotions behind those words (Picard, 2000). Many methods for this recognition are either expensive or intrusive, such as galvanic skin responses, facial recognition, or voice analysis (Katerina and Nicolaos, 2018; Nahin et al., 2014, inter alia). Keystroke pattern analysis provides an unobtrusive and low-cost methodology for detecting user emotions.

Importantly for the many applications of my research, keystroke dynamics can capture multiple dimensions of emotion. As I demonstrated in Study 2, and has also been detected in prior research such as López-Carral et al. (2019), keystrokes can capture not only whether a user is experiencing a positive or negative emotion, but also the level of arousal, i.e. whether a user is mildly happy or ecstatic. This is one feature that sets keystroke analysis apart from a biometric such as facial expression recognition, where the latter often requires classification between a set of discrete basic emotions, rather than degrees of an emotion (Fasel and Luettin, 2003). The ability to identify emotion on a continuous scale is equally, if not more important, than simply identifying *what* emotion a user is displaying.

Further, as shown in Study 3, keystrokes can also capture complex emotions such as rapport. Rapport is made up of many fundamental emotions, and so trying to directly ask about it or creating a single measure is difficult and less effective. Similar to Raj Prabhu et al. (2020), which measured multi-dimensional "conversation quality," I asked multiple questions after a conversation to construct a nuanced measure of rapport. Keystrokes were shown to be sensitive to this nuanced measure, demonstrating that keystrokes could also be used to detect complex, multi-dimensional emotions.

Ethical implications

Before discussing any practical applications of my research, it is imperative to discuss its ethical implications. Most crucially, keystrokes are a biometric, like a heart rate, a fingerprint or a voice pattern. Just as these are considered private or personal, keystrokes are also considered private or personal. Because of this my experiments required full final consent from participants, so that we had permission to collect and study their keystroke patterns. In addition, I received approval from Northwestern's IRB.

In addition to being a biometric for personal identification, keystroke patterns can also be used to identify demographics such as gender or education level (Brizan et al., 2015, inter alia). However most major internet browsers make it easy for developers to collect keystrokes (Acien et al., 2021). Thus, it is important to consider ethical implications when collecting keystroke information, as this information is both private and highly informative.

Any practical application discussed below could also be extended to more invasive and privacy-violating methods. The notion of identifying underlying intentions is very close to the idea of "Thought Police" (Orwell, 1949). As a recent example of the misuse of biometrics, many law enforcement agencies are using vocal analysis of 911 calls to determine if the caller is the actual perpetrator of the crime, despite the fact that this method has been debunked and multiple convictions have been overturned (Murphy, 2022). Along these lines, it's not difficult to imagine a scenario akin to the movie *Minority Report* where law enforcement officers arrest people who they have determined *intend* to commit a crime, but before any crime has been committed. The "intentions" I identify in my thesis are somewhat ambiguous, and could be interpreted in many ways.

A dystopian use of keystroke analysis is already in place within the domain of employee surveillance. As more jobs become remote it becomes more difficult for employers to monitor employees. One method becoming increasingly popular for productivity surveillance is keystroke monitoring (Indiparambil, 2019; Tham and Holland, 2022). If an employee stops typing or is typing on a social media website, then the employer is notified. This has even been extended to using keystroke analysis to monitor employees' levels of motivation, and notify employers when employees are not motivated and likely not outputting high-quality work (Ball et al., 2021).

This also ties into the observation by some scholars that users sometimes *choose* to use text-based communication, even when other modalities are available, because they do not want their true intentions or mindset to be fully evident in a conversation. For example, Scissors and Gergle (2013) and Scissors et al. (2014) found that romantic couples would switch communication modalities to deescalate heightened emotions and avoid conflict. Similar, these studies found that users with low self-esteem find text-based CMC appealing because it mediates interactions, by lowering "face threat" and instead encouraging more "distancing" behavior.

Given findings such as those above, it is important to take into consideration that a user may not always *want* their complete mindset to be communicated via text-based CMC, e.g. by creating a visual representation of the emotions conveyed by typing patterns. Because of this, in any application of my findings in the real world, users would need the option to disable a visualization or anything in addition to the text appearing on the screen.

A more ethical approach to using keystrokes to monitor employee motivation has recently been undertaken at Microsoft Howe et al. (2022). Rather than reporting low motivation to a manager, a system that senses low motivation or stress suggests to that employee that they should take a break and perhaps perform a stress-reduction exercise. In this way, mental states are not reported to a supervisor at all.

One advantage that keystroke analysis has above other forms of monitoring is that it can allow for anonymity while still extracting useful information about typing patterns from subsets of keystrokes. Monaco and Tappert (2016) presents a method for systematically obfuscating keystrokes so that a typist cannot be identified or impersonated. However, the anonymization does not negatively impact the typing experience: certain traits called "soft biometrics" can still be extracted from typing patterns, without even knowing the lexical content, and the keystroke obfuscation process was not noticeable or distracting to the user. The findings in Study 3 of my thesis also provide hope on this front since the findings demonstrate that rapport ratings can be extracted just using a subset of data rather than a full typing session. This could help to protect a user from sharing enough information that they can be personally identified.

Nonetheless, despite these advancements, privacy invasion is a very real concern when using keystroke analysis. This must be kept in mind when researchers or engineers use keystroke patterns to better understand a user.

Practical applications

The research in my thesis can be applied to any domain that utilizes text-based interaction. These include areas such as text-based telehealth, remote work on a Slack-like interface, or online chats with customer service. The common denominator in all of these scenarios is that it is critically important for one party to understand the mindset of the other party. Branigan et al. (2011) shows that we always try to infer the mindset in a conversation, whether a speaker believes they are interacting with another human or with a computer agent. However, inferring this mindset, or building a mental model of a partner, is difficult regardless of whether the partner is a human or computer agent (Gero et al., 2020; Yan et al., 2020).

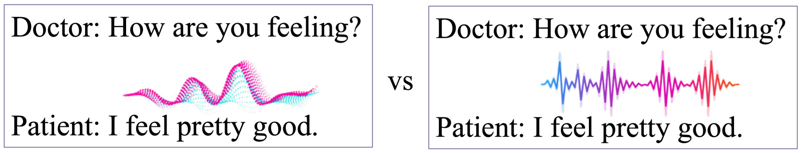

Doctor: How are you feeling?
Patient: I feel pretty good.
vs
Doctor: How are you feeling?
Patient: I feel pretty good.

Figure 8.1
A toy example of different typing patterns visualized with the same lexical content

Typing data can tell us not just *what* is said, but *how* it's said. The style of delivery can be just as important as lexical content when trying to understand underlying mindsets of a language producer (Cowan et al., 2015). For instance, it has been well established that lexical alignment between speakers is a sign of positive affect between the partners. However, typing patterns are partially reflective of cognitive load (Brizan et al., 2015), and so typing patterns could inform observers about the level of cognitive effort underlying lexical alignment. If cognitive effort is significantly different across instances of lexical alignment, then perhaps not all lexical alignment signifies the same relationship quality between partners.

Thus, as an immediate extension of my thesis research, we can use important typing features to create a visual representation of the typing style behind the words transmitted, so as to inform a partner about the mindset behind the words. As an example, Figure 8.1 shows a fictional interaction between a doctor and patient. In this scenario, the colored waves accompanying the text are a visual representation of the patient's typing patterns, where the smooth wave on the left represents a consistent or consistent typing style, while the jagged wave on the right represents an inconsistent, hesitant typing style. Although the lexical output is identical, it is clear that the typing patterns that produced them are very different. As a result, a doctor should interpret "pretty good" very differently in each case.

In reality, though, a visual encoding schema such as that in Figure 8.1 would need significant refinement before being put into production. While some visual encodings might inherently convey meaning, these can be difficult to consistently produce (Iliinsky and Steele, 2011). More likely, users of an application that visualizes typing patterns would need to be explicitly instructed as to what each visualization means, as well as what that typing pattern means about the typist's mindset.

A CMC system that is aware of keystroke patterns could also assist human-to-human CMC by mediating a conversation. For example, Gergle et al. (2004) shows how the amount of chat history displayed affects the collaborative success of an interaction. If a computer agent moderating a conversation used typing patterns to informs its decisions, the computer agent could dynamically decide how much chat history to display in order to improve collaboration. If the computer agent determines via typing patterns that an utterance was produced under duress or during a moment of anger, than it may be conducive to collaboration if this utterance is quickly removed from the visible chat history. Conversely, if a backward-facing dialogue act was produced, asking for additional clarification, then a computer moderator could keep that utterance in the conversation's visible history for longer. Since study 2 of my thesis showed different typing patterns in forward- vs backward-facing dialogue acts, this is an area where typing information could be utilized.

Pommeranz et al. (2012) shows that people are willing to spend more effort if the feedback mechanism enables them to be more expressive. A feedback mechanism that takes into account production patterns, such as typing patterns signaling sentiment or opinions, would allow a user to be more expressive. Thus they may put more effort into their interaction. If this took place, it would also lend support to a channel expansion theory (Walther, 2011; Walther and Parks, 2002).

Implicit Prosody Hypothesis

While my thesis is not intended to prove or disprove any specific cognitive theories, it does use cognitive science as a basis for its investigations. Specifically, the features used in all of my studies are based on features of spoken prosody, in order to test the Implicit Prosody Hypothesis (IPH Fodor, 2002a).

Spoken prosody has proved to be useful in many HCI settings such as trust development (Beňuš et al., 2018), communication efficiency and emotional engagement (Suzuki and Katagiri, 2007), naturalness in an interaction (Bell et al., 2003), and mental models of a computer conversational partner (Cowan et al., 2015). Spoken prosody has also been shown to be useful in DA classification, sentiment analysis, and rapport detection (see previous Related Work chapters).

Prior studies have found pause locations in keystrokes to be similar to pause locations in spoken language. Plank (2016) and Goodkind and Rosenberg (2015) found that typists pause for longer at the boundaries of syntactic units, which is also a feature of speaking. In their research, Plank (2016) was actually able to use keystroke timing as a crude syntactic parser.

In my own studies, keystroke features based on prosodic features improved the accuracy of identifying underlying motivations and mindsets of participants engaging in dialogue, which are also commonly signaled in spoken interaction using speech prosody. While this does not constitute incontrovertible proof that a user is utilizing the exact same prosodic cognitive pathways when typing as they are when speaking, it does provide a possible association and an intriguing foundation for future research.

Since spoken prosody is at least partially determined by a speaker's relationship with the audience they are addressing (audience design, Horton, 2017), typing patterns could also be informative about a typist's relationship with their audience. Future research should look for more direct evidence of parallels, which would open the door for using better-studied features of speech prosody to be co-opted in keystroke dynamics.

Typing metrics as an independent and dependent variable

In my respective experiments, I used keystrokes both as a response variable and a predictor. As I explain below, this speaks to the distinctiveness of keystroke patterns in different situation, as well as their consistency within a situation. Using the various definitions of a "non-verbal cue" provided in Kalman (2007), it seems that the consistently different properties of keystrokes that I have identified would make them a proper cue.

In Study 1, Exp 1a treated a dialogue act binary as the dependent variable with a set of keystroke metrics as independent predictors, but Exp 1b treated a single keystroke metric as a dependent variable with dialogue acts as a single independent predictor with multiple levels. In Study 2, Exp 2a treated a sentiment level binary as the dependent variable with a set of keystroke metrics as independent predictors, but Exp 2b treated a single keystroke metric as a dependent variable with sentiment level and opinion ratings, respectively, as independent predictors. Given this, a natural question to ask is why keystrokes are both predictors as well as response variables.

In Studies 1 and 2, when a set of keystroke metrics were independent predictors, the question at hand was whether keystrokes, collectively, could be used to differentiate between a binary distinction. The question was not whether a single typing pattern could be used as a differentiator, as it has been well-established that a mix of keystroke metrics provide for a better overall model. In these experiments, the set of keystroke metrics had to produce unique values for each level of the dependent variable.

On the other hand, when a single keystroke metric was used as a response variable, the question at hand was whether each level of the predictor(s) had a robust timing signature for that keystroke metric or resulted in a consistent change. The signature at each level did not need to be unique, but rather needed to be consistent.

Limitations and shortcomings

The nature of my study design, data collection methods, and recruited population imposed limitations on how generalizable the results reported in my thesis will be. This is not to say that my study was flawed; rather, every study sets boundaries for specificity and these impose limitations on the scope of the findings.

The limitations imposed by the study design can be broken into three facets: 1) the experimental interface, 2) the experimental prompts, and 3) the recruited population.

As seen in Figure 4.7, the experimental interface used a very generic aesthetic, which looked very dissimilar to a modern chat client such as Slack or Google Messenger. Studies such as Branigan et al. (2011) have shown that humans interact differently with a computer agent depending on whether the aesthetics of the interface make the agent look sophisticated. It is possible that this extends to human-human conversations. Since the interface did not look anything like a chat interface on a participant's phone or computer, the participant would be more consciously aware of the interface and how different it looked from the interfaces they are used to. As such, the typing patterns collected would not be as naturalistic as the typing patterns collected when in Slack or on an iPhone.

In addition, I disabled autocorrect, autocompletion, use of emojis, and the ability to change aspects of the font. Users of modern chat interfaced are likely very used to autocorrect and autocompletion, and so imposing the necessity of typing out an entire word and being more self-conscious of correct spelling also made for a less naturalistic experience. In addition, studies such as Liebman and Gergle (2016b) show that emoticon use is used in text-based communication to mediate interpersonal affinity. By depriving my subjects of the ability to use emojis or convert emoticons to emojis, I also deprived them of a tool used to convey emotion in a text-based environment.

My experimental prompts also provide limitations for my collected data. My goal was to strike a balance between a tightly-controlled but less naturalistic conversation and completely free-wheeling conversations that were different lengths and contained completely different content. In order to achieve this goal, my prompts included the role that each participant would play ("receiver" versus "provider" of recommendations), the general genre of movies/TV shows to discuss, and some pointers about what the conversation should look like, i.e. short messages akin to a text message conversation with a friend.

Very few real-world interactions would have all of the constraints I imposed. Certainly, a typing analysis used in a real-life setting would need to be able to accommodate very different content, and messages of very different lengths. In addition, a real-life scenario would involve implicit role assignment, but these assignments would not be explicitly delineated. For example, if a user logs onto a telehealth platform for medical advice, they would implicitly take on the role of recipient. However, during a conversation these roles will naturally fluctuate, e.g. the patient *providing* medical history to their doctor. As a result, while Study 3 of my thesis found a high degree of accuracy using the Receiver subset, in the real world it would not be trivial to partition this data.

An additional limitation was the recruited population. Because I recruited exclusively from a crowdsourcing platform, where users were required to have a fair amount of experience with online experiments, it is safe to assume that the participants had familiarity with computer interfaces, and typed more fluidly on average. However, an application deployed in the real world will have uses of varying typing competence. As such, it would need to accommodate a naturally slow or inaccurate typist, but understand that this is not necessarily indicative of high cognitive load or hesitation.

Similarly, one data collection methodological issue was that participants used different types of computers alongside different internet connection speeds. If a typing application was deployed in the wild, the developers would need to account for different keystroke latencies across devices/connections. Whether these timing differences are perceptible by humans should be studied in the future. However, if fine-grained keystroke timing differences are necessary for an analysis, then this needs to be taken into account.

A major experiment methodological limitation was that the conversations in my study lacked a tangible goal. While exchanging recommendations was a "goal," there was no objective measure of success. For example, in Kalman et al. (2010) participants played a trading game where success depended on collaboration and the outcome of the game partially determined compensation. Because of this, participants were motivated to collaborate, and the number of successful trades was used to operationalize collaborative success. In my studies, there was no way to measure a successful outcome and participants had no enticements to collaborate.

Despite this, while it is not an objective measure, the high rapport ratings in study 3 show that participants were willing to invest effort into making high-quality movie and TV show recommendations. This occurred despite that my experiment did not involve additional incentives for good recommendations. This speaks to features of CMC such as *fluidity*, where ease of communication induces better communication; Faraj et al. (2011) cites as a reason for free labor in online communities such as Wikipedia (Rafaeli and Ariel, 2008). Because a text-based chat interface provides few barriers to knowledge sharing, participants are more willing to contribute effort to providing high-quality recommendations.

In addition, my experimental methodology was perhaps too naturalistic and not controlled enough to get stronger experimental results and establish a more firm foundation. As keystroke analysis of dialogue is a relatively unstudied area, it is possible that a semi-spontaneous conversation did not provide enough controls for a foundational study. That being said, the full dataset, called the Keystrokes in Dialogue (KiD) Corpus, is available at https://github.com/angoodkind/KiDcorpus. This data should be of great value to the larger HCI community and allow for the foundational studies in my thesis to be expanded upon.

Conclusion

Despite the shortcomings of the work in my thesis, the research does provide important contributions to the fields of HCI and cognitive science. Typing, especially in interactions, is a fascinating modality to explore because it combines together elements of language production, language comprehension, and a dyadic relationship. For example, when a message is being sent its composition is only viewable by the producer, but its final product is shared with a partner. Thus, typing production in text-based messages allows us to understand an interaction both as an internal process as well as a shared process. Findings from my thesis shed light on theories of why we communicate in the way we do, and how this effect is received by those we interact with.

Finally, the vast majority of typing studies have been performed on typing in isolation. My thesis research studies typing, though, as an interactive process, where "language is used for doing things (Clark, 1996, p. 3). My hope is that my thesis research, in turn, can be used to improve human communication and allow us to collaborate more successfully via text-based computer-mediated communication.

References

Abadi, E. and Hazan, I. (2020). Improved Feature Engineering for Free-Text Keystroke Dynamics. In Markantonakis, K. and Petrocchi, M., editors, *Security and Trust Management*, Lecture Notes in Computer Science, pages 93–105, Cham. Springer International Publishing.

Acien, A., Morales, A., Monaco, J. V., Vera-Rodriguez, R., and Fierrez, J. (2021). TypeNet: Deep Learning Keystroke Biometrics. *arXiv:2101.05570 [cs]*.

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K. M., Likens, A. D., and McNamara, D. S. (2016). {ENTER} ing the time series {SPACE}: Uncovering the writing process through keystroke analyses. *International Educational Data Mining Society*.

Altman, D. G. and Bland, J. M. (1994). Statistics notes: Diagnostic tests 2: predictive values. *Bmj*, 309(6947):102.

Anderson, R. P. and Anderson, G. V. (1962). Development of an Instrument for Measuring Rapport. *The Personnel and Guidance Journal*, 41(1):18–24.

Ashby, J. and Clifton Jr., C. (2005). The prosodic property of lexical stress affects eye movements during silent reading. *Cognition*, 96(3):B89–B100.

Auerswald, M. and Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological methods*, 24(4):468.

Baaijen, V. M., Galbraith, D., and de Glopper, K. (2012). Keystroke Analysis: Reflections on Procedures and Measures. *Written Communication*, 29(3):246–277.

Bajaj, S. and Kaur, S. (2013). Typing speed analysis of human for password protection (based on keystrokes dynamics). *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 3(2):2278–3075.

Ball, K. et al. (2021). Electronic monitoring and surveillance in the workplace. *European Commission Joint Research Centre*.

Ballier, N., Pacquetet, E., and Arnold, T. (2019). Investigating Keylogs as Time-Stamped Graphemics. In *Graphemics in the 21st Century*, pages 353–365, Brest. Fluxus Editions.

Banerjee, S. P. and Woodard, D. (2012). Biometric Authentication and Identification Using Keystroke Dynamics: A Survey. *Journal of Pattern Recognition Research*, 7(1):116–139.

Barghouthi, H. (2009). Keystroke dynamics: How typing characteristics differ from one application to another. Master's thesis, Gjovik University College.

Barnett, M., Sawyer, J., and Moore, J. (2020). An experimental investigation of the impact of rapport on Stroop test performance. *Applied Neuropsychology: Adult*, pages 1–5.

Barnett, M. D., Parsons, T. D., Reynolds, B. L., and Bedford, L. A. (2018). Impact of rapport on neuropsychological test performance. *Applied Neuropsychology: Adult*, 25(3):258–265.

Barrett, M., González-Garduño, A. V., Frermann, L., and Søgaard, A. (2018a). Unsupervised induction of linguistic categories with records of reading, speaking, and writing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2028–2038.

Barrett, M., González-Garduño, A. V., Frermann, L., and Søgaard, A. (2018b). Unsupervised induction of linguistic categories with records of reading, speaking, and writing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2028–2038.

Bazillon, T., Esteve, Y., and Luzzati, D. (2008). Manual vs assisted transcription of prepared and spontaneous speech. In *LREC*, page 5.

Bell, L., Gustafson, J., and Heldner, M. (2003). Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS*, volume 3, pages 833–836. Citeseer.

Benus, S., Enos, F., Hirschberg, J. B., and Shriberg, E. (2006). Pauses in deceptive speech. *Speech Prosody (ISCA)*.

Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., and Levitan, R. (2018). Prosodic entrainment and trust in human-computer interaction. In *Proceedings of the 9th International Conference on Speech Prosody*, pages 220–224. International Speech Communication Association Baixas, France.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047, Austin, Texas. Association for Computational Linguistics.

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, 14(4):359–375.

Blacfkmer, E. R. and Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39(3):173–194.

Bobicev, V. and Sokolova, M. (2017). Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102.

Borj, P. R. and Bours, P. (2019). Detecting Liars in Chats using Keystroke Dynamics. In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications - ICBEA 2019*, pages 1–6, Stockholm, Sweden. ACM Press.

Bos, N., Olson, J., Gergle, D., Olson, G., and Wright, Z. (2002). Effects of four computer-mediated communications channels on trust development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 135–140.

Bothe, C., Magg, S., Weber, C., and Wermter, S. (2017). Dialogue-Based Neural Learning to Estimate the Sentiment of a Next Upcoming Utterance. In Lintas, A., Rovetta, S., Verschure, P. F., and Villa, A. E., editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, volume 10614, pages 477–485, Cham. Springer International Publishing.

Brandt, D. (2014). *The Rise of Writing: Redefining Mass Literacy*. Cambridge University Press.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57.

Breen, M. (2014a). Empirical investigations of the role of implicit prosody in sentence processing. *Language and Linguistics Compass*, 8(2):37–50.

Breen, M. (2014b). Empirical Investigations of the Role of Implicit Prosody in Sentence Processing. *Language and Linguistics Compass*, 8(2):37–50.

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

Brizan, D. G., Goodkind, A., Koch, P., Balagani, K., Phoha, V. V., and Rosenberg, A. (2015). Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies*, 82:57–68.

Bruce, G. and Touati, P. (1990). On the analysis of prosody in spontaneous dialogue. *Working papers/Lund University, Department of Linguistics and Phonetics*, 36:37–55.

Bukeer, A., Roffo, G., and Vinciarelli, A. (2019). Type Like a Man! Inferring Gender From Keystroke Dynamics in Live-Chats. *AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS*, page 9.

Buker, A. A. and Vinciarelli, A. (2021). I feel it in your fingers: Inference of self-assessed personality traits from keystroke dynamics in dyadic interactive chats. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Carlson, J. R. and Zmud, R. W. (1994). Channel expansion theory: A dynamic view of medial and information richness perceptions. In *Academy of management proceedings*, pages 280–284. academy of management Briarcliff Manor, NY 10510.

Carlson, J. R. and Zmud, R. W. (1999). Channel expansion theory and the experiential nature of media richness perceptions. *Academy of management journal*, 42(2):153–170.

Carney, D. R., Colvin, C. R., and Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072.

Cascone, L., Nappi, M., Narducci, F., and Pero, C. (2022). Touch keystroke dynamics for demographic classification. *Pattern Recognition Letters*, 158:63–70.

Chafe, W. and Tannen, D. (1987). The Relation between Written and Spoken Language. *Annual Review of Anthropology*, 1:383–407.

Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133.

Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., and Caruana, R. (2021a). How interpretable and trustworthy are gams? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 95–105.

Chang, H.-C., Li, J., Wu, C.-S., and Stamp, M. (2021b). Machine Learning and Deep Learning for Fixed-Text Keystroke Dynamics. *arXiv:2107.00507 [cs].*

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61:1–36.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, R., Levy, R., and Eisape, T. (2021). On factors influencing typing time: Insights from a viral online typing game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Choe, H. (2018). Type your listenership: An exploration of listenership in instant messages. *Discourse Studies*, 20(6):703–725.

Chung, C. K. and Pennebaker, J. W. (2014). Using Computerized Text Analysis to Track Social Processes. In Holtgraves, T. M., editor, *The Oxford Handbook of Language and Social Psychology*. Oxford University Press.

Clark, H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Clark, H. H. (1996). *Using Language*. Cambridge university press.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on Socially Shared Cognition.*, pages 127–149. American Psychological Association, Washington.

Clark, H. H. and Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in Psychology*, volume 9, pages 287–299. Elsevier.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to Discourse. *Cognitive Science*, 13(2):259–294.

Cohen, J. (1988). The effect size index: d. statistical power analysis for the behavioral sciences. *Abingdon-on-Thames: Routledge Academic*.

Conijn, R. (2020). *The Keys to Writing: A Writing Analytics Approach to Studying Writing Processes Using Keystroke Logging*. PhD thesis, Tilburg University.

Conijn, R., Roeser, J., and Van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, 32(9):2353–2374.

Convertino, G., Mentis, H. M., Rosson, M. B., Carroll, J. M., Slavkovic, A., and Ganoe, C. H. (2008). Articulating common ground in cooperative work: content and process. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1637–1646.

Coover, J. E. (1923). A method of teaching typewriting based on a psychological analysis of expert typing. In *National Education Association Addresses and Proceedings*, volume 61, pages 561–567.

Core, M. G. and Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme. In *In Proc. Working Notes AAAI Fall Symp. Commun. Action in Humans*.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue. *International Journal of Human-Computer Studies*, 83:27–42.

Cowen, A. S., Elfenbein, H. A., Laukka, P., and Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698.

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied linguistics*, 11(2):183–199.

Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571.

Dagum, P. (2018). Digital biomarkers of cognitive function. *NPJ digital medicine*, 1(1):1–3.

Dahlmann, I. and Adolphs, S. (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions - MWE '07*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv:1106.3077 [physics]*.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Association for Computational Linguistics (NAACL)*.

Dhakal, V., Feit, A. M., Kristensson, P. O., and Oulasvirta, A. (2018). Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–12, Montreal QC, Canada. ACM Press.

Dhillon, R. (2008). Using pause durations to discriminate between lexically ambiguous words and dialog acts in spontaneous speeech. *The Journal of the Acoustical Society of America*, 123(5):3425–3425.

D'Onofrio, A. (2020). Personae in sociolinguistic variation. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(6):e1543.

Edelsky, C. (1981). Who's got the floor? *Language in society*, 10(3):383–421.

Edlund, J., Heldner, M., and Gustafson, J. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, pages 576–587.

Epp, C., Lippold, M., and Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, page 715, Vancouver, BC, Canada. ACM Press.

Erkens, G., Jaspers, J., Prangsma, M., and Kanselaar, G. (2005). Coordination processes in computer supported collaborative writing. *Computers in Human Behavior*, 21(3):463–486.

Faggioli, G., Ferrante, M., Ferro, N., Perego, R., and Tonellotto, N. (2021). Hierarchical dependence-aware evaluation measures for conversational search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1935–1939.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145.

Faraj, S., Jarvenpaa, S. L., and Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization science*, 22(5):1224–1239.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275.

Flynn, J. (2022). 25 Trending Remote Work Statistics [2022]: Facts, Trends, And Projections – Zippia.

Fodor, J. D. (2002a). Prosodic Disambiguation In Silent Reading. *North East Linguistics Society (NELS)*, 32:21.

Fodor, J. D. (2002b). Psycholinguistics Cannot Escape Prosody. In *International Conference in Speech Prosody*, page 7.

Forsyth, E. N. (2007). *Improving Automated Lexical and Discourse Analysis of Online Chat Dialog*. PhD thesis, Naval Postgraduate School, Monterey, CA.

Fujie, S., Kobayashi, T., Yagi, D., and Kikuchi, H. (2004). Prosody based attitude recognition with feature selection and its application to spoken dialog system as para-linguistic information. In *Eighth International Conference on Spoken Language Processing*.

Fulk, J., Schmitz, J., and Ryu, D. (1995). Cognitive elements in the social construction of communication technology. *Management Communication Quarterly*, 8(3):259–288.

Fulk, J., Steinfield, C. W., Schmitz, J., and Power, J. G. (1987). A social information processing model of media use in organizations. *Communication research*, 14(5):529–552.

Galbraith, D. and Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In *Observing Writing*, pages 306–325. Brill.

Ganesan, A., Varadarajan, V., Mittal, J., Subrahmanya, S., Matero, M., Soni, N., Guntuku, S. C., Eichstaedt, J., and Schwartz, H. A. (2022). WWBP-SQT-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258, Seattle, USA. Association for Computational Linguistics.

Garrod, S. (1999). The challenge of dialogue for theories of language processing. *Language processing*, pages 389–415.

Garrod, S. and Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, 1(2):292–304.

Gergle, D. (2017). Discourse Processing in Technology-Mediated Environments. In *The Routledge Handbook of Discourse Processes*. Routledge.

Gergle, D., Millen, D. R., Kraut, R. E., and Fussell, S. R. (2004). Persistence matters: Making the most of chat in tightly-coupled work. In *Proceedings of the 2004 Conference on Human Factors in Computing Systems - CHI '04*, pages 431–438, Vienna, Austria. ACM Press.

Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D. R., Campbell, M., Kumaravel, S., and Zhang, W. (2020). Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Honolulu HI USA. ACM.

Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Gidron, M., Sabag, M., Yarmolovsky, J., and Geva, R. (2020). Participant–experimenter rapport in experimental settings: A test case of executive functions among children with ADHD. *Journal of Experimental Psychology: General*, 149(9):1615.

Gill, A. J., Gergle, D., French, R. M., and Oberlander, J. (2008). Emotion rating from short blog texts. In *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, page 1121, Florence, Italy. ACM Press.

Gnjatović, M. and Delić, V. (2013). Electrophysiologically-inspired evaluation of dialogue act complexity. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 167–172. IEEE.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520.

Goodkind, A., Brizan, D. G., and Rosenberg, A. (2017). Utilizing overt and latent linguistic structure to improve keystroke-based authentication. *Image and Vision Computing*, 58:230–238.

Goodkind, A. and Rosenberg, A. (2015). Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 87–95, Denver, Colorado. Association for Computational Linguistics.

Grahe, J. E. and Bernieri, F. J. (2002). Self-awareness of judgment policies of rapport. *Personality and Social Psychology Bulletin*, 28(10):1407–1418.

Gravano, A. and Hirschberg, J. (2009). Turn-Yielding Cues in Task-Oriented Dialogue. In *Proceedings of the SIGDIAL 2009 Conference*, pages 253–261, London, UK. Association for Computational Linguistics.

Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Gravano, A., Levitan, R., Willson, L., Beňuš, Š., Hirschberg, J., and Nenkova, A. (2011). Acoustic and prosodic correlates of social behavior. In *Twelfth Annual Conference of the International Speech Communication Association*.

Gregory, M., Johnson, M., and Charniak, E. (2004). Sentence-internal prosody does not help parsing the way punctuation does. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 81–88.

Hara, K., Adams, A., Milland, K., Savage, S., Hanrahan, B. V., Bigham, J. P., and Callison-Burch, C. (2019). Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–6.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., and Zimmermann, R. (2018). Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132.

Heath, M. (2021). No need to yell: A prosodic analysis of writing in all caps. *University of Pennsylvania Working Papers in Linguistics*, 27(1):10.

Hegselmann, S., Volkert, T., Ohlenburg, H., Gottschalk, A., Dugas, M., and Ertmer, C. (2020). An evaluation of the doctor-interpretability of generalized additive models with interactions. In *Machine Learning for Healthcare Conference*, pages 46–79. PMLR.

Heldner, M. and Edlund, J. (2010a). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Heldner, M. and Edlund, J. (2010b). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Herzig, J., Feigenblat, G., Shmueli-Scheuer, M., Konopnicki, D., and Rafaeli, A. (2016). Predicting customer satisfaction in customer support conversations in social media using affective features. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 115–119.

Hieronymus, J. L. and Williams, B. J. (1991). A comparison of the prosody in read speech and directed monologue in british english. In *Phonetics and Phonology of Speaking Styles*.

Hirschberg, J. B., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., et al. (2005). Distinguishing deceptive from non-deceptive speech. *Interspeech*.

Horton, W. S. (2017). Theories and Approaches to the Study of Conversation and Interactive Discourse. *The Routledge Handbook of Discourse Processes*.

Horton, W. S. and Gerrig, R. J. (2016). Revisiting the Memory-Based Processing Approach to Common Ground. *Topics in Cognitive Science*, 8(4):780–795.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Housen, A. and Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, 30(4):461–473.

Howe, E., Suh, J., Bin Morshed, M., McDuff, D., Rowan, K., Hernandez, J., Abdin, M. I., Ramos, G., Tran, T., and Czerwinski, M. P. (2022). Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Hutto, C. J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Iliinsky, N. and Steele, J. (2011). *Designing data visualizations: Representing informational Relationships*. " O'Reilly Media, Inc.".

Indiparambil, J. J. (2019). Privacy and beyond: Socio-ethical concerns of 'on-the-job'surveillance. *Asian Journal of Business Ethics*, 8(1):73–105.

Ivanovic, E. (2005). Dialogue act tagging for instant messaging chat sessions. In *Proceedings of the ACL Student Research Workshop on - ACL '05*, page 79, Ann Arbor, Michigan. Association for Computational Linguistics.

Janssen, D. and Carradini, S. (2021). Generation z workplace communication habits and expectations. *IEEE Transactions on Professional Communication*, 64(2):137–153.

Jokinen, K. (2010). Hesitation and uncertainty as feedback. In *DiSS-LPSS Joint Workshop 2010*.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. *University of Colorado at Boulder & SRI International*.

Kalman, Y. M. (2007). *Silence In Text-Based Computer Mediated Communication: The Invisible Component*. PhD thesis, University of Haifa.

Kalman, Y. M. and Gergle, D. (2009). Repeats as Cues in CMC Letter and Punctuation Mark Repeats as Cues in Computer-Mediated Communication. In *5th Annual Meeting of the National Communication Association*, Chicago, IL.

Kalman, Y. M., Scissors, L. E., and Gergle, D. R. (2010). Chronemic Aspects of Chat and Their Relationship to Trust in a Virtual Team. *Proceedings of the Fifth Mediterranean Conference on Information Systems: Professional Development Consortium*.

Kalman, Y. M., Scissors, L. E., Gill, A. J., and Gergle, D. (2013a). Online chronemics convey social information. *Computers in Human Behavior*, 29(3):1260–1269.

Kalman, Y. M., Scissors, L. E., Gill, A. J., and Gergle, D. (2013b). Online chronemics convey social information. *Computers in Human Behavior*, 29(3):1260–1269.

Kasher, A. (1972). Sentences and Utterances Reconsidered. *Foundations of Language*, 8(3):313–345.

Katerina, T. and Nicolaos, P. (2018). Mouse behavioral patterns and keystroke dynamics in end-user development: What can they tell us about users' behavioral attributes? *Computers in Human Behavior*, 83:288–305.

Khawaja, M. A., Chen, F., and Marcus, N. (2014). Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction*, 30(5):343–368.

Khawaji, A., Chen, F., Zhou, J., and Marcus, N. (2014). Trust and Cognitive Load in the Text-Chat Environment: The Role of Mouse Movement. *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design*, page 4.

Killourhy, K. S. and Maxion, R. A. (2012). Free vs. transcribed text for keystroke-dynamics evaluations. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results - LASER '12*, pages 1–8, Arlington, Virginia. ACM Press.

Kołakowska, A. (2013). A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 548–555.

Kołakowska, A. (2015). Recognizing emotions on the basis of keystroke dynamics. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 291–297.

Kołakowska, A. (2018). Usefulness of Keystroke Dynamics Features in User Authentication and Emotion Recognition. In Hippe, Z. S., Kulikowski, J. L., and Mroczek, T., editors, *Human-Computer Systems Interaction: Backgrounds and Applications 4*, Advances in Intelligent Systems and Computing, pages 42–52. Springer International Publishing, Cham.

Koudenburg, N., Postmes, T., and Gordijn, E. H. (2017). Beyond content of conversation: The role of conversational form in the emergence and regulation of social structure. *Personality and Social Psychology Review*, 21(1):50–71.

Krauss, R. M. and Fussell, S. R. (1996). Social psychological models of interpersonal communication. In *Social Psychology: Handbook of Basic Principles*. Guilford Publications.

Kuhn, M. and Silge, J. (2022). *Tidy Modeling with R*. " O'Reilly Media, Inc.".

Kuzminykh, I., Ghita, B., and Silonosov, A. (2020). Impact of network and host characteristics on the keystroke pattern in remote desktop sessions. *arXiv preprint arXiv:2012.03577*.

LaBahn, D. W. (1996). Advertiser Perceptions of Fair Compensation, Confidentiality and Rapport: The Influence of Advertising Agency Cooperativeness and Diligence. *Journal of Advertising Research*, 36.

Lai, C. (2012). *Rises all the way up: The interpretation of prosody, discourse attitudes and dialogue structure*. University of Pennsylvania.

Leach, M. J. (2005). Rapport: A key to treatment success. *Complementary therapies in clinical practice*, 11(4):262–265.

LeDell, E. and Poirier, S. (2020). H2o automl: Scalable automatic machine learning. In *Proceedings of the AutoML Workshop at ICML*, volume 2020.

Lee, P.-M., Tsui, W.-H., and Hsiao, T.-C. (2014). The influence of emotion on keyboard typing: An experimental study using visual stimuli. *BioMedical Engineering OnLine*, 13(1):81.

Lee, P.-M., Tsui, W.-H., and Hsiao, T.-C. (2015a). The Influence of Emotion on Keyboard Typing: An Experimental Study Using Auditory Stimuli. *PLOS ONE*, 10(6):e0129056.

Lee, P.-M., Tsui, W.-H., and Hsiao, T.-C. (2015b). The Influence of Emotion on Keyboard Typing: An Experimental Study Using Auditory Stimuli. *PLOS ONE*, 10(6):e0129056.

Leijten, M. and Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30(3):358–392.

Leinonen, J., Ihantola, P., and Hellas, A. (2017). Preventing Keystroke Based Identification in Open Data Sets. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, pages 101–109, Cambridge Massachusetts USA. ACM.

Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Li, Y., Ishi, C. T., Ward, N., Inoue, K., Nakamura, S., Takanashi, K., and Kawahara, T. (2017a). Emotion recognition by combining prosody and sentiment analysis for expressing reactive emotion by humanoid robot. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1356–1359. IEEE.

Li, Y., Ishi, C. T., Ward, N., Inoue, K., Nakamura, S., Takanashi, K., and Kawahara, T. (2017b). Emotion recognition by combining prosody and sentiment analysis for expressing reactive emotion by humanoid robot. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1356–1359, Kuala Lumpur. IEEE.

Liebman, N. and Gergle, D. (2016a). Capturing Turn-by-Turn Lexical Similarity in Text-Based Communication. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, pages 552–558, San Francisco, California, USA. ACM Press.

Liebman, N. and Gergle, D. (2016b). It's (Not) Simply a Matter of Time: The Relationship Between CMC Cues and Interpersonal Affinity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, pages 569–580, San Francisco, California, USA. ACM Press.

Lindgren, E., Westum, A., Outakoski, H., and Sullivan, K. P. (2019). Revising at the leading edge: Shaping ideas or clearing up noise. In *Observing Writing*, pages 346–365. Brill.

Ling, R. and Baron, N. S. (2007). Text messaging and im: Linguistic comparison of american college data. *Journal of language and social psychology*, 26(3):291–298.

Linzen, T., Chrupała, G., and Alishahi, A. (2018). Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Locklear, H., Govindarajan, S., Sitová, Z., Goodkind, A., Brizan, D. G., Rosenberg, A., Phoha, V. V., Gasti, P., and Balagani, K. S. (2014). Continuous authentication with cognition-centric text production and revision features. In *IEEE International Joint Conference on Biometrics*, pages 1–8.

Logan, G. D. and Crump, M. J. (2011). Hierarchical Control of Cognitive Processes. In *Psychology of Learning and Motivation*, volume 54, pages 1–27. Elsevier.

López-Carral, H., Santos-Pata, D., Zucca, R., and Verschure, P. F. (2019). How you type is what you type: Keystroke dynamics correlate with affective content. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–5.

Lovric, N. (2003). *Implicit Prosody in Silent Reading: Relative Clause Attachment in Croatian*. PhD thesis, City University of New York (CUNY Graduate Center).

Lubold, N. and Pon-Barry, H. (2014). Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12, Istanbul Turkey. ACM.

Maalej, A. and Kallel, I. (2020). Does keystroke dynamics tell us about emotions? a systematic literature review and dataset construction. In *2020 16th International Conference on Intelligent Environments (IE)*, pages 60–67. IEEE.

Malhotra, G., Waheed, A., Srivastava, A., Akhtar, M. S., and Chakraborty, T. (2022). Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 735–745.

Manandhar, R., Wolf, S., and Borowczak, M. (2019). One-class classification to continuously authenticate users based on keystroke timing dynamics. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 1259–1266. IEEE.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

Matsumoto, S. and Araki, M. (2016). Scoring of response based on suitability of dialogue-act and content similarity. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*.

McCauley, S. M., Isbilen, E. S., and Christiansen, M. H. (2017). Chunking ability shapes sentence processing at multiple levels of abstraction. In *Proceeding of CogSci*.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Medimorec, S. and Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, 30(6):1267–1285.

Meta Platforms (2022). React – a javascript library for building user interfaces. Accessed: 2022-03-06.

Michalsky, J. and Schoormann, H. (2017). Pitch Convergence as an Effect of Perceived Attractiveness and Likability. In *Interspeech*, pages 2253–2256.

Microsoft (2021). The next great disruption is hybrid work: Are we ready? https://www.microsoft.com/en-us/worklab/work-trend-index/hybrid-work.

Mijic, I., Sarlija, M., and Petrinovic, D. (2017). Classification of cognitive load using voice features: A preliminary investigation. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000345–000350, Debrecen. IEEE.

Miller, G. (1956a). Human memory and the storage of information. *IRE Transactions on Information Theory*, 2(3):129–137.

Miller, G. (1956b). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Mohammad, S. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Monaco, J. V. and Tappert, C. C. (2016). Obfuscating keystroke time intervals to avoid identification and impersonation. *arXiv preprint arXiv:1609.07612*.

Monaco, J. V. and Tappert, C. C. (2017). Obfuscating Keystroke Time Intervals to Avoid Identification and Impersonation. *arXiv:1609.07612 [cs]*.

Monrose, F. and Rubin, A. (1997a). Authentication via keystroke dynamics. In *Proceedings of the 4th ACM Conference on Computer and Communications Security - CCS '97*, pages 48–56, Zurich, Switzerland. ACM Press.

Monrose, F. and Rubin, A. (1997b). Authentication via keystroke dynamics. In *Proceedings of the 4th ACM Conference on Computer and Communications Security - CCS '97*, pages 48–56, Zurich, Switzerland. ACM Press.

Morales, A., Acien, A., Fierrez, J., Monaco, J. V., Tolosana, R., Vera-Rodriguez, R., and Ortega-Garcia, J. (2020). Keystroke Biometrics in Response to Fake News Propagation in a Global Pandemic. *arXiv:2005.07688 [cs]*.

Moroney, L. (2017). The firebase realtime database. In *The Definitive Guide to Firebase*, pages 51–71. Springer.

Muir, K., Joinson, A., Cotterill, R., and Dewdney, N. (2017). Linguistic Style Accommodation Shapes Impression Formation and Rapport in Computer-Mediated Communication. *Journal of Language and Social Psychology*, 36(5):525–548.

Müller, P., Huang, M. X., and Bulling, A. (2018). Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces*, pages 153–164, Tokyo Japan. ACM.

Murphy, B. (2022). They called 911 for help. police and prosecutors used a new junk science to decide they were liars. *propublica.org*.

Murray, G., Carenini, G., and Ng, R. (2010). Interpretation and transformation for abstracting conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 894–902.

Mushin, I., Stirling, L., Fletcher, J., and Wales, R. (2003). Discourse Structure, Grounding, and Prosody in Task-Oriented Dialogue. *Discourse Processes*, 35(1):1–31.

Nahin, A. N. H., Alam, J. M., Mahmud, H., and Hasan, K. (2014). Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology*, 33(9):987–996.

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

Nottbusch, G., Weingarten, R., and Sahel, S. (2007). From written word to written sentence production. *Studies in Writing*, pages 30–53.

Novotney, S. and Callison-Burch, C. (2010). Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, Los Angeles, California. Association for Computational Linguistics.

Ojamaa, B., Jokinen, K., and Muischenk, K. (2015). Sentiment analysis on conversational texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 233–237.

Olive, T., Alves, R. A., and Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, 21(5):758–785.

Olsen, J. K. and Finkelstein, S. (2017). Through the (Thin-Slice) Looking Glass: An Initial Look at Rapport and Co-Construction Within Peer Collaboration. *International Society of the Learning Sciences*, page 8.

Ondobaka, S., Kilner, J., and Friston, K. (2017). The role of interoceptive inference in theory of mind. *Brain and cognition*, 112:64–68.

Ooms, J. (2022). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. https://docs.ropensci.org/hunspell/ (docs), https://github.com/ropensci/hunspell (devel) https://hunspell.github.io (upstream).

Orwell, G. (1949). *Nineteen Eighty-Four*. Secker and Warburg.

Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Park, Y., Cho, J., and Kim, G. (2018). A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801.

Pecune, F., Chen, J., Matsuyama, Y., and Cassell, J. (2018). Field Trial Analysis of Socially Aware Robot Assistant. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, page 9.

Pecune, F., Murali, S., Tsai, V., Matsuyama, Y., and Cassell, J. (2019). A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 135–143, Kyoto Japan. ACM.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4):1643–1662.

Peng, W., Hu, Y., Xing, L., Xie, Y., Sun, Y., and Li, Y. (2022). Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *IJCAI 2022*.

Pennebaker, J. (2011). *The Secret Life of Pronouns: What Our Words Say about Us*. Bloomsbury Press.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1):547–577.

Pew Research Center (2019). Americans favor mobile devices over desktops and laptops for getting news. https://www.pewresearch.org/fact-tank/2019/11/19/americans-favor-mobile-devices-over-desktops-and-laptops-for-getting-news/.

Pew Research Center (2020). How Coronavirus Has Changed the Way Americans Work. https://www.pewresearch.org/social-trends/2020/12/09/how-the-coronavirus-outbreak-has-and-hasnt-changed-the-way-americans-work/.

Picard, R. W. (2000). *Affective Computing*. MIT press.

Pierrehumbert, J. and Hirschberg, J. B. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. *Intentions in Communication*.

Pinet, S. and Nozari, N. (2021). The role of visual feedback in detecting and correcting typing errors: A signal detection approach. *Journal of Memory and Language*, 117:104193.

Pinet, S., Ziegler, J. C., and Alario, F.-X. (2016). Typing is writing: Linguistic properties modulate typing execution. *Psychonomic bulletin & review*, 23(6):1898–1906.

Pinet, S., Zielinski, C., Mathôt, S., Dufau, S., Alario, F.-X., and Longcamp, M. (2017). Measuring sequences of keystrokes with jsPsych: Reliability of response times and interkeystroke intervals. *Behavior research methods*, 49(3):1163–1176.

Plank, B. (2016). Keystroke dynamics as signal for shallow syntactic parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 609–619, Osaka, Japan. The COLING 2016 Organizing Committee.

Poesio, M. and Rieses, H. (2010). Completions, coordination, and alignment in dialogue. *Dialogue & Discourse*, 1(1).

Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.-P., and Jonker, C. M. (2012). Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction*, 22(4):357–397.

Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions. *ISSCO WorkingPaper*, 62:1–46.

Pradhan, A., Findlater, L., and Lazar, A. (2019). " phantom friend" or" just a box with information" personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Priva Cohen, U. (2010). Constructing Typing-Time Corpora: A New Way to Answer Old Questions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, page 7.

Qi, Z. and Li, M. (2014). Mining domain-dependent noun opinion words for sentiment analysis. In *Multimedia and Ubiquitous Engineering*, pages 165–171. Springer.

Rafaeli, S. and Ariel, Y. (2008). Online motivational factors: Incentives for participation and contribution in wikipedia. *Psychological aspects of cyberspace: Theory, research, applications*, 2(08):243–267.

Raj Prabhu, N., Raman, C., and Hung, H. (2020). Defining and Quantifying Conversation Quality in Spontaneous Interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 196–205, Virtual Event Netherlands. ACM.

Rauch, G. (2013). Socket.io-the cross-browser websocket for realtime apps.

Reichman, R. (1978). Conversational coherency. *Cognitive science*, 2(4):283–327.

Reid, F. J., Ball, L. J., Morley, A. M., and Evans, J. S. B. (1997). Styles of group discussion in computer-mediated decision making. *British Journal of Social Psychology*, 36(3):241–262.

Riordan, M. A. (2011). *The Use of Verbal and Nonverbal Cues in Computer-Mediated Communication: When and Why?* Ph.D., The University of Memphis, United States – Tennessee.

Roffo, G., Giorgetta, C., Ferrario, R., Riviera, W., and Cristani, M. (2014). Statistical Analysis of Personality and Identity in Chats Using a Keylogging Platform. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, pages 224–231, Istanbul, Turkey. ACM Press.

Rozumowski, A. V., Kotowski, W., and Klaas, M. (2020). Resistance to Customer-driven Business Model Innovations: An Explorative Customer Experience Study on Voice Assistant Services of a Swiss Tourism Destination. *ATHENS JOURNAL OF TOURISM*, 7(4):191–208.

Rumelhart, D. E. and Norman, D. A. (1982). Simulating a Skilled Typist: A Study of Skilled Cognitive-Motor Performance. *Cognitive Science*, 6(1):1–36.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. Elsevier.

Saevanee, H., Clarke, N. L., and Furnell, S. M. (2012). Multi-modal Behavioural Biometric Authentication for Mobile Devices. In Gritzalis, D., Furnell, S., and Theoharidou, M., editors, *Information Security and Privacy Research*, volume 376, pages 465–474. Springer Berlin Heidelberg, Berlin, Heidelberg.

Schegloff, E. A. and Sacks, H. (1973). *Opening up closings*. Walter de Gruyter.

Schilperoord, J. (2002). On the cognitive status of pauses in discourse production. In *Contemporary Tools and Techniques for Studying Writing*, pages 61–87. Springer.

Scissors, L. E. and Gergle, D. (2013). " back and forth, back and forth" channel switching in romantic couple conflict. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 237–248.

Scissors, L. E., Gill, A. J., Geraghty, K., and Gergle, D. (2009). In CMC we trust: The role of similarity. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, page 527, Boston, MA, USA. ACM Press.

Scissors, L. E., Roloff, M. E., and Gergle, D. (2014). Room for interpretation: The role of self-esteem and cmc in romantic couple conflict. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3953–3962.

Scott, C. R. (2009). A Whole-Hearted Effort to Get It Half Right: Predicting the Future of Communication Technology Scholarship. *Journal of Computer-Mediated Communication*, 14(3):753–757.

Seide, F., Li, G., Chen, X., and Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 24–29. IEEE.

Selkirk, E. (1995). Sentence prosody: Intonation, stress, and phrasing. *The handbook of phonological theory*, 1:550–569.

Seo, S. H., Griffin, K., Young, J. E., Bunt, A., Prentice, S., and Loureiro-Rodríguez, V. (2018). Investigating People's Rapport Building and Hindering Behaviors When Working with a Collaborative Robot. *International Journal of Social Robotics*, 10(1):147–161.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press.

Shon, S., Brusco, P., Pan, J., Han, K. J., and Watanabe, S. (2021). Leveraging Pre-trained Language Model for Speech Sentiment Analysis. *arXiv:2106.06598 [cs, eess].*

Short, J., Williams, E., and Christie, B. (1976). *The Social Psychology of Telecommunications*. Toronto; London; New York: Wiley.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154.

Sisman, B., Yamagishi, J., King, S., and Li, H. (2021). An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Snow, D. (1994). Phrase-Final Syllable Lengthening and Intonation in Early Child Speech. *Journal of Speech, Language, and Hearing Research*, 37(4):831–840.

Stivers, T. (2012). Sequence Organization. In Sidnell, J. and Stivers, T., editors, *The Handbook of Conversation Analysis*, pages 191–209. John Wiley & Sons, Ltd, Chichester, UK.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., Meteer, M., Ries, K., Taylor, P., and Ess-Dykema, C. V. (1998). Dialog Act Modeling for Conversational Speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

Stragapede, G., Delgado-Santos, P., Tolosana, R., Vera-Rodriguez, R., Guest, R., and Morales, A. (2022). Mobile keystroke biometrics using transformers. *arXiv preprint arXiv:2207.07596*.

Stranc, S. and Muldner, K. (2019). Learning from videos showing a dialog fosters more positive affect than learning from a monolog. In *International Conference on Artificial Intelligence in Education*, pages 275–280. Springer.

Su, M.-H., Wu, C.-H., and Chen, L.-Y. (2019). Attention-based response generation using parallel double q-learning for dialog policy decision in a conversational system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:131–143.

Suzuki, N. and Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science*, 19(2):131–141.

Svec, J. G. and Granqvist, S. (2010). Guidelines for Selecting Microphones for Human Voice Production Research. *American Journal of Speech-Language Pathology*, 19(4):356–368.

Svennevig, J. (2014). Direct and indirect self-presentation in first conversations. *Journal of Language and Social Psychology*, 33(3):302–327.

Swerts, M. and Geluykens, R. (1994). Prosody as a Marker of Information Flow in Spoken Discourse. *Language and Speech*, 37(1):21–43.

Tannen, D. (1984). *Conversational Style: Analyzing Talk Among Friends*. Oxford University Press.

Teevan, J., Baym, N., Butler, J., Hecht, B., Jaffe, S., Nowak, K., Sellen, A., and Yang, L. (2022). Microsoft new future of work report 2022. Technical report, Technical Report. Microsoft Research Tech Report MSR-TR-2022–3.

Tegos, S., Demetriadis, S., Psathas, G., and Tsiatsos, T. (2020). A configurable agent to advance peers' productive dialogue in moocs. In *International Workshop on Chatbot Research and Design*, pages 245–259. Springer.

Tham, T. L. and Holland, P. (2022). Electronic monitoring and surveillance: The balance between insights and intrusion. In *The Emerald Handbook of Work, Workplaces and Disruptive Issues in HRM*, pages 493–512. Emerald Publishing Limited.

Tickle-Degnen, L. and Rosenthal, R. (1990). The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, 1(4):285–293.

Tidwell, L. C. and Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human communication research*, 28(3):317–348.

Tolles, J. and Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, 316(5):533–534.

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., and Bernstein, A. (2021). Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, pages 77–87.

Trott, S., Reed, S., Ferreira, V., and Bergen, B. (2019). Prosodic cues signal the intent of potential indirect requests. In *Proceedings of CogSci 2019*, page 7.

Tsimperidis, I. and Arampatzis, A. (2020). The Keyboard Knows About You: Revealing User Characteristics via Keystroke Dynamics. *International Journal of Technoethics*, 11:34–51.

Twenge, J. M. and Farley, E. (2021). Not all screen time is created equal: Associations with mental health vary by activity and gender. *Social Psychiatry and Psychiatric Epidemiology*, 56(2):207–217.

Van Waes, L. and Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to l1 and l2. *Computers and Composition*, 38:79–95.

Vicsi, K. and Szaszák, G. (2010). Using prosody to improve automatic speech recognition. *Speech Communication*, 52(5):413–426.

Villani, M., Tappert, C., Ngo, G., Simone, J., Fort, H., and Cha, S.-H. (2006). Keystroke Biometric Recognition Studies on Long-Text Input under Ideal and Application-Oriented Conditions. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 39–39.

Vizer, L. M. (2009). Detecting cognitive and physical stress through typing behavior. In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '09*, page 3113, Boston, MA, USA. ACM Press.

Vizer, L. M. and Sears, A. (2017). Efficacy of personalized models in discriminating high cognitive demand conditions using text-based interactions. *International Journal of Human-Computer Studies*, 104:80–96.

Walther, J. B. (1992). Interpersonal Effects in Computer-Mediated Interaction: A Relational Perspective. *Communication Research*, 19(1):52–90.

Walther, J. B. (2011). Theories of Computer- Mediated Communication and Interpersonal Relations. In *The Handbook of Interpersonal Communication*, volume 4, pages 443–479. Sage.

Walther, J. B. (2018). The Emergence, Convergence, and Resurgence of Intergroup Communication Theory in Computer-Mediated Communication. *Atlantic Journal of Communication*, 26(2):86–97.

Walther, J. B. and Parks, M. R. (2002). Cues filtered out, cues filtered in: Computer-mediated communication and relationships. *Handbook of interpersonal communication*, 3:529–563.

Wang, Z., Wang, L., Ji, Y., Zuo, L., and Qu, S. (2022). A novel data-driven weighted sentiment analysis based on information entropy for perceived satisfaction. *Journal of Retailing and Consumer Services*, 68:103038.

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., and Lestantyo, P. (2019). Cross-validation metrics for evaluating classification performance on imbalanced data. In *2019 international conference on computer, control, informatics and its applications (IC3INA)*, pages 14–18. IEEE.

Wei, K., Knox, D., Radfar, M., Tran, T., Müller, M., Strimel, G. P., Susanj, N., Mouchtaris, A., and Omologo, M. (2022). A neural prosody encoder for end-to-end dialogue act classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7047–7051. IEEE.

Welch, C., Pérez-Rosas, V., Kummerfeld, J. K., and Mihalcea, R. (2019). Learning from personal longitudinal dialog data. *IEEE Intelligent systems*, 34(4):16–23.

Willemyns, M., Gallois, C., and Callan, V. (2003). Trust me, I'm your boss: Trust and power in supervisor–supervisee communication. *The International Journal of Human Resource Management*, 14(1):117–127.

Wilson, T. P., Wiemann, J. M., and Zimmerman, D. H. (1984). Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3(3):159–183.

Wood, S. (2022). Package 'mgcv'. *R package version 1.8*.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Yamaguchi, M., Crump, M. J. C., and Logan, G. D. (2013). Speed–accuracy trade-off in skilled typewriting: Decomposing the contributions of hierarchical control loops. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):678–699.

Yan, M., Tan, H., Jia, L., and Akram, U. (2020). The antecedents of poor doctor-patient relationship in mobile consultation: A perspective from computer-mediated communication. *International Journal of Environmental Research and Public Health*, 17(7):2579.

Yang, L. and Qin, S.-f. (2021). A review of emotion recognition methods from keystroke, mouse, and touchscreen dynamics. *IEEE Access*.

Yeh, S.-L., Lin, Y.-S., and Lee, C.-C. (2019). An Interaction-aware Attention Network for Speech Emotion Recognition in Spoken Dialogs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689, Brighton, United Kingdom. IEEE.

Yu, G. (2010). Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*, 31(2):236–259.

Yuasa, M., Saito, K., and Mukawa, N. (2006). Emoticons convey emotions without cognition of faces: an fmri study. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1565–1570.

Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., and Zhou, G. (2019). Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421.

Zhang, Y., Tiwari, P., Song, D., Mao, X., Wang, P., Li, X., and Pandey, H. M. (2021). Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis. *Neural Networks*, 133:40–56.

Zhao, R., Papangelis, A., and Cassell, J. (2014). Towards a Dyadic Computational Model of Rapport Management for Human-Virtual Agent Interaction. In Bickmore, T., Marsella, S., and Sidner, C., editors, *Intelligent Virtual Agents*, volume 8637, pages 514–527. Springer International Publishing, Cham.

Zhao, R., Sinha, T., Black, A. W., and Cassell, J. (2016). Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218–233. Springer.

Zhu, Q. (2020). On the performance of matthews correlation coefficient (mcc) for imbalanced dataset. *Pattern Recognition Letters*, 136:71–80.

Appendix A

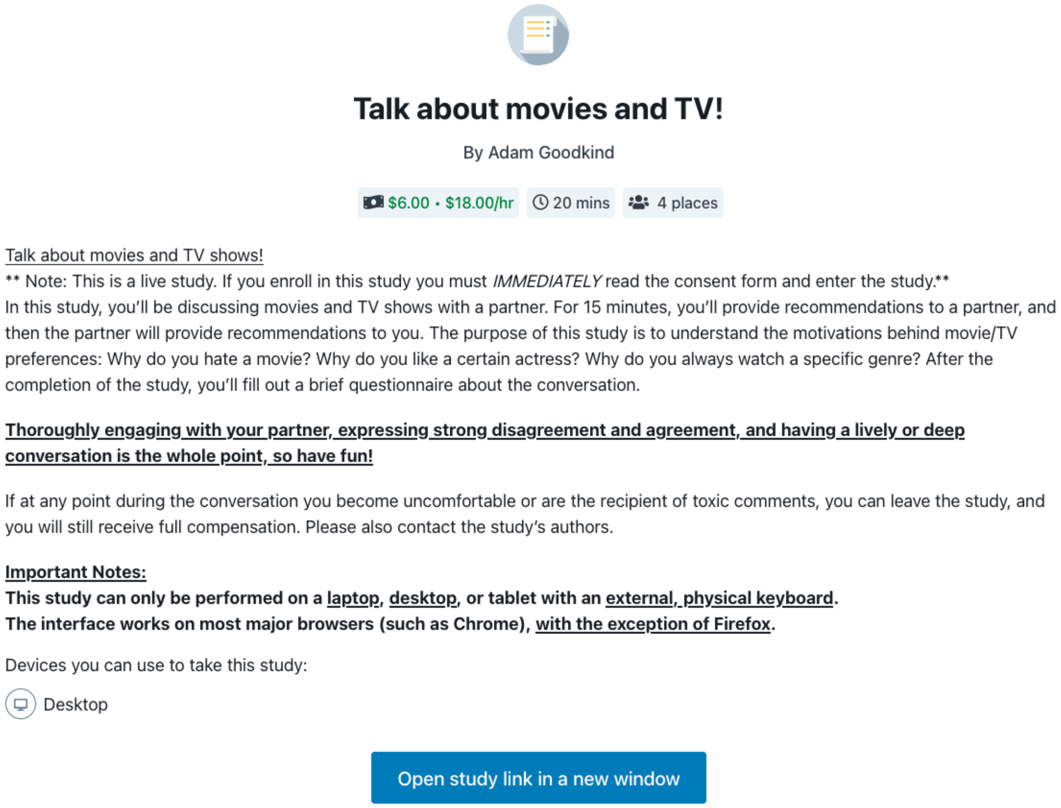
IRB Detail

Prior to IRB approval, a pilot study consisting of ten 10-minute dialogues was collected. This was used for testing purposes to ensure that the experimental apparatus was accurately collecting all of the data necessary for my studies. Further, I solicited comments from pilot participants in order to improve the experimental setup. Because the pilot study was conducted prior to IRB approval and used a different setup, all of the data was discarded after testing, and is not included in any of my thesis studies.

The IRB required consent before participants began the experiment. In a pre-experiment consent form, participant were told what the experiment would entail, and the approximate amount of time it would take. Although the subject matter they would be discussing was relatively innocuous (movie and TV preferences), participants were informed that if they felt uncomfortable with the conversation or it turned toxic, they could leave the experiment and still receive compensation.[1] Only after this initial consent was submitted could a participant enter the experiment apparatus.

In a post-experiment consent form, full disclosure was provided as to the true objective of the experiment, and exactly what data was collected and why. It was explained that keystrokes are considered a soft biometric, and can be personally revealing, akin to an iris scan (Banerjee and Woodard, 2012). However participants were also reminded that their conversation would be

[1] In my full data collection, consisting of over 200 participants, only one participant left the experiment for this reason. However, it should be noted that their specific objections were unclear, and I am unsure if they were mentally stable or intoxicated. In any case, the data from this experiment was thrown out and not included.



Talk about movies and TV!

By Adam Goodkind

Talk about movies and TV shows!
** Note: This is a live study. If you enroll in this study you must *IMMEDIATELY* read the consent form and enter the study.**
In this study, you'll be discussing movies and TV shows with a partner. For 15 minutes, you'll provide recommendations to a partner, and then the partner will provide recommendations to you. The purpose of this study is to understand the motivations behind movie/TV preferences: Why do you hate a movie? Why do you like a certain actress? Why do you always watch a specific genre? After the completion of the study, you'll fill out a brief questionnaire about the conversation.

Thoroughly engaging with your partner, expressing strong disagreement and agreement, and having a lively or deep conversation is the whole point, so have fun!

If at any point during the conversation you become uncomfortable or are the recipient of toxic comments, you can leave the study, and you will still receive full compensation. Please also contact the study's authors.

Important Notes:
This study can only be performed on a laptop, desktop, or tablet with an external, physical keyboard.
The interface works on most major browsers (such as Chrome), with the exception of Firefox.

Devices you can use to take this study:

Figure A.1
The advertisement for the experiment, sent out by Prolific to potential participants.

manually anonymized, and that only the minimal demographics provided by Prolific would be used. Upon reading this, participants had the option to contact the researchers and ask to opt-out (while still receiving compensation). In the actual data collection, none of the participants chose to exercise this option after the experiment was complete, though.

Appendix B

Prolific Advantages and Disadvantages

Prolific provides a number of key benefits over Amazon's Mechancial Turk:

1. Prolific performs more stringent screening of participants, so researchers are less likely to encounter scammers or bots, and more likely to receive trustworthy and engaged participants.

2. Prolific encourages fair compensation for participants, encouraging a healthier environment for research (e.g. Hara et al., 2019). Before running an experiment a researcher must declare an hourly rate, and the number of participants required. Given this, they then deposit the necessary funds to pay every participant their quoted rate. Upon completion of the experiment, Prolific pays all participants based on the *actual* average time taken to complete the experiment. For example, if a researcher estimates that their experiment will only take five minutes, but in reality most participants take ten minutes, then all participants will be paid for ten minutes of work.

3. Prolific takes privacy very seriously: from initial recruitment through final payment a research only ever sees a participant's ID number and limited demographics, rather than names, credit card numbers, etc.

4. Pre-screening is free and easy on Prolific, which was essential for ensuring that all participants currently lived in the United States (per the IRB). We also pre-screened for other criteria, which will be enumerated in Section 4.1. On the other hand, MTurk charges for pre-screening.

One downside to Prolific is that their pool of participants is smaller than that of MTurk's. That being said, even after applying my pre-screening criteria for my data collection, I still had a pool of approximately 28,000 participants that were eligible to receive notification of my experiment. Of these, approximately 20,000 identified as female, and 8,000 identified as male. Because of this imbalance, Prolific also provides the option to distribute an experiment to a gender-balanced sample. I selected this option for a few batches, but it did not make much difference, i.e. my population was usually well-balanced anyway.

Prolific has also made their peak times for experiments available on their website. Because I needed to quickly pair participants for my dialogues, I chose to do small batches on a near-daily basis at the peak times (usually between 10 am and 12 pm Central Standard Time). Over the course of one month, I ran 13 batches of data collection, collecting 10 dialogues in each batch. Usually all 10 dialogues were collected within 45 minutes, since not all participants would join the moment a study was sent out.

Appendix C

Prosodic Parallels in Speech and Typing

Tables C.1 and C.2 below outline prosodic features of spoken language and then describes speech prosody's analogs in keyboard typing.

Feature in speech	Manifestation in speech	Manifestation in typing
Pauses	Pauses are usually only measured between words. Pauses between phonemes within a word are often difficult to impossible to measure. In fact, some phoneme transitions do not have any delimiting characteristics; rather, a speaker produces them in contiguous succession.	Pauses are relatively trivial to measure. Pauses can be measured in many ways, as seen in Figure 2.1. Pauses between intra-word keystrokes are typically measured in the same way as pauses in between-word keystrokes.
Energy/ intensity/ loudness	Speakers consciously and unconsciously choose how much energy to use in producing speech. These choices are usually directly perceivable by the listener.	A typist can choose to alter the visual attributes of their message, such as all capital letters or bold font. Evidence shows that if a typist is (silently) producing more intense language, this is manifested as increased typos, revisions, or longer keypresses (dwell time) (Lee et al., 2015a).
Length of sound/ duration	Syllable lengthening in speech is learned early in development and implemented for many different reasons (Snow, 1994). Measuring or at least comparing syllable duration is relatively robust in speech science.	Repeated letters are employed frequently in typing. Kalman and Gergle (2009) finds evidence for many uses of letter repetition, which parallel uses in spoken prosody.

Table C.1

A comparison of parallel features in spoken prosody and keystroke dynamics (continued in Table C.2)

Feature in speech	Manifestation in speech	Manifestation in typing
Speech rate	Speakers speed up and slow down for a large variety of reasons. The production rate of language, per se, can encode significant information about the intended message.	If a message is only transmitted upon completion, then the typing rate within that message is not necessarily known. If a number of messages are transmitted rapidly, it can be inferred by the receiver that the language is being produced rapidly. A real-time typing environment, which is less common today, would also facilitate awareness of production rate.
Pitch/ fundamental frequency	Humans continuously alter the pitch of their voice, e.g. high tones and low tones. These alterations can convey significant amounts of information about the affective or emotional properties of the speaker.	Aside from inferences drawn by the receiver from altered language production, pitch cannot be conveyed in typing production.
Timbre	Timbre is difficult to define succinctly, but it represents the quality of sound that makes a particular voice have a different sound from another, even when producing the same phoneme.	In CMC, the messaging medium outputs uniform text styling. Hand-written communication, such as shaky or sloppy text, could possibly be considered a parallel for voice timbre.

Table C.2

A comparison of parallel features in spoken prosody and keystroke dynamics (continued from Table C.1)

Appendix D

Experimental iterations

Minor elements of the experiment were improved upon. Because the IRB was ruled exempt from further review (see Section 4.4.1), these minor changes did not require IRB approval. In addition, the changes did not alter the nature of the experiment in any significant way. The following subsections (within Section 4.4.2) are included in order to explain the rationale for certain features of the experiment.

Timer and Timing

One of the most significant changes I made was actually adding a countdown timer to the experiment. Although I had included a "warning" during the experiment saying "One Minute Left," many participants felt this was not sufficient and that they easily lost track of time during the experiment. The pilot study was also only 10 minutes long, and many participants asked for additional time. In a way, I took both of these as positive signs that my experimental prompts were engaging and that a naturalistic conversation was taking place, which fully occupied the attention of the participants.

Adding a timer had the additional benefit of allowing the conversations to follow a more natural trajectory, especially towards the end of the experiment. Since participants could see the amount of time remaining, they could more properly "wrap up" their conversations, rather than being caught off guard when the experiment ended.

Since Study 1 (Chapter 5) looks at dialogue acts, which also facilitate the function of changing the direction of the conversation, it was important to also have data where utterances were intended to conclude a conversation, reflecting on what had taken place, rather than launching in to a new topic.

Age Constraints

A significant change that I made to the experiment setup was removing age constraints after the second batch of data collection. Initially I had limited the experiment to participants between ages 25-40, encapsulating the "Millennial" generation. My initial intention was that by constraining the age range, participants would be more likely to have cultural awareness of the same movies in general, even if they did not share the same preferences.

However, the downside to an age constraint was that it seemed to foster *too much* agreement. In other words, while my experiment was intended to evoke both agreement and disagreement along with positive and negative sentiment, implementing an age constraint led to very little disagreement and negative sentiment. This takeaway was gleaned from participants' comments on my experiment after the pilot study and first two batches of data collection. Almost without exception, every participant said only positive comments about the conversation itself as well as what they thought of their partner. For example, one participant in the pilot study said "[My partner was] very informative and thoughtful. I will absolutely be looking into one of the movies they recommended."

In this situation, the changes I made to my experiment did not come from explicit feedback. Rather, participant feedback was qualitatively analyzed and adjustments were made based on this analysis.

Removing the age constraint did not materially effect my experiment. While in the first two batches, participants could not be more than fifteen years apart in age, throughout the entire data collection process a large number of pairs of participants were within fifteen years of each other. As an illustration of how removing age constraints did not affect the experiment, see Figure 4.3.

However, after collecting batches three and four (with age constraints removed), I tested whether larger age differences had a significant effect on participants' subjective enjoyment of the conversation. If age difference did significantly change the nature of the experiment, then I would not be able to include batches one and two with the rest of the data. As can be seen in Figure 4.3 though, larger age differences did not substantially increase or decrease enjoyment of the conversation.

Explicit Instructions and Prompt Wording Modification

The final change I made came from an online discussion with participants who frequently use Prolific. My initial experimental prompts and instructions were based on previous experiments that were run in-person primarily with undergraduate students enrolled in relevant courses, and participating in experiments as a course requirement. In my case, though, experiments were being run online with a large and diverse population, where participants are primarily financially motivated.

Given the heterogeneity of my participant population, it was also necessary to more explicitly write out my experiment instructions and prompts. As seen in the instructions below, I learned that it was necessary to instruct participants to enter the experiment immediately, since they needed to be paired with a partner at the same time. This is distinct from most online experiments which are asynchronous and involve stimuli much as surveys that can be taken at any time. Most likely this instruction limited who participated in my experiments, since they needed to have 15-20 minutes available at the moment the experiment was sent out.

In addition, many online participants take part in experiments as an additional source of income, and therefore want to complete as many experiments as possible as quickly as possible (Chandler and Kapelner, 2013). On the other hand, an undergraduate coming into a lab is prepared to begin an experiment immediately, and is not trying to maximize their earnings by taking part in multiple experiments. For this reason I added into both the initial instructions below as well as in the conversational prompts that participants should have fun, and that disagreements are completely acceptable.

I also observed that the pauses between utterances in my pilot study were unusually long in some places. It is also possible that online participants will be undertaking multiple experiments at one time on other platforms, and thereby take advantage of the conversational nature of my experiment by taking long pauses to either rest or work on other studies. For this reason I also added to my prompts, “Please make sure to make FULL use of ALL 8 minutes. Keep the conversation active and lively...” a a passive-aggressive way to discourage slow responses.

Appendix E

Study 1 details

Study 1a details

Model building

As a base model, I looked at the typing patterns of the entire utterance, where the model was calculated to predict whether the dialogue act was a Non-opinion or Opinion statement. The base predictors were chosen because they seemed more fundamental to the typing process, and some also had more obvious spoken prosody parallels. The base predictors were:

1. Utterance typing speed (keystrokes/utterance duration)
2. Utterance average inter-keystroke interval (IKI)
3. The interaction of speed and IKI
4. The pause between the previous message being sent and the beginning of the current utterance (pre-utterance gap)
5. The edit count (pressing BACKSPACE or DELETE)

The reason that the interaction of typing speed and IKI is included is because speed is more sensitive to word length, whereas IKI is not. While in future work I will find a more precise way to measure typing rate, it seems that the interaction of the two terms, and the variance absorbed by this interaction, improve the predictive power of the other predictors.

In model testing (see Table E.1), the pre-utterance pause, utterance speed, and edit count predictors were found to significantly predict whether the utterance was a Non-opinion or Opinion ($p < 0.05$). Because standardization was performed by-subject, the model coefficients are less meaningful and do not directly represent any definite "unit" of measurement. They are similar to a z-score, though. The model showed that the pre-utterance gap was shorter for Non-opinions, typing speed was faster when typing Opinions, and that opinion utterances contained significantly fewer edits. Using the variance inflation factor (VIF) test, none of the predictors exhibited collinearity.

The model was then extended to see if in addition to the average speed of typing, the variability of typing speed was a significant factor. To accomplish this, the standard deviation of typing speed was added as a predictor. Importantly, I only looked at the standard deviation of the IKI at the beginning of the word. This was because that gap reflects lexical recall, which could be expected to vary between different types of dialogue acts. On the other hand, intra-word intervals reflect motor skills, which are expected to be more consistent across a typing session (Logan and Crump, 2011).

While the additional variability predictor was not significant in and of itself ($p = 0.45$), adding variability increased the significance of the other predictors. In addition, the log-likelihood and BIC of the model also improved. The coefficient showed that variability is greater when typing an opinion utterance than when typing a statement.

In addition to typing patterns spanning the entire utterance, an additional goal was to see if typing patterns in the initial part of an utterance are also sensitive to dialogue act type. A new model was tested that incrementally added: the typing speed of the first word, the typing speed of the second word, and the duration of the gap between the first two words. Without any utterance-spanning predictors, none of these factors improved the model's accuracy or fit above a baseline model. However, the interaction of the gap before and after the first word was significant ($p = 0.05$), and the addition of the terms improved the fit of the model as well as the BIC. The rate at which the first and second words were typed did not improve the predictive power of the model.

One of the most surprising findings (not reported) was that adding a crossed or nested random effect of last sender did not improve the model. I had assumed that an utterance, especially a pre-utterance pause, would be effected by whether the message was a response to the partner, or a continuation of a turn by the same sender. The fact that conversation position explained so much variance is one reason for to the necessity of Study 2, which will look at the trends of dyadic pairs during the progression of a conversation.

Study 1b details

See the tables below for details of each model.

	Dependent variable: Non-opinion vs Opinion				
	Model				
Covariate	Base (Fixed)	Base (Mixed)	+ Typing Variability	+ Gaps	+ Word Speeds + intx
(Intercept)	-0.54^{***} (0.04)	-0.54^{***} (0.05)	-0.54^{***} (0.05)	-0.54^{***} (0.05)	-0.54^{***} (0.05)
Pre-utterance gap	0.05^{**} (0.03)	0.05^{**} (0.03)	0.05^{**} (0.03)	0.05^{**} (0.03)	0.05^{**} (0.03)
Interkey-interval (IKI)	0.07 (0.05)	0.08^{*} (0.05)	0.07 (0.05)	0.07 (0.05)	0.06 (0.05)
Utterance speed	0.07^{**} (0.04)	0.08^{**} (0.04)	0.08^{**} (0.04)	0.08^{**} (0.04)	0.08^{**} (0.04)
IKI:speed	0.02 (0.02)	0.02 (0.02)	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)
Edit Count	-0.01^{**} (0.004)	-0.01^{**} (0.004)	-0.01^{**} (0.004)	-0.01^{**} (0.004)	-0.01^{**} (0.004)
Speed variability (sd)			0.02 (0.03)	0.02 (0.03)	0.02 (0.03)
Word 1-2 gap				0.002 (0.03)	0.001 (0.03)
Pre-utt-gap:word 1-2 gap				-0.05^{*} (0.03)	-0.05^{*} (0.03)
Word 1 speed					-0.03 (0.03)
Word 2 speed					-0.004 (0.03)
Word 1:word 2 speed					-0.04 (0.03)
Observations	2,965	2,965	2,965	2,965	2,965
Log Likelihood	-1,744.21	-1,742.88	-1,742.66	-1,740.75	-1,739.17
AIC	3,500.42	3,499.76	3,501.32	3,501.50	3,504.33
BIC		3,541.73	3,549.27	3,561.44	3,582.26
Note:					*p<0.1; **p<0.05; ***p<0.01

Table E.1

Effects of adding covariates to a model predicting non-opinion vs opinion binary dialogue acts

	Dependent variable:
	Utterance speed
(Intercept)	−0.07
	(0.05)
Acknowledge	−0.15*
	(0.06)
Closing	0.10
	(0.10)
Directive	0.39*
	(0.16)
Negative-Answer	−0.30^{+}
	(0.17)
Non-opinion	0.07
	(0.05)
Opening	−0.53**
	(0.17)
Opinion	0.15**
	(0.05)
Question	0.26***
	(0.05)
Word count	−0.002
	(0.002)
Observations	4,108
R^2	0.02
Adjusted R^2	0.01
Residual Std. Error	0.95 (df = 4099)
F Statistic	8.55*** (df = 8; 4099)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.2

Results of whether dialogue acts controlling for word count can predict the speed at which the utterance was produced.

	Dependent variable:
Variable	Edit count
(Intercept)	0.64^{*}
	(0.27)
Acknowledge	0.37
	(0.36)
Closing	−0.46
	(0.60)
Directive	−0.84
	(0.90)
Negative-Answer	−0.05
	(0.99)
Non-opinion	−0.45
	(0.27)
Opening	2.14^{*}
	(0.99)
Opinion	−0.48
	(0.30)
Question	−0.22
	(0.31)
Word count	0.34^{***}
	(0.01)
Observations	4,108
R^2	0.24
Adjusted R^2	0.24
Residual Std. Error	5.44 (df = 4099)
F Statistic	164.62^{***} (df = 8; 4099)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.3

Results of whether dialogue acts can predict edit counts, controlling for word count.

	Dependent variable:
Variable	Typing speed variability (sd)
(Intercept)	-0.12^{**}
	(0.05)
Acknowledge	0.17^{**}
	(0.06)
Closing	-0.16
	(0.11)
Directive	-0.24
	(0.16)
Negative-Answer	0.16
	(0.18)
Non-opinion	-0.03
	(0.05)
Opening	0.27
	(0.18)
Opinion	-0.03
	(0.05)
Question	-0.13^{*}
	(0.06*Z*)
Word count	0.01^{***}
	(0.002)
Observations	4,108
R^2	0.02
Adjusted R^2	0.02
Residual Std. Error	0.96 (df = 4099)
F Statistic	12.37^{***} (df = 8; 4099)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.4

Results of whether dialogue acts predict variation in typing speed, when controlling for word count.

	Dependent variable:	
	Pre-utterance gap	
Variables	Base model	+ last sender
(Intercept)	0.09^{*}	-0.11^{*}
	(0.05)	(0.05)
Acknowledge	-0.01	0.09
	(0.06)	(0.06)
Closing	0.02	0.04
	(0.10)	(0.10)
Directive	-0.08	-0.18
	(0.16)	(0.15)
Negative-Answer	-0.05	0.04
	(0.17)	(0.17)
Non-opinion	-0.05	-0.08^{+}
	(0.05)	(0.05)
Opinion	0.04	-0.003
	(0.05)	(0.05)
Question	0.12^{*}	0.09^{+}
	(0.05)	(0.05)
Word count	-0.01^{***}	-0.003
	(0.002)	(0.002)
Last sender		0.49^{***}
		(0.03)
Observations	4,069	4,069
R^2	0.01	0.06
Adjusted R^2	0.01	0.06
Residual Std. Error	0.97 (df = 4061)	0.95 (df = 4060)
F Statistic	6.02^{***} (df = 7; 4061)	34.64^{***} (df = 8; 4060)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001	

Table E.5

Results of whether dialogue acts can predict pre-utterance gaps, when controlling for word count, and word count + who the last sender was.

	Dependent variable		
Variables	Word 1 speed	Word 2 speed	Word 1-2 gap
	(1)	(2)	(3)
(Intercept)	−0.05	0.03	−0.04
	(0.05)	(0.05)	(0.05)
DA:Acknowledge	−0.05	-0.17^{**}	0.13^{*}
	(0.06)	(0.06)	(0.06)
DA:Closing	−0.04	0.20^{+}	-0.19^{+}
	(0.11)	(0.11)	(0.11)
DA:Directive	−0.05	0.10	−0.01
	(0.16)	(0.16)	(0.16)
DA:Negative-Answer	−0.12	0.14	0.02
	(0.18)	(0.18)	(0.18)
DA:Non-opinion	0.14^{**}	0.01	−0.01
	(0.05)	(0.05)	(0.05)
DA:Opening	−0.01	-0.44^{*}	0.11
	(0.18)	(0.18)	(0.18)
DA:Opinion	0.10^{+}	0.02	0.002
	(0.05)	(0.05)	(0.05)
DA:Question	0.02	0.13^{*}	−0.05
	(0.06)	(0.06)	(0.06)
Word count	−0.002	−0.0004	0.004^{*}
	(0.002)	(0.002)	(0.002)
Observations	4,108	4,108	4,108
R^2	0.005	0.01	0.005
Adjusted R^2	0.003	0.01	0.003
Residual Std. Error (df = 4099)	0.98	0.97	0.96
F Statistic (df = 8; 4099)	2.43^{*}	3.90^{***}	2.36^{*}

Note: + p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.6

Results of whether dialogue acts can predict the typing speed of word 1, word 2, and the length of time between words 1 and 2.

Appendix F

Manual Sentiment Analysis Guidelines

These guidelines were compiled by my research assistant, based on the guidelines set forth in Mohammad (2016).

1 - Negative

- There is an explicit or implicit clue suggesting that the speakers is conveying a negative opinion or is in a negative state
- Displeasure, anger, frustration, irritation, dislike, etc.

2 - Somewhat Negative

- There is evidence to suggest that the speaker is conveying a slightly negative opinion or is in a negative state, but is not conveying strong negativity
- On the cusp of negative feelings listed above
- Not a neutral statement due to some suggestion of negativity, but not outright negative in sentiment

3 - Neutral

- The statement is neither explicitly positive or negative. No emotional state or opinion is conveyed through the statement

4 - Somewhat Positive]

- There is evidence to suggest that the speaker is conveying a slightly positive opinion or is in a positive state, but is not conveying strong positivity
- On the cusp of positive feelings listed below
- Not a neutral statement due to some suggestion of positivity, but not outright positive in sentiment

5 - Positive

- There is an explicit or implicit clue suggesting that the speakers is conveying a positive opinion or is in a positive state
- Pleasure, optimism, joy, relaxed, admiring, like/interest, excitement, etc.