A Practical Guide to Conversation Research:

How to Study What People Say to Each Other

Michael Yeomans[1], F. Katelynn Boland[2], Hanne K. Collins[3], Nicole Abi-Esber[3]

& Alison Wood Brooks[3]

Corresponding Author: Michael Yeomans (m.yeomans@imperial.ac.uk)

[1]Imperial College London

[2] Columbia Business School

[3] Harvard Business School

Abstract

Conversation—a verbal interaction between two or more people—is a complex, pervasive, and consequential human behavior. Conversations have been studied across many academic disciplines. However, advances in recording and analysis techniques over the last decade have allowed researchers to more directly and precisely examine conversations, in natural contexts and at a larger scale than ever before, and these advances open new paths to understand humanity and the social world. Existing reviews of text analysis and conversation research have focused on text generated by a single author (e.g. product reviews, news articles, and public speeches), and thus leave open questions about the unique challenges presented by interactive conversation data (i.e., dialogue). In this article, we suggest approaches to overcome common challenges in the workflow of conversation science, including recording and transcribing conversations, structuring data (to merge turn-level and speaker-level datasets), extracting and aggregating linguistic features, estimating effects, and sharing data. This practical guide is meant to shed light on current best practices and empower more researchers to study conversations more directly—to expand the community of conversation scholars and contribute to a greater cumulative scientific understanding of the social world.

Keywords: Natural language processing; text analysis; conversation; social interaction; open science

1. Introduction

Conversation is one of the most pervasive of all human behaviors—people talk to each other all the time, all over the world (Dunbar, Marriott & Duncan, 1997). Most interpersonal relationships develop through a series of conversations over time—time spent talking and not talking, together and apart. Though a frequent and familiar task, each conversation is complex—it requires (and enables) people to coordinate their behavior and beliefs about the world (Clark et al., 2011; Misyak et al., 2014; Jaques et al., 2019; Rossignac-Milon et al., 2021). Conversations are consequential, allowing people to pursue a wide array of informational and relational goals (Yeomans, Schweitzer & Brooks, 2022) in the short term and over the long term—spanning each individual conversation and longer-term relationships (Fitzsimons & Finkel, 2018). Indeed, the amount and quality of social interaction is one of the most enduring predictors of human well-being (Collins et al., 2022; Diener & Seligman, 2002; Epley & Schroeder, 2014; Mehl, Vazire, Holleran, & Clark, 2010; Sun, Harris, & Vazire, 2019; Quoidbach et al., 2019).

It is no surprise that researchers are increasingly interested in studying conversations, the contextual factors that surround them, and the short- and long-term effects of having them. This practical guide argues for the relevance of this work *now*, the benefits and challenges researchers should expect from studying conversations, and how to analyze conversation data, pair transcripts with surveys, and share results, as we move toward a cumulative science of conversation (see Figures 1 and 2).

1.1 Why now?

At least three developments have enabled a recent boom in conversation research. First, conversations have become increasingly mediated through technology, as a consequence of the Digital Revolution and Information Age of the 20th century and the social media era of the 21st

century (Rainie, & Wellman, 2012), shifts that were accelerated during to the COVID-19 pandemic (Nguyen et al., 2020). These mediated communication technologies allow for the recording of text, audio, and/or video, and thus preserve a rich source of conversation data for analysis. Second, there have been many advances in natural language processing (NLP)—an interdisciplinary subfield at the intersection of linguistics, computer science, and artificial intelligence that seeks to learn, parse, and understand human language content using quantitative techniques (Hirshberg & Manning, 2015). This field develops computational tools that turn raw conversations into behavioral data—words into numbers—especially at scale (Jurafsky & Martin, 2017). Finally, the value of larger-scale analyses has been underscored by the recent revolution in research practices (Nelson, Simmons, & Simonsohn, 2018). Taken together, these cultural and methodological developments offer wide promise for the study of conversation across a variety of academic disciplines.

1.2 Measuring Conversation

Although conversations are common and consequential, they are also complicated—no two are identical. Researchers have dealt with the complexity of conversation with a wide range of approaches aimed to simplify and isolate different aspects of a conversation. In exchange for simplicity, these approaches can make conversations less natural and more abstract. For example, researchers often study dialogue indirectly by having participants: talk to a trained confederate, respond to hypothetical vignettes, make evaluations of carefully-selected transcript segments, recall a previous conversation from memory, or offer holistic evaluations of a conversation after it is over. These approaches constitute creative and generative ways to study conversations and were particularly useful when conversation technology was nascent.

These approaches allow researchers to simplify and study conversations, but they also suffer from several well-known biases. For instance, confederate simulations rely on faithful execution of researchers' instructions; hypothetical and recall methods suffer from errors in forecasting and memory; self-report measures suffer from social desirability bias, hindsight bias, and demand effects; and experimenter-generated stimuli remove the conversational context in which they would occur in the real world. Conversation is a complex, contextual, and improvisational environment, and these kinds of simplifications can result in a misunderstanding between the assumed, perceived, and actual goals and psychological experiences of the speakers (Stokoe, 2021).

On the other hand, many researchers have taken on the daunting task of studying natural, contextualized conversational behavior, beginning with study of "ordinary language" as early as the mid-twentieth century (e.g., Garfinkel, 1956; Schegloff, 1968; Schegloff & Sacks, 1973; Sacks, Schegloff, & Jefferson, 1974; Goffman, 1981; Pomerantz, 1990; Heritage, 2008; Stivers et al, 2010; Stivers & Sidnell, 2012). This work has typically prioritized attention to descriptive detail in natural settings by scrutinizing isolated portions of transcripts, at the expense of scalability and controlled measures of outcomes and effects. Further, linguistic inquiry often assumes rationality on the part of speakers (e.g., Grice, 1975; Misyak, Mekonyan, Zeitoun & Chater, 2014; Goodman & Frank, 2016), and infers intent based on outcomes. This assumption can be limiting. People constantly deviate from rational behavior (Kahneman, 2002), so it is important to measure both intentions and outcomes to see whether speakers are making wise choices, enacting good behaviors, or making mistakes (Yeomans, Brooks, & Schweitzer, 2022).

More recent work has conceptualized conversation as a diagnostic window into variables like health status, personality, and well-being (e.g., Robbins et al., 2011; Conner & Mehl, 2015;

Collins et al., 2022; Collins et al., 2018; de Barbaro, 2019; Jaidka et al., 2020). This simplification abstracts away from the details of each particular conversation and focuses instead on person-level variables. The focus is on what speakers' behavior says about themselves, rather than the effects of their behavior on their partner, and ignores the specific goals and outcomes of individual conversations.

Many prior papers have compared conversation behavior to context and outcome data (e.g., Word, Zanna, & Cooper, 1974; Weingart et al, 2004). But this work usually relies on human annotation to quantify conversational behavior from recordings or transcripts. While these insights are useful, they are costly to scale, and often do not give a transparent or interpretable definition of how a measure is calculated (see Section 4.2 below for more on this point). Similarly, speakers are often asked to quantify the content of the conversation themselves, using retrospective survey measures. Again, these measures are convenient but opaque, and suffer from the same self-report and memory biases of other survey methods.

In this paper, we highlight how recent technological advances provide researchers with novel capabilities to combine the best aspects of these research approaches, and directly measure conversation behavior in more natural contexts at scale. The tools for conversation science are rapidly improving—both for *recording* conversations and for *analyzing* them, leading to an emerging boom of conversation research in a wide range of contexts, across a wide range of academic disciplines (see Table 1 for a review). Modern workflows have made it easier than ever for researchers to combine detailed transcript analysis with algorithmic tools to scale up their insights and obtain robust measures of context and outcome variables surrounding conversational choices.

In this practical guide, we aim to make data, tools, and methods more accessible to a wider group of researchers by describing common challenges that face behavioral researchers who wish to study conversations, and by suggesting approaches that address those challenges. This review is aimed at researchers across disciplines who are looking to incorporate conversation research methodology into their work for the first time or expand on a body of conversation research by incorporating new methods and techniques.

1.3 The Scope of this Paper: A Focus on Transcripts

Conversations can include a wide array of psychological and behavioral content, including verbal features—what words are uttered, by whom, and in what order—as well as nonverbal features—tone of voice, gesture, posture, facial expressions, and so on (via visual and/or audio inputs). We focus primarily on *verbal* content for three key reasons. First, every conversation includes verbal content, whereas nonverbal cues are not present in many conversations (e.g., emails and phone calls). Second, verbal content presents common challenges for conversation research, no matter what other types of cues are also present. The decisions, beliefs, and consequences that stem from the verbal content of conversation are only beginning to be rigorously understood. Finally, while nonverbals can inflect the meaning of the words spoken, it is the words themselves that form most of the meaning: they define the topics of conversation and what is being said about them. Indeed, verbal content has an overwhelming effect on how nonverbals are interpreted (Lapakko, 1997).

For these reasons, the scope of this paper focuses on the aspects of conversations that can be captured in a transcript. This includes conversations conducted through sound, and through

writing.[1] Transcript data primarily includes all words and phrases uttered by the speakers, the relative order and timing in which they are produced, and who produced them. Additionally, transcript data can include some paralinguistic features (e.g., laughter, back-channel feedback like "yea" "uh huh", and disfluencies like "um," "uhhh"). Likewise, written conversation sometimes includes features intended to represent nonverbal information (e.g., emojis or emoticons).

However, this scope excludes data that are present in many types of conversations. Primarily, this excludes paralinguistic (acoustic) information, such as the tone, pitch, and volume of voice, as well as visual nonverbal information, such as the speakers' facial expressions, hand gestures, and body posture. We also focus almost exclusively on monolingual English conversations, as the complexities of conversing in two or more languages simultaneously are too manifold for us to address properly herein. Most common NLP tools are available in many languages, though in cases where researchers are studying dialogue from under-resourced languages or other complex sources (e.g., slang, jargon, multiple languages at once), they may want to rely more heavily on expert human annotation.

In cases where these other sources of information are relevant to the research question, we urge researchers to take a more tailored approach, rather than rely only on our simplified workflow. For example, we note that these other types of conversational content can be added or annotated within transcript data (which we address in section 3) and can be quite important in some cases.

[1] For simplicity, we will use the term "speaker" to refer generically to all conversation participants throughout.

As a final statement of scope, we have also avoided research on dialogue generation (i.e., building models that can converse autonomously, sometimes called "dialogue agents," "dialogue systems," or "chatbots"). Transcripts of conversations that include bots can be analyzed in essentially the same way as conversations that only include humans, but building the chatbots themselves is more arduous. A practical reason to avoid this topic is because it is an especially fast-moving field. For example, between our initial submission of this paper and its final acceptance, ChatGPT was released (OpenAI, 2022), followed by a rush of similarly impressive language generation models. Although the future remains uncertain, we do anticipate that novel and enhanced models will emerge and become available in the coming years, and will become increasingly important in the field of conversation analysis.

At this point in time, effective chatbots in the real world tend to be task-specific (e.g., customer service phone trees, smart home assistants), or serve narrow roles, such as a conversation facilitator for human speakers (e.g., Adamson et al., 2014; Traeger et al., 2020). When chatbots participate in broader conversations, they often have problems with listening, consistency, factuality, and other basic skills, although this may improve in the near future (Huang, Zhu, & Gao, 2020).

2. Leveraging the Predictable Structure of Conversation

Conversation is constructed jointly by (at least) two people, each of whom has their own independent goals, preferences, beliefs, perceptions, traits, and choices, often intertwined in an interdependent relational system (Fitzsimons & Finkel, 2018; Yeomans, Schweitzer & Brooks, 2022). In light of the difficult coordination puzzle that conversation presents, it is a wonder that humans manage to communicate at all. Remarkably, they do figure out how to understand each other (Grice, 1975; Misyak, Mekonyan, Zeitoun & Chater, 2014; Goodman & Frank, 2016). In

fact, the predictable and intuitive structure of conversation—a pattern humans learn to recognize and produce from a very young age—facilitates information flow between speakers. The raw data of conversation is carefully structured by the participants themselves. For example, conversation partners alternate turns, jointly establish topics as a common frame of reference, and ask and answer questions (Pickering & Garrod, 2004; Schegloff, 2007).

However, from a researcher's perspective, conversational transcript data is difficult to analyze quantitatively, involving many steps (see Figure 1). First, sounds sometimes need to be converted into words (e.g., "uh-huh" or "[laughter]"). Then, all words need to be arranged in sentences and turns. Once the transcript is generated, researchers will notice how conversation data is high-dimensional—no two conversations are exactly alike. Within one conversation, every possible turn branches into an exponentially large decision tree containing what could be said next, in quick, recursive cycles across multiple speakers. Though researchers can take advantage of the predictable aspects of conversational structure, they must also sift through the exponential complexity—they must make many judgment calls to determine which features are counted, and how to do so from raw text.

2.1. The Distinctiveness of Dialogue v. Single-Voice Text

For our purposes, conversations consist of dialogue[2] generated between two or more people over a series of turns. This definition primarily serves to distinguish conversations from documents authored from a single perspective, including speeches, essays, newspaper and magazine articles, books, product reviews, legal documents, social media posts. While the "great bulk" of language use is conversational (Levinson, 2016), single-voice documents have been the dominant source material in applied text analysis, and many review papers in related fields have

[2] To clarify a common misconception, dialogue refers to any number of speakers—not just two. It is derived from the prefix "dia", meaning "through" (e.g., diagonal), not the prefix "di", meaning "two" (e.g., dichotomy).

focused only on single-voice documents (e.g., Pennebaker, Mehl & Niederhoffer, 2003; Grimmer & Stewart, 2013; Hirschberg & Manning, 2015; Berger et al., 2020; Benoit, 2020; Gentzkow, Kelly & Taddy, 2019; Boyd & Schwartz, 2021; Jackson et al., 2021). Many of the techniques developed for single-voiced documents are also useful for studying conversations. However, the differences between single-voiced text and dialogue motivate different ways that researchers should capture and analyze conversation data.

First, unlike single-voiced text, conversations include multiple interchanging contributors. Each person's contribution to the full conversation must be disambiguated (e.g., who said what?). Second, conversations are generated on the spot, and responsively, which puts a special priority on understanding the *sequence* of what is said, when it is said, and how it relates to adjacent conversational turns. Third, conversations are usually less thoroughly edited than single-voice documents, often because the turns are spontaneously composed. This means conversation entails looser sentence structure, as well as breakdowns in the coordination of common ground, including more interruptions, cross-talk, silence, repairs, repetitions, misarticulations, clarifications, back-channels, conflicts, slurs, and jargon (Fox Tree, 2010). Lack of editing means conversations tend to have more spelling and grammatical errors, as well as disfluencies (e.g., "umm", "uh-huh"). Fourth, conversation often covers many topics and goals (Yeomans, Schweitzer & Brooks, 2022), whereas most single-voiced documents focus on one or a small number of topics (e.g., product reviews or news articles). These complications of conversation pose many novel challenges (and opportunities) for researchers, even for those familiar with text analysis of single-voiced text data.

2.2 Managing Conversation Datasets: Analyzing Turn-Level and Speaker-Level Data Simultaneously

Managing conversation data requires researchers to handle two distinct datasets: a turn-level dataset (to examine conversational behavior) and a speaker-level dataset (to compare that conversational behavior to pre- or post-conversation data, such as individual differences, experimental conditions, or outcomes). Thus, conversation analysis calls for analytical software (such as R or Python) that allows researchers to efficiently manage multiple datasets at once.

2.2.1 Turn-Level Dataset: The Contents of Conversation. Most conversations can be discretized into a series of speaker "turns," much like a screenplay or script. This data can be represented as a transcript, where each row contains information about a single turn—specifically, who was speaking, the words spoken during the turn, and timestamps indicating when the turn started and ended. This data structure requires the turn-level dataset to have unique identifiers for every conversation or group (e.g., group 1, 2, 3, 4…), every turn in each conversation (e.g., turn number 1, 2, 3, 4, 5…), *and* each speaker in the group (e.g., speaker A, B, C…). We provide an example of a turn-level dataset in Table 2.

In general, the boundaries of each turn are determined by the time during which a single speaker is talking. Every new turn will involve a different speaker than the turn prior. Linguists distinguish the concept of a turn from that of an "utterance," defined as a single continuous expression by a speaker. A turn can be composed of multiple utterances. For example, a speaker could send several messages in a row before their partner responds. In that case, as a simplifying assumption, researchers typically collapse multiple consecutive utterances from a single speaker into a single turn.

2.2.2 Speaker-Level Dataset: Data from Outside the Conversation. In a speaker-level dataset, there is a unique row for each speaker. Each conversation will have multiple rows (one for each speaker in that conversation) and speakers who joined multiple conversations will have

multiple rows (one for each conversation). The unique identifiers for the conversation (or group) and speaker included in the turn-level dataset can be used to connect the speakers' conversational behaviors to the speaker-level dataset, which also contains the conversation (or group) and speaker identifiers, in addition to other variables recorded before the conversation (e.g., random assignment(s), time of day, context, demographics) and after the conversation (e.g., self-reported survey items, negotiated outcomes). We provide an example of a speaker-level dataset in Table 3.

Many researchers will conduct their final analyses in the speaker-level dataset since many research questions focus on variation at the person or context level. When this is the case, the turn-level dataset is used to generate measures of conversational behaviors (e.g., the number of questions or interruptions), which are then summarized at the person-level, and tallied in the speaker-level dataset (e.g., Speaker A in group 4 asked 41 questions, 5 hedges, and interrupted 3 times during the conversation). We provide further detail on this topic in Section 5.

3. Capturing Conversation Data

There are considerable challenges involved in coercing conversation data into the datasets described above, and they vary based on modality. We focus on the two most common conversational modalities, in which words are either written as text or spoken out loud. Each of these major modalities presents unique challenges and opportunities for speakers and researchers (e.g., Berry, 2013; Boland et al., 2021; Meredith & Stokoe, 2014).

In either case, the fixed cost of structuring a conversation dataset is not trivial. Once it is done, a good dataset can benefit many subsequent research projects (and, possibly, many different researchers). Thus, we encourage researchers to explore whether it is possible to pilot test their research ideas in datasets from past research, including in archives purpose-built for

conversation data (e.g., Liberman & Cieri, 1998; Miller et al., 2017; Chang et al., 2020; Reece et al., 2022). For similar reasons, we also encourage researchers to share their own data after they have structured it (see Section 6 below).

3.1. Text-Only Conversations

Research on text-only conversation has proliferated in part because of the availability of text data, which is easy to record and store. It is often produced in massive internet forums (e.g., Wikipedia, Twitter) or in catalogued archives (e.g., newspaper articles, books, legal documents, earnings calls) where records are public and accessible to researchers (Hirschberg & Manning, 2015), or scraped using one of many available software tools that can scrape text content from webpages. Additionally, people often have records of conversations conducted by chat or email. Accordingly, some researchers use software that allows consenting participants to extract and share their own text or social media conversations (e.g. Stillwell & Kosinski, 2004). Researchers also collect their own text conversations within controlled experiments, with emerging technologies like ChatPlat (www.chatplat.com; Huang et al., 2017), iDecisionGames (www.idecisiongames.com), Smartriqs (Molnar, 2019), and survconf (Brodsky et al., 2022).

Text conversation can be easier to analyze than spoken conversation because the words are already transcribed during the conversation itself (by the speakers). The style of conversation conducted via text is also different; compared to voice conversation, text-only conversation tends to be more asynchronous, with more time for cognitive preparation, reflection, and processing within and between turns, clearer sentence structures, and fewer disfluencies (Berry, 2013; Meredith & Stokoe, 2014). Still, text conversation data presents unique challenges for researchers.

3.1.1. Turn Boundaries. The time course of text-only conversation can be tricky to pinpoint, as transcripts often only include one timestamp per turn: when a message is "sent" or "posted." If the conversation is more synchronous (e.g., instant messages), the lag time between these stamps may be a useful signal of the time spent reading the last message or composing the next one. If the conversation is more asynchronous (e.g., email), the lag time may not be as informative.

Additionally, in text conversations, people can compose their turns simultaneously, which can lead to multiple disjointed threads. When topics overlap, researchers must disentangle them by hand (or else accept some measurement error). Further, most text platforms allow a single person to send multiple messages in a row, essentially replying to themselves. This can be simplified by combining consecutive messages from the same person into discrete, alternating turns—or by considering each message as separate turns.

3.1.2. Standardizing Typing. In text-based conversation, people type their own transcripts. Writing style differs across people, cultures, languages, and time, and spelling and grammatical errors are common. There is a range of unique spellings in modern written language, including emojis (e.g., ":-)"), variants (e.g., "oh nooo", "woot!"), representations of sounds (e.g., "jajajaja", "haha") and acronyms (e.g., "tbh", "lmk", "lol", "tldr", "wtf").

In many analyses, variants are simply ignored, especially if they are rare. However, some research questions might require attention to variants (e.g., grouping different kinds of typed laughter, or unpacking emoji valence to detect emotional sentiment). Clear writing errors can be more pernicious, as most feature extraction systems rely on correct spelling and grammar. To address this, we strongly recommend that a person looks through each text at least once, perhaps assisted with spell-checking software, to fix obvious errors.

3.2. Voice Conversations

Research on spoken conversations usually requires additional steps because spoken words are expressed in continuous sound waves that must be discretized into words, sentences, and turns. Some high-stakes audio conversations are routinely transcribed (e.g., interviews, conference calls, government proceedings) and some academic papers examine such documents (e.g., Berry et al., 1997; Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012; Chen et al., 2018; Hansen, McMahon, & Prat, 2018). However, the burden of accurately transcribing conversations often falls on researchers themselves. With technological advances, automatic speech recognition (ASR) and speaker disambiguation have improved (Park et al., 2022), but they are still not nearly as good at parsing speech as human transcribers (Errattahi et al., 2018; Meier et al., 2021), and this is likely to remain true for some time. Furthermore, these automated tools are often trained on convenience data samples, so they may be most inaccurate for speakers from underrepresented groups, that may use an accent or vocabulary that is not well-represented in the training data (Dehghani et al., 2015).

We urge researchers to put serious effort into assuring data quality, both through preparation before the conversations happen *and* after they have been recorded. Here, we suggest a series of steps and several tips to capture research-quality voice conversations.

3.2.1. Record. Researchers often underestimate the importance of audio recording quality. This is especially critical when researchers have complete control over the recording protocol (e.g., recording participants speaking to each other inside a behavioral lab). However, there are cases where researchers have less control, for example the Electronically Activated Recorder (EAR; e.g., Mehl et al., 2001; Mehl, 2017; Kaplan et al, 2020), the Language Environment Analysis System (LENA; Ganek & Eriks-Brophy, 2016), and other experience-

sampling methods that require people to carry microphones with them throughout the day. Further, online experiments may have people conversing through their own home computers, which researchers do not have control over. Nevertheless, each of these protocols involves different considerations and constraints to optimize audio quality. Across all these study designs, we urge researchers to test their recording set-up in advance.

High-quality audio recordings will lead to higher quality transcriptions later. If you are having trouble hearing words when listening to a recording, your transcriber (human or ASR) will certainly struggle. Make sure you can clearly identify what words are being said, and by whom. Some of the main factors to consider include: microphone quality (such as sensitivity, internally generated noise, distortion and directional characteristics), speaker clarity, background noise, distance from the microphones, and reverb. Ideally, researchers should rely on solutions that do not place a burden on the speakers; for example, a change in microphone placement will be a more reliable fix than asking speakers to enunciate more clearly.

One common decision point for researchers is the number of audio recordings per conversation: should the entire conversation be captured in one file, or should each individual be recorded separately? A single recording may seem easier to set up, but may complicate the analysis later, as audio transcription services often struggle with speaker differentiation, especially when two speakers have similar-sounding voices. With only a single recording, transcribers must determine whether the person talking is (a) different from the previous turn (did the speaker change?), and (b) the same as any of the previous turns (has this person spoken before?). This task is especially difficult when speakers have similar speaking styles, vocal registers, and as the number of speakers increases. Video recordings can help, although we have found that professional human transcription services often do not look at videos.

When possible, we recommend collecting separate audio recordings for each speaker. This makes speaker differentiation simple and improves audio quality by moving microphones closer to each speaker. Fortunately, virtual meeting services (e.g., Zoom) record separate audio streams from each computer, which automatically differentiate speakers (if each person has their own computer). Some services automatically combine these separate streams into a single turn-by-turn transcript (including Zoom and Microsoft Teams). If separate recordings are set up manually, they must then be combined and sorted into the correct order using the timestamps for each turn.

In order to connect the speaker-level data to the data collected outside the conversation (e.g., demographics and survey data), each speaker and each conversation must have a unique identifier that can be used to link the turn-level and speaker-level datasets. As a safety measure, researchers may consider reading the conversation identifier out loud at the beginning (or end) of the audio recording, and use the identifier as the name of the audio file as well. Likewise, each speaker in the conversation should say their unique speaker identifier as one of their first turns in the recording, so their voices can be unmistakably matched to their conversation-level and speaker-level data.

We recommend conducting a few test recordings, which run through as much of the workflow as possible. The researchers should check to see that the file records well (that the audio is clear, and that the spacing of microphones and speakers is appropriate), that it can be played back properly, that it is saved in a format that is compatible with the intended transcription method, and that the researcher can match each recording and each speaker to the metadata. Finally, don’t forget to press “record.”

3.2.2. Transcribe. Transcriptions of the audio recordings will form the foundation of the turn-level dataset. There are several approaches to generate transcriptions from audio files. Most commonly, researchers pay traditional transcription services, which hire trained humans to type words while they listen to audio recordings. However, this approach is often inadequate (and expensive)—the quality is inconsistent, typos are inevitable, and transcribers use different formatting methods (even within the same company). Some researchers hire research assistants to transcribe. Although this affords more control over formatting, the training can be long, and the work can be arduous and inefficient. Others use automated speech recognition software. While software will never be as accurate at recognizing words as the best trained humans, they produce the most precise timestamps, and they deliver consistent formatting and spelling.

We strongly recommend a hybrid approach, combining automated speech recognition software with trained humans, which is both accurate and cost-effective. First, automatic speech recognition software can generate a low-cost first draft transcription, tackling the easiest sections of the transcript quickly and producing transcripts with consistent formatting and reliable timestamps. Then, this initial draft of the transcript can then be edited by a human, who can focus their time and attention on the more difficult tasks, such as speaker differentiation and correcting any passages with low-quality audio.

It is important to establish consistent formatting conventions early. Many transcription services (human and machine) export their data in text documents (e.g., Microsoft Word, PDF) rather than tabular files (e.g., Microsoft Excel, CSV). However, as long as all files have a consistent format, researchers can write code to parse the text files into an analyzable tabular format. Subtitle file formats (.VTT files) are also common for mapping utterances to timestamps, and these files can be processed into tabular formats automatically in R (Knight, 2023).

There are many automatic transcription services available today (e.g., Otter, Temi, Amberscript, Descript, Trint, Sonix, Happy Scribe, Wreally, Ebby, Scribie; Table A1), with new services and iterations rapidly emerging (e.g., OpenAI's Whisper tool, which was released during the revision process of this paper). In September 2020, we systematically tested 10 of the most popular transcription services available. Each service transcribed the same series of audio recordings, and we evaluated the services along the following dimensions: 1) transcription accuracy, 2) speaker differentiation, 3) incorporation of timestamps, 4) user friendliness, and 5) pricing. We summarize our findings in Appendix A.

This review is not meant to be definitive. Rather, its primary purpose is to demonstrate how researchers might test and compare various transcription tools. Automatic speech recognition products and services have been rapidly evolving over time. Thus, we strongly encourage readers to conduct their own contemporaneous search at the time they require these services, evaluating their options based on the dimensions we list above. Researchers' needs may also vary depending on what is best for their project(s), so there is not a single best transcription service for everyone. However, we believe a hybrid transcription approach—automated transcription followed by human correction—is and will remain the most cost-efficient way to produce accurate, research-quality transcripts, at least in the near term.

3.2.3. Check. Automated transcription services have become more accurate over time, but they are not perfect (and neither are human transcribers). We strongly recommend asking people to listen to the audio recording while reading through the transcript, fixing any mistakes, and ensuring that formatting conventions are consistent throughout.

For example, transcription services have different policies about how to demarcate inaudible moments. Many will simply skip over this moment and leave a blank, while others will

flag this with “[inaudible]”, sometimes with a timestamp including duration. Our preference is typically to use the “[inaudible]” flag, which can be removed as needed; either way, it is essential to be consistent throughout. Further, there are many paralinguistic features that may be ignored by some transcription services. Common examples of these are laughter (“[laughter]”, “[laughs]”, or “[laughing]”) and interruptions (“[interruption]”, “[interposing]” or “--” at the start of an interrupting turn). Similar approaches are taken for other paralinguistic cues, like sighing, singing, crying, yelling, whispering, or cross-talk. Research question(s) should inform your approach: if laughter is important, make sure you annotate it, and do so consistently.

Checking transcripts can also uncover errors in the timestamps. One common error is typos from human transcribers—large errors can often be detected in later analyses (e.g., typos often result in negative or very long inter-turn pauses), though smaller errors also happen. When speakers are recorded separately, their timestamps may be aligned to different benchmarks in each recording (e.g., if the recordings start at different times). In this case, timestamps must be realigned to a common reference time before the transcripts from each recording are merged.

It is often useful to have human coders fix errors made by the speakers themselves, too, unless those errors are of research interest (e.g., self- and other-initiated repairs are important conversational phenomena). For example:

- Include and standardize the spelling of back-channels (e.g., “yeah”, “uh-huh”, “oh”).
- Remove erroneously repeated words (e.g., “I thought you...thought you were ready”).
- Include punctuation (e.g., question marks, periods, commas, ellipses).
- Change “gonna”, “sorta”, “dunno”, etc. to “going to”, “sort of”, “don’t know”, etc.
- Correct misspoken words, where the intended meaning is clear (e.g., “nice to mate you”).

There can be subtle but important differences in meaning among non-standard variations (e.g., “yes”, “yup”, “yasss”). However, there is a trade-off between specificity and statistical power. In general, differentiation could be reasonable if there is an adequate sample size of each variation, and if the distinctions matter for the research question(s) at hand. Otherwise, it may be best to aim for consistency (e.g., “yes” to study linguistic affirmation broadly).

Although transcript checking can be monotonous, the process can be designed efficiently. We typically find it easier for research assistants to complete all tasks for one document at a time, rather than completing one task for all documents before moving to the next task. However, to batch tasks like this, you must plan your checking needs in advance. For more efficiency, error checking can also be batched with human feature annotation (see Section 4.1. below).

4. Extracting Features from Text

Perhaps the most daunting task for conversation researchers is to decide which features to extract from the transcripts. Each “feature” can be thought of as a measure of one behavior in the transcript (e.g., the number of first-person pronouns; the percentage of words that mention food; the average length of pauses). There are a large (and increasing) number of tools available to researchers for this task, and researchers are presented with a wide array of options, even for measuring the same underlying construct (Yeomans, 2021; Schweinsberg et al., 2021).

We offer a brief review of common techniques, with a special focus on the challenges of studying dialogue data (as opposed to single-voice documents). While tools for these steps are available in several software environments, we will point readers to tools in the R software language. However, we note that Python also has many excellent tools for NLP (ConvoKit, in particular; Chang et al., 2020). Notably, both Python and R allow users to manage the two

datasets—turn-level and speaker-level data—simultaneously. This means researchers can integrate their feature extraction code with their analysis code (see section 5).

4.1. Feature Extraction Objectives

Before we introduce common feature extraction methods below, we first describe the important dimensions on which these methods can differ. This is important because there is no one "correct" approach. Instead, researchers must choose techniques based on their own idiosyncratic objectives and constraints, which are determined by their skillset, audience, research goals, resources, deadlines, and so on. Each of these dimensions should be considered when choosing a feature extraction method.

4.1.1. Accuracy. First and foremost, researchers should hope the features they extract from text data are valid, accurate measures of the underlying behavior or intention. Thankfully, accuracy can be evaluated empirically, within a validation dataset that has labels that can be treated as "ground truth" for comparison. For example, a turn-by-turn measure of question-asking should correlate as highly as possible with the true number of questions in each turn.

However, accuracy is not an inherent property of any method—it can only be defined within a particular population of interest. For instance, a model trained to label different types of questions in a doctor's office may not be as valid for labelling question types in a job interview. Researchers should be explicit about their intended populations, and the boundary conditions of their results (Simons, Shoda & Lindsay, 2017). They should also routinely conduct tests of "transfer learning" (Weiss et al., 2016; Yeomans, 2021), by explicitly testing how well their methods perform when they are developed in one context and applied to data from a different context.

4.1.2. Fairness. Bias is a concern shared by both humans and artificial intelligence (AI) systems. Just as humans are prone to unconscious biases (Greenwald & Banaji, 1995), AI models can exhibit algorithmic bias (Kordzadeh & Ghasemaghaei, 2022). Mitigating this bias is essential to ensuring the accuracy and fairness of research outcomes, regardless of the initial source.

Given that language models learn about the world from data used to train them, anything that learns from biased language data may unwittingly generate models that reinforce and codify prejudice, stereotypes, or other unsavory aspects of human judgment (Caliskan, Bryson & Narayanan, 2017). And, as is often the case with historical (and present-day) datasets, the speakers in the training data may themselves be biased or prejudiced. Sometimes this bias is the subject of research inquiry itself; however, if the focus is on other aspects of human behavior, this bias can undermine the goals of the research. This is especially true when a model or estimate is used to make decisions that affect real people. Consider, for example, an algorithm used to match job candidates to job postings based on similarity to exemplars in past training data. If that training data reflects a past in which some demographic groups (e.g., women, minorities) were excluded or discouraged from leadership roles, then the model on which it is trained may unwittingly reinforce that bias going forward. For example, an algorithm employed for recruitment at Amazon was later shown to be unwittingly discriminating against female applicants, based on the data it learned from showing that most leaders tended to be male (Dastin, 2018).

The accuracy of a model can thus vary across social groups in ways that may have biased consequences for the outcomes of those group members. Models trained on only one kind of speech, such as data from the most commonly studied sources (e.g., from demographic majority groups, from American-English speech), may be much less accurate when they parse speech

from groups that are historically underrepresented, from speakers from non-American countries, or for other reasons not included in the training data (Koenecke et al., 2020). This is an issue for all kinds of slang, jargon, and other language that is contextually—or socially—determined, and this type of language is very common in conversation.

There are no surefire techniques that can ensure a model is unbiased. One approach that has grown more common in recent years is to conduct an "algorithm audit," in which AI systems are evaluated to ensure they work as expected and do so without bias or discrimination (Brown et al., 2021; Koshiyama et al., 2022). Moreover, transfer learning tests, as described in 4.1.1, are very useful—by comparing how well a model's accuracy varies across different populations, researchers can evaluate whether particular groups may be adversely affected. When transfer learning tests are not possible, researchers should explicitly acknowledge the limitations of their training data, so that their tools are not misused by others. To improve the model itself, researchers should try to find training data that best represents the people involved, perhaps even oversampling from less numerous groups so that they are accounted for in the model. Above all, we recommend not taking model outputs as ground truth; instead, researchers should try to interpret and understand their models as much as possible, evaluate the contents based on their own domain expertise, and be as thorough as possible in making sure the model is behaving as expected.

4.1.3. Interpretability. Behavioral scientists are rarely concerned only with prediction accuracy. We also seek to understand and explain how people behave, which means we also need to understand what drives the results of our statistical models. Interpretability allows researchers to scrutinise their models so that they might improve them, and think about how well they might generalise to new contexts (Bianchi & Hovy, 2021). Improving interpretability can

also improve fairness, by allowing users (including regulatory bodies) to evaluate the model's strengths and failings in detail (Doshi-Velez et al., 2019), and users generally trust models more when they understand them (Gilpin et al., 2018; Yeomans et al., 2019). We recommend a similar skepticism from researchers - so-called "black box" methods which are not explained should not be relied upon to provide scientific insights.

While interpretability is almost universally desirable, it is difficult to define or quantify it precisely (Lipton, 2018). But generally speaking, models can be made more interpretable along two dimensions. First, the methods themselves should be transparent. Their exact content, code, and training procedure should shared, and benchmarked against related models across diverse contexts (Mitchell et al., 2019). However, transparency is necessary but not sufficient - many modern NLP models are still too complex to scrutinize, even by experts (Bender et al., 2021). More troublingly, this information is often not shared, due to expediency and to prioritize individual success over progress as a field (Belz et al., 2021). For example, the DICTION software package provides only broad generalities about how its features are scored, or how its formulae were determined and validated (Hart, 2000), even though its license fee is much higher than open-source models that are much more transparent.

In addition to transparency, models can be made more interpretable by generating additional outputs, in addition raw feature scores. One approach is to use the model scores to find excerpts from the dialogue that highlight contrasting levels of a given measure (e.g., high vs. low warmth; follow-up vs switch question). Often, they can also extract coefficients directly from the model to that reveal which features most affect a model's output (e.g., Voigt et al., 2017; Huang et al., 2017). Even when researchers must rely on an uninterpretable model due to their high accuracy (e.g., human annotators or black box NLP), they should still try to understand its

workings. One approach is to train a simpler model that approximates the predictions of the more complex one, and interpret that simpler one instead (Ribeiro, Singh & Guestrin, 2016; Madsen, Reddy & Chandar, 2021).

4.1.4. Scalability. Researchers usually need to anticipate the costs of calculating and extracting features at a large scale. All feature extraction methods involve direct resource costs. These costs come in the form of upfront investment (e.g., learning how to use a new software package, or developing an annotation scheme) and in the marginal cost of applying a method to new data (e.g., computation or annotation time). There are other limitations that affect the costs of implementing different methods. For example, when data is proprietary, identifiable, or otherwise sensitive, some methods (e.g., human annotators reading raw text) may come under more intense scrutiny from stakeholders than other, less-invasive methods (e.g., computing average turn length).

4.1.5. Complexity. Many of these objectives are related to the complexity of a feature extraction method, even though complexity is not itself an objective. Complex features tend to be costlier to implement, but this extra effort is typically justified on the basis of improved accuracy, fairness, or interpretability. Conversation is itself complex, so a perfectly accurate feature extractor would have to be correspondingly complex. Instead, researchers often settle on a trade-off between acceptable effort and acceptable accuracy, and this can be done iteratively: simpler measures can be used first, and if that is insufficient, then more complex measures. To borrow an idiom: before investing in a more complex method, researchers should first consider if "the juice is worth the squeeze."

Complexity is often related to the scope of information needed from the transcript to identify a single feature, whether it's responsiveness, warmth, question types, expressions of

gratitude, disfluency, or inter-turn pause length. The simplest and most common methods treat a person's turns as a block of static text, as if they were single-voice documents (see Section 4.3). This allows researchers to draw on the large toolkit from single-voice document analysis. However, this ignores the features of text that make conversation unique. For example, some features incorporate the timestamps from the transcripts (Section 4.4). Many other features look at consecutive sequences of turns to understand the structure of how speakers are interacting (Section 4.5). We illustrate these different input scopes in Figure 2.

4.2. NLP versus Human Annotation

Before computational tools were available, researchers traditionally annotated conversations, scoring various features in transcripts by hand. In theory, any annotation task done by a human could be attempted with an algorithm instead, and vice versa. Thus, it is tempting to see NLP as a potential substitute for human labor, to automate simple workloads and reduce time spent reading.

However, we argue the opposite: researchers should consider NLP as a *complement* to human work. These algorithms make close reading more powerful because they can be used to scale up and interpret human insights. Humans can develop typologies and provide labels to train supervised algorithms. Researchers themselves can read their corpora, to guide their intuitions on which algorithms might be the best fit for their data and context.

4.2.1. Advantages of Humans. Human and algorithmic feature extraction have contrasting strengths and weaknesses. For example, many conversational phenomena are too complex for current tools to automatically detect with sufficient accuracy. In these cases, trained human annotators usually produce more accurate labels, and can be used as the gold standard for evaluating NLP performance (Bommasani et al., 2021). Human annotators can use their

knowledge about the social context of a conversation to frame their responses, while an algorithm typically applies the same scoring rule regardless of context. For example, humans use their knowledge about speakers and context to infer sarcasm, while algorithms are typically built to take all of a speaker's words at face value. Humans are better at understanding nuanced meaning amidst social exchange.

4.2.2. Limitations of Humans. People can be inconsistent from day to day, and between one another—annotators almost always have some amount of disagreement. Furthermore, their thought processes may be hard to know or interpret (Nisbett & Wilson, 1977). Annotators often do not—or cannot—give precise reasons for their judgments. While the exact protocols used to train the annotators can be shared, this does not guarantee that human annotators followed them, or followed them in the same way. Thus, algorithms are not the only "black box" feature extractors used in research—humans can be black boxes, too.

Humans can suffer from many of the same problems that algorithms do. Accuracy within and across domains is always a concern. When humans annotators perform poorly, it can be hard to know if the task is inherently difficult, human judgment is too subjective, or they are lacking the right training. Human annotators can treat people unfairly, due to historical bias and prejudice or inexperience in the domain, among other reasons (Denton et al., 2021). All of the tools available to interpret algorithmic judgments should be used to scrutinize human annotations for unintended biases or blind spots.

4.2.3. Costs of Human Annotation. The costs of using human annotators are typically higher than using an algorithm. Much of this difference lies in the marginal costs of annotating new data—annotator time scales linearly with the amount of data, while the marginal cost of automatically processing more data is trivial once an algorithm is built. However, there are

upfront fixed costs for both. For humans, researchers must establish clear definitions and protocols for assigning labels. Annotators then practice until they reach sufficient agreement on training cases. Researchers may revise their protocols during training, as their definitions are applied to edge cases in real data. This process is iterative: drafting a scheme, then testing it individually and via group discussion, revising the scheme, and re-testing. These details are usually context-specific, and researchers should work with domain experts to develop their annotation schemes.

Often, researchers try to reduce annotation costs by crowdsourcing label generation to pools of online workers (e.g., from Mechanical Turk). However, crowdsourced workers have their own problems. They are hard to train, do not provide good feedback during protocol development, and can be inattentive. The task must be cleverly allocated across many workers since each one can only label part of the dataset (e.g., Benoit et al., 2016; Kiritchenko & Mohammad, 2017). Accuracy concerns are less relevant for simple tasks and can be mitigated in part by averaging over many annotators (though, this reduces their cost advantage).

In general, we have found that if annotation tasks are sufficiently complex, a pair of in-house research assistants can produce more accurate labels than a larger pool of crowdsourced workers. Moreover, in-house annotators can complete the necessary checking and cleaning tasks described above (Section 3.2.3 “checking”).

4.2.4. Human-Algorithm Hybrids. As with transcription, a hybrid approach may be useful during feature extraction. Human annotations can be used to train interpretable algorithms that reproduce human judgments. This approach identifies the linguistic features that are driving the humans’ judgments. A side benefit to this hybrid approach is that if the resulting algorithm is accurate, it can be directly applied on new data without having to recruit new human annotators.

Additionally, rough algorithmic approaches can be used as a first pass, to focus the efforts of human annotators.

We used this workflow ourselves in Huang et al. (2017) when we wanted humans to annotate different question types. First, we applied a simple algorithm to identify turns that included a question (to assist the humans' search through the transcript). Then human research assistants coded these questions as one of several question types. After the human annotations were collected, the consensus labels were then fed back into a supervised learning algorithm, to train a question type detector. The final model included both the initial search filter and the supervised model, so that it could reproduce the human annotators' judgments at scale. It was trained on 4,209 annotated question turns within 368 conversations from a lab experiment, and then applied to an observational dataset with 987 conversations and 19,321 question turns.

4.3. Static Text Features

There are many review articles covering different methods for extracting features from single-voice documents. For brevity, we will review the most common methods, with a focus on why they may function differently in dialogue. These methods treat turn content as though it were from a single author document, like a news article. However, individual turns vary wildly in word count. In practice this means many turns from one speaker are collapsed into a single piece of text (this is discussed in detail in Section 5.1).

4.3.1. Counting Words. A common, straightforward approach to analyze text is the "bag of words" approach: count each word that occurs at least once, ignoring order. This can produce a very large feature set (perhaps thousands of different words in a single conversation). There are many preprocessing steps commonly used to smooth out the raw counts; this includes: reducing

words to their stems, expanding contractions, removing rare words, removing common “stop words”, and constructing “n-grams” (two- or three-word phrases).

These techniques improve models, but they should be considered in light of the specific research questions that are being addressed (Denny & Spirling, 2018). Conversation has a lot of stylistic and structural language, which tends to be determined by the more common function words—pronouns (“you”, “they”), adpositions (“to”,”from”), determiners (“the”, “your”), and adverbs (“mostly”). For example, question words (“who”, “what”, “where”, “when”, “why”, “how”, “which”) are essential for determining what types of questions people are asking (Huang et al., 2017; Zhang et al., 2017). However, these words tend to get removed by most off-the-shelf stop word lists, which were typically built for single-voiced text.

4.3.2. Dictionaries. Dictionaries are lists of words generated by expert human annotators, which give scores to words that group them into simpler dimensions of meaning. For example, a “food” dictionary would give all the in words relating to food (e.g., "pizza", "broccoli") a score of one, and the rest of the words (e.g., "bicycle", "reading", "heavenly") a score of zero. Other dictionaries assign each word a score on a continuous scale based on average ratings (e.g., concreteness, Coltheart, 1981; Warriner, Kuperman & Brysbaert, 2013). To calculate the summary score for the whole text, the scores of the individual words within it are averaged. For binary dictionaries, this score is the percentage of words that come from a dictionary.

Dictionaries are common and accessible. The Linguistic Inquiry Word Count (LIWC) is probably the most often-used NLP tool in psychology (Tausczik & Pennebaker, 2010), because it requires no special skill to conduct analyses and many features are simple to understand (e.g., first person pronouns, words about music). While dictionaries can be quite useful, users should be aware of their limitations. Most obviously, dictionaries (like bag of words) ignore the order

of words, sentences, phrases, and topics—how verbal content unfolds in sequence. For example, most dictionaries do not account for negations (“not bad” versus “bad”) or relative magnitude (“very bad” versus “bad” verses “terrible”; though see Hutto & Gilbert, 2014). Furthermore, the interpretation of dictionary results is often lacking. While it is tempting to simply take the title of a dictionary at face value, its meaning should be determined from the actual words it contains, and the procedure by which it was created and validated. Sometimes these details are not shared publicly.

Furthermore, authors should make sure the dictionary is capturing what is intended in their context, by comparing texts from their data to the dictionary’s scores, perhaps starting with texts that get especially high or low scores. Most dictionaries implicitly assume domain-generality— that the contained words each have a single, stable meaning (Hamilton et al., 2016). This is not always true in conversation (Eichstaedt et al., 2020; Boyd & Schwartz, 2021; Yeomans, 2021). For example, even something simple like emotional sentiment (e.g., positive words minus negative words) can fail to measure closely related concepts like the experience of happiness or well-being of the speaker (Beasley & Mason, 2015; Sun et al., 2019; Kross et al., 2019; Jaidka et al., 2020) or the nuances of how a business or product is being described (Frankel, Jennings & Lee, 2022; Rocklage et al., 2022). While domain-specific dictionaries can help these concerns (e.g., Loughran & McDonald, 2016), the boundary for what is in- versus out-of-domain is not always clear, and researchers are usually best off conducting their own in-domain validation.

4.3.3. Sentence Structure. Modern NLP tools can extract not just the words themselves, but the underlying structure of sentences—that is, the grammatical parsing of sentences into subjects, verbs, objects, modifiers, clauses, and so on. This improves the features extracted from

a typical bag of words model, by making use of structures that determine meaning—for example, negations ("bad" vs. "not bad"), named entities ("apple" the company vs. the fruit) and homonyms ("like" the positive-valence verb vs. "like" the valence-neutral adposition). Researchers can use pre-trained neural network models (Manning et al., 2014; Honnibal & Johnson, 2015; Manning et al., 2020) to generate grammar tags for each word and then build features based on the tagged set.

These tools have been effectively applied to measure markers of politeness from individual turns (Danescu-Niculescu-Mizil et al., 2013; Voigt et al., 2017; Yeomans, Kantor & Tingley, 2018; Yeomans et al., 2020). In conversational text, politeness features often succeed at capturing the robust dimensions of how a speaker structures their conversational turns—agreement, disagreement, acknowledgement, hedging, gratitude, subjectivity, apologies, greetings, and goodbyes. Models trained on these dimensions have generalized well across multiple domains, as they focus on structural and stylistic features, rather than the main content features that tend to define a domain (e.g., specific nouns and verbs). Figure 3 provides an example of politeness features extracted from a dataset to show the differences in linguistic style that result from a randomized pre-conversation assignment to condition.

4.3.4. Embeddings. A common approach to detecting semantic content is to use pre-trained "embedding spaces" that represent words and sentences as vectors within a space of meaning (e.g., Landauer & Dumais, 1997; Mikolov et al., 2013). Most modern embedding models are extracted from small neural networks, trained to estimate which words tend to have the same neighbors (Bhatia, Richie & Zou, 2019). To solve this problem, the inner layer of the neural network groups words with similar meanings close to one another within the space. These embeddings are particularly useful for tasks that involve a similarity calculation—for example,

measuring the semantic similarity of two texts (Arora et al., 2017) or improving dictionaries. Rather than using a dictionary to count words in a binary sense (i.e., presence/absence), authors can compute the similarity of a whole document to the dictionary as a continuous measure (e.g., Sagi & Dehghani, 2014; Garten et al., 2018).

Embedding models have several advantages over raw word counts. These models group words with similar meanings into a common dimension, whereas a word count model treats each word as its own dimension, reducing the feature space considerably. While word count models typically remove rare words to simply the estimation, embedding models are pre-trained on large data where a high frequency of words are seen often enough to be included in the model.

However, embedding spaces are difficult to interpret—the dimensions themselves do not directly correspond to meaningful concepts, and researchers must use other tools to interpret what the model is doing. Additionally, many common pre-trained embedding models are mapped to individual words, which means that they ignore the order of words spoken in conversation, and other sources of contextual variation in meanings. Still, newer models of embeddings are able to encode entire sentences within an embedding space (e.g., Devlin et al., 2018) and can be fine-tuned to incorporate some contextual differences in meaning if the researchers have enough data. This is a frontier of constant progress in the NLP community.

4.4. Timing Features

In this section, we review several conversation-specific features that can be derived from timestamps. Many types of conversation features are particularly prevalent in some parts of the conversation (see Figure 4 for an example). Furthermore, the impact of some features of language may vary in meaning or effect depending on *when* they are said during a conversation (e.g., Li, Packard & Berger, 2022). The most common use of time stamps is to organize other

features of text, and to select features from certain parts of the conversation for analysis. This is relevant for causal vs. predictive inference (see Section 5.2).

4.4.1. Pauses. Typically, there is some amount of pause between turns, measured as the difference between one turn's end timestamp and the next turn's start timestamp. Pauses tend to be longer in asynchronous and text conversations, and shorter in synchronous and spoken conversations. Interestingly, teleconference conversations tend to be somewhere in the middle of the two (Boland et al., 2022). Within a particular dataset, pauses of various lengths can be counted as turn-level features (Templeton et al., 2022). Some researchers simply dichotomize each turn into pause or no pause, based on a threshold and will show that results are robust over a range of thresholds (e.g., Curhan et al., 2021). It is more difficult to define within-turn pauses, where someone picks up after their own silence, and the relevant timestamps are not included in a turn-level dataset. Transcribers (human or algorithmic) can be instructed to indicate a mid-turn pause as a non-verbal (e.g., "so anyways… [pause] did you see them at the wedding?"), which can be counted or removed as needed.

4.4.2. Interruptions. Sometimes speakers do not leave any time in between their turns, or even talk over one another. This often happens when the first speaker is interrupted by the second, and this type of interruption is often given a special annotation in transcripts (e.g., a single dash at the beginning or end of a turn) as well as a zero or negative gap between the end time of the previous turn and the start time of the interruption. The meaning of these interruptions is the subject of scholarly study—as a signal of disrespect or authority in formal settings (Li et al., 2004; Mendelberg & Karpowitz, 2016); as a sign of excited, enjoyable discourse (Li et al., 2004; Yeomans & Brooks, 2023); or as a signal that one person was merely filling dead air until their partner was ready to take their turn. The content of the interrupter's

turn also distinguishes different types of interruptions, such as backchannels, questions, and arguments (Shi, Yeomans, Truong & Fast, 2022).

4.4.3. Speaking Time. Timestamps can also be used to measure speech patterns over longer periods. For example, speaking time (i.e., "participation" or "airtime") is commonly measured as the percentage of the total time that is used by a particular speaker. When timestamps are not available, airtime can be approximated using the number of words spoken by each speaker as a percentage of the total words spoken (though this does not account for when no one is speaking). Comparing turn length to the time stamps will give an estimate of the person's speaking speed (i.e., "cadence").

4.5. Interactive Features

4.5.1. Backchannels. During conversation, listeners often insert a brief utterance to signal they understand (e.g., "yeah", "ok", "mm-hmm") while someone else is talking. Different definitions have been used, and it varies based on context (e.g., audio vs. text chat). Typically, backchannels are treated as a single turn within the flow of conversation, with zero-time gap between the preceding and subsequent turn. This may unnaturally divide the longer turn of the backchannel recipient into two separate turns, which could interfere with sentence-level features. Some researchers have avoided this by considering backchannels as features of the turn receiving the backchannel. Then, each turn has a feature counting the number of backchannels it receives from other speakers (Reece et al., 2022).

4.5.2. Dialogue Acts. Most of what is said in conversation imposes a structure on what is said in subsequent turns: asking different types of questions; stating facts, opinions, or feelings; making requests or commands; signaling understanding, agreement, or disagreement; or initiating repair. These "dialogue acts" are essential to understand how speakers are

communicating with one another (Stolcke et al., 2000; Bunt et al., 2010). Other theoretical frameworks (e.g., speech acts; Searle, 1965) capture roughly the same idea, which is that conversational turns are usually more than just statements of fact about the world. Rather, they communicate a speaker's intentions and give structure to the response they expect to receive.

Some dialogue acts can be reasonably approximated with features extracted from individual turns by the politeness package (e.g., gratitude, apologies, acknowledgement; Yeomans, Kantor & Tingley, 2018; see Figure 3). However, many other dialogue acts are difficult to identify without information from other turns. For example, adjacency pairs (e.g., consecutive turns such as question/answer, offer/acceptance, misunderstanding/repair) often demarcate essential decisions in a conversation.

There is no universally accepted, domain-general list of dialogue acts. Instead, the set of relevant dialogue acts will change depending on the conversational context (e.g., the modality of exchange, the goals of the speakers, etc.). For example, consider the sequence of formal offers within a negotiation. Specific offers are among the most important dialogue acts, so the impact of measurement error on these features would be considerable. In fact, most negotiation platforms (e.g., iDecisionGames or eBay) require that formal offers be made separately from the unstructured stream of conversation, so that the speakers themselves can understand their partners. Algorithms may be able to parse the offers in simple negotiations (Lewis et al., 2017), but if the negotiation involves multiple complicated issues, automatic extraction may not be possible. The same treatment may be necessary for other dialogue where particular turns have formal significance—for example, voting during a meeting or generating creative ideas (Brucks & Levav, 2022).

4.5.3. Accommodation. One of the most common and reliable results in conversation analysis is accommodation—the tendency of one speaker to mirror the linguistic features of the previous speaker (Giles, Coupland & Coupland, 1991). Several models of accommodation have been proposed. The most common measure combines the entire transcript of each person separately, and then calculates the similarity of those two documents (Ireland et al, 2011). However, this ignores order and directionality (e.g., which of the speakers is doing the accommodating?). Other models are purpose-built for conversation, and explicitly identify accommodation from one turn to the next (Danescu-Niculescu-Mizil, Gammon & Dumais, 2011; Demszky et al., 2021; Doyle & Frank, 2016), and this can be aggregated as a feature of one or several turns.

Researchers have considered several feature sets over which accommodation should be measured. Some papers focus on mirroring of content (e.g., if I talk about my dog, will you talk about your dog?; Fusaroli et al., 2012; Babcock, Ta & Ickes, 2014), others focus on stylistic categories (e.g., if I use more quantifiers, will you do the same?; Danescu-Niculescu-Mizil, Gammon & Dumais, 2011) or syntactic structure (e.g., if I use short, clipped sentences, will you? Boghrati et al., 2018). Other papers include a wide range of features, combining content and style (Niederhoffer & Pennebaker, 2002; Srivastava et al., 2018). In truth, it is not clear whether conversational style and content can be cleanly separated, and the two often correlate with one another—in essence, some types of content naturally pair with particular styles. This is a subject of ongoing research.

4.5.4. Topics. Conversations are very often broken into discrete topics (e.g., the weather, then work, then cooking, and so on) based on speakers' varied intentions (Passonneau & Litman, 1993). There are well-known NLP algorithms that focus on extracting topical content from text

(i.e., topic modeling). The most common approach, Latent Dirichlet Allocation (LDA), assumes that each text document is a mixture of a small number of topics, and that each word's presence is attributable to one of the document's topics (Blei, Jordan & Ng, 2003; Roberts, Stewart & Tingley, 2019).

Alas, conversation data is not well-suited for topic models built for single-voice text. Topic models focus on the distinctive words that demarcate content, typically remove common words, like pronouns (e.g., "I", "you", "it", "she", "they"). However many turns contain no topic-relevant information (e.g. "why is that?" could be asked in almost any topic), and most turns are too short to reliably estimate word co-occurrence. Instead, blocks of turns must be segmented into topics for analysis, and dividing dialogue into segments is arguably even harder than assigning a topic to a particular segment (Purver, 2011; though see Hearst, 1997; Nguyen et al., 2014). Furthermore, in both single voice and dialogue it can be hard to choose the number of topics and interpret the words within each topic (Chang et al., 2009; Boyd-Graber, Mimno & Newman, 2014). Still, topic models may be a useful tool for rough exploration and descriptions of the main themes of a body of dialogue.

If the topical structure is important to measure precisely, we suggest researchers avoid relying on an unsupervised algorithm, but instead develop their own categories based on their knowledge of the domain and their exploration. For example, conversations can have a list of pre-assigned topics, which makes ex-post segmentation much easier (e.g., Yeomans & Brooks, 2023). Many conversations that are repeated often—such as sales calls, customer service, doctor-patient interactions, police interviews, parole hearings—have explicit or accepted dialogue scripts, which speakers have been trained to follow as a progression through a series of stages. These scripts can be used to develop domain-specific rules to segment individual transcripts into

discrete topics or stages (e.g. Takanobu et al., 2018). This is a subject of ongoing work, and NLP researchers have made progress in tracking topics shifts within dialogue (e.g. Xu et al., 2021; Xing & Carenini, 2021).

5. Model Construction

Most conversation research does not just examine transcripts. Instead, conversational behavior from transcripts is compared to data from outside the conversation, such as the speakers' gender, when the conversation took place, the terms they negotiated, or how they felt about each other when the conversation ended. This means that feature counts in the turn-level dataset (their words) need to be aggregated and merged with the speaker-level dataset (other measures outside the conversation). Then a statistical model must be estimated and interpreted. Finally, the results must be reported and benchmarked.

5.1. Aggregating Conversation Features

While many conversation features are observed at the turn-level, other variables of interest may be measured at a higher level, such as at the level of the conversation, the individual, dyad, group, organization, or society. Usually, these are measured once per conversation, either as context variables before the conversation (e.g., mood, location, preferences, random assignment to an experimental condition) or outcome variables after it (e.g., enjoyment, learning, negotiated outcomes). However, they can also be measured once per speaker (e.g., demographics) or after multiple conversations in a relationship.

To estimate the links between conversation features and these higher-level measures, turn-level features should be aggregated in some form (e.g., count, average, sum, standard deviation). These aggregations can then be merged to the speaker-level dataset using the speaker-

and conversation-level unique identifiers (the dplyr R package makes this process easier; Wickham et al., 2019).

5.1.1. Aggregation Window. Researchers should almost always separate the features of each person in the conversation before analysis (e.g., how many questions did Mary ask?), rather than across the entire transcript (e.g., how many questions did everyone ask?). This is necessary any time speaker-level variables vary within a conversation, such as occupying different roles, experimental conditions, and demographics.

Additionally, researchers may only want to aggregate features from a subset of the conversation. For example, they may remove greetings, off-topic chatter, or final decisions from analysis of a task-focused conversation. In other cases, they may only aggregate features from the beginning of the conversation, in order to focus on each person's behavior before they are influenced by their partner's manner of speech or because the meaning of a feature changes at different times (e.g., Li, Packard & Berger, 2022).

5.1.2. Controlling for Speaking Time. Researchers should be clear about counts versus rates. The total word count of each turn, and each conversation, is used in many analyses—it is a common and simple benchmark to use for prediction tasks. Other times, feature counts are transformed into feature rates to control for the length of each text (e.g., feature count per minute, or per 100 words, which is the default in the LIWC dictionary approach). Analyses are simplest when word counts are relatively similar across texts. When word count differences are large, researchers must decide whether the difference is endogenous (i.e., controllable) or not. For example, if someone is studying a mix of 30- and 60-minute meetings, then total feature counts would be mainly driven by the pre-scheduled meeting length. Thus, controlling for the total word count would make it easier to compare language across the two time frames.

Sometimes total speaking time is an outcome. For example, when people are told to ask more questions, their partner speaks more and enjoys the conversation more (Huang et al., 2017). This is not a confound—one reason there is an increase in talking is due to the amount of questions asked. Furthermore, enjoyment early in the conversation can increase talking as the conversation continues. In these cases, it may be better to focus only on the early part of the conversation, before differences in speaking time emerge (Shi, Yeomans, Truing & Fast, 2022). Otherwise, researchers should look at both *what* and *how much* is said as two distinct outcomes.

5.2. Model Estimation

While a review of the rich existing literature on model estimation (i.e., constructing a statistical model to test a hypothesis) is outside of the scope of this paper, we will briefly touch on several challenges that are particularly common in conversation research.

5.2.1. Units of Observation. Although each speaker is given their own row in the speaker-level dataset, these are not independent observations. There is often some shared variance with their partner in the context and outcomes. There is also shared variance when a speaker is present in multiple conversations (e.g., in a round-robin design or when tracking relationships over time), or when outcomes are measured multiple times per conversation (e.g., once per topic). This is commonly addressed by using heteroskedasticity-robust standard errors (e.g., through the estimatr R package, Zeileis, Koll & Graham, 2020). Researchers who ignore these issues can end up overstating the precision of their estimates, and overfit models that are too complex to be estimated well by their datasets (Bertrand, Duflo & Mullainathan, 2004; Yeomans et al., 2019).

5.2.2. Interpreting Effects. The time course of conversation complicates the interpretation of estimated effects. In particular, we distinguish between *causal* relationships

("what is the effect of X?"), *predictive* relationships ("will X happen next?"), and *descriptive* relationships ("did X happen?"). All three have some practical value (Kleinberg et al., 2015; Mullainathan & Spiess, 2017), but it is important to know the difference. This is especially difficult in interpersonal interaction, because there are many possible third variables that could confound any estimate: someone's mid-conversation behavior could either affect their outcomes directly, be correlated with something that affects outcomes, or be an outcome of something that happened earlier in the conversation.

The gold standard for causal estimation is a randomized experiment, in which at least one speaker is randomly assigned to an intervention that affects some part of their conversational behavior (e.g. try to interrupt a lot v. try not to interrupt at all) or outcomes they or their partner will report (e.g. come with as many ideas as you can v. choose one idea to pursue). In lieu of experimental control, some empirical approaches can help make causal interpretations more plausible. If speakers have stable conversational tendencies across conversations (e.g., some people always laugh more frequently, or have a penchant for arguing), then the random assignment of speakers to their partners can be used as an instrumental variable (Zhang et al., 2020). Researchers have also sharpened their interpretations by focusing on conversation features (as in Section 4.3.1) from the beginning of conversations, before speakers are deeply influenced by their partner (e.g., Curhan & Pentland, 2007; Voigt et al., 2017; Zhang et al., 2018). Other common causal inference strategies (e.g., controlling for pre-conversation variables, matching, event studies) may also be useful (Angrist & Pischke, 2008).

5.3. Reporting Results

Only a subset of a researcher's analyses will end up in a final publication. The low cost of additional analyses can be harnessed to produce a variety of benchmark models, alternative

specifications, and robustness checks. While it is often tempting to report only the positive results, these other analyses are often more useful when they produce negative results, as they highlight limitations and boundary conditions.

While not all of these additional analyses need to make the main body of the paper, online appendices often have no word limit. Additionally, researchers who share their analysis code and data can encourage their readers to explore alternative models themselves. At the very least, researchers should conduct and report basic sanity checks—for instance, that their results cannot be obtained using simpler text analysis, such as word count or sentiment analysis.

5.3.1. Benchmarks. Often researchers are focused on a particular variable (e.g., question-asking), and they may want to demonstrate that the variable has a uniquely strong relationship with the outcome of interest. However, because conversation data is complex, there are many potential comparisons that can be constructed.

Instead, researchers should always give context to their focal model with some reasonable set of benchmark models (e.g., Eichstadt et al., 2020; Yeomans, 2021). For example, computer science papers routinely include tables comparing the performance of many models on the same dataset. Since conversation data is rich, benchmarks could be drawn from contextual data or from other features of the transcript. Another approach, to check the importance of a single feature, is called an "ablation test." There, a feature is removed from a more complex model—if the performance of the new model decreases, then the removed feature is considered essential for the original model.

Similar concerns arise when selecting control variables. There are many ways to define a model specification using conversation data, and researchers may find value in estimating alternative models to demonstrate robustness - sometimes called a "multiverse" or "specification

curve" analysis (Simohnson, Simmons & Nelson, 2020; Schweinsberg et al., 2021). The most reliable results will hold not only across individual specifications within a dataset, but across datasets and contexts.

5.3.2. Confirmatory vs Exploratory Results. The high dimensions of text allow for near-infinite researcher degrees of freedom (Yeomans, 2021). This means the standard concerns about p-hacking, data-dependent modeling choices, and non-replicability should be especially important for conversation research. Best practices include pre-registering NLP analyses whenever possible—including exact analysis code, and detailed information on what data is collected, and how the sample will be determined (Nelson, Simmons, & Simonsohn, 2018). Likewise, researchers should be wary of assuming generalizability for models that have only been tested in one dataset, or one context. However, exploratory results can be tremendously useful (Moore, 2016; Collins, Whillans & John, 2021). Thus, we recommend a balanced approach that prioritizes pre-registered results where possible, as a complement to (rather than to the exclusion of) well-grounded exploratory work.

When researchers publish results that have not been pre-registered, they can still take steps to enhance the credibility of their findings. For example, they can separate validation analyses from their extraction and estimation strategies, using cross validation or split samples within their dataset (Poldrack, Huckins & Varoquaux, 2020). While a common default for these validation checks assigns data into training and testing folds randomly, researchers may find added value from non-random splits (Weiss et al., 2016). For instance, they could assign data to training and testing at the level of conversations (so that all speakers within a single conversation are all in the same fold together) or the level of speakers (so that when a speaker appears in multiple conversations, all of their conversations are grouped into the same fold together). This is

also relevant when researchers have data across a large time span. For example, researchers who want to forecast stock prices from CEO interviews might train on data from 2010-2020 and then test their model on 2021-2022, so that their model is tested on a simulation of its eventual application: seeing into the future. Other examples might be training and testing on different company types, countries, or CEO characteristics (e.g., gender). These non-random splits allow researchers to make stronger claims about the robustness and generalizability of their conclusions.

6. Data Sharing

Collecting and cleaning conversation data for academic research can be costly in terms of time and money. This can make conversation research prohibitive for early-career scholars, and privilege scholars from well-resourced institutions. Moreover, costs may lead individual researchers to be reluctant to share their data with others who did not bear those costs themselves. However, we think this reluctance could be holding conversation science back—it is the costliness of collecting conversation data that makes its sharing especially valuable and productive. The field will be better off if we establish norms for researchers to share their materials, data, and code openly. We hope to encourage a more cumulative, inclusive, and collaborative research community. To this end, in our own work, we have shared as much of our conversation data as we can. Further, our own research has directly benefited from the generosity of others who were willing to share their data and analyses (e.g., Ranganath et al., 2009; Huang et al., 2017).

Open science practices are important (NASEM, 2018), and we think they are especially important for conversation science (Reece et al., 2022). First, conversation is so multifaceted that the same dataset can be used to answer many research questions, beyond the scope of the initial

research question of the researchers who collected the data. Second, the upfront costs of collecting and cleaning large-sample conversation data are immense and may be prohibitive for some researchers. Third, the upfront costs of the analysis are also quite high, so researchers can quickly build on one another's work by publishing reproducible code that can be shared and improved. Finally, individual hypotheses can be more robustly tested if analyses and results can be replicated over multiple datasets that may have been collected in different contexts.

6.1. Data Privacy

There are barriers to openly sharing data. In our view, the most common and legitimate concern is privacy. Many common privacy issues are exacerbated in conversation research because conversation datasets include identifiable data (Rubinstein & Hartzog, 2016; Cychosz et al., 2020). When conversations are recorded on video and/or audio, these rich media make it easier for subjects to be identified. Furthermore, even the transcripts of conversations can contain revealing details about a person that could be identifiable, either individually or in combination (Sweeney, 2002). These are essential questions for researchers to grapple with, and while there are more extensive treatments of the relevant issues (e.g., Robbins, 2017; Meyer, 2018), we intend to highlight the main concerns.

6.1.1. Preventative measures. The most important step in accounting for privacy is to obtain explicit consent from participants. In practice, we have found that researchers often fail to anticipate future data sharing needs and are not clear in asking for permission to store and to share de-identified data. Participants and Institutional Review Boards (IRB) rarely blanch at these requests in consent forms, as it is increasingly an essential part of the research process. Furthermore, an explicit warning about sharing may prompt participants not to share anything truly private.

It is worth assessing the importance of individuating information for the research question. For example, if researchers are studying performance during a negotiation simulation, in which the particulars are assigned at random in the case materials, then the speakers' true persona (including names, demographics, and location) are irrelevant to many research questions. In these cases, researchers should directly ask participants to refrain from providing any identifying information before the conversation begins. However, this restriction can interfere with some research questions. Consider two examples—doctor-patient conversations and speed-dating conversations—where personal information is essential to the goals of the speakers. In these cases, researchers cannot reasonably ask speakers not to share personal information.

6.1.2. De-Identification. It is best practice to anonymize conversation datasets, when possible. This is especially important for conversation data, since it is open-ended: during a conversation people can say virtually anything. If data are to be shared for public use (which we encourage), it is essential that the text be completely de-identified. Many feature extraction techniques automatically remove identifying information. For example, if an n-gram model is used and all n-grams that occur less than 1% of the time are removed, this will mechanically remove any individuating information (as long as no individual makes up more than 1% of the data).

Anonymizing raw text is more challenging. This can be done manually—by a human coder reading through each transcript and removing any identifiers—or automatically. For example, there are software packages that can de-identify most data by replacing named entities (e.g., specific names, addresses, etc.) with generic tags, although no algorithmic method is perfect (Mendels, 2018, Kleinberg, 2023). Like transcription, the best approach may be hybrid—

using an algorithm as a first pass at anonymization, followed by a human check to handle the most difficult-to-detect identifiable information.

Some conversation data is especially difficult to anonymize (e.g., audio or video data). We are not aware of any robust method for automatically de-identifying video or audio data, it may be better to simply focus on sharing transcriptions, and turn-level extracted features (metadata), rather than the complete or raw data. Likewise, even transcripts can be difficult to anonymize. For example, a real estate negotiation will likely reveal identifying features of the property in question, which can then be linked to other public records. In these cases, we still encourage researchers to share the turn-level dataset with the text removed, leaving only the unique identifiers and the extracted features. Note, however, that this is not always a guarantee of de-identification. It is possible that text or demographic variables (e.g., gender) could be reconstructed from the feature counts. This is primarily a risk for very elaborate feature extraction (e.g., sentence embeddings) whereas it is exceedingly unlikely to be an issue with simpler features (e.g., counts of pauses or questions).

We encourage researchers to scrutinize the identifiability of the metadata they collect outside the conversation (e.g., demographics). If there is a concern about these data, they can be de-identified. Common solutions include coarsening variables to broad categories (e.g., reporting age buckets, rather than exact age; Samarati & Sweeney, 1998) or perturbing variables by adding noise (e.g., reporting age +/- 5 years; Kargupta et al., 2003). This is especially important when researchers combine publicly available text data with non-public data, for example, if text from someone's (public) Twitter account is paired with their (private) school transcripts. Because the text can be searched, this risks identification of each participants' entire record.

6.1.3. Handling Sensitive Data. There are unique privacy concerns that arise in many common conversation data settings. Imagine conversations between financial advisors and their clients, or professors and their students. In these cases, researchers must prioritize their responsibilities to protect the rights of the speakers, and to uphold the norms of the context in which they were speaking. For example, consent is not always possible to collect from the speakers themselves, and speakers may not be aware of how their data will end up being used.

Many organizations establish their own policies around data sharing. For example, a company may have permission from its users to share data but may not want to make the raw data public because they consider that information proprietary. We strongly encourage researchers to be proactive about this topic when exploring collaborations with outside organizations. Many of the anonymization techniques mentioned above, such as extracting aggregated linguistic features using open source software (e.g. Yeomans, Kantor & Tingley, 2018) and metadata rather than raw transcript data, can be initiated before researchers see any of the data, so that no raw text ever leaves the organization.

Depending on their capabilities, an organization may be able to execute analysis code that a researcher writes without ever seeing more than a small example of their internal data. Many feature extraction algorithms remove identifying information from text (e.g., counts of politeness features). The resulting turn-level feature counts could then be analyzed by researchers and shared publicly, along with the code that was used to tally the features.

There are also unique concerns when dealing with text collected from publicly available sources (e.g., social media data or online forums), because there is also a heightened risk that it can be re-identified. If the dataset includes metadata that is not publicly available, this creates potential risks for the speakers. For example, if a researcher shares the exact turn-level word

embeddings or word counts of entire conversations, that information, though ostensibly anonymized, may be enough to reverse-search and uncover the source of the data. In these cases, researchers may want to increase the anonymity by adding noise to the extracted feature counts and/or the metadata.

7. Conclusion

This is an exciting time to be studying conversation, a fundamental activity of our social world. With technological advances, it is becoming easier to collect and analyze large-scale conversation data, and to pair turn-level conversation data with speaker-level data containing more traditional survey and behavioral measures. Still, collecting and analyzing text data, and combining turn-level and speaker-level datasets presents unique challenges. The complexities of this domain provide opportunities for researchers to build a community of inquiry that shares methods, tools, and data, and strives for an ever-growing, cumulative science of conversation.

8. Acknowledgements

This paper was much improved by helpful comments on earlier drafts from many other researchers, including (among others) Ken Benoit, Ryan Boyd, Gus Cooney, Morteza Dehghani, Grant Donnelly, Bennett Kleinberg, Andrew Knight, Celia Moore, James Pennebaker, Gillian Sandstrom, Martin Schweinsberg, Lyle Ungar, and Simine Vazire.

References

Adamson, D., Dyke, G., Jang, H., & Rosé, C. P. (2014). Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education, 24(1),* 92-124.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton university press.

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In International conference on learning representations.

Ashokkumar, A., & Pennebaker, J. W. (2022). Tracking group identity through natural language within groups. *PNAS Nexus, 1*(2), pgac022.

Babcock, M. J., Ta, V. P., & Ickes, W. (2014). Latent semantic similarity and language style matching in initial dyadic interactions. *Journal of Language and Social Psychology, 33*(1), 78-88.

Backus, M., Blake, T., Pettus, J., & Tadelis, S. (2020). Communication and bargaining breakdown: An empirical analysis (No. w27984). National Bureau of Economic Research.

Beasley, A., & Mason, W. (2015, June). Emotional states vs. emotional words in social media. In Proceedings of the ACM web science conference (pp. 1-10).

Belz, A., Agarwal, S., Shimorina, A., & Reiter, E. (2021). A systematic review of reproducibility research in natural language processing. *arXiv preprint arXiv:2103.07929*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278-295.

Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and explaining political sophistication through textual complexity. American Journal of Political Science, 63(2), 491-508.

Benoit, K. (2020). Text as data: An overview. The SAGE Handbook of Research Methods in Political Science and International Relations, SAGE Publishing, London (forthcoming).

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing, 84(1),* 1-25.

Berry, M. (2013). Towards a study of the differences between formal written English and informal spoken English. *Systemic functional linguistics: Exploring choice*, 365-83.

Berry, D. S., Pennebaker, J. W., Mueller, J. S., & Hiller, W. S. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin, 23*(5), 526-537.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *The Quarterly journal of economics, 119(1),* 249-275.

Bianchi, F., & Hovy, D. (2021, August). On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3895-3901).

Boghrati, R., Hoover, J., Johnson, K. M., Garten, J., & Dehghani, M. (2018). Conversation level syntax similarity metric. *Behavior research methods, 50*(3), 1055-1073.

Boland, J. E., Fonseca, P., Mermelstein, I., & Williamson, M. (2022). Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, *151*(6), 1272.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21-41.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications,* 225-255.

Brodsky, A., Lee, M.J., Leonard, B. (2022). Discovering new frontiers for dyadic and team interaction studies: Current challenges and an open-source solution—survconf—for increasing the quantity and richness of interactional data. *Academy of Management Discoveries.*

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352*(6282), 220-224.

Brucks, M. S., & Levav, J. (2022). Virtual communication curbs creative idea generation. Nature, 605(7908), 108-112.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31-36.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352(6282),* 220-224.

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society, 8(1), 2053951720983865.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J. W., Fang, A. C., Hasida, K., ... & Traum, D. (2010, May). Towards an ISO standard for dialogue act annotation. In Seventh conference on International Language Resources and Evaluation (LREC'10).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Carter, A.J., Croft, A., Lukas, D., Sandstrom, G.M. (2018). Women's visibility in academic seminars: Women ask fewer questions than men. *PLOS One,* 14(2).

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020, July). ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 57-60).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22.

Chen, J. V., Nagar, V., & Schoenfeld, J. (2018). Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies, 23(4),* 1315-1354.

Clark, H., Kirsh, D., Goldin-Meadow, S., & Rogers, Y. (2011). Interactivity and thought. *CogSci 2011*, 20-23.

Clarke, J. S., Cornelissen, J. P., & Healey, M. P. (2019). Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *Academy of Management Journal*, *62*(2), 335-360.

Collins, G., Poleski, J., Mehl, M., Tackman, A., Reyes, R., Kraft, A., ... & Casebeer, W. (2018). Building a Cognitive Profile with a Non-Intrusive Sensor: How Speech and Sounds Map onto our Cognitive Worlds. *Frontiers in Human Neuroscience*, 12.

Collins, H.K., Hagerty, S.F., Quoidbach, J., Norton, M.I., Brooks, A.W. (2022). Relational diversity in social portfolios predicts well-being. Proceedings of the National Academy of Sciences, 119(43).

Collins, H. K., Whillans, A. V., & John, L. K. (2021). Joy and rigor in behavioral science. Organizational Behavior and Human Decision Processes, 164, 179-191.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497-505.

Conner, T. S., & Mehl, M. R. (2015). Ambulatory assessment: Methods for studying everyday life. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource, 2015*, 1-15.

Cooney, G., Mastroianni, A. M., Abi-Esber, N., & Brooks, A. W. (2020). The many minds problem: disclosure in dyadic versus group conversation. *Current Opinion in Psychology*, *31*, 22-27.

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., ... & Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods, 53*(2), 467-486.

Curhan, J. R., Overbeck, J. R., Cho, Y., Zhang, T., & Yang, Y. (2021). Silence is golden: Extended silence, deliberative mindset, and value creation in negotiation. *Journal of Applied Psychology, in press.*

Curhan, J. R., & Pentland, A. (2007). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology, 92(3),* 802.

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., ... & Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior research methods, 52*(5), 1951-1969.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011, March). Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World Wide Web* (pp. 745-754).

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012, April). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web* (pp. 699-708).

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013, May). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 307-318).

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A Computational Approach to Politeness with Application to Social Factors. In *51st Annual Meeting of the Association for Computational Linguistics* (pp. 250-259). ACL

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Ethics of Data and Analytics (pp. 296-299). *Auerbach Publications.*

de Barbaro, K. (2019). Automated sensing of daily activity: A new lens into development. *Developmental psychobiology, 61*(3), 444-464.

Dehghani, M., Khooshabeh, P., Nazarian, A., & Gratch, J. (2015). The subtlety of sound: Accent as a marker for culture. *Journal of Language and Social Psychology, 34*(3), 231-250.

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., ... & Parmar, N. J. (2017). TACIT: An open-source text analysis, crawling, and interpretation tool. *Behavior research methods, 49*(2), 538-547.

Dehghani, M., & Boyd, R. L. (Eds.). (2022). *Handbook of Language Analysis in Psychology*. Guilford Publications.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. *Arxiv.*

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis, 26(2), 168-189.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., & Rosen, R. (2021). Whose ground truth? accounting for individual and collective identities underlying dataset annotation. arXiv preprint arXiv:2112.04554.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*

Diener, E., & Seligman, M. E. (2002). Very happy people. *Psychological science*, *13*(1), 81-84.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134.

Doyle, G., Goldberg, A., Srivastava, S., & Frank, M. C. (2017, July). Alignment at work: Using language to distinguish the internalization and self-regulation components of cultural fit in organizations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 603-612).

Doyle, G., & Frank, M. C. (2016, August). Investigating the sources of linguistic alignment in conversation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 526-536).

Drew, P., & Holt, E. (1998). Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in society, 27*(4), 495-522.

Dunbar, R. I., Marriott, A., & Duncan, N. D. (1997). Human conversational behavior. *Human nature*, *8*(3), 231-246.

Dupas, P., Modestino, A. S., Niederle, M., & Wolfers, J. (2021). *Gender and the dynamics of economics seminars* (No. w28494). National Bureau of Economic Research.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., ... & Ungar, L. H. (2020). Closed-and open-vocabulary approaches to text analysis: a review, quantitative comparison, and recommendations. *Psychological Methods*.

Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. Journal of Experimental Psychology: General, 143(5), 1980.

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, *128*, 32-37.

Fitzsimons, G.M. & Finkel, E.J. (2018). Goal transactivity. In Van Lange, P.A.M., Kruglanski, A.W., & Higgins, E.T. (Eds.), *Social Psychology: Handbook of Basic Principles* (3rd edition). New York: Guilford.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465.

Fox Tree, J. E. (2010). Discourse markers across speakers and settings. *Language and linguistics compass, 4(5),* 269-281.

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods. Management Science, 68(7), 5514-5532.

Fu, L., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Tie-breaker: Using language models to quantify gender bias in sports journalism. arXiv preprint arXiv:1607.03895.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. Psychological science, 23(8), 931-939.

Galley, M., McKeown, K., Fosler-Lussier, E., & Jing, H. (2003). Discourse segmentation of multi-party conversation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 562-569).

Ganek, H., & Eriks-Brophy, A. (2016, November). The Language ENvironment Analysis (LENA) system: a literature review. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 24-32).

Garfinkel, H. (1956). Conditions of successful degradation ceremonies. *American journal of Sociology, 61*(5), 420-424.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods, 50*(1), 344-361.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57(3),* 535-74.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics, 10*.

Goffman, E. (1981). *Forms of talk.* University of Pennsylvania Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818-829.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychological review, 102(1), 4.

Greenwood, C. R., Schnitz, A. G., Irvin, D., Tsai, S. F., & Carta, J. J. (2018). Automated language environment analysis: A research synthesis. *American Journal of Speech-Language Pathology, 27*(2), 853-867.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58). Brill.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis, 21(3),* 267-297.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing, 595.*

Hansen, S. & Ash, E. (2023). Text Algorithms in Economics. *Annual Review of Economics*, in press.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics, 133(2),* 801-870.

Hart, R. P. (2001). Redeveloping DICTION: theoretical considerations. Progress in communication sciences, 43-60.

He, H., Chen, D., Balakrishnan, A., & Liang, P. (2020, January). Decoupling strategy and generation in negotiation dialogues. In *2018 Conference on Empirical Methods in Natural Language Processing, (pp. 2333-2343).* Association for Computational Linguistics.

Hearst, M. A. (1997). Text Tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics, 23*(1), 33-64.

Heritage, J. (2008). Conversation analysis as social theory. *The new Blackwell companion to social theory*, 300-320.

Hirschberg, J., & Manning, C.D. (2015) Advances in natural language processing. *Science, 349*(6245), 261-266.

Honnibal, M., & Johnson, M. (2015, September). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373-1378).

Huang, K., Yeomans, M., Brooks, A. W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology, 113(3),* 430.

Huang, M., Zhu, X., & Gao, J. (2020). Challenges in building intelligent open-domain dialog systems. ACM Transactions on Information Systems (TOIS), 38(3), 1-32.

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science, 22*(1), 39-44.

Jackson, J. C., Watts, J., List, J. M., Puryear, C., Drabble, R., & Lindquist, K. A. (2021). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science.*

Jacobi, T., & Schweers, D. (2017). Justice, interrupted: The effect of gender, ideology, and seniority at Supreme Court oral arguments. *Va. L. Rev.*, *103*, 1379.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D. J., ... & De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 3040-3049.

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences, 117(19),* 10165-10171.

Jeong, M., Minson, J., Yeomans, M., & Gino, F. (2019). Communicating with warmth in distributive negotiations is surprisingly counterproductive. *Management Science*, *65*(12), 5813-5837.

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Jurafsky, D., & Martin, J. H. (2017). Speech and language processing (Vol. 4). London: Pearson

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel prize lecture, 8(1),* 351-401..

Kaplan, D. M., Rentscher, K. E., Lim, M., Reyes, R., Keating, D., Romero, J., ... & Mehl, M. R. (2020). Best practices for Electronically Activated Recorder (EAR) research: A practical guide to coding and processing EAR data. *Behavior Research Methods, 52(4),* 1538-1551.

Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2003, November). On the privacy preserving properties of random data perturbation techniques. In *Third IEEE international conference on data mining* (pp. 99-106). IEEE.

Kiritchenko, S., & Mohammad, S. (2017, July). Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 465-470).

Kleinberg, B. (2023). Textwash. Software.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491-95.

Knight, A. (2023). zoomGroupStats. R Package.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409.

Koshiyama, A., Kazim, E., & Treleaven, P. (2022). Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. Computer, 55(4), 40-50.

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., ... & Jonides, J. (2019). Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. *Emotion, 19(1),* 97.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.

Lapakko, D. (1997). Three cheers for language: A closer examination of a widely cited study of nonverbal communication. *Communication Education, 46(1),* 63-67.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. Trends in cognitive sciences, 20(1), 6-14.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125.*

Li, H.Z., Krysko, M., Desroches, N.G., DEagle, G. (2004). Reconceptualizing interruptions in physician-patient interviews: Cooperative and intrusive. *Communication and Medicing, 1(2),* 145-157.

Li, Y., Packard, G., & Berger, J. (2022). Conversational Dynamics: When Does Employee Language Matter? *Working paper.*

Liberman, M., & Cieri, C. (1998). The creation, distribution and use of linguistic data: the case of the linguistic data consortium. In LREC (pp. 159-166).

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54(4),* 1187-1230.

Madsen, A., Reddy, S., & Chandar, S. (2021). Post-hoc interpretability for neural nlp: A survey. arXiv preprint arXiv:2108.04840.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046-30054.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Mastroianni, A. M., Gilbert, D. T., Cooney, G., & Wilson, T. D. (2021). Do conversations end when people want them to? *Proceedings of the National Academy of Sciences, 118*(10).

McCabe, R., & Healey, P. G. (2018). Miscommunication in doctor–patient communication. *Topics in cognitive science*, *10*(2), 409-424.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. Journal of Personality and Social Psychology, 84, 857–870.

Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological science*, *21*(4), 539-541.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily

activities and conversations. *Behavior research methods, instruments, & computers, 33*(4), 517-523.

Mehl, M. R. (2017). The electronically activated recorder (EAR) a method for the naturalistic observation of daily social behavior. *Current directions in psychological science, 26*(2), 184-190.

Meier, T., Boyd, R. L., Mehl, M. R., Milek, A., Pennebaker, J. W., Martin, M., ... & Horn, A. B. (2021). (Not) lost in translation: Psychological adaptation occurs during speech translation. *Social Psychological and Personality Science, 12*(1), 131-142.

Mendelberg, T., & Karpowitz, C. F. (2016). Power, gender, and group discussion. *Political Psychology*, *37*, 23-60.

Mendels, O., Peled, C. Levy, N.V., Rosenthal, T. & Lahiani, L. (2018). Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images.

Meredith, J., & Stokoe, E. (2014). Repair: Comparing Facebook 'chat' with spoken interaction. *Discourse & communication, 8*(2), 181-207.

Meyer, M. N. (2018). Practical tips for ethical data sharing. Advances in methods and practices in psychological science, 1(1), 131-144.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Miller, A. H., Feng, W., Fisch, A., Lu, J., Batra, D., Bordes, A., ... & Weston, J. (2017). Parlai: A dialog research software platform. arXiv preprint arXiv:1705.06476.

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*, *18*(10), 512-519.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Molnar, A. (2019). SMARTRIQS: a simple method allowing real-time respondent interaction in qualtrics surveys. *Journal of Behavioral and Experimental Finance*, *22*, 161-169.

Moore, D. A. (2016). Preregister if you want to. American Psychologist, 71(3), 238.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

National Academies of Sciences, Engineering, and Medicine. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research.* National Academies Press.

Nelson, L., Simmons, J., & Simonsohn (2018). Psychology's Renaissance, *Annual Review of Psychology, 69,* 511-534.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*(4), 337-360.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2014). Modeling topic control to detect influence in conversations using nonparametric topic models. Machine Learning, 95(3), 381-421.

Nguyen, M., Gruber, J., Fuchs, J., Marler, W., Hunsaker, A., & Hargittai, E. (2020). Changes in Digital Communication During the COVID-19 Global Pandemic: Implications for Digital Inequality and Future Research. *Social Media+ Society, 6*(3).

OpenAI. (2022). GPT-3.5 architecture [Computer software]. Retrieved from https://openai.com

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language, 72, 101317.

Passonneau, R. J., & Litman, D. (1993). Intention-Based Segmentation: Human Reliability and Correlation With Linguistic Cues. In *31st Annual Meeting of the Association for Computational Linguistics (pp. 148-155).*

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90-93.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*(1), 547-577.

Pennebaker, J. W., & Slatcher, R. B. (2006). How do I love thee? Let me count the words: the social effects of expressive writing. *Psychological Science*, 17, 660-664.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169-190.

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, *77*(5), 534-540.

Pomerantz, A. (1990). Conversation analytic claims. *Communications Monographs, 57*(3), 231-235.

Purver, M. (2011). Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, 291-317.

Quoidbach, J., Taquet, M., Desseilles, M., de Montjoye, Y. A., & Gross, J. J. (2019). Happiness and social behavior. *Psychological science*, *30*(8), 1111-1122.

Rainie, H., & Wellman, B. (2012). *Networked: The new social operating system* (Vol. 10). Cambridge, MA: MIT Press.

Ranganath, R., Jurafsky, D., & McFarland, D. (2009, August). It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 334-342).

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., ... & Marin, S. (2022). Advancing an interdisciplinary science of conversation: Insights from a large multimodal corpus of human speech. arXiv preprint arXiv:2203.00674.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

Robbins, M. L. (2017). Practical suggestions for legal and ethical concerns with social environment sampling methods. Social Psychological and Personality Science, 8(5), 573-580.

Robbins, M. L., Mehl, M. R., Holleran, S. E., & Kasle, S. (2011). Naturalistically observed sighing and depression in rheumatoid arthritis patients: a preliminary study. *Health Psychology, 30*(1), 129.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(1), 1-40.

Rocklage, M. D., He, S., Rucker, D. D., & Nordgren, L. F. (2022). EXPRESS: Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language. *Journal of Marketing Research*, 00222437221134802.

Rogers, T., Ten Brinke, L., & Carney, D. R. (2016). Unacquainted callers can predict which citizens will vote over and above citizens' stated self-predictions. *Proceedings of the National Academy of Sciences*, *113*(23), 6449-6453.

Romero, D. M., Uzzi, B., & Kleinberg, J. (2016, April). Social networks under stress. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 9-20).

Rossignac-Milon, M., Bolger, N., Zee, K. S., Boothby, E. J., & Higgins, E. T. (2021). Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology, 120*(4), 882.

Rubinstein, I. S., & Hartzog, W. (2016). Anonymization and risk. *Wash. L. Rev.*, *91*, 703.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation." *Language 50*: 696–735.

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social science computer review, 32*(2), 132-144.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Schegloff, E. A. (1968). "Sequencing in Conversational Openings." *American Anthropologist* 70: 1075–95.

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

Schegloff, E. A., and Sacks, H. (1973). "Opening Up Closings." *Semiotica 8*: 289–327.

Searle, J. R. (1965). What is a speech act. Perspectives in the philosophy of language: a concise anthology, 2000, 253-268.

Seih, Y. T., Beier, S., & Pennebaker, J. W. (2017). Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology, 36*(3), 343-355.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12(6),* 1123-1128.

Sommers, S. R. (2006). On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations. *Journal of personality and social psychology*, 90(4), 597.

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. (2018). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science, 64(3),* 1348-1364.

Stillwell, D. J., & Kosinski, M. (2004). myPersonality project: Example of successful utilization of online social networks for large-scale social research. American Psychologist, 59(2), 93-104.

Stivers, T., Enfield, N. J., & Levinson, S. C. (2010). Question-response sequences in conversation across ten languages: an introduction. *Journal of Pragmatics, 42*, 2615-2619.

Stivers, T., & Sidnell, J. (Eds.). (2012). The handbook of conversation analysis. John Wiley & Sons.

Stokoe, E. (2010). 'I'm not gonna hit a lady': Conversation analysis, membership categorization and men's denials of violence towards women. *Discourse & Society, 21*(1), 59-82.

Stokoe, E. (2021). The sense of a conversational ending.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics, 26*(3), 339-373.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364.

Sun, J., Harris, K., & Vazire, S. (2019). Is well-being associated with the quantity and quality of social interactions? *Journal of Personality and Social Psychology*.

Swaab, R. I., Lount Jr, R. B., Chung, S., & Brett, J. M. (2021). Setting the stage for negotiations: How superordinate goal dialogues promote trust and joint gain in negotiations between teams. *Organizational Behavior and Human Decision Processes, 167,* 157-169.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C., Van Assen, M. A., ... & Schulte-Mecklenbeck, M. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organizational Behavior and Human Decision Processes, 165, 228-249.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 557-570.

Takanobu, R., Huang, M., Zhao, Z., Li, F. L., Chen, H., Zhu, X., & Nie, L. (2018, July). A Weakly Supervised Method for Topic Segmentation and Labeling in Goal-oriented Dialogues via Reinforcement Learning. In *IJCAI* (pp. 4403-4410).

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web* (pp. 613-624).

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology, 29*(1), 24-54.

Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., & Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences, 117(12),* 6370-6375.

Turmunkh, U., Van den Assem, M. J., & Van Dolder, D. (2019). Malleable lies: Communication and cooperation in a high stakes TV game show. *Management Science*, *65*(10), 4795-4812.

van Werven, R., Bouwmeester, O., & Cornelissen, J. P. (2019). Pitching a business idea to investors: How new venture founders use micro-level rhetoric to achieve narrative plausibility and resonance. *International Small Business Journal*, *37*(3), 193-214.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, *114*(25), 6521-6526.

Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., Zhang, M. & Zhou, G. (2020, April). Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 9177-9184).

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society*.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191-1207.

Weingart, L. R., & Olekalns, M. (2004). Communication processes in negotiation: Frequencies, sequences, and phases. *The handbook of negotiation and culture*, 143-157.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686.

Word, C. O., Zanna, M. P., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of experimental social psychology, 10(*2), 109-120.

Xing, L., & Carenini, G. (2021, July). Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 167-177).

Xu, Y., Zhao, H., & Zhang, Z. (2021, May). Topic-aware multi-turn dialogue modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 16, pp. 14176-14184)..

Yeomans, M., Kantor, A., & Tingley, D. (2018). The politeness Package: Detecting Politeness in Natural Language. *R Journal*, *10*(2).

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. Journal of Behavioral Decision Making, 32(4), 403-414.

Yeomans, M., Brooks, A. W., Huang, K., Minson, J., & Gino, F. (2019). It helps to ask: The cumulative benefits of asking follow-up questions.

Yeomans, M., Minson, J., Collins, H., Chen, F., & Gino, F. (2020). Conversational receptiveness: Improving engagement with opposing views. Organizational Behavior and Human Decision Processes, 160, 131-148.

Yeomans, M. & Brooks, A.W. (2023). Topic Preference Detection. *Working Paper.*

Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, *162*, 81-94.

Yeomans, M., Schweitzer, M. E., & Brooks, A. W. (2022). The Conversational Circumplex: Identifying, Prioritizing, and Pursuing Informational and Relational Motives in Conversation. *Current Opinion in Psychology.*

Yeomans, M., Stewart, B. M., Mavon, K., Kindel, A., Tingley, D., & Reich, J. (2018). The civic mission of MOOCs: Engagement across political differences in online forums. *International journal of artificial intelligence in education*, *28*(4), 553-589.

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software, 95*(1), 1-36.

Zhang, J., Spirling, A., & Danescu-Niculescu-Mizil, C. (2017, September). Asking too much? The rhetorical role of questions in political discourse. In *Proceedings of the 2017* Conference on Empirical Methods in Natural Language Processing (pp. 1558-1572).

Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., & Thain, N. (2018, July). Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1350-1361).

Zhang, J., Mullainathan, S., & Danescu-Niculescu-Mizil, C. (2020). Quantifying the Causal Effects of Conversational Tendencies. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-24.

Zheng, W., Yan, L., Gou, C., Zhang, Z. C., Zhang, J. J., Hu, M., & Wang, F. Y. (2021). Pay attention to doctor-patient dialogues: Multi-modal knowledge graph attention image-text embedding for COVID-19 diagnosis. *Information Fusion.*

Figure 1. A workflow for researchers to collect and analyze conversation data.

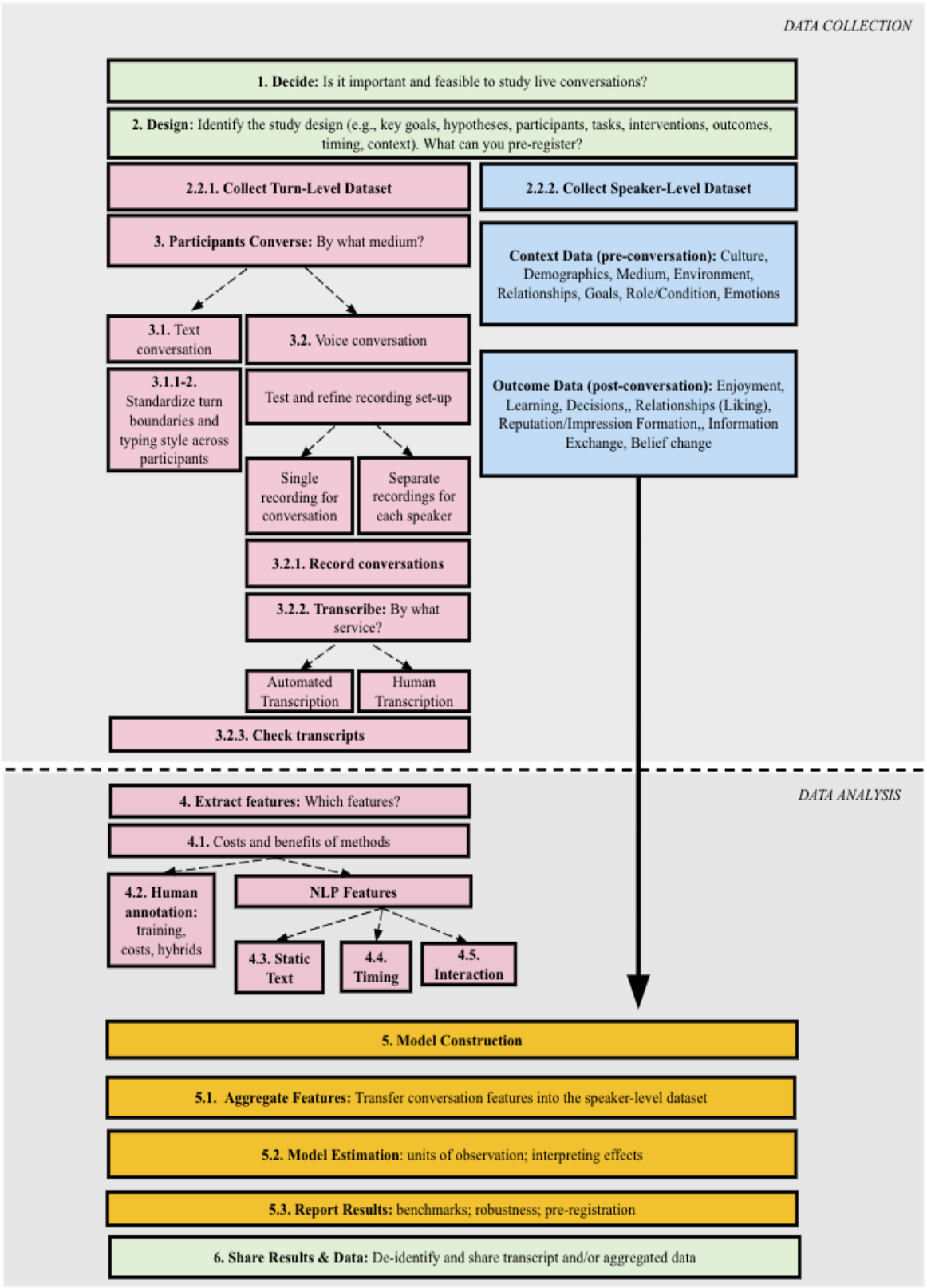
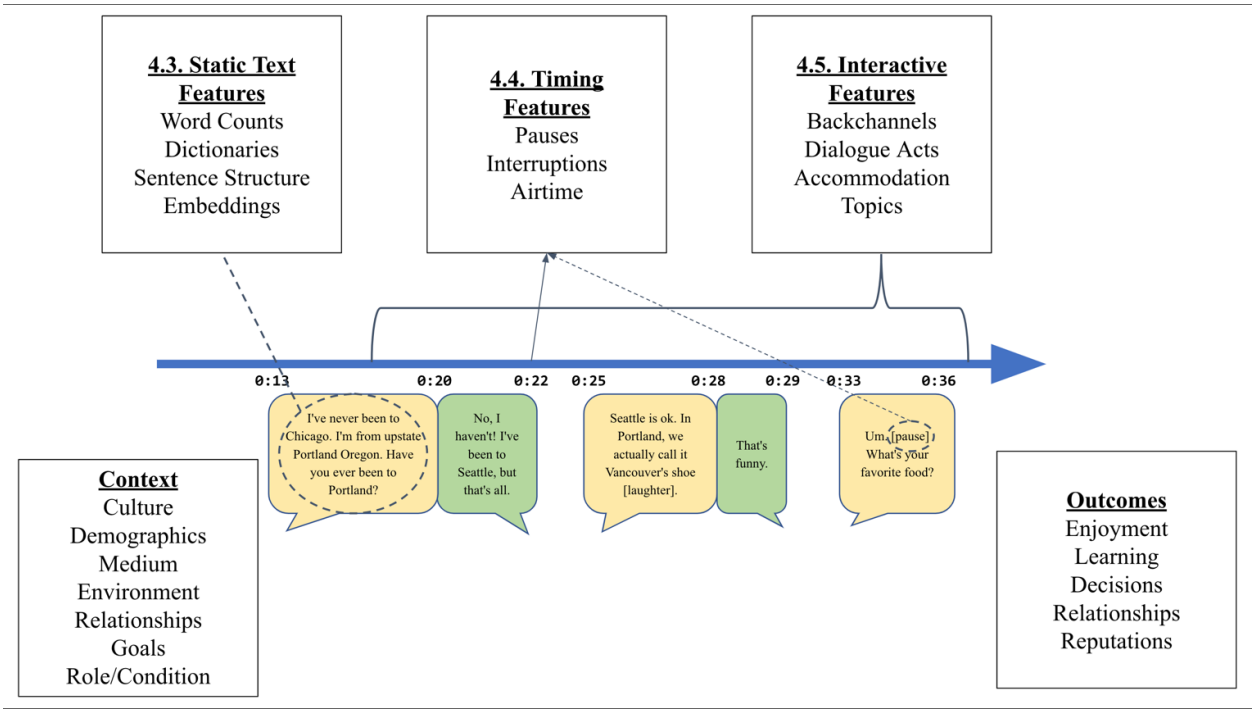


Figure 2. A map of data sources for conversation research. From the transcript itself, features can be extracted using static text methods, as well as relying on timestamps and interactivity. These conversation features are then compared to pre-conversation context variables, and post-conversation outcomes.



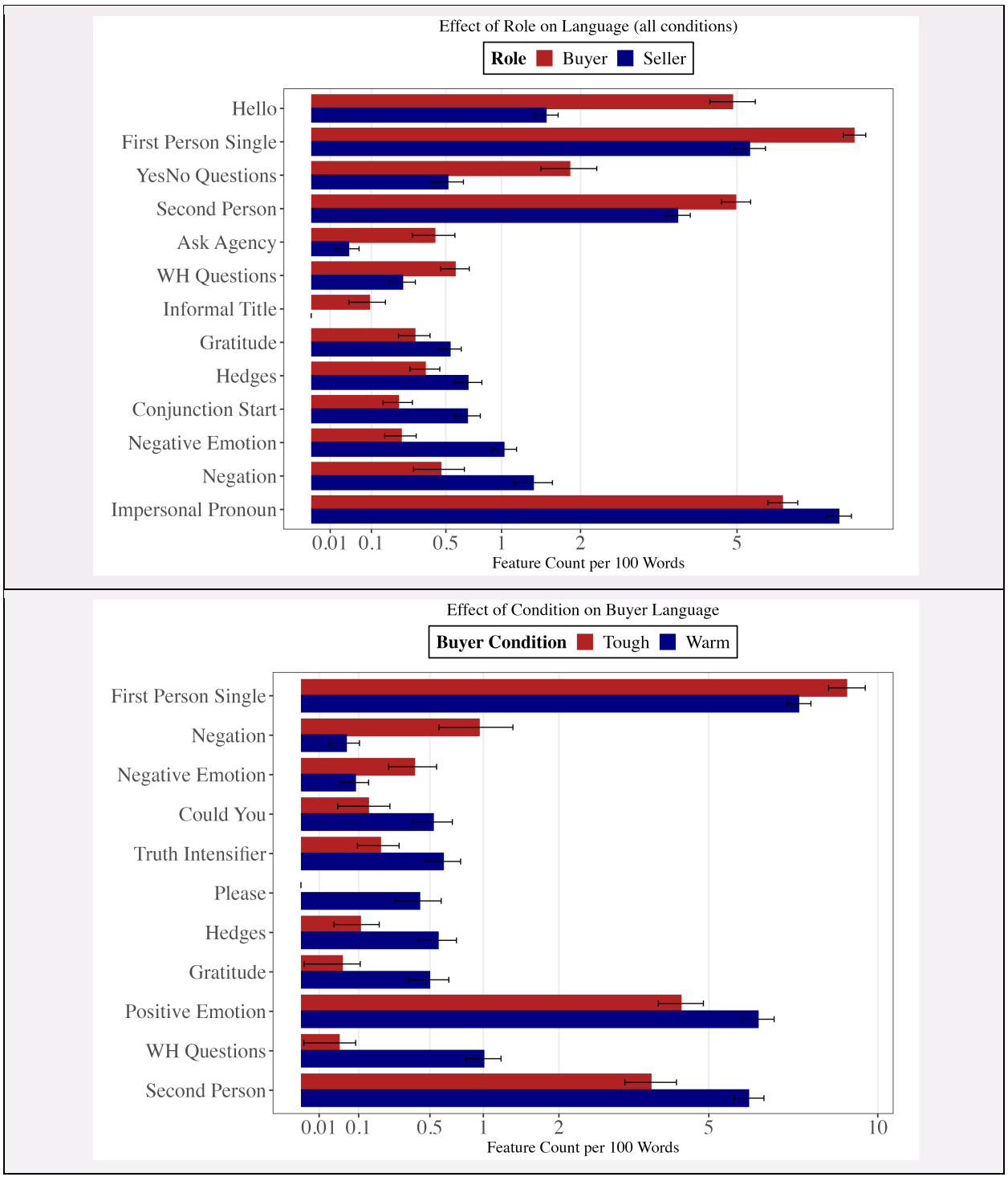


Figure 3. An example graph showing dialogue features extracted from negotiation transcripts (Jeong et al., 2018) using the politeness R package (Yeomans, Kantor & Tingley, 2018). The top panel compares the feature usage between buyers and sellers; the bottom panel compares the feature usage of buyers instructed to be warm and friendly, versus tough and firm. All bars show

group means and standard errors. Note that plots show feature counts per 100 words, because buyers (especially buyers instructed to be warm) use many more words than sellers.

Figure 4. An example conversational time-series graph, showing frequency of question types asked over the course of approximately 300 conversations between strangers (from Huang et al., 2017).

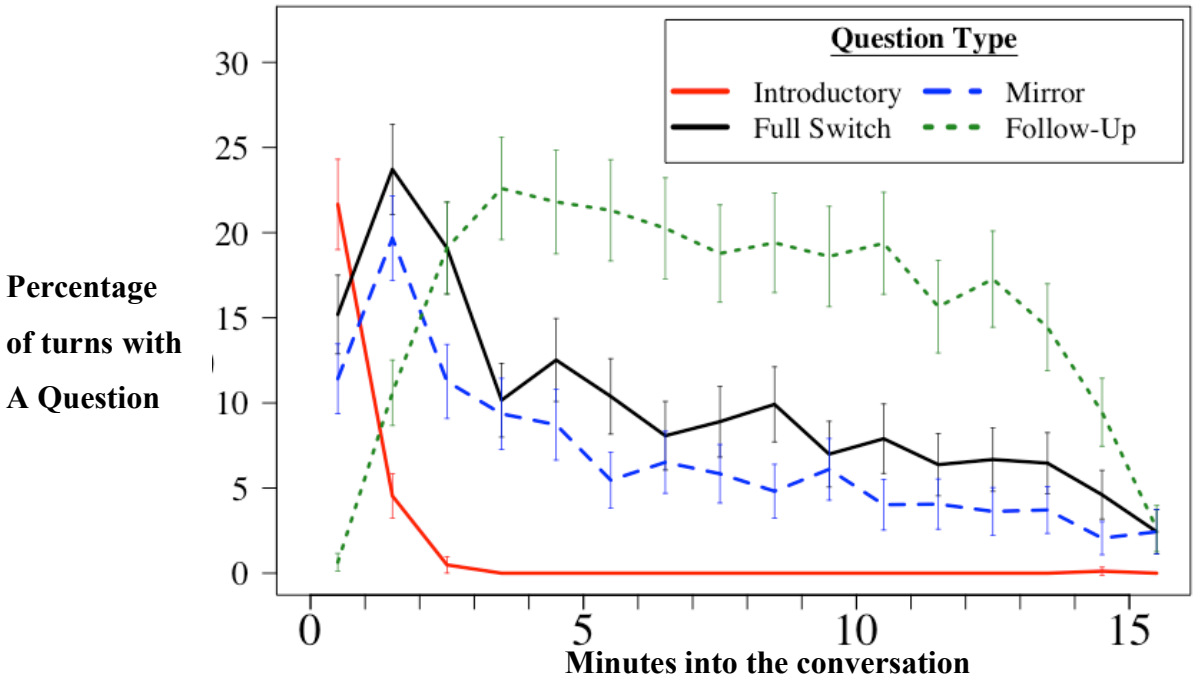


Table 1. A non-exhaustive list of recent research that analyzes transcript data across behavioral domains, conducted across academic disciplines.

Behavioral Domain	Paper Title	Application	Citation
Negotiations	"Communicating with Warmth in Distributive Negotiations Is Surprisingly Counterproductive"	The authors trained a natural language processing algorithm to quantify the difference between how people enact warm and friendly versus tough and firm communication styles in a distributive negotiation.	Jeong et al., 2019
	"Communication and Bargaining Breakdown: An Empirical Analysis"	The authors used text analysis to show that repeat players learn how to use communication in bargaining, and that the messaging strategies of experienced sellers are correlated with successful bargaining.	Backus et al., 2020
	"Setting the stage for negotiations: How superordinate goal dialogues promote trust and joint gain in negotiations between teams"	The authors used structured dialogues to identify the boundary conditions in negotiations that shape when superordinate goal dialogues are most likely to increase joint gain, as well as when they will not be effective.	Swaab et al., 2021
Work emails	"Social Networks Under Stress"	The authors analyzed instant messages among the decision-makers in a large hedge fund and their network of outside contacts to investigate the link between price shocks, network structure, and change in the affect and cognition of decision-makers in the network.	Romero, Uzzi & Kleinberg, 2016
	"Alignment at Work: Using Language to Distinguish the Internalization and Self-Regulation Components of Cultural Fit in Organizations"	The authors developed a measure of cultural fit based on linguistic alignment and used this measure to find that patterns of alignment in the first six months of employment are predictive of individuals downstream outcomes, especially involuntary exit.	Doyle, Goldberg, Srivastava, & Frank, 2017
Work meetings	"Virtual communication curbs creative idea generation"	The authors randomly assign work teams to conduct team meetings in person or on zoom, and study how that affects idea generation and decision quality	Brucks, M. S., & Levav, J. (2022).
Interviews	"Tie-breaker: Using language models to quantify gender bias in sports journalism"	The authors proposed a language-model-based approach to quantify differences in questions posed to female vs. male athletes and applied it to tennis post-match interviews.	Fu et al., 2016

Behavioral Domain	Paper Title	Application	Citation
Entrepreneurial pitches	"Pitching a business idea to investors: How new venture founders use micro-level rhetoric to achieve narrative plausibility and resonance"	The authors analyzed micro-level arguments underpinning pitch narratives of entrepreneurs who joined a business incubator and discerned four rhetorical strategies that these entrepreneurs used to achieve narrative plausibility and resonance.	van Werven, Bouwmeester & Cornelissen, 2019
	"Actions Speak Louder than Words: How Figurative Language and Gesturing in Entrepreneurial Pitches Influences Investment Judgments"	The authors identified distinct pitching strategies entrepreneurs use, involving different combinations of verbal tactics and gesture, and examined the impact of these strategies on investors' propensity to invest.	Clarke, Cornelissen & Healey, 2019
Quarterly earnings calls	"Manager-analyst conversations in earnings conference calls"	The authors conduct sentiment analysis to look at how well the questions asked (and their associated answers) predict changes in stock prices following quarterly earnings calls by publicly-traded companies.	Chen, Negar & Schonfeld, 2018
	"Disclosure Sentiment: Machine Learning vs. Dictionary Methods"	The authors find that machine learning methods are better at detecting disclosure sentiment than dictionary methods, in 10-K filings and earnings calls.	Frankel, Jennings & Lee, 2022
Medical Conversations	"Miscommunication in Doctor–Patient Communication"	The authors used conversation analysis to explore the effectiveness of medical treatment and shared understanding between patient and clinician in the context of psychiatric consultations.	McCabe & Healey, 2018
	"Naturalistically observed sighing and depression in rheumatoid arthritis patients: a preliminary study."	This study tested the degree to which naturalistically observed sighing in daily life is a behavioral indicator of depression and reported physical symptoms in rheumatoid arthritis (RA) patients.	Robbins et al., 2011
Police investigations	"'I'm not gonna hit a lady': Conversation analysis, membership categorization and men's denials of violence towards women"	The authors used British police interrogation materials and conversation analysis to shed light on the location and design of, and responses to, suspects' 'category-based denials' that they are not 'the kind of men who hit women'.	Stokoe, 2010
	"Language from police body camera footage shows racial disparities in officer respect"	The authors presented a systematic analysis of officer body-worn camera footage, using computational linguistic techniques to automatically measure the respect level that officers display to community members.	Voigt et al., 2017

Behavioral Domain	Paper Title	Application	Citation
Courtrooms	"On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations"	The authors examined the effects of racial diversity on group decision making and extended previous findings that racial issues, in the form of jury selection questions, increase leniency toward a Black defendant on trial.	Sommers, 2006
	"Echoes of power: language effects and power differences in social interaction"	The authors proposed an analysis framework based on linguistic coordination that they then use to study how conversational behavior can reveal power relationships in discussions among Wikipedians and arguments before the United States Supreme Court.	Danescu-Niculescu-Mizil et al., 2012
	"Justice, Interrupted: The Effect of Gender, Ideology, and Seniority at Supreme Court Oral Arguments"	The authors studied how the Justices of the United States Supreme Court compete to have influence at oral argument by examining the extent to which the Justices interrupt each other and how advocates interrupt the Justices, contrary to the rules of the Court.	Jacobi & Schweers, 2017
Central bank meetings	"Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach"	The authors used computational linguistics algorithms to explore the effect of transparency on monetary policy makers' deliberations.	Hansen, McMahon & Prat, 2018
Voter turnout drives	"Unacquainted callers can predict which citizens will vote over and above citizens' stated self-predictions"	The authors used conversation analysis to find that strangers can use nonverbal signals to improve predictions of follow through on self-reported intentions.	Rogers, Ten Brinke & Carney, 2016
Game shows	"Malleable Lies: Communication and Cooperation in a High Stakes TV Game Show"	The authors conducted an empirical analysis that showed that statements that carry an element of conditionality or implicitness are associated with a lower likelihood of cooperation and confirmed that malleability is a good criterion for judging the credibility of cheap talk.	Turmunkh et al., 2019
Government debates	"Asking Too Much? The Rhetorical Role of Questions in Political Discourse"	The authors used an unsupervised methodology for extracting surface motifs that recur in questions, and for grouping them according to their latent rhetorical role.	Zhang et al., 2017
Online forums	"No country for old members: user lifecycle and linguistic change in online communities"	The authors proposed a framework for tracking linguistic change in online communities and for understanding how specific users react to these evolving linguistic norms.	Danescu-Niculesu-Mizil et al., 2013

Behavioral Domain	Paper Title	Application	Citation
	"Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions"	The authors used discussions from an online community on Reddit to study and understand the mechanisms behind persuasion.	Tan et al., 2016
	Tracking group identity through natural language within groups.	The authors developed and validated a language-based metric of group identity strength and demonstrated its potential in tracking identity processes over time in reddit communities,	Ashokkumar & Pennebaker, 2022
Classrooms	"Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains"	The authors adopted a content analysis approach to analyze students' cognitively relevant behaviors in a massive open online course (MOOC) discussion forum and further explored the relationship between the quantity and quality of that participation with their learning gains.	Wang et al., 2015
	"The Civic Mission of MOOCs: Engagement across Political Differences in Online Forums"	The authors collected measures of students' political ideology and observed student behavior in the course discussion boards to find that students hold diverse political beliefs, participate equitably in forum discussions, directly engage with students holding opposing beliefs, and converge on a shared language rather than talking past one another.	Yeomans et al., 2018
Academic seminars	"Women's visibility in academic seminars: Women ask fewer questions than men"	The authors quantified women's visibility through the question-asking behavior of academics at seminars using observations and an online survey.	Carter et al., 2018
	"Gender and the Dynamics of Economics Seminars"	The authors collected data on every interaction between presenters and their audience in hundreds of research seminars, summer conferences, and job market talks across most leading economics departments to find that women presenters are treated differently than their male counterparts.	Dupas et al., 2021
Speed dates	"It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates"	The authors created a flirtation detection system which uses paralinguistic, dialogue, and lexical features to detect a speaker's intent to flirt on a speed-date with up to 71.5% accuracy, outperforming both the baseline and the human interlocuters.	Ranganath et al., 2009

Behavioral Domain	Paper Title	Application	Citation
	“It doesn’t hurt to ask: Question-asking increases liking.”	The authors trained a natural language processing algorithm as a “follow-up question detector” and applied it to speed-dating data to find that speed daters who ask more follow-up questions during their dates are more likely to elicit agreement for second dates from their partners, a behavioral indicator of liking.	Huang et al., 2017
Customer service calls	"Conversational Dynamics: When Does Employee Language Matter? "	The authors investigate the warmth-competence trade-off in customer service agents. They find that warm language is most common during the beginning and ends of successful calls, compared to the middle of those calls.	Li, Packard & Berger, 2022
Door-to-door campaigns	“Durably reducing transphobia: A field experiment on door-to-door canvassing”	The authors showed that a single 10-minute conversation that actively encouraged taking the perspective of others markedly reduces prejudice for at least 3 months.	Broockman & Kalla, 2016

Table 2: Example of a turn-level dataset. The column labels are: Group ID, used to distinguish between conversations; Turn, an index for each turn in the conversation in order; Start Time and End Time, indicating the time span of the turn; Speaker ID, indicating the speaking participant; Text, what was said in the turn; Question, a code for whether the turn contained a question; Laughter, code for whether the turn contained laughter; Word Count, a count of words spoken during the turn.

Group ID	Turn	Start Time	End Time	Speaker ID	Text	Question	Laughter	Word Count
1	1	0:00:01	0:00:03	A1	Hey, how are you? My name is [name] but my friends call me [name].	1	0	14
1	2	0:00:04	0:00:06	B1	Nice to meet you, [name]. I'm [name]. Where are you from?	1	0	11
1	3	0:00:06	0:00:12	A1	Thanks for asking! I'm from a small town outside of Chicago actually, you probably haven't heard of it. What about you?	1	0	21
1	4	0:00:13	0:00:20	B1	Probably not [laughter]. I've never been to Chicago. I'm from upstate Portland Oregon. Have you ever been to Portland?	1	1	19
1	5	0:00:20	0:00:22	A1	No, I haven't! I've been to Seattle, but that's all.	0	0	10
1	6	0:00:25	0:00:28	B1	Seattle is ok. In Portland, we actually call it Vancouver's shoe [laughter].	0	1	12
1	7	0:00:28	0:00:29	A1	That's funny.	0	0	3
1	8	0:00:33	0:00:36	B1	Um. [pause]. What's your favorite food?	1	0	5
1	9	0:00:37	0:00:55	A1	Hmm. That's a hard question. [pause] I really like all different foods. I made this really good stew the other day that I think might be the best thing I've eaten lately. But I'm always partial to a good hamburger.	0	0	39
1	10	0:00:56	0:00:59	B1	Cool. What was in your stew?	1	0	6

Table 3: Example of speaker-level dataset, with round-robin design. The column labels are: Group ID, used to distinguish between conversations; Speaker ID and Partner ID, used to distinguish between participants in a conversation; Age, the age of the participant; Gender and Partner Gender, the gender of the participants in the conversation; Condition assignment; Liking and Partner Liking, self-reported measures; Questions, total number of questions the speaker asked in that conversation; Laughter, total amount of speaker laughter in that conversation; Turn, total number of turns in the conversation; Word Count, the word count of the speaker in that conversation.

Group ID	Speaker ID	Partner ID	Age	Gender	Partner Gender	Condition	Liking	Partner Liking	Questions	Laughter	Turns	Word Count
1	A1	B1	24	1	2	1	5	6	2	1	5	87
2	A1	B2	24	1	1	1	2	7	3	1	4	60
3	A1	B3	24	1	2	1	7	6	1	0	3	54
1	B1	A1	34	2	1	1	6	5	4	2	5	53
2	B1	A2	34	2	1	1	6	2	0	3	6	102
3	B1	A3	34	2	2	1	5	4	3	1	7	131
1	A2	B2	57	1	1	2	2	5	0	0	2	45
2	A2	B3	57	1	2	2	1	7	1	1	4	75
3	A2	B1	57	1	2	2	2	6	1	0	5	64
1	B2	A2	23	1	1	2	5	2	1	0	4	24
2	B2	A3	23	1	2	2	7	5	3	3	5	33
3	B2	A1	23	1	1	2	7	2	4	2	6	98
1	A3	B3	55	2	2	1	3	4	2	1	3	112
2	A3	B1	55	2	2	1	4	5	5	1	4	33
3	A3	B2	55	2	1	1	5	7	1	2	2	16
1	B3	A3	19	2	2	2	4	3	1	0	3	47
2	B3	A1	19	2	1	2	6	7	0	0	4	87
3	B3	A2	19	2	1	2	7	1	0	1	6	101

APPENDIX A

Comparative Analysis of 10 Popular Transcription Services

There are many automatic transcription services available, and each service varies in effectiveness on different dimensions. In September 2020, we tested 10 of the most popular transcription services on the market, and rated them along five dimensions: (1) transcription accuracy, (2) speaker differentiation, (3) incorporation of timestamps, (4) user friendliness, and (5) pricing details.

The transcription services that we tested in September 2020 were Otter, Temi, Amberscript, Descript, Trint, Sonix, Happy Scribe, Wreally, Ebby, and Scribie (see Table A1). These services were determined by aggregating the top-rated auto-generated transcription services identified by The New York Times, PC Magazine, TechRadar, and Poynter.[3,4,5,6]

Transcription Accuracy: To assess transcription accuracy, we first manually transcribed the audio files ourselves. These transcriptions were considered the ground-truth. We then generated transcriptions from each service, using the same audio files. Each automatic transcription service that we tested was able to generate the recordings relatively quickly (within 30 minutes). We then systematically submitted the ground-truth transcriptions along with the matching automatically generated transcriptions to a plagiarism website called CopyLinks. By comparing the two files, we were able to determine the accuracy of the service: the calculated "text accuracy" score represents the percentage of words that overlap between the human-generated and auto-generated transcriptions. The higher the percentage, the better the service was at correctly transcribing the audio file. Otter and Temi were the most accurate from our testing.

Speaker Differentiation: Speaker differentiation is the ability of the transcription service to separate different speakers in the audio recording. Unfortunately, all of the services struggled with auto-generated speaker differentiation, but most allowed you to manually edit and correct this using the service's main platform. None of the services were able to differentiate by speaker in a consistently accurate way; all required human correction.

Incorporation of Timestamps: Incorporation of timestamps refers to whether each transcription service included the time when each speaker began their turn. Each of the top 5 services that we identified included timestamps at each turn, and most of the services allowed for easy adjustment in the service's main platform. When speaker differentiation was manually corrected in each service's platform, most of the timestamps were automatically updated to

---

[3] The New York Times. (2018, October 15). The best transcription services. The New York Times. https://www.nytimes.com/wirecutter/reviews/best-transcription-services/.

[4] Moore, B. (2018, August 22). The best transcription services. PCMAG. https://www.pcmag.com/picks/the-best-transcription-services.

[5] DeMuro, J. P., & Turner, B. (2021, March 30). Best transcription services IN 2021: Transcribe audio and video into text. TechRadar. https://www.techradar.com/best/best-transcription-services.

[6] LaForme, R. (2018, November 16). The best automatic transcription tools for journalists. Poynter. https://www.poynter.org/tech-tools/2017/the-best-automatic-transcription-tools-for-journalists/.

accurately mark the beginning or end of a turn. Trint was one of the best services at providing accurate timestamps, with the possibility to apply a 120-frame rate.

User Friendliness: User friendliness refers to how easy each service's platform was to use and navigate (upload files, edit within the auto-generated transcripts, export final transcripts, etc.). The final data format ended up being especially important. Most transcription services export transcripts as document files (i.e., Microsoft Word, PDF, or text format)[7], which all require an additional conversion to a tabular file (i.e., CSV or Microsoft Excel) for analysis. Usually, it is not hard to write cleaning code to parse the Word, PDF, or text files, although this depends on the exact formatting of the transcription service. Subtitle file formats (.VTT files) have also been used by services like Zoom to map transcribed utterances to timestamps, and there is an R package that can process these files into tabular formats automatically (Knight, 2021).

Pricing Details: Pricing details refers to how much each service costs to use. There is quite a bit of variation among the pricing models for each service—some charge a monthly membership fee while others charge per minute transcribed. In general, those that offered a membership fee tended to have a better overall platform, and those that charged by the minute were more cost-effective.

Approach

We started with all 10 auto-transcription services and 5 audio files gathered from YouTube. For the first round of testing, we only used audio files that involved two speakers and lasted approximately 10 minutes. The videos contained a range of accents, ages, and other characteristic differences between the speakers (see Table A2).

After generating the transcriptions of each of the 5 videos across all 10 platforms, it was clear that some of the transcription services out-performed the others across all dimensions. We narrowed down our list to the top 5 services, and then ran a second round of testing. In this round, we used 5 different audio files (selected following the same criteria as the first round of files), except that they involved two or more speakers. From this second round of testing, we were then able to decipher our top 5 transcription services.

Researchers' choice of transcription service will vary based on their needs, as well as technological improvements and additional services that enter the market.

[7] At the time of writing, Trint was the only automatic transcription service we tested that allowed researchers to export transcriptions as an Excel spreadsheet or CSV file.

Table A1: Transcription Services, Best 5 services indicated with an asterisk (*)

Service Name	Transcription Accuracy (Average Text Comparison)	Speaker Differentiation	Inclusion of Timestamps	User Friendliness
Otter*	79.47% match	Must edit speakers in the main service platform (unless paired with Zoom).	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will automatically populate based on your edits.	Otter is a great service, especially whe paired with Zoom (where it is extreme accurate in speaker differentiation). It allows for easy edits within the docum The options to export are with a Word document, PDF, text file, SRT, or clip There are a decent number of options terms of platform capabilities as well.
Temi*	75.21% match	Must edit speakers in the main service platform.	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will automatically populate based on your edits.	Temi is relatively cost-effective to use the output is not only one of the most accurate, but it allows for easy edits w the document. The options to export a with a Word document, PDF, or text f There are not as many options in term the platform itself, but we think this is definitely a top runner for ease and ba needs.

Service Name	Transcription Accuracy (Average Text Comparison)	Speaker Differentiation	Inclusion of Timestamps	User Friendliness
Amberscript*	71.57% match	Must edit speakers in the main service platform. Distinguishing speakers is moderately better when you provide the amount of people in the interaction.	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will populate based on your edits.	Amberscript is a good service. It allow easy edits within the document. The o to export are with a Word document, J or text file. There are not as many opti terms of the platform, but it covers the basics.
Descript*	70.89% match	Must edit speakers in the main service platform.	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will populate based on your edits. This service also has the most options with regard to timestamps.	Descript is a great service as well. It a for easy edits within the document, an some of the best capabilities. The opti export are with a Word document or R file. You download this service onto y computer and the platform is pretty straightforward to use.

Service Name	Transcription Accuracy (Average Text Comparison)	Speaker Differentiation	Inclusion of Timestamps	User Friendliness
Trint*	68.13% match	Must edit speakers in the main service platform.	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will populate based on your edits. This service also offers the most accurate timestamp (120 frame rate).	Trint is a great service as well. It allow easy edits within the document. This s has one of the most options to export: document, SRT, VTT, STL, EDL, HT XML, CSV, or text file. There are also decent number of options in terms of platform capabilities.
Sonix	75.86% match	Must edit speakers in the main service platform.	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will populate based on your edits.	Sonix is a little bit more accurate than and it has a similar number of options terms of platform capabilities. It allow easy edits within the document and is straightforward. We rank Trint higher because the timestamps tend to be a bi better. This service has one of the mos options to export: Word document, PD SRT, VTT, SESX, XML, FCPHML, c file.

Service Name	Transcription Accuracy (Average Text Comparison)	Speaker Differentiation	Inclusion of Timestamps	User Friendliness
Happy Scribe	74.6% match	Must edit speakers in the main service platform.	Timestamps are present at each turn, and are pretty close as well (time frame looks high). It is possible to edit within the platform and the timestamps will populate based on your edits.	HappyScribe is another good option, b was simply not as accurate as the othe services. It has a similar number of op in terms of platform capabilities. It all for easy edits within the document and accurate timestamps. This service has the most options to export: Word docu PDF, SRT, VTT, SESX, XML, FCPH STL, HTML, XML, JSON, or text file
Wreally	79.5% match	Must edit speakers in the main service platform.	Timestamps are present at determined intervals. It is possible to add a timestamp, but they tend to not be as accurate.	Wreally is one of the most accurate se with text comparison, but it is not a gr service overall. It allows for edits with document, but it is not as clear and eas with other services. The options to exp are only with a text file. There are a li number of options in terms of platform capabilities.

Service Name	Transcription Accuracy (Average Text Comparison)	Speaker Differentiation	Inclusion of Timestamps	User Friendliness
Ebby	67.58% match	Distinguishing speakers is not great; can edit within the document but automatic generation is not great.	Timestamps are present at each turn. It is possible to edit within the platform and the timestamps will populate based on your edits.	Ebby is one of the least accurate servi with text comparison, although it does a decent amount of platform capabiliti allows for edits within the document. options to export are: Word document SRT, VTT, HTML, or text file. This is decent service, but just not as accurate others.
Scribie	58.36% match	Distinguishing speakers is not great; can edit within the document but automatic generation is not great.	Timestamps are present only where you place them. It is possible to add a timestamp, but they tend to not be as accurate.	Scribie is one of the least accurate ser with text comparison, and does not ha many platform capabilities. It allows f edits within the document with its onli editor. The options to export are: Wor document, PDF, SRT, VTT, ODT, or file. This is a decent service, but just n accurate as the others.

Table A2: YouTube Videos

Name	Details	Round	Length	Num Parti
Anthony Rizzo On Chicago Cubs Rivalries & Baseball Superstitions While Eating Spicy Wings | Hot Ones	Hot Ones is a YouTube series where host Sean Evans asks celebrity guests questions while they eat chicken wings coated in ever-spicier hot sauce.	1	10:17	2 par
Christian Woman Interview-Shannon	Interview and portrait of Shannon, a Christian woman in West Virginia who shares her life story.	1	10:36	2 par
Penelope Cruz | Times Talks with the New York Times	From The New York Times: Penélope Cruz talks with NYT contributor Logan Hill about her latest film, "Mama," as well as her many wide–ranging roles and challenges & opportunities for women in film.	1	9:13	2 par
Unorthodox: Deborah Feldman's Escape from Brooklyn to Berlin | DW Interview	From DW News: Deborah Feldman tells the story of her life as an ultra-orthodox Jew and her rejection of Hasidic traditions.	1	11:58	2 par
Shah Rukh Khan, Bollywood Star | Journal Interview	From DW News: Interview of Shah Rukh Khan on Bollywood: Illusion and Reality in India's Dream Factory.	1	9:16	2 par

Name	Details	Round	Length	Num Parti
Suranne Jones' greatest weaknesses are coffee & Hugh Jackman | The Graham Norton Show	The Graham Norton show is a British talk show. Guests on this episode are: Suranne Jones, Hugh Jackman, Zack Efron, and Zendaya.	2	10:27	5 par
From child bride to global voice for women's empowerment, Dr. Tererai Trent shares her journey	From CBS This Morning: Dr. Tererai Trent, the author of "The Awakened Woman: Remembering & Reigniting Our Sacred Dream," shares her incredible journey from child bride in Zimbabwe to achieving her doctorate.	2	12:35	2 par
Pimp and Prostitute Interview- Master J and Little Mama	Soft White Underbelly interview and portrait of "Master J" and "Little Mama" in South Central Los Angeles.	2	16:32	3 par
German Soldier Remembers WW2 | Memoirs of WWII #15	From the YouTube series, Memoirs of WWII: Gert Schmitz talks about his life as a German soldier in WWII.	2	13:45	2 par
What's It Like Being a Foreigner in Korea? | ASIAN BOSS	ASIAN BOSS is a South Korea-based media company and in this video, they ask people walking by to answer questions.	2	10:49	5+ pa