

TOWARDS ACTIONABLE UNDERSTANDINGS
OF CONVERSATIONS:
A COMPUTATIONAL APPROACH

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Justine Zhang

August 2021

© 2021 Justine Zhang
ALL RIGHTS RESERVED

TOWARDS ACTIONABLE UNDERSTANDINGS OF CONVERSATIONS:
A COMPUTATIONAL APPROACH

Justine Zhang, Ph.D.

Cornell University 2021

Conversations are central to many consequential settings. Understanding how conversationalists navigate through them could unlock great improvements in domains like mental health, where the provision of social support is crucial. Such domains also present a promising opportunity for research: many interactions are recorded in large collections of transcripts, facilitating systematic analyses. In this dissertation, we take up this opportunity: we consider computational approaches to analyzing conversations, that can arrive at descriptively rich and prescriptively informative accounts of how conversationalists interact.

We start by proposing methodology to model two particular conversational phenomena. In the British House of Commons, we consider the wide range of rhetorical roles encompassed by the questions that legislators ask, and develop an unsupervised method to infer types of rhetorical roles given a dataset of questions and answers. In the context of a crisis counseling service, we develop a method to model how counselors orient the flow of complex and high-stakes interactions with people in mental health crises. We apply these methods to analyze the respective domains, drawing correspondences between interactional dynamics and broader aspects of the setting, such as a legislator's political standing or the effectiveness of a counseling conversation.

We then describe a general approach, the Expected Conversational Context

Framework, for modeling utterances in terms of their roles in a conversation. The framework's key idea is that we can derive a range of characteristics of an utterance by accounting for its *expected conversational context*—i.e., the distribution of preceding or subsequent utterances that could occur next to it in a conversation. Via a series of empirical explorations, we illustrate how the framework is generative of a variety of characterizations and analyses, including and beyond those proposed in our initial studies.

We end with a critical appraisal of the extent to which such approaches can arrive at actionable understandings. Drawing on a broad range of literature, ranging from sociological studies of interaction to causal inference, we consider the various complexities of conversations and the challenges they raise for methods such as ours.

BIOGRAPHICAL SKETCH

Justine Zhang was born in Calgary, Canada. She received her B.S. in Computer Science from Stanford University, and completed her Ph.D. in Information Science at Cornell University. Along the way, she's spent time in Palo Alto, Berlin, Saarbrücken, Ithaca, New York City, and Seattle.

Justine feels that the whole exercise of writing a biographical sketch is like an awkwardly mundane out-of-body experience. She suspects that the sketch might read better in the alternate universes where she became a paleontologist, a pianist, or a number theorist instead. Seriously, dinosaurs! Her seven-year-old self must be furious. Though—she supposes—this universe has its perks. There's ice cream, coffee shops, and plenty of places to take long walks.

To my fellow conversationalists.

ACKNOWLEDGEMENTS

It was during the long and peripatetic walks that things clicked into place, the fog was lifted, the ideas were turned over in my head until they became familiar and strange. It was during those walks that I got lost, and lost in conversation with others. I'm thinking of the times I circumnavigated Ithaca, skipped down the steps along Cascadilla Gorge, strolled around Beebe Lake looking for ducklings. I'm thinking, also, of when we walked from Seattle's Capitol Hill all the way to U District (a mostly flat journey), wandered around the hills and streets and museums of Berkeley, Florence, Manhattan and Montreal, through the beautiful mountain meadows of Banff National Park. I'm grateful for those walks and for the people I took them with.

I'd like to thank my advisor, Cristian Danescu-Niculescu-Mizil, for many things. Where to start? At the end, with his continued guidance, all the insurmountable obstacles transformed into rewarding journeys. In the middle, when the winds picked up into a storm, and when they were frustratingly still, I'm grateful for his encouragement and wisdom in all sorts of weather. And at the beginning, when I was an undergraduate who thought of her future as a big empty space, who had no idea how to do research or write or think, I'm grateful he took a chance on me. Along the way, he taught me how to do research and write and think, and so much more. We had some hard and valuable and wonderful conversations. I'd also like to think we had a lot of fun, and that there are a few more open parentheses lurking in our whiteboard scribbles.

I'd also like to thank my committee members, Jon Kleinberg and Lillian Lee. From taking their classes, going to their office hours, and receiving their careful, generous feedback on papers and talks, I learned a lot about the craft of doing research and chewing on problems. I remember and treasure miserable, rainy,

post-paper-rejection days that were brightened up by a conversation with them.

I'm extremely grateful for the opportunity to work with Crisis Text Line. It was during this collaboration that I really came to appreciate the extent to which conversational behaviours could be richly crafted and deeply impactful. I'd especially like to thank Christine Morrison and Jaclyn Weiser for their help, support and insight.

I had the great fortune of working with many fantastic collaborators. In particular, I had a wonderful, formative experience wandering through the research weeds with Will Hamilton. Arthur Spirling pointed us to the parliamentary questions setting, guided us through it, and reminded me that academic writing can make you smile. Sendhil Mullainathan has been a constant source of wisdom during and beyond our causal inference adventures, and continually reminds me that adventures are fun.

I met many of the characters in this dissertation during internships at Facebook and at Microsoft Research. Thanks to Sean Taylor for introducing me to causal inference and for cheering me on through an ambitious and tangly project. Thanks to Susan Dumais, Eric Horvitz (cc) Jamie Pennebaker for pointing me to much of the literature that informed Chapters 6 and 7, for telling me stories about your emails, and for your inspiring intellectual energy.

Deciding to do my Ph.D. at Cornell Information Science was one of the best decisions I've ever made. I'm grateful for the support and wisdom from—and random hallway/kitchen/Zoom chats with—Yoav Artzi, Claire Cardie, Paul Ginsparg, Malte Jung, Karen Levy, David Mimno and Phoebe Sengers.

Many fantastic friends made my Ph.D. experience rich and full of laughter. Thanks to Maria Antoniak, Su Lin Blodgett, Emily Tseng and Angela Zhou for inverted reading group, a beacon in a pandemic year. Thanks to Jonathan

Chang, Liye Fu, Jack Hessel, Hajin Lim, Xanda Schofield and Laure Thompson for being amazing officemates, excellent feedback-givers, and willing commiserators. Thanks to Xiao Ma for brunch, cocktails, and your invigorating curiosity in everything. Thanks to Angelina Garron for the postcards and the sweets. Thanks to Vlad Niculae for teaching me so much about Python, machine learning, indie music, and being a good person, for being my first friend in Ithaca, and for continuing to exchange weird Internet things with me.

I'm infinitely grateful to my parents, Xue Yan and Hongwen Zhang, who taught me how to read, ski, and do algebra, and who ferried me between home, piano lessons, the airport, and the mountains. They instilled in me a foundational love of learning, and they are the harbour I return to. I'd also like to thank my brother, Albert, for being in cahoots with me all the time.

Finally, I'd like to thank Ali Alkhatib for introducing me to many words, like *peripatetic* and *praxis*—not just so I can say they *technically* made it into the dissertation. Thank you, also, for being all the things that make you the last person I name in these acknowledgements.

I am grateful to have been supported by a Cornell Information Science Fellowship and a Microsoft Research Ph.D. Fellowship. The collaboration with Crisis Text Line was supported by the Robert Wood Johnson Foundation; the views expressed here do not necessarily reflect the views of the foundation. The work in this dissertation was also supported in part by a Discovery and Innovation Research Seed Award from the Office of the Vice Provost for Research at Cornell, a Google Faculty Award, NSF Grant SES-1741441 and NSF CAREER Award IIS1750615.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	viii
List of Tables	xi
List of Figures	xiv
1 Introduction	1
1.1 Organization and contributions	2
1.2 An overview of motivations	5
I Phenomena	9
2 Modeling the rhetorical role of questions in parliamentary discourse	10
2.1 Overview	10
2.2 Introduction	11
2.3 Related work	13
2.4 Setting: Parliamentary question periods	15
2.5 Inferring latent question types	17
2.6 Application to question periods data	20
2.6.1 Inferred question types	21
2.6.2 Validation: Labeled data	23
2.6.3 Validation: Relation to party affiliation	26
2.6.4 Relation to answerer's department	29
2.6.5 Relation to career trajectory	31
2.7 Discussion	34
3 Modeling the orienting role of utterances in counseling conversations	36
3.1 Overview	36
3.2 Introduction	37
3.3 Setting: Counseling conversations	39
3.4 Background and related work	40
3.5 Measuring orientation	43
3.5.1 High-level sketch	43
3.5.2 Operationalization	47
3.6 Application to counseling data	50
3.6.1 Validation: Counseling strategies	52
3.6.2 Validation: Conversation structure	53
3.6.3 Alternative measures	55
3.6.4 Relation to conversation effectiveness	56
3.6.5 Relation to message construction	61
3.7 Discussion	61

II	Framework	63
4	The Expected Conversational Context Framework	64
4.1	Overview	64
4.2	Conceptual description	65
4.2.1	Conversational context	65
4.2.2	Expected context	66
4.2.3	Inference from conversation data	68
4.2.4	Embedding-based approach	69
4.2.5	Derived characterizations	70
4.3	Related work	73
4.4	Methodology	77
4.4.1	Specifying the input data	77
4.4.2	Deriving latent term representations	79
4.4.3	Aggregating from term to utterance-level representations	85
4.4.4	Particular implementation choices	87
5	Exploring framework output	88
5.1	Overview	88
5.1.1	Examples of nearby vector representations	90
5.2	Exploring latent representations	94
5.2.1	Characterizing answers in parliamentary discourse	95
5.2.2	Characterizing messages in counseling conversations	102
5.2.3	Other choices of conversational context: next turn	110
5.3	Exploring derived properties	113
5.3.1	Expectation strengths	113
5.3.2	Characterizing expected shifts	116
5.3.3	Shifts to next turn	120
5.4	Utterance- vs. conversation-based representations	121
5.5	Comparing expected and actual replies	131
5.5.1	Properties of the unexpectedness measure	133
5.5.2	Analysis of parliamentary question periods	139
5.5.3	Analysis of counseling conversations	142
5.6	Application to other datasets	147
5.6.1	Forwards representations in Wikipedia discussions	147
5.6.2	Orientation in US Supreme Court oral arguments	149
5.6.3	Exploration of the Switchboard Dialog Act Corpus	151
5.7	Discussion	156
III	Action?	159
6	Towards actionable understandings?	160
6.1	Overview	160

6.2	Conversational complexities	161
6.2.1	Conversational context	162
6.2.2	Situational context	164
6.2.3	Implications for descriptive accounts	165
6.3	Challenges for deriving prescriptive conclusions	168
6.3.1	Analysis of the causal inference task	171
6.3.2	Empirical demonstration	183
6.4	Discussion	188
7	Conclusion and future work	192
7.1	Epistemological tensions	192
7.2	Future directions	195
7.2.1	Conversations as processes	195
7.2.2	Conversations as parts of broader tasks	197
7.2.3	Conversational professions	198
A	Further methodological details	199
A.1	UK parliamentary question periods	199
A.2	Crisis counseling conversations	201
A.3	Other datasets	203
A.3.1	Wikipedia talk page discussions	203
A.3.2	US Supreme Court oral arguments	204
A.3.3	Switchboard Dialog Act Corpus	205
B	Further examples	206

LIST OF TABLES

2.1	Representative examples of question and answer terms, and (abbreviated) question-answer pairs, for each question type inferred from the parliamentary question periods data. Bolded terms are present in the original text.	22
3.1	Example terms and sentences with labeled strategies from crisis counselors' messages, at varying orientations: backwards-oriented (from the bottom 25% of Ω), middle, and forwards-oriented (from top 25%).	51
3.2	Counseling strategies and representative examples derived from the training material. The number of sentences (out of 400) assigned to each label is shown in parentheses (11 were not labeled as any action).	52
4.1	Overview of characterizations of terms w and utterances a derived via the Expected Conversational Context Framework, and empirically examined in the indicated dissertation chapters. . . .	71
4.2	Notation and key equations used in computationally operationalizing the Expected Conversational Context Framework. . .	78
5.1	Example parliamentary question terms w , with question and answer terms whose latent representations are close to forwards-representation $\vec{\phi}(w)$	91
5.2	Example counselor terms w , with counselor and texter terms whose latent representations are close to forwards-representation $\vec{\phi}(w)$ or backwards-representation $\overleftarrow{\phi}(w)$	91
5.3	Example parliamentary questions a , with questions and answers whose latent representations are close to forwards-representation $\vec{\Phi}(a)$	92
5.4	Example counselor messages a , with counselor and texter messages whose latent representations are close to forwards-representation $\vec{\Phi}(a)$ or backwards-representation $\overleftarrow{\Phi}(a)$	93
5.5	Representative examples of answer and question terms, and question-answer pairs, for each answer type inferred from the parliamentary question periods data.	97
5.6	Examples of counselor (C) and texter (T) terms, and message-reply pairs, for each inferred forwards type.	103
5.7	Examples of counselor (C) and texter (T) terms, and message-predecessor pairs, for each inferred backwards type.	104

5.8	Representative examples of counselor terms and terms in subsequent counselor messages (C and C', respectively), and pairs of successive counselor messages, for each skip type inferred from the counseling conversation data.	112
5.9	Examples of counselor terms with low and high δ (in the bottom and top 25%). For examples w with low δ , we show texter terms whose representations are close to $\overleftarrow{\phi}(w)$ and $\overrightarrow{\phi}(w)$. For w with high δ , we show examples of texter terms which are close to $\overleftarrow{\phi}(w)$, contrasting them with examples that are close to $\overrightarrow{\phi}(w)$. . .	118
5.10	Examples of parliamentary question terms with low and high d (in the bottom and top 25%), comparing forwards - and LSA representations. For w with low d , we show other question terms whose representations are close to $\overrightarrow{\phi}(w)$ and $\overleftarrow{\phi}(w)$. For w with high d , we show examples of terms which are close to $\overrightarrow{\phi}(w)$, contrasting with examples close to $\overleftarrow{\phi}(w)$	126
5.11	Examples of counselor terms with low and high d (in the bottom and top 25%), comparing forwards - and LSA representations. For w with low d , we show other counselor terms whose representations are close to both $\overrightarrow{\phi}(w)$ and $\overleftarrow{\phi}(w)$. For w with high d , we show examples of terms that are close to $\overrightarrow{\phi}(w)$, contrasting with examples close to $\overleftarrow{\phi}(w)$	127
5.12	Examples of counselor terms with low and high d (in the bottom and top 25%), comparing backwards - and LSA representations. For w with low d , we show other counselor terms whose representations are close to both $\overleftarrow{\phi}(w)$ and $\overrightarrow{\phi}(w)$. For w with high d , we show examples of terms that are close to $\overleftarrow{\phi}(w)$, contrasting with examples close to $\overrightarrow{\phi}(w)$	128
5.13	Examples of question-answer pairs in the parliament setting with low or high unexpectedness.	134
5.14	Examples of counselor message-texter reply pairs in the counseling setting with low or high unexpectedness.	135
5.15	Examples of terms and comments, for each comment type inferred from the Wikipedia talk page discussions data.	147
5.16	Example terms and sentences from utterances of Supreme Court justices which are more backwards- or more forwards-oriented (bottom and top 25% of Ω).	150
B.1	Representative examples of question terms and questions, for each parliamentary question type.	207
B.2	Representative examples of answer terms and answers, for each parliamentary answer type.	208

B.3	Representative examples of counselor (C) terms and messages, for each forwards type.	209
B.4	Representative examples of counselor (C) terms and messages, for each backwards type.	210

LIST OF FIGURES

2.1	Positive pointwise mutual information statistics between question types (columns) and annotated question labels (rows). Darker squares denote types and labels that are more associated with each other. \uparrow indicates the label occurs significantly more in-type than overall (Fisher’s exact test $p < 0.05$, Bonferroni-corrected in the number of types), \downarrow indicates less.	25
2.2	Log-odds ratios of questions of each type asked by government compared to opposition MPs.	27
2.3	Mean propensities for each question type, for MPs who switch from being in the opposition to being in government (top), and vice versa (bottom) after an election; the left and right points in each type denote propensities before and after the switch, while \bullet and \square denote propensities for government and opposition-affiliated MPs, respectively. Stars indicate statistically significant differences at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) levels (Bonferroni-corrected Wilcoxon test).	27
2.4	Positive pointwise mutual information between question type and the department of the minister answering questions; darker squares indicate that the corresponding type is overrepresented in the department, relative to random chance. Departments are shown in descending order in terms of number of questions asked. \uparrow indicates the type occurs significantly more often in-department vs. overall (Bonferroni-corrected Fisher’s $p < 0.05$), \downarrow indicates less often.	29
2.5	Median asker tenures over each question type, for government (\bullet) and opposition (\square) askers. Overall median tenures are also shown for reference (solid blue line for government, dashed red line for opposition).	32
2.6	Mean propensities for each question type among MPs at various career stages. For government MPs, \bullet , \blacksquare , \blacktriangle denote propensities in their first five years, fifth to tenth year, and after their tenth year, respectively. For opposition MPs, \bullet and \blacktriangle denote propensities in their first five years, and after their fifth year. * and \wedge indicate statistically significant differences at the $p < 0.05$ (*, \wedge), $p < 0.01$ (**, $\wedge\wedge$) and $p < 0.001$ (***, $\wedge\wedge\wedge$) levels (Wilcoxon test); * compares the first two stages for both affiliations, \wedge compares the next two stages for government MPs.	32

3.1	Two possible exchanges in a counseling conversation, illustrating key objectives that a counselor must balance: in c_1 , the counselor aims to <i>advance</i> the conversation towards a discussion of possible confidants; in c_2 , they aim to <i>address</i> the emotion underlying the preceding utterance.	38
3.2	Words representative of replies and predecessors for utterances with two example terms, as observed in training data. Top row: observed replies to utterances with w_1 span a narrower range than observed predecessors (relative sizes of red and blue circles); w_1 thus has smaller <i>forwards-range</i> $\vec{\sigma}_{w_1}$ than <i>backwards-range</i> $\overleftarrow{\sigma}_{w_1}$ (i.e., it is forwards-oriented, $\Omega_{w_1} > 0$). Bottom row: observed predecessors to utterances with w_2 span a narrower range than replies; w_2 thus has smaller $\overleftarrow{\sigma}_{w_2}$ than $\vec{\sigma}_{w_2}$ (i.e., it is backwards-oriented $\Omega_{w_2} < 0$).	45
3.3	Outline of steps to compute the orientation Ω_w of term w , as described in Section 3.5.2. Panels A-D show the procedure for computing forwards-range $\vec{\sigma}_w$; the procedure for backwards-range $\overleftarrow{\sigma}_w$ is analogous.	48
3.4	Validating the orientation measure and comparing to alternatives. A: Leftmost: Mean Ω per counseling strategy label (vertical line denotes $\Omega = 0$). Next three: same for other measures. B: Mean Ω^{\max} and Ω^{\min} per segment for risk-assessed (orange) and non-risk-assessed (black) conversations. Both: Solid circles indicate statistically significant differences (Wilcoxon $p < 0.01$, comparing within-counselor).	54
3.5	Mean naive distance, backwards-range ($\overleftarrow{\sigma}$), and % of messages with questions, per segment for risk-assessed (orange) and non-risk-assessed (black) conversations; solid circles indicate statistically significant differences (Wilcoxon $p < 0.01$, comparing conversation types within counselor).	54
3.6	Relation between orientation and conversational effectiveness. A: Mean Ω^{\min} and Ω^{\max} in conversations rated as helpful (green) or unhelpful (grey) (macroaveraged per conversation). Differences in both measures are significant (Mann Whitney U test $p < 0.001$). B, C: Mean Ω^{\min} and Ω^{\max} of conversations with varying lengths (in # of messages). Both plots: Error bars show 95% bootstrapped confidence intervals.	58
4.1	Sketch of procedure to compute latent term representations, as outlined in Sections 4.2.3 (left) and 4.2.4 (right).	68

5.1	Positive pointwise mutual information statistics between question types (rows) and answer types (columns). Darker red denotes that an answer of that type is more likely to follow a question of that type, relative to random chance. \uparrow and \downarrow indicate significant differences in each direction (Fisher's $p < 0.05$, Bonferroni-corrected in the number of possible type pairs).	99
5.2	Log-odds ratios of answers of each type given to askers who are affiliated with the government, versus the opposition.	100
5.3	Positive pointwise mutual information statistics between answer types (columns) and annotated labels of answers (rows). Darker squares denote types and labels that are more associated with each other. \uparrow and \downarrow indicate statistical significance.	101
5.4	Positive pointwise mutual information statistics between annotated labels and (left) forwards types or (right) backwards types of sentences written by counselors. Darker red denotes types and labels that are more associated with each other. \uparrow and \downarrow indicate statistical significance; white squares indicate the type and label did not co-occur in the data.	106
5.5	Proportion of sentences occurring in each segment of the conversation for (left) forwards types and (right) backwards types. Darker squares indicate segments of the conversation where each type occurs more often.	107
5.6	Positive pointwise mutual information statistics between forwards types (rows) and backwards types (columns) for sentences in counselor messages. Darker squares indicate that sentences are more likely than chance to be of a particular forwards type and backwards type.	109
5.7	Distributions of $\vec{\Sigma}$ of questions for each parliamentary question type. Points correspond to median values, while error bars denote bottom and top quartiles. Median $\vec{\Sigma}$ over all questions is shown as the dotted line.	115
5.8	Distributions of $\vec{\Sigma}$ and $\overleftarrow{\Sigma}$ for counselor sentences per (top) forwards and (bottom) backwards type. Rows correspond to $\vec{\Sigma}$, binned into tertiles; columns correspond to $\overleftarrow{\Sigma}$. Bottom left corner corresponds to utterances with low forwards- and backwards-range; upper right corner corresponds to utterances with high forwards- and backwards-range.	116
5.9	Distributions of Δ for sentences of each (left) forward type and (right) backward type in the counseling conversation setting, shown as box plots indicating quartiles. Median Δ over all sentences is shown as the dotted lines in each plot.	119

5.10	Distributions of Δ' for sentences of each skip type in the counseling conversation setting, shown as box plots. Median Δ' over all sentences is shown as the dotted line.	121
5.11	Histogram of d for each question term in the parliament setting. Dotted grey line indicates median d over all terms.	124
5.12	Histogram of d for each counselor term, comparing LSA representations with (left) forwards- and (right) backwards-representations of terms. Dotted grey lines indicates median d over all terms, for each comparison.	124
5.13	Distributions of d for parliamentary questions of different types, shown as box plots. Median d over all questions is indicated by the dotted line.	129
5.14	Distributions of d for counselor sentences of different (left) forwards and (right) backwards types, shown as box plots. Median d over all sentences is indicated by the dotted lines.	129
5.15	Log-odds ratios of comment types exhibited in the first and second comments of conversations that turn awry, versus those that stay on track. Δ and \square denote log-odds ratios in the first and second comments, respectively; points are solid if they reflect significant (two-tailed binomial test $p < 0.05$) differences. * denotes statistical significance at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) levels for the first comment; + denotes corresponding p -values for second comment.	149
6.1	Graphical representations of the key dependencies underlying the inference task, between tendency \mathcal{T} , outcome Y , behaviour B and context C . Our goal is to estimate the effect of tendency on outcomes (blue path), however the contexts in which the behaviours and outcomes are observed confound this estimation (red arrows).	173
6.2	Graphical representations of the dependence between assignment A and outcome Y through behaviour B and context C that result from the observational nature of our analyses, giving rise to the selection bias exposed in (6.3). I : the problematic pathways from A to Y ; II : an idealized setting where conversations are randomly assigned to agents, in which the dependency is trivially broken; III : a scenario where assignment is governed by a set of observable <i>selection variables</i> S	176
6.3	Graphical representations of the entanglement between conversational context C , behaviours B and outcomes Y , giving rise to the bias in (6.6). I : dependencies in a non-interactional setting; II : problematic dependencies when B interacts with contexts C that also shape Y ; III : our approach, observing behaviours and outcomes on different splits of data.	181

6.4 Relation between counselor-level behavioural tendencies and outcomes, measured as Kendall's tau correlations, for different estimation approaches: \triangle correlates counselor behaviour and outcome propensity; \square computes this correlation across temporally-interleaved splits of conversations; \circ further controls for shift time, thus reflecting the allocation effect formulated in Equation 6.1 while accounting for the inference challenges we described. Error bars show bootstrapped 95% confidence intervals; shapes are filled for bootstrapped and Bonferroni-corrected $p < 0.01$. Abbreviated citations indicate studies that have demonstrated correlational relationships between the respective behaviours and outcomes. 187

CHAPTER 1

INTRODUCTION

Consider the following example: in a crisis counseling service, people in mental health crises have one-on-one conversations with trained counselors. During such an interaction, the counselor aims to guide a distressed individual towards a calmer mental state, through empathetically exploring their situation, and through working together to find ways to cope. In short, in the span of a conversation, the counselor must build a supportive, meaningful connection with a total stranger—with a lot at stake.

We highlight two key aspects of this scenario. First, the counseling service is an example of a *conversational setting*, premised on having rich, complex and often challenging conversations. Second, the conversationalists in this setting use conversations to accomplish a consequential task, such as helping someone in crisis find their footing. Taken in conjunction, these aspects suggest an important scientific opportunity: by analyzing conversations, we could arrive at *actionable understandings*—that inform ways to help conversationalists more effectively have conversations to accomplish broader tasks. The counseling example, in particular, underscores the potential impact of such a research agenda.

Note that there are many other consequential, conversational settings, in which analyzing conversations could lead to actionable implications. Consider legislators engaging in policy discussions and holding governments to account [Thomas et al., 2006, Eggers and Spirling, 2014]; interactions among students and teachers in a classroom [Nystrand et al., 2003, Demszky et al., 2021], exchanges between justices, lawyers and witnesses in a courtroom [Atkinson and Drew, 1979, Danescu-Niculescu-Mizil et al., 2012], between physicians and patients [Beckman and Frankel, 1984, Ford et al., 1996], between police officers

and community members at traffic stops [Epp et al., 2014, Voigt et al., 2017]. As an example of particular relevance when this dissertation was written, consider the job of contact tracers [Akam, 2020, Becker, 2020], who talk to community members to glean information about the spread of an infectious disease. In all of these settings, the interactions between the conversationalists involved play key roles in broader, societally important causes. Analyzing these interactions could help conversationalists, or help to hold them to account.

This dissertation focuses on a methodological possibility: in many conversational settings, including several listed above, interactions are recorded in large collections of transcripts. By capturing and examining aspects of conversations in these transcripts, computational methods could yield descriptions of conversational dynamics that are grounded in, and that are able to account for systematicities and variations across these vast records. An increasing body of computational work has used methods from natural language processing (NLP) to model and analyze these interactions. The central question we consider, as we draw on and add to these efforts, is this: how can computational approaches account for the inherent complexities of conversations, and arrive at actionable understandings of them?

1.1 Organization and contributions

At a high level, we address this question in terms of two criteria: an actionable understanding must somehow be *descriptively* meaningful and *prescriptively* informative. In other words, we'd like to arrive at sufficiently rich accounts of how conversationalists interact, and we'd also like to rigorously point to the effects of these interactional dynamics, and of policies that intervene on them.

We start with the descriptive problem. First, we consider two particular conversational phenomena that are salient in two particular settings, and that point to questions of substantive importance in those domains. For each phenomenon, we present a computational method to model it, and apply the method to examine how it occurs in large conversation datasets. At a high level, both phenomena reflect some aspect of how an utterance fits into an interaction:

- In parliamentary venues, such as the British House of Commons, weekly *question periods* are held in which legislators ask questions to, and theoretically receive answers from, government ministers. These questions play a wide range of rhetorical roles, from narrow requests for information, to pointed criticism, to praise. Systematically examining these varied question-asking behaviours could provide insights into how legislators engage in political discourse. In Chapter 2, we detail an unsupervised method for inferring types of rhetorical roles spanned by questions, and apply it to analyze a dataset of parliamentary question periods.
- Counselors in a crisis counseling conversation must balance between multiple imperatives in trying to help a person in crisis—they must nudge them towards potential solutions without rushing the conversation, and must patiently address what’s being disclosed with empathy without stalling the interaction. Examining how counselors strike this balance could provide insights into the process of counseling that might point to ways of helping them deal with such conversational challenges. In Chapter 3, we describe a method to capture the degree to which an utterance is intended to direct the flow of a conversation forwards or backwards, and apply it to analyze a dataset of crisis counseling conversations.

In Chapter 4, we describe a broader computational framework to model utterances. The framework’s key idea is that we can arrive at a range of characteristics of an utterance’s interactional role by accounting for its *expected conversational context*—i.e., the range of replies or preceding utterances that could plausibly occur next to it in a conversation. Our technical contribution is a method to account for and quantify aspects of this expected context, given conversation data. We ground the framework in ideas from discourse and conversation analysis, and show how it complements or builds on other computational methods for studying conversations; in particular, we illustrate that the methods presented in Chapters 2 and 3 can be seen as particular instances of the broader approach. In Chapter 5, we also show how the framework generates a variety of utterance characterizations, which we explore on the parliament and counseling datasets, along with other settings.¹

In the remainder of the dissertation, we critically appraise the extent to which such approaches can arrive at actionable understandings. In Chapter 6, we discuss the framework’s limitations, in terms of the extent to which it addresses key complexities of conversations. By noting its shortcomings, and the ways in which these shortcomings show through in our empirical analyses, we approach ideas familiar to literature in sociology and anthropology: that conversations are deeply embedded in and informed by the particular interactional and situational contexts in which they arise. In contrast, we suggest that computational frameworks like the one we’ve presented don’t adequately address the contextual and particular nature of conversations (even if they nominally “account for context”). We then illustrate how these conversational complex-

¹Code implementing the framework, and demonstrating its use on the public datasets we considered in this dissertation, is available via the ConvoKit library [Chang et al., 2020], at <https://convokit.cornell.edu/>.

ities directly lead to difficulties in establishing prescriptive insights. Drawing on the causal inference literature, we translate these fundamental qualities of a conversation to mathematical challenges that analysts must address in order to rigorously establish that a policy enacted in a conversational setting will actually lead to a good outcome. We also take note of remaining nuances that the causal analysis leaves unaddressed.

In sum, we end the dissertation leaving our central question open. In Chapter 7, we raise some epistemological questions that have emerged in this research. We also suggest some future directions for computational work that reiterate the rich and complex nature of conversational settings, and that could enrich our descriptive and prescriptive understandings of conversations.

1.2 An overview of motivations

In aiming at actionable understandings, the work in this dissertation reflects a wide range of motivations, found in a wide range of literature. Here, we provide a brief overview of the varied ideas we draw on, illustrating the multifaceted nature of our overall goal.

Scholarship across human-computer interaction, social psychology and NLP has sought to support or augment conversations and the broader tasks they enable. Such work has tackled a wide variety of problems, including fostering better online discussions [Zhang, 2018], facilitating successful collaborations [Tausczik and Pennebaker, 2013, Cao et al., 2020], improving the provision of emotional support [Choudhury and Kıcıman, 2017, Yang et al., 2019], and informing more effective mental health conversations [Althoff et al., 2016, Pérez-Rosas et al., 2018]. A central idea in much of this work, and ours, is that ana-

lyzing data of interactions can inform policies to improve them; a key question is how analyses can then be translated to actionable insights. Here, we look to the causal inference literature, which is in large part concerned with statistically bridging this gap to “help decision makers make better decisions” [Hernán and Robins, 2020]. We more substantively examine this idea in Chapter 6.

A large body of NLP work has been concerned with identifying linguistic signals of social variables, drawing on and extending ideas from the social sciences [Nguyen et al., 2015]. The scope of this work is quite vast, ranging from studies of linguistic and communicative factors like politeness [Danescu-Niculescu-Mizil et al., 2013a], sentiment [Pang et al., 2002], empathy [Pérez-Rosas et al., 2017, Sharma et al., 2020], persuasiveness [Tan et al., 2016, Zhang et al., 2016], deceptiveness [Ott et al., 2012], and linguistic coordination [Danescu-Niculescu-Mizil et al., 2012], to attributes like status [Gilbert, 2012, Prabhakaran and Rambow, 2013], role [Yang et al., 2015, 2019] or standing in a community [Danescu-Niculescu-Mizil et al., 2013b, Hamilton et al., 2017]. Methodologically, what ties these studies together is a focus on data derived from, or at least intended to emulate real-world interaction, and analytic approaches that aim to associate linguistic indicators and social factors by way of statistical correlations or prediction tasks. We draw on this paradigm as a way of interpreting and probing the social significance of the characterizations we derive in Chapters 2, 3 and 5: by showing that a quantitative property reflects a social attribute like career trajectory or conversation quality, we provide some evidence that the quantity is somehow descriptively meaningful.

In seeking to systematically describe conversations, we echo the goals espoused in sociological approaches—in particular, in conversation analysis [Schiffrin, 1994, Hoey and Kendrick, 2017]. Such work has focused on exam-

ining everyday interactions, as well as more organizational and institutional exchanges like those found in medical or legal domains [Heritage and Clayman, 2011].² Its aim is describing the various processes that social interaction is comprised of, rather than analyzing interaction as a way to access other variables; as informative precursors, consider also Goffman’s work on rituals and face acts as examining “the traffic rules of interaction” [Goffman, 1955] or Austin’s theory of how people “do things with words” [Austin, 1962]. As Sacks [1989b] puts it, the focus is on “[constructing] the objects that get used to make up ranges of activities, and then [seeing] how it is these objects get used.” In practice, this is accomplished by examining recurring instances of a conversational phenomenon, and then coming up with a formal account of the varied ways in which it occurs. We can view the framework and analyses we present in Chapters 4 and 5 as a rough computational parallel to these ideas: we use our method to extract and analyze instances of utterances that exhibit certain patterns, in how they are “expected to relate to” surrounding turns. We elaborate on further connections and notable contrasts in later chapters.

Our work has methodological roots in NLP research, which has conventionally been driven by other primary motivations: understanding human language use [Allen, 1995] and generating humanlike language [Weizenbaum, 1983, Brown et al., 2020]. We do not directly comment on generation, though note that our analyses and the caveats we later raise could fruitfully inform the development and deployment of natural language generation engines. In terms of natural language understanding, we draw a tentative distinction between the goal of achieving human-level understanding of language, and our goal of

²As an interesting connection, much of the foundational work in conversation analysis sprang out of Sacks’ studies of suicide hotline conversations, a setting closely related to the crisis counseling service we later examine [Sacks, 1992].

understanding—as analysts—how humans use language to interact. In practice, this means that the focus of this work is on analyzing interactions, rather than seeking to computationally approximate formal definitions of meaning and understanding. Of course, this distinction is not so clear-cut. Methodologically, understanding what utterances mean (in a computational sense) is a precursor to computationally analyzing how utterances get used. Several challenges have also been raised on the extent to which natural language understanding can be accomplished in a scientifically principled and socially just way [Winograd, 1980, Bender and Koller, 2020, Bisk et al., 2020, Bender et al., 2021], without accounting for factors exogenous to the language—notably, its social use.

Part I

Phenomena

CHAPTER 2
MODELING THE RHETORICAL ROLE OF QUESTIONS IN
PARLIAMENTARY DISCOURSE

2.1 Overview

Questions play a prominent role in social interactions, performing rhetorical functions that go beyond that of simple information exchange. The surface form of a question can be informative of the person asking it and their intention, as well as the nature of their relation with the interlocutor. While the informational nature of questions has been extensively examined in the context of question-answering applications, their rhetorical aspects have been largely understudied in the computational literature.

In this chapter we introduce an unsupervised method for characterizing and grouping questions according to their latent rhetorical role, allowing us to derive a typology of questions in terms of their rhetorical function. By applying this method to a dataset of question periods in the UK parliament, we show that the resulting typology reflects key aspects of the political discourse—such as the bifurcation in questioning behaviour between government and opposition parties—and reveals new insights into the relation between a legislator’s participation in political discourse, and their tenure and career ambitions.

Note on source material. This chapter was originally published in Zhang et al. [2017b], with Arthur Spirling and Cristian Danescu-Niculescu-Mizil. The work presented in this dissertation reflect the following updates. First, in the original paper, we represent questions as *motifs*—complex lexico-syntactic patterns. To simplify the discussion and to make our later generalization to the broader

framework in Chapter 4 smoother, we instead represent questions in terms of terms extracted from their dependency parses. Additionally, we made some minor improvements to how the latent representations of questions are computed and clustered to infer types, to match our general formulation. Together, these changes mean that the question types inferred and analyzed in this section are similar, but not identical, to those presented in the paper.

We also made some changes to the analyses; notably, we considered a larger subset of questions, reflecting slightly more permissive data filtering decisions. We added some discussion on the relation between question types and ministers' departments, and elaborated on how the question types correspond to labels from the annotated dataset we compare with; we also fixed some processing issues in the annotated data. In examining the relation between question type and tenure, we present a new, within-legislator analysis.

2.2 Introduction

Why do we ask questions? Perhaps we are seeking factual information or requesting a favour. Alternatively, we could be simply making a rhetorical point at the start of a dissertation chapter.

Questions play a prominent role in social interactions [Goffman, 1976], performing a multitude of rhetorical functions that go beyond mere factual information gathering [Kearsley, 1976]. While the informational component of questions has been well-studied in the context of question-answering applications, there is relatively little computational work addressing the rhetorical and social role of these basic dialogic units.

One domain where questions have a particularly important role is politics. The ability to question the actions and intentions of governments is a crucial part of democracy [Pitkin, 1967], particularly in parliamentary systems. Consequently, scholars have studied parliamentary questions in detail, in terms of their origins [Chester and Bowring, 1962], their institutionalization [Eggers and Spirling, 2014] and their importance for oversight [Proksch and Slapin, 2011]. In particular, the United Kingdom’s House of Commons, renowned for theatrical questions periods, has been examined in some depth, though those accounts are largely qualitative in nature [Bull and Wells, 2012, Bates et al., 2014].

The present work: methodology. We introduce an unsupervised framework to structure the space of questions according to their rhetorical role. Our key intuition is that this role can be inferred from the type of *answer* a question receives. To operationalize this intuition we construct a latent question-answer space in which terms within questions that trigger similar answers are mapped to the same region (Section 2.5).

The present work: application. We apply this general framework to analyze the discourse that occurs during parliamentary question sessions in the British House of Commons (Section 2.4). Our framework extracts intuitive question types ranging from narrow factual queries to pointed criticisms disguised as questions (Section 2.5, Table 2.1). We validate our framework by aligning these types with prior understandings of parliamentary proceedings from the political science literature (Section 2.6). In particular, previous work [Bates et al., 2014] has categorized questions asked in Parliament according to the intentions of the asker (e.g., to help the answerer, or to adversarially put them on the spot); we find interpretable correspondences between these expert-coded categories and the induced typology. We further show that the types of questions specific

legislators tend to ask vary with whether they are part of the governing or opposition party, consistent with well-established accounts of partisan differences [Cowley, 2005, Spirling and McLean, 2007, Eggers and Spirling, 2014]: government legislators exhibit a preference for overtly friendly questions, while the opposition slants towards more aggressive question types.

We then apply our methodology to further explore questioning behaviour. In particular, we provide new insights into how a legislator’s questioning behaviour varies throughout their career. The pressures faced by legislators at various stages in their career are cross-cutting, and multiple possible hypotheses emerge. Younger, more enthusiastic legislators may be motivated to ask harder-hitting questions, but risk being passed over for future promotion if they are too combative [Cowley, 2005]. Older legislators, whose opportunities for promotion are largely behind them and hence have “less to lose”, may act more aggressively [Benedetto and Hix, 2007]; or simply seek a quiet path to retirement. By enabling large-scale longitudinal analyses of legislators’ questioning behaviours, our method provides evidence for the latter hypothesis.

2.3 Related work

Questions have been examined in several sociological accounts, which point to their foundational nature in structuring and setting up subsequent interactions [Goffman, 1976, Sacks, 1989a, inter alia]. Here, we briefly survey computational approaches to analyzing questions and other interactional dynamics.

Question-answering. Computationally, questions have received considerable attention in the context of question-answering (QA) systems; for a survey, see Gupta and Gupta [2012]. Techniques have been developed to categorize ques-

tions based on the nature of these information needs, such as in the context of the TREC QA challenge [Harabagiu et al., 2003, Harabagiu, 2008], and to identify questions asking for similar information [Jeon et al., 2005, Shtok et al., 2012, Zhang et al., 2017c]. Questions have also been classified by topic [Cao et al., 2010] and quality [Treude et al., 2011, Ravi et al., 2014]. In contrast, our work is not concerned with the information need central to QA applications, and instead focuses on the rhetorical aspect of questions.

Question types. To facilitate retrieval of frequently-asked questions, Lytinen and Tomuro [2002] manually developed a typology of surface question forms (e.g., what- and why-questions) starting from Lehnerts’ conceptual question categories [Lehnert, 1977]. Dialog-based typologies have also been developed, distinguishing between yes-no, wh-, open-ended and rhetorical questions [Jurafsky et al., 1997, Core and Allen, 1997, Dhillon et al., 2004]. These typologies have been used to hand-annotate datasets, to the ends of automated categorization of questions. Complementary to this line of work, we introduce a completely unsupervised methodology, enabling analysts to derive domain-tailored question typologies and bypassing the need for human annotation.

Pragmatic dimensions. One key pragmatic dimension of questions that has been previously studied computationally is their level of politeness [Danescu-Niculescu-Mizil et al., 2013a, Aubakirova and Bansal, 2016]; in the context of making requests, politeness was shown to correlate with the social status of the asker. Previous research has also been directed at identifying rhetorical questions [Bhattachali et al., 2015], understanding the motivations of their “askers” [Ranganath et al., 2016], and distinguishing requests from general conversation [Sachdeva and Kumaraguru, 2017]. Using the relationship between questions and answers, our work examines the rhetorical and social aspect of questions

without committing to a particular pragmatic dimension or relying on labeled data. We also complement these efforts by analyzing other situations where questions may be posed without an information-seeking intent.

Political discourse. Finally, our work contributes to a rapidly growing area of NLP applications to political domains [Monroe et al., 2008, Gonzalez-Bailon et al., 2010, Grimmer et al., 2012, Grimmer and Stewart, 2013, Iyyer et al., 2014, Niculae et al., 2015b, Card et al., 2016, inter alia]. Particularly relevant examples have considered discourse in congressional and parliamentary settings [Thomas et al., 2006, Boydston et al., 2014, Rheault et al., 2016].

2.4 Setting: Parliamentary question periods

We focus on the questions asked and responses given during parliamentary question periods in the British House of Commons. Below, we provide a brief overview of key features of this political system in general, as well as a description of the question period setting.

Parliamentary systems. Legislators in the House of Commons (*Members of Parliament*, henceforth *MPs* or *members*) belong to two main voting and debating *affiliations*: a *government* party that controls the executive, and a set of *opposition* parties.¹ The executive is headed by the Prime Minister (PM) and run by a cabinet of *ministers*, high-ranking government MPs responsible for various departments such as finance and education.

Question periods. The House of Commons holds weekly, moderated *question periods*, in which MPs of all affiliations take turns to ask questions to (and the-

¹We use *affiliation* to refer broadly to the government and opposition roles, independent of the identity of the current government and opposition parties. In subsequent analyses we focus on the largest, “official” opposition party.

oretically receive answers from) government ministers for each department regarding their specific domains. Such events are a primary way in which legislators hold senior policy-makers responsible for their decisions. In practice, beyond narrow requests for information about specific policy points, MPs use their questions to critique or praise the government, or to promote their own agendas; indeed, certain sessions, such as Questions to the Prime Minister, have gained renown for their partisan clashes, often fueled by the (mis)handling of a current crisis. The following question, asked to the Prime Minister by an opposition MP about contamination of the meat supply in 2013, encapsulates this varied mix of purposes:

“The Prime Minister is rightly shocked by the revelations that many food products contain 100% horse. Does he share my concern that, if tested, many of his answers may contain 100% bull?”²

The moderated, relatively rigid format of question periods, along with the multifaceted array of underlying incentives and interpersonal relationships, results in a structurally controlled yet socially rich venue. This makes question periods a particularly fruitful setting in which to extend our understanding of questions beyond factual queries, and to study their social role.

Dataset description. Our dataset covers question periods from May 1979 to December 2016, encompassing six different Prime Ministers. For each question period, we extract all question-answer pairs, along with the identity of the asker and answerer. Because our focus here is on how questions are posed in a social setting we ignore questions which were pre-registered prior to the session; we also ignore any follow-up questions from the asker.³ We augment this collection

²MPs almost always address each other in 3rd person.

³Follow-up questions and extended dialogues occur very infrequently in question periods, and are generally restricted to a few specific askers, such as the Leader of the Opposition.

with further information about each asker and answerer, including their political party and the time when they first took office. Such information is used to validate our methodology, interpret our results, and perform further analyses, as described in Section 2.6.

In total there are 216,894 question-answer pairs in our data, occurring over 4,776 days and 6 prime-ministerships. The questions cover 1,975 different askers, 1,066 different answerers, and a variety of government departments with responsibilities ranging from defense to transportation.⁴

2.5 Inferring latent question types

Our framework aims to characterize questions according to their functional roles. The method starts by extracting features of the surface form of a question that encapsulate its functional nature. Ultimately, however, we would like to draw analogies between questions with similar rhetorical functions, even if their surface forms are different.

Our main intuition is that the nature of the *answer* that a question receives provides a good indication of the question’s intention. Therefore, if two questions are phrased differently but answered in similar ways, the parallels exhibited by their answers should reflect commonalities in their askers’ intentions. To operationalize this intuition, we derive a latent space based on answers, and then map terms within questions to the same space. Using the resultant latent representations, we can then cluster questions in terms of their rhetorical functions, beyond similarities or differences in surface form.

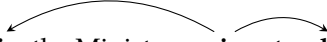
⁴The data can be accessed at <https://convokit.cornell.edu/documentation/parliament.html>.

Question terms. We start by extracting the key terms within a question that encapsulate its functional nature. Following the intuition that the bulk of this functional information is contained in the root of a question’s dependency parse along with its outgoing arcs [Iyyer et al., 2014], we take the terms of a question to be the root of its parse tree, along with each root-child pair. To capture cases when the operational word in the question is not connected to its root (as in wh-questions, e.g., “What...” or “Why...”), we also consider the initial unigram and bigram of a question as terms.

Because our goal is to capture rhetorical commonalities, agnostic to the topic of a question, we ignore all terms that contain a noun phrase (NP) or pronoun. NP subtrees are identified based on their outgoing dependencies from the root;⁵ in the event that an NP starts with a WH-determiner (WDT), we consider (root, WDT) to be a fragment and drop the remainder of the NP.⁶ Finally, we note that some questions consist of multiple sub-questions (“What does the Minister think [...], *and* why [...]?”). For such questions, we recursively extract terms from each child subtree in the same manner, starting from their roots.

Per our method, the following question has 5 terms: *going*, *is going* and *going do* (root-child pairs from the dependency parse); and *what* and *what is* (the initial unigram and bigram).

(1) **What is** the Minister **going** to **do** about ... ?



Constructing a space of answers. In line with our focus on functional characterizations, we extract the terms from each sentence of an *answer*, defined in

⁵We consider NPs as subtrees connected to the root with the following: nsubj, nsubjpass, dobj, iobj, pobj, attr.

⁶In this particular dataset, removing NPs also removes conventional, partisan address terms (e.g. “my hon. Friend”).

the same way as question terms. We then construct a term-document matrix \mathcal{A} , where columns correspond to answer terms, and rows correspond to documents, i.e., individual answers in the corpus. We tf-idf reweight the rows of this matrix and scale to unit norm, producing a term-answer matrix \mathcal{A} . We perform singular value decomposition on \mathcal{A} and obtain a low-rank representation $\mathcal{A} \approx \hat{\mathcal{A}} = UsV^T$, for some rank d , where rows of U correspond to answers and rows of V correspond to answer terms.

Latent representations of question terms. We can draw an intuitive correspondence between a question term w^q and answer term w^a if w^q occurs in a question whose answer contains w^a . We build on this idea to compute representations of question terms in the same space as $\hat{\mathcal{A}}$.

Concretely, we construct a tf-idf reweighted question-term matrix Q , where columns correspond to terms and rows correspond to questions. Importantly, Q and \mathcal{A} are aligned, in the sense that the i th row of Q represents a question whose answer is represented as the i th row of \mathcal{A} . To represent Q in the latent answer space, we solve for \hat{Q} in $Q = Us\hat{Q}^T$ as $\hat{Q} = Q^TUs^{-1}$, scaling rows to unit norm.⁷ Row j of \hat{Q} then gives a d -dimensional representation of term j .

Latent representations of questions. To represent a question q^* in the latent answer space, we first transform it to a tf-idf reweighted vector \bar{q}^* , whose j th entry corresponds to the weight of term w_j^q in the question. We project it into the latent space as $\hat{q}^* = \bar{q}^*\hat{Q}s^{-1}$.

Grouping similar questions. Finally, we identify *question types*. If two questions q^1 and q^2 have vectors \hat{q}^1 and \hat{q}^2 that are close together in the latent space, this means their constituent terms elicit answers that are close in the latent space; we

⁷Here, we use the fact that SVD derives orthonormal matrices U and V —such that $U^{-1} = U^T$ —and a diagonal matrix of singular values, s .

therefore infer that they are functionally similar. Formally, we use the K-Means algorithm [Macqueen, 1967] to cluster latent question vectors into k clusters; these clusters then constitute the desired typology of questions. Note that by assigning question term representations (rows of \hat{Q}) to these inferred types, we can interpret question types with respect to representative questions as well as typical terms.⁸

Since answers and answer terms have been mapped to the same latent space as well (as rows of U and V), we can also assign answers to question types. This further facilitates interpretability, in that we can inspect the answers commonly triggered by a particular type of question.

2.6 Application to question periods data

We now apply our framework to our dataset of parliamentary question periods, structuring the space of questions posed within these sessions according to their rhetorical function. We represent questions, answers and terms in a 25-dimensional latent space and induce a typology of 8 question types, choosing this number to capture a rich array of questions represented in this space while preserving interpretability. We include further implementation details in the appendix (Section A.1).

To validate the induced typology, we show that it recovers askers' intentions as labeled in an expert-coded dataset (Section 2.6.2); we also show that it qualitatively aligns with established understandings of parliamentary dynamics in

⁸Note that types could be inferred via clustering question representations (and then assigning terms to clusters), or via clustering *term* representations (and then assigning questions to clusters). We use the former approach here, and note that previous versions of the method, found in Zhang et al. [2017b] and Zhang et al. [2018] use the latter version.

the political science literature (Section 2.6.3). We then use the framework to further explore political discourse in Parliament, examining how questioning behaviour varies across different government departments (Section 2.6.4) and with a member's tenure in the institution (Section 2.6.5).

2.6.1 Inferred question types

Below, we outline the question types captured by our framework. Our descriptions draw heavily on interpretations provided by our collaborator on the original work, Arthur Spirling, a political scientist with domain expertise in the UK parliamentary setting. In Table 2.1, we show examples of questions, answers and terms per type, along with the frequency each type appears in the data; further examples are included in the appendix (Table B.1).

0. Demand for account. An aggressive demand for the minister to explain themselves or account for a perceived policy failure. Answers involve some amount of pushing back at the question, or voicing incredulity that the asker would put forth such an idea.

1. Shared concerns. A straightforward question with no strong ideological underpinnings; answers are typically vague and involve explaining that the government takes it seriously, will continue to do so and will consult with the relevant stakeholders.

2. Agreement. Airing a laudatory remark about a policy that the minister and MP clearly already agree on. Often these questions effectively serve as attempts to curry favour with the minister and bolster their (mutual) party.

<p>Demand for account (11.1% questions, 18.4% terms) Question terms: <i>can [you] explain, why does</i> Answer terms: <i>am surprised, is wrong</i> Q: Why does the Minister not admit that [...] there was never any evidence to support the decision? A: The hon. Gentleman is entirely wrong about the [contents] of the report [...]</p>
<p>Shared concerns (14.4% questions, 14.9% terms) Question terms: <i>will [you] ensure, agree [to] meet</i> Answer terms: <i>am interested, is obviously</i> Q: Will [the PM] ensure that any reduction applies to farmers across Europe, not just those in the UK? A: Obviously, our aim is [this] significant cut [...]</p>
<p>Agreement (12.1% questions, 9.9% terms) Question terms: <i>does [the Minister] agree, is important</i> Answer terms: <i>certainly agree, agree with</i> Q: Does the [PM] agree that one of the best ways to improve the trade balance is to continue the Government's strong policies? A: I agree with my hon. Friend [...]</p>
<p>Issue update (9.3% questions, 12.4% terms) Question terms: <i>what [will you] do, work with</i> Answer terms: <i>are supporting, had recently</i> Q: Will my hon. Friend work with employers to try to incentivise [truck driving] as a career choice for young people? A: Yes, we had a discussion recently about this being an excellent opportunity [...]</p>
<p>Questioning premises (10.3% questions, 8.9% terms) Question terms: <i>is [it] true, does [the Minister] think</i> Answer terms: <i>am disappointed, is impossible</i> Q: Is it not true that the Prime Minister has failed Britain? A: I am disappointed in what the hon. Gentleman says [...]</p>
<p>Request for assurance (13.3% questions, 10.4% terms) Question terms: <i>can [you] give, will [you] assure</i> Answer terms: <i>am concerned, can assure</i> Q: Will the Minister assure me that the [logistics problem] will never happen again? A: I am concerned about the nature of that matter [...]</p>
<p>Prompt for comment (11.5% questions, 9.9% terms) Question terms: <i>will [you] tell, can [you] confirm</i> Answer terms: <i>can tell, am able</i> Q: Can [the PM] confirm that the aspiration to a united Ireland will still be permitted? A: I can tell the House that in a poll, a majority of both communities wanted to try and find an accommodation [...]</p>
<p>Accept and propose (18.1% questions, 15.1% terms) Question terms: <i>will [you] accept, would [it] be</i> Answer terms: <i>be difficult, am certain</i> Q: Will the Minister accept that [this issue] [...] is a responsibility for the Government? A: [We have been] trying to repair the damage done by that privatisation; I am certain that we will have a better rail system soon.</p>

Table 2.1: Representative examples of question and answer terms, and (abbreviated) question-answer pairs, for each question type inferred from the parliamentary question periods data. Bolded terms are present in the original text.

3. Issue update. Requests for information or updates about a current event, issue or policy. Typically the policy refers to a genuinely “national” concern, rather than a partisan issue for which the major parties may have differing views. Answers tend to provide or at least address the requested update.

4. Questioning premises. Asking a minister to respond to a fact or premise, often with the implication that the government has been incompetent and has failed to address this information. Contrast to **demand for account** questions, which seem to demand accounts of existing failings rather than of the ideological premises called into question here.

5. Request for assurance. Asking the minister to provide assurance that they are seeing to a generally uncontentious issue, often serving as an indirect way to voice a concern for the MP or their constituents.

6. Prompt for comment. Requests for comments, especially on information that would not normally be immediately accessible to MPs, such as the contents of meetings or reports.

7. Accept and propose. Asking the minister to comment on a premise and its proposed consequences: if the minister accepts a certain premise, would they speculate on a related idea? Often used to suggest an alternate policy that, per the asker, better addresses the premises voiced in the question.

2.6.2 Validation: Labeled data

We compare our output to a dataset of 1,413 questions asked to various Prime Ministers, from Bates et al. [2014]. Each question in this data is annotated by a domain expert with one of three labels indicating the rhetorical intention of

the asker: compared to *standard* questions, denoting straightforward factual queries, *helpful* questions serve as prompts for the PM to talk favorably about their government, while *unanswerable* questions are effectively vehicles for delivering criticisms that the PM cannot respond to.⁹ If our framework is able to capture meaningful rhetorical information, we expect a given label to be over-represented in some of our inferred types, and underrepresented in others. We include further details about the labeled dataset in the appendix.

Even though our typology of questions is generated in an unsupervised fashion without any guidance from the coded rhetorical roles, we see several clear correspondences between our types and those annotations. In particular, helpful questions are highly associated with the **agreement** type (constituting 28% of questions of that type compared to 15% over the entire dataset, Fisher’s exact test $p < 0.001$, Bonferroni-corrected in the number of types), reinforcing our interpretation that this type captures MPs cheerleading their own government. Conversely, unanswerable questions are frequently of the **demand for account** type (19% in-type vs. 12% overall, Bonferroni-corrected $p < 0.05$). We visualize these correspondences in Figure 2.1, depicting positive pointwise mutual information statistics between types and annotated labels (we use this statistic to adjust for the variation in frequencies of types and labels).¹⁰

We note that our inferred typology also offers complementary information to the hand-coded labels, often making finer distinctions between questions. For instance, we find that along with the **agreement** type, helpful questions are also somewhat associated with the **issue update** type (comprising 24% of questions

⁹Responses to questions are also labeled according to the extent to which they answered the questions; we examine those labels in Chapter 5.2.1).

¹⁰For the remainder of this dissertation, the order in which question types appear in figures corresponds to the relative extent to which a type is asked by a government vs. an opposition MP, as detailed in Section 2.6.3.

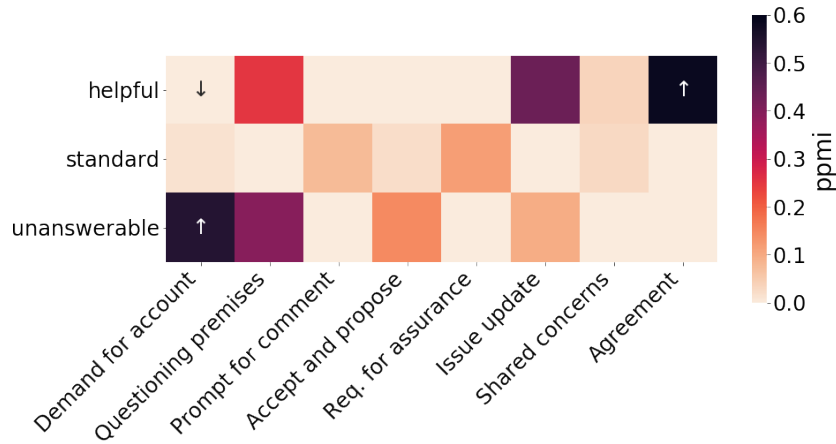


Figure 2.1: Positive pointwise mutual information statistics between question types (columns) and annotated question labels (rows). Darker squares denote types and labels that are more associated with each other. ↑ indicates the label occurs significantly more in-type than overall (Fisher’s exact test $p < 0.05$, Bonferroni-corrected in the number of types), ↓ indicates less.

of that type, $p = 0.07$). Via manual inspection, we find that questions we identified as **agreement** and as **issue update** exhibit differences in how they convey and prompt statements favourable to the government. **Agreement** questions often serve as prompts for exchanging laudatory or partisan remarks:

Q: Does [the PM] agree that one of the saddest legacies the Government inherited is the fact that one child in five grows up in a household in which nobody is in work?

A: [That] is entirely right [...] that is what our welfare reforms, so scandalously neglected by the previous Government, have set out to achieve.

In comparison, **issue update** questions make more specific inquiries on matters like government policies, prompting responses that more substantively discuss a course of action:

Q: The Government’s policy of helping lone parents get into work has assisted thousands of families [...] what further steps will the Prime Minister take to help those parents [...]?

A: Today I met with a group of employers [...] that wish to employ lone parents [...]

2.6.3 Validation: Relation to party affiliation

We also compare the question-asking activity of government and opposition-affiliated MPs—as viewed through the question types—to established characterizations of these affiliations in the political science literature. In particular, prior accounts have considered the bifurcation in behaviour between government and opposition members, in their differing focus on various issues [Louwerse, 2012], and in settings such as roll call votes [Cowley, 2005, Spirling and McLean, 2007, Eggers and Spirling, 2014]. Since government MPs are elected on the same party ticket and manifesto as the government, they primarily act to support the government’s various policies and bolster the status of their cabinet, seldom airing disagreements publicly. In contrast, opposition members tend to offer trenchant partisan criticism of government policies, seeking to destabilize the government’s relationship with its MPs and create negative press in the country at large. In characterizing the question-asking activity of government and opposition MPs, this friendly vs. adversarial behaviour should also be reflected in a rhetorical typology of questions.

We quantify the relative extent to which a particular question type t is asked by government MPs by computing the log-odds ratio of type t questions asked by government MPs, compared to opposition MPs. For the subsequent analyses, we focus on a subset of 80,907 questions for which we had information on asker and answerer affiliations (see the appendix for further details).

Figure 2.2 shows the resultant log-odds ratios of each question type. Notably, we see that **agreement**-type questions are significantly more likely to originate

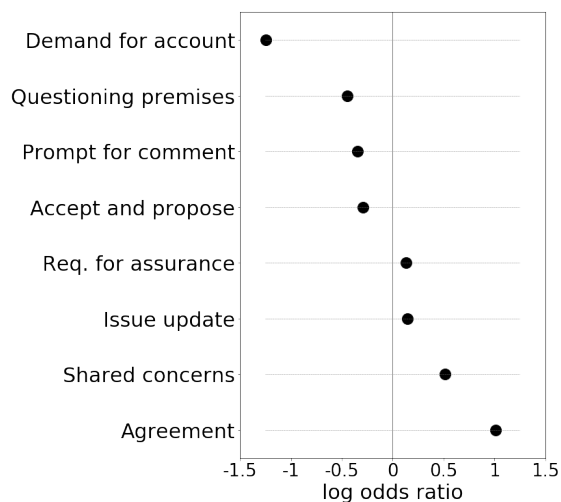


Figure 2.2: Log-odds ratios of questions of each type asked by government compared to opposition MPs.

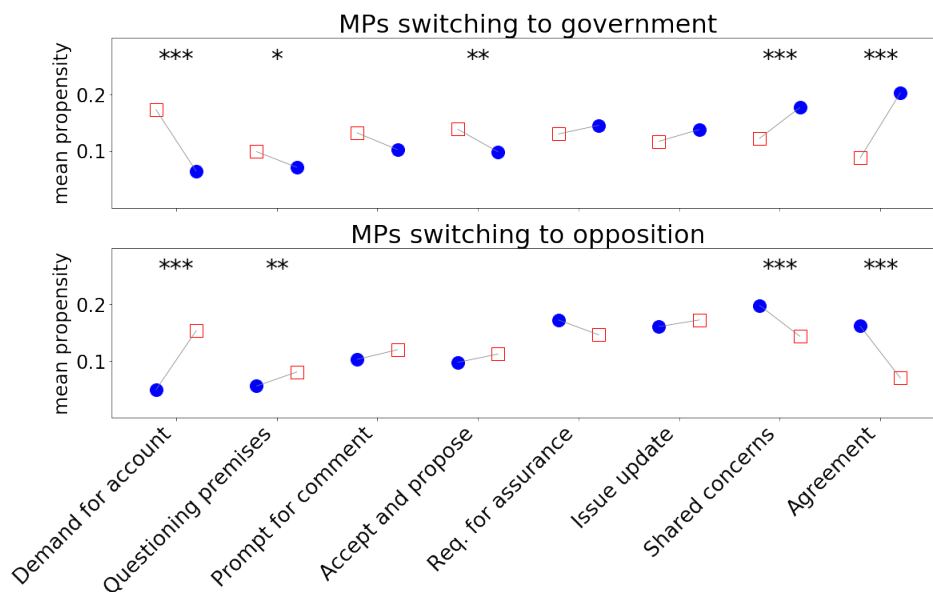


Figure 2.3: Mean propensities for each question type, for MPs who switch from being in the opposition to being in government (top), and vice versa (bottom) after an election; the left and right points in each type denote propensities before and after the switch, while ● and □ denote propensities for government and opposition-affiliated MPs, respectively. Stars indicate statistically significant differences at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) levels (Bonferroni-corrected Wilcoxon test).

from government than from opposition MPs, while the opposite holds for **demand for account** and **questioning premises** questions (Fisher’s exact $p < 10^{-4}$ for each, comparing within-type to overall proportions of questions in each affiliation). Much weaker slants are exhibited in the **request for assurance** and **issue update** types, suggesting that such questions tend to serve as informational queries about relatively non-partisan issues. These results strongly cohere with the “textbook” accounts of parliamentary activity in the literature, as well as our interpretation of these types as bolstering or antagonistic.

Moreover, we find that the *same MP* shifts in their propensity for different question types as their affiliation changes. When a new political party is elected into office, MPs who were previously in the opposition now belong to the government party, and vice versa. Such a switch occurs within our data between the Brown and Cameron governments (Labour to Conservative, 2010). For both switches, we consider all MPs who asked at least 5 questions both before and after the switch, resulting in 107 members who *became* government MPs and 112 who became opposition MPs.

For an MP M we compute $P_{M,t}$, their *propensity* for a question type t , as the proportion of questions they ask which are from t . Comparing $P_{M,t}$ before and after a switch, we replicate the key differences observed above—for instance, we find that former opposition MPs who become government MPs decrease in their propensity for **demand for account** questions, while newly opposition MPs move in the other direction (Wilcoxon $p < 0.001$, Figure 2.3). This suggests that the general trends we observed before are driven by the shift in affiliation, and hence parliamentary role, of *individual* MPs.¹¹

¹¹In particular, this means that MPs change in their propensity to ask different types of questions over their career, underlining that the question types we infer reflect rhetorical categories across the parliamentary institution, and are not unique to particular members.

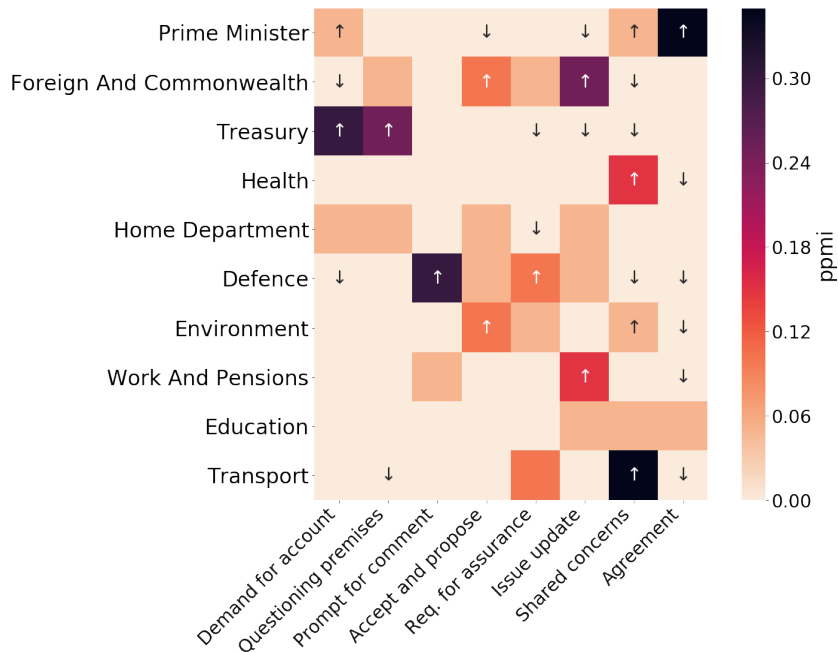


Figure 2.4: Positive pointwise mutual information between question type and the department of the minister answering questions; darker squares indicate that the corresponding type is overrepresented in the department, relative to random chance. Departments are shown in descending order in terms of number of questions asked. \uparrow indicates the type occurs significantly more often in-department vs. overall (Bonferroni-corrected Fisher’s $p < 0.05$), \downarrow indicates less often.

2.6.4 Relation to answerer’s department

We now apply our framework to explore the nature of political discourse in Parliament, in terms of the types of questions asked. Here, we focus on the relation between question type and the government department represented by the minister answering questions, which broadly determines the topic of the questions asked. We restrict the following analyses to the ten departments with the most questions asked, comprising 58% of all questions. These departments are listed in Figure 2.4 (we include “Prime Minister”, which encompasses questions asked directly to the PM).

Since we've designed our method to be topic-agnostic, we expect each question type to be represented across all departments. Indeed, we see that no type is entirely absent from any of the departments: each type occurs in at least 6% of the questions asked in each department. Likewise, no type entirely dominates, with each type comprising at most 22% of the questions per department.

While questions of each type are asked across departments, we also expect the distribution of types to somewhat vary—contrast departments that are responsible for particularly contentious vs. fairly uncontroversial policies. To explore the relative prevalence of question types per department, we compute the pointwise mutual information between question type and answerer department, shown as a heatmap in Figure 2.4.

We find certain type-department pairs where the type is particularly associated with the department, relative to random chance (shown as dark squares in the figure). For instance, relatively aggressive **demands for account** and **questioning premises** questions are overrepresented among questions directed at the Treasury department, perhaps pointing to the contentious nature of economic policy discussions. In contrast, **issue update** questions are relatively prevalent in questions directed at the Foreign and Commonwealth department, which deals with British interests worldwide.¹² We also find that **agreement** questions are overrepresented in questions asked to the PM, perhaps reflecting the relatively partisan and performative nature of these sessions.

¹²Note that Brexit-related question periods, which we expect to contain more acrimonious question-asking, are listed under separate departments in the data.

2.6.5 Relation to career trajectory

Finally, we investigate how questioning behaviour varies with a member's tenure in the institution. Two alternative hypotheses arise: younger MPs may be more vigorously critical compared to older members out of enthusiasm, but are potentially tempered by their stake in future promotion prospects [Cowley, 2005, 2012]. Alternatively, older MPs who have less at stake in terms of prospects of further promotion may ask more antagonistic questions.¹³ In order to examine the extent to which young or old members contribute a specific type of question, for each question type t we compute the median tenure of askers of each question in t , and compare the median tenures of different question types, for each affiliation (Figure 2.5).¹⁴

We see that among government MPs, more aggressive questions (in the **demand for account** and **questioning premises** types) originate more from older members, reflected in significantly higher median tenures (Mann Whitney U test $p < 0.001$ comparing within-type median tenure with outside-type median tenure). Notably, MPs are directing such questions towards their *own* government. This supports the “less to lose” intuition, offering a rhetorical parallel to previous findings about the increased tendency to vote contrary to party lines from MPs with little chance of ministerial promotion [Benedetto and Hix, 2007]. In contrast, less confrontational **issue update** and **agreement** questions tend to come from younger members ($p < 0.001$, both affiliations).

To discount the possibility of these trends being solely driven by a few very prolific young or old MPs, we also make comparisons within-MP, shown in Fig-

¹³Throughout, *young* and *old* refer to tenure—i.e., how many years someone has served as an MP—rather than biological age.

¹⁴Median tenures for opposition members are generally higher; winning an election tends to result in more newly-elected and therefore younger MPs [Webb and Farrell, 1999].

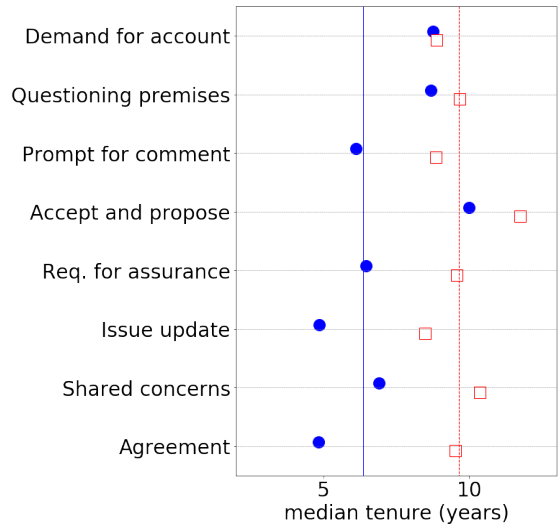


Figure 2.5: Median asker tenures over each question type, for government (●) and opposition (□) askers. Overall median tenures are also shown for reference (solid blue line for government, dashed red line for opposition).

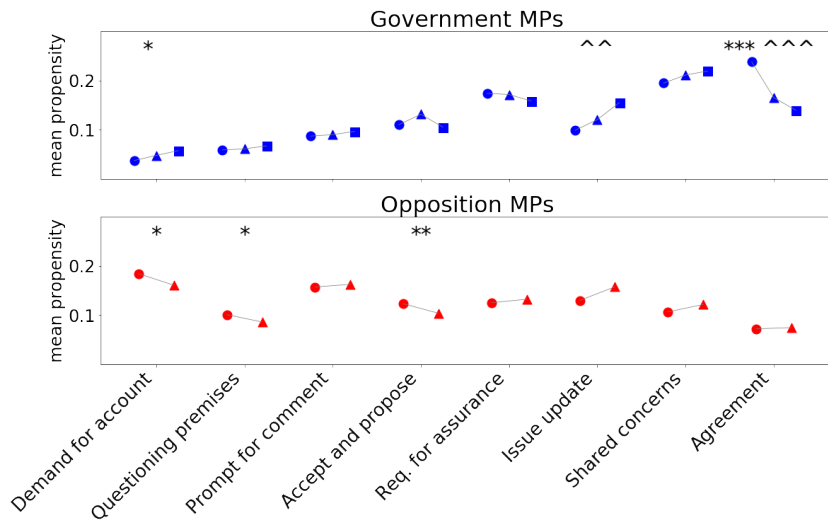


Figure 2.6: Mean propensities for each question type among MPs at various career stages. For government MPs, ●, ■, ▲ denote propensities in their first five years, fifth to tenth year, and after their tenth year, respectively. For opposition MPs, ● and ▲ denote propensities in their first five years, and after their fifth year. * and ^ indicate statistically significant differences at the $p < 0.05$ (*, ^), $p < 0.01$ (**, ^^) and $p < 0.001$ (***, ^^) levels (Wilcoxon test); * compares the first two stages for both affiliations, ^ compares the next two stages for government MPs.

ure 2.6. For MPs affiliated with the government, we consider the subset of 73 MPs who asked at least 5 questions in their first five years in office, and who asked at least 5 questions after their tenth year. We compare the type propensities of those MPs during their first five years (●) and their fifth to tenth year (■), as well as after their tenth year (▲); throughout, we consider only the questions they asked while affiliated with the government. For opposition MPs, we consider the subset of 102 MPs who asked at least 5 questions in their first five years, and at least 5 questions after their fifth year, while affiliated with the opposition, comparing the type propensities in these career stages (● and ▲, respectively).¹⁵

Among government-affiliated askers, we see that the propensity for **agreement** questions steadily decreases as MPs get more tenured: between their first five and next five years, 74% MPs decrease in propensity, and 66% further decrease in propensity after their tenth year (Wilcoxon $p < 0.001$ between each career stage considered). In contrast, the propensity for **demand for account** questions increases among 66% of MPs between their first and next five years ($p < 0.05$), with 57% exhibiting a further increase after their tenth year ($p = 0.09$). This echoes the effects we observed previously, suggesting that the differences in median tenure across these types indeed reflects behavioural changes at the level of individual MPs.

Interestingly, we note that while askers of **issue update** questions have a relatively low median tenure, the propensity to ask such questions actually increases with tenure under our controlled analysis (among 64% of MPs after their tenth year, $p < 0.01$). This may reflect that such questions tend to be asked both

¹⁵The difference in time periods considered between government and opposition MPs reflects the relative frequency of questions asked at different career stages. Due to the low number of MPs considered in this more controlled analysis, we lose some statistical power. To point out notable effects, we *do not* Bonferroni correct the p values computed, and supplement these less rigorous significance tests by reporting effect sizes, in the proportion of MPs who increased or decreased in propensity for a question type between two career stages.

by newer MPs and by experienced members who become invested in a particular cause. Additionally, we see that opposition MPs are more prone to asking aggressive **demand for account** and **questioning premises** questions earlier in their careers (60% and 64% of MPs decrease in propensity to ask each type, respectively, after their fifth year; $p < 0.05$ for both). While further work is needed to fully explain these differences, we speculate that they may reflect strategic attempts by younger opposition MPs to signal traits that could facilitate future promotion, such as partisan loyalty [Kam, 2009].

2.7 Discussion

In this chapter we introduced an unsupervised framework for structuring the space of questions according to their rhetorical role, enabling us to derive a typology of questions from a dataset. We instantiated and validated our approach in a dataset of UK parliamentary question periods, and revealed new interactions between questioning behaviour and career trajectories.

From a technical standpoint, future work could augment the representation of questions and answers presently used in our framework, beyond our heuristic of using root arcs without noun phrases. Richer linguistic representations, as well as more judicious ways of weighting different terms, could enable us to capture a wider range of possible surface and rhetorical forms, especially in settings where language use is less structured by institutional conventions. Additionally, as with most unsupervised methods, users of our approach must use their own discretion to hand-select parameters such as the number of clusters, and manually interpret the typology's output. Having annotations of these corpora could better motivate the methodology and these parameter choices, and

enable further evaluation and interpretation.

We note that our methodology is not tied to a particular domain, and it would also be interesting to explore the method's potential in a variety of other domains where questions likewise play a crucial role. In particular, we've also applied the approach to characterize comments used to start discussions on Wikipedia Talk Pages [Zhang et al., 2018], showing that the latent comment representations it derives can signal whether an initially-civil discussion later derails into overtly hostile behaviour (see Chapter 5.6.1 for further details). To suggest a less structured setting, examining how interviewers in high-profile media settings (e.g., Frost on Nixon) use questions to elicit responses from influential people would aid us in the broader normative goal of holding elites to account, by gaining a better understanding of what and how to ask, and what (not) to accept as an answer.

CHAPTER 3
MODELING THE ORIENTING ROLE OF UTTERANCES IN
COUNSELING CONVERSATIONS

3.1 Overview

Throughout a conversation, participants make choices that can orient the flow of the interaction. Such choices are particularly salient in the consequential domain of crisis counseling, where a difficulty for counselors is balancing between two key objectives: advancing the conversation towards a resolution, and empathetically addressing the crisis situation.

In this chapter, we present an unsupervised methodology to quantify how counselors manage this balance. Our main intuition is that if an utterance can only receive a narrow range of appropriate replies, then its likely aim is to advance the conversation forwards, towards a target within that range. Likewise, an utterance that can only appropriately follow a narrow range of possible utterances is likely aimed backwards at addressing a specific situation within that range. By applying this intuition, we map each utterance to a continuous *orientation* axis that captures the degree to which it is intended to direct the flow of the conversation forwards or backwards.

This unsupervised method allows us to characterize counselor behaviours in a large dataset of crisis counseling conversations, where we show that known counseling strategies intuitively align with this axis. We also illustrate how our measure can be indicative of a conversation's progress and effectiveness.

Note on source material. This section was originally published in Zhang and Danescu-Niculescu-Mizil [2020]. For this dissertation, we modified the method

description to align more closely with our description of the broader framework in Chapter 4. We’ve also added some discussion on how orientation varies over the sentences within a message.

3.2 Introduction

Participants in a conversation constantly shape the flow of the interaction through their choices. In psychological crisis counseling conversations, where counselors support individuals in mental distress, these choices arise in uniquely complex and high-stakes circumstances, and are reflected in rich conversational dynamics [Sacks, 1992]. As such, counseling is a valuable context for computationally modeling conversational behaviour [Atkins et al., 2014, Althoff et al., 2016, Pérez-Rosas et al., 2018, Zhang et al., 2019]. Modeling the conversational choices of counselors in this endeavour is an important step towards better supporting them.

Counselors are driven by several objectives that serve the broader goal of helping the individual in distress. Two key objectives are exemplified in Figure 3.1.¹ The counselor must *advance* a conversation towards a calmer state where the individual is better equipped to cope with their situation [Mishara et al., 2007, Sandoval et al., 2009]: in c_1 , the counselor prompts the individual to brainstorm options for social support. The counselor must also empathetically *address* what was already said, “coming to an empathic understanding” of the individual [Rogers, 1957, Hill and Nakayama, 2000]: in c_2 , the counselor validates feelings that the individual has just shared.

¹These examples are derived from material used to train counselors in our particular setting, detailed in Section 3.3.

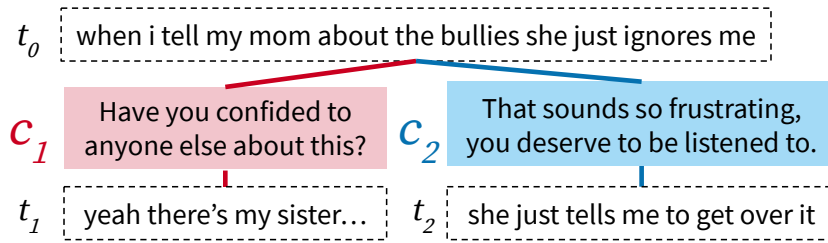


Figure 3.1: Two possible exchanges in a counseling conversation, illustrating key objectives that a counselor must balance: in c_1 , the counselor aims to *advance* the conversation towards a discussion of possible confidants; in c_2 , they aim to *address* the emotion underlying the preceding utterance.

Balancing both objectives is often challenging, and overshooting in one direction can be detrimental to the conversation. A counselor who leans too much on advancing *forwards* could rush the conversation at the expense of establishing an empathetic connection; a counselor who leans too much *backwards*, on addressing what was already said, may fail to make any progress.

In this work, we develop a method to examine counselor behaviours as they relate to this balancing challenge. We quantify the relative extent to which an utterance is aimed at advancing the conversation, versus addressing existing content. We thus map each utterance onto a continuous backwards-forwards axis that models the balance of these objectives, and refer to an utterance's position on this axis as its *orientation*.

At an intuitive level, our approach considers the *range* of content that is expected to follow or precede a particular utterance. For an utterance like c_1 that aims to advance the conversation towards an intended target, we would expect a narrow range of appropriate replies, concentrated around that target (e.g., suggestions of possible confidants). We would likewise expect an utterance like c_2 that aims to address a previously-discussed situation to only be an appropriate

reply for a narrow range of possible utterances, concentrated around that specific type of situation (e.g., disclosures of frustrating scenarios). Starting from this intuition, we develop an unsupervised method to quantify and compare these expected forwards and backwards ranges for any utterance, yielding our orientation measure.

Using this measure, we characterize counselor behaviours in a large collection of text-message conversations from a crisis counseling service, which we accessed in collaboration with the service and with the participants' consent. We show how orientation distinguishes between key conversational strategies that counselors are taught during their training. We also show that our measure tracks a conversation's progress and can signal its effectiveness, highlighting the importance of balancing the advancing and addressing objectives, and laying the basis for future inquiries in establishing potential causal effects.

In summary, we develop an unsupervised methodology that captures how counselors balance the conversational objectives of advancing and addressing (Section 3.5), apply and validate it in a large dataset of counseling conversations (Section 3.6), and use it to investigate the relation between a counselor's conversational behaviour and their effectiveness (Section 3.6.4). While our method is motivated by a salient challenge in counseling, we expect similar balancing problems to recur in other settings where conversationalists must carefully direct the flow of the interaction, such as court trials and debates (Section 3.7).

3.3 Setting: Counseling conversations

We develop our method in the context of Crisis Text Line, a crisis counseling platform that provides a free 24/7 service for anyone in mental crisis—

henceforth *texters*—to have one-on-one conversations via text message with affiliated counselors. We accessed a dataset of over 1.5 million conversations, in collaboration with the platform and with the consent of the participants. The data was scrubbed of personally identifiable information by the platform. The extensive ethical and privacy considerations, and policies accordingly implemented by the platform, are detailed in Pisani et al. [2019].²

In each conversation, a crisis counselor’s high-level goal is to guide the *texter* towards a calmer mental state. These conversations are quite substantive, averaging 25 messages with 29 and 24 words per counselor and *texter* message, respectively. All counselors first complete 30 hours of training provided by the platform, which draws on existing literature in counseling to recommend best practices and conversational strategies. The author of this dissertation also completed the training to gain familiarity with the domain.

While the platform offers guidance to counselors, their task is inevitably open-ended, given the emotional complexity of crisis situations, and the particular concerns of each *texter*. As such, the counselors are motivated by an explicit goal that structures the interaction, but they face a challenging flexibility in choosing how to act.

3.4 Background and related work

We now elaborate on the conversational challenge of balancing between advancing the conversation forwards or addressing what was previously said. Our description of the challenge and our computational approach to studying

²The data was accessed via a fellowship program. The service’s present data access policy is detailed at <https://www.crisistextline.org/data-philosophy/>.

it are informed by literature in counseling, the platform's training material and informal interviews with its staff.

A conversational balance. A crisis counselor must fulfill multiple objectives in their broader goal of helping a texter. One objective is guiding the texter through their initial distress to a calmer mental state [Mishara et al., 2007, Sandoval et al., 2009], as in Figure 3.1, message c_1 . Various strategies that aim to facilitate this *advancing* process are taught to counselors during training: for instance, a counselor may prompt a texter to identify a goal or coping mechanism [Rollnick and Miller, 1995]. As such, they attempt to move the conversation *forwards*, towards its eventual resolution.

The counselor must also engage with the texter's concerns [Rogers, 1957, Hill and Nakayama, 2000], as in message c_2 , via strategies that empathetically *address* what the texter has already shared [Rollnick and Miller, 1995, Weger et al., 2010, Bodie et al., 2015]. For instance, counselors are taught to *reflect*, i.e., reframe a texter's previous message to convey understanding, or draw on what was said to affirm the texter's positive qualities. In doing so, the counselor looks *backwards* in the conversation.

Past work has pointed to the importance of fulfilling both objectives [Mishara et al., 2007]. However, as the training acknowledges, striking this balance is challenging. Overzealously seeking to advance could cut short the process of establishing an empathetic connection. Conversely, focusing on the conversation's past may not help with eventual problem solving [Bodie et al., 2015], and risks stalling it. A texter may start to counterproductively *ruminate* on their concerns [Nolen-Hoeksema et al., 2008, Jones et al., 2009]; indeed, prior accounts in psychology have highlighted the thin line between productive reflection and rumination [Rose et al., 2007, Landphair and Preddy, 2012].

Orientation. To examine this balancing dynamic, we model the choices that counselors make at each turn in a conversation. Our approach is to derive a continuous axis spanned by advancing and addressing. We refer to an utterance’s position on this axis, representing the relative extent to which it aims at either objective, as its *orientation* Ω . We interpret a *forwards-oriented* utterance with positive Ω as aiming to advance the conversation, and a *backwards-oriented* utterance with negative Ω as aiming to address what was previously brought up. In the middle, the axis reflects the graded way in which a counselor can balance between aims—for instance, using something the texter has previously said to help motivate a problem-solving strategy.

Related characterizations. We view the orientation measure as a complement to other characterizations of conversational behaviours in varied settings.

Prior work has also considered how utterances relate to the surrounding discourse [Webber, 2001]. Frameworks like centering theory [Grosz et al., 1995] aim at identifying referenced entities, while we aim to more abstractly model interlocutor choices. Past work has examined how interlocutors mediate a conversation’s trajectory through taking or ceding control [Walker and Whittaker, 1990], shifting topic [Nguyen et al., 2014], or taking up what another interlocutor has said [Demszky et al., 2021]; Althoff et al. [2016] considers the rate at which counselors in our setting advance across stages of a conversation. While these actions can be construed as forwards- or backwards-oriented, we focus more on the interplay between forwards- and backwards-oriented actions. A counselor’s objectives may also cut across these concepts: for instance, the training stresses the need for empathetic reflecting across all stages and topics.

Orientation also complements prior work on dialogue acts, which consider various roles that utterances play in discourse [Mann and Thompson, 1988, Core

and Allen, 1997, Ritter et al., 2010, Bracewell et al., 2012, Rosenthal and McKeown, 2015, Prabhakaran et al., 2018, Wang et al., 2019]. In counseling settings, such approaches have highlighted strategies like reflection and question-asking [Houck, 2008, Gaume et al., 2010, Atkins et al., 2014, Can et al., 2015, Tanana et al., 2016, Pérez-Rosas et al., 2017, 2018, Park et al., 2019, Lee et al., 2019, Cao et al., 2019, Sharma et al., 2020]. Instead of modeling a particular strategy or taxonomy of strategies, we model how counselors balance among the underlying objectives; we later relate orientation to these strategies (Section 3.6). Additionally, most of these approaches use annotations or pre-defined labeling schemes, while our method is unsupervised.

3.5 Measuring orientation

We now describe our method to measure orientation, discussing our approach at a high level before elaborating on our particular operationalization.

3.5.1 High-level sketch

The orientation measure compares the extent to which an utterance aims to advance the conversation forwards with the extent to which it aims backwards. Thus, we must somehow quantify how the utterance relates to the subsequent and preceding interaction.

Naive attempt: direct comparison. As a natural starting point, we may consider a similarity-based approach: an utterance that aims to address its preceding utterance, or *predecessor*, should be similar to it; likewise, an utterance that aims to advance the conversation should be similar to the reply that it prompts. In prac-

tice, having to make these direct comparisons is limiting: an automated system could not characterize an utterance in an ongoing conversation by comparing it to a reply it has yet to receive.

This approach also has important conceptual faults. First, addressing preceding content in a conversation is different from recapitulating it. For instance, counselors are instructed to *reframe* rather than outright restate a texter's message, as in Figure 3.1, c_2 . Likewise, counselors need not advance the conversation by declaring something for the texter to simply repeat; rather than giving specific recommendations, counselors are instructed to prompt the texters to come up with coping strategies on their own, as in c_1 . Further, texters are not bound to the relatively formal linguistic style counselors must maintain, resulting in noticeable lexical differences. Measuring orientation is hence a distinct task from measuring similarity.

Second, a speaker's *intent* to advance need not actually be realized with their utterance. A counselor's cues may be rebuffed or misunderstood [Thomas, 1983, Schegloff, 1987]: a texter could respond to c_1 by continuing to articulate their problem with t_2 . Likewise, a counselor may intend to address a texter's concerns but misinterpret them. To model the balance in objectives that a counselor is aiming for, our characterization of an utterance cannot be contingent on its actual reply and predecessor.

Our approach: characterizing expectations. We instead consider the range of replies we *expect* an utterance to receive, or the range of predecessors that we expect it follow. Intuitively, an utterance with a narrow range of appropriate replies aims to direct the conversation towards a particular target, moreso than an utterance whose appropriate replies span a broader range. For instance, consider leading versus open-ended questions: when people ask leading questions,

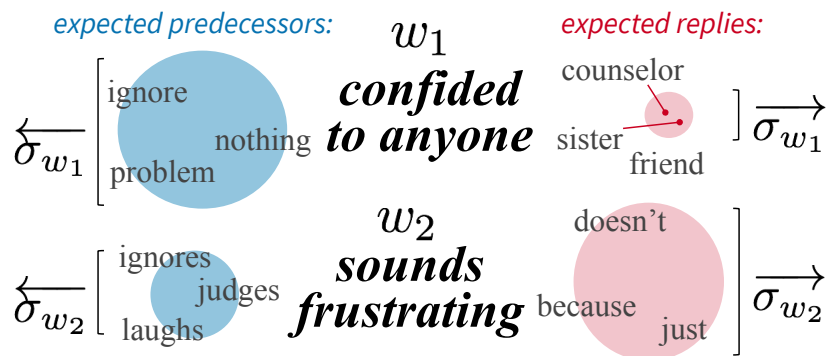


Figure 3.2: Words representative of replies and predecessors for utterances with two example terms, as observed in training data. Top row: observed replies to utterances with w_1 span a narrower range than observed predecessors (relative sizes of red and blue circles); w_1 thus has smaller *forwards-range* $\overrightarrow{\sigma}_{w_1}$ than *backwards-range* $\overleftarrow{\sigma}_{w_1}$ (i.e., it is forwards-oriented, $\Omega_{w_1} > 0$). Bottom row: observed predecessors to utterances with w_2 span a narrower range than replies; w_2 thus has smaller $\overleftarrow{\sigma}_{w_2}$ than $\overrightarrow{\sigma}_{w_2}$ (i.e., it is backwards-oriented $\Omega_{w_2} < 0$).

they intend to direct the interaction towards specific answers they have in mind; when people ask open-ended questions, they are more open with respect to what answers they receive and where the interaction is headed. Similarly, an utterance that is an appropriate reply to only a narrow range of possible predecessors likely aims to address a particular situation. Operationally, we draw on unlabeled data of past conversations to form our expectations of these ranges, and build up our characterizations of utterances from their constituent terms, e.g., words or dependency-parse arcs.

The intuition for our approach is sketched in Figure 3.2. From our data, we observe that utterances containing *confided to anyone* generally elicited replies about potential confidants (e.g., *sister*, *friend*), while the replies that followed utterances with *sounds frustrating* span a broader, less well-defined range. As such, we have a stronger expectation of what a reply prompted by a *new* utter-

ance with *confided to anyone* might contain than a reply to a new utterance with *sounds frustrating*.

More generally, for each term w , we quantify the strength of our expectations of its potential replies by measuring the range spanned by the replies it has already received in the data, which we refer to as its *forwards-range* $\vec{\sigma}_w$. We would say that *confided to anyone* has a smaller $\vec{\sigma}_w$ than *sounds frustrating*, meaning that its observed replies were more narrowly concentrated; this is represented as the relative size of the red regions on the right side of Figure 3.2.

In the other direction, we observe in our data that *sounds frustrating* generally follows descriptions of frustrating situations (e.g., *ignores*, *judges*); the range of predecessors to *confided to anyone* is broader. We thus have a stronger expectation of the types of situations that new utterances with *sounds frustrating* would respond to, compared to utterances with *confided to anyone*. For a term w , we quantify the strength of our expectations of its potential predecessors by measuring its *backwards-range* $\overleftarrow{\sigma}_w$, spanned by the predecessors we've observed. As such, *sounds frustrating* has a smaller $\overleftarrow{\sigma}_w$ than *confided to anyone*, corresponding to the relative size of the blue regions on the left side of Figure 3.2.

The relative strengths of our expectations in either direction then indicate the balance of objectives reflected by the utterance. If we have a stronger expectation of w 's replies than of its predecessors—i.e., smaller $\vec{\sigma}_w$ than $\overleftarrow{\sigma}_w$ —we would infer that utterances with w aim to advance the conversation towards a particular reply more than they aim to address a particular situation. Conversely, if we have stronger expectations of w 's predecessors—i.e., smaller $\overleftarrow{\sigma}_w$ —we would infer that utterances with w aim to address the preceding interaction, rather than trying to drive the conversation towards a target.

We thus measure orientation by comparing a term’s forwards- and backwards-range. The expectation-based approach allows us to circumvent the shortcomings of a direct comparison-based approach; we may interpret our method as modeling a counselor’s *intent* in advancing and addressing at each utterance [Moore and Paris, 1993, Zhang et al., 2017b].

3.5.2 Operationalization

We now detail the steps of our method, which are outlined in Figure 3.3. Formally, our input consists of a set of utterances from counselors $\{c_i\}$, and a set of utterances from texters $\{t_i\}$, which we’ve observed in a dataset of conversations (Figure 3.3A). We note that each texter utterance can be a reply to, or a predecessor of, a counselor utterance (or both). We use this unlabeled “training data” to measure the forwards-range $\vec{\sigma}_w$ (Figures 3.3B-D), the backwards-range $\overleftarrow{\sigma}_w$, and hence the orientation Ω_w of each term w used by counselors (Figure 3.3E). We then aggregate to an utterance-level measure.

For each counselor term w , let \vec{T}_w denote the subset of texter utterances that are replies to counselor utterances containing w (Figure 3.3A). As described above, the forwards-range $\vec{\sigma}_w$ quantifies the spread among elements of \vec{T}_w , which we measure by comparing vector representations of these utterances \vec{U}_w (Figure 3.3B, detailed below) to a central reference point \vec{u}_w (Figures 3.3C and 3.3D).³ Likewise, $\overleftarrow{\sigma}_w$ quantifies the similarity among elements of \overleftarrow{T}_w , the set of predecessors to counselor utterances with w ; we compute $\overleftarrow{\sigma}_w$ by comparing each corresponding vector in \overleftarrow{U}_w to a central point \overleftarrow{u}_w .

³Using a central reference point to calculate the forwards-range, as opposed to directly computing pairwise similarities among replies in \vec{U}_w , allows us to account for the context of w in the utterances that prompted these replies (via tf-idf weighting, as subsequently discussed).

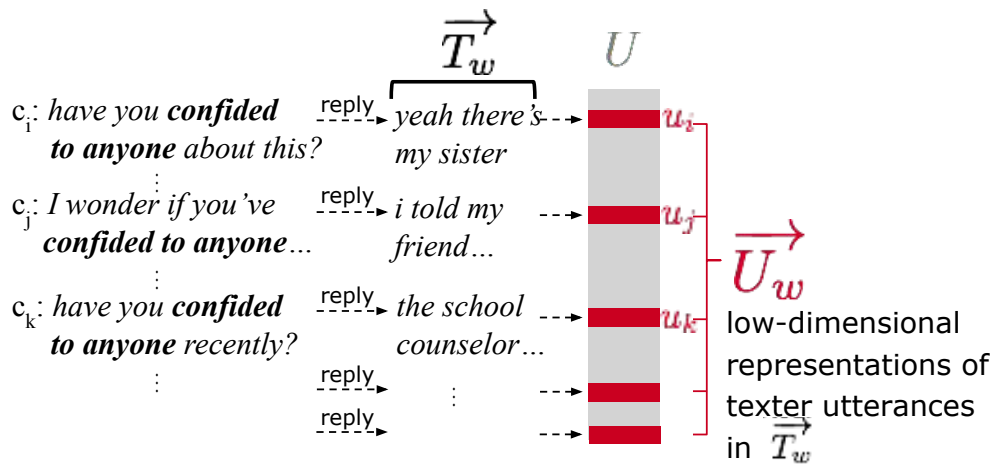
**A. Input: observed
texter replies to
counselor utterances**

example w :
confided to anyone

**B. Derive vector
representations of
texter utterances**

$$\mathcal{X} \approx^{SVD} U S V^T$$

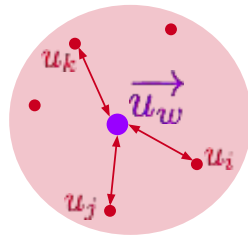
where \mathcal{X} is a
tf-idf reweighted
term-document matrix
of all texter utterances



C. Derive central points

$$\vec{u}_w = W_w^T U_w S^{-1}$$

where w_w^i is the
tf-idf weight
of w in c_i



**D. Compute
forwards-range**

$$\vec{\sigma}_w = \text{avg}(\bullet \leftrightarrow \bullet)$$

$\forall u \in U_w$

where $\bullet \leftrightarrow \bullet$
is the cosine distance
between \vec{u}_w and u

E. Compute orientation: $\Omega_w = \overleftarrow{\sigma}_w - \vec{\sigma}_w$

Figure 3.3: Outline of steps to compute the orientation Ω_w of term w , as described in Section 3.5.2. Panels A-D show the procedure for computing forwards-range $\vec{\sigma}_w$; the procedure for backwards-range $\overleftarrow{\sigma}_w$ is analogous.

Deriving vector representations (Figure 3.3B). To obtain vectors for each texter utterance, we construct \mathcal{X} , a tf-idf reweighted term-document matrix where rows represent texter utterances and columns represent terms used by texters. To ensure that we go beyond lexical matches and capture conceptual similarities (e.g., among possible confidants or frustrating situations), we use singular value decomposition to get $\mathcal{X} \approx UsV^T$. Each row of U is a vector representation u_i of utterance t_i in the induced low-dimensional space \mathbb{T} . \vec{U}_w then consists of the corresponding subset of rows of U (highlighted in Figure 3.3B).

Deriving central points (Figure 3.3C). For each w , we take its corresponding central point \vec{u}_w to be a weighted average of vectors in \vec{U}_w . Intuitively, a texter utterance t_i with vector u_i should have a larger contribution to \vec{u}_w if w is more prominent in the counselor utterance c_i that preceded it. We let w_w^i denote the normalized tf-idf weight of w in c_i , and use w_w^i as the weight of the corresponding vector u_i . To properly map the resultant weighted average $\sum w_w^i u_i$ into \mathbb{T} , we divide each dimension by the corresponding singular value in s . As such, if w_w is a vector of weights w_w^i , we can calculate the central point \vec{u}_w of \vec{U}_w as $\vec{u}_w = w_w^T \vec{U}_w s^{-1}$. In the other direction, we likewise compute $\overleftarrow{u}_w = w_w^T \overleftarrow{U}_w s^{-1}$.

Forwards- and backwards-ranges (Figure 3.3D). We take the forwards-range $\vec{\sigma}_w$ of w to be the average cosine distance from each vector in \vec{U}_w to the central point \vec{u}_w , and $\overleftarrow{\sigma}_w$ to be the average distance from each vector in \overleftarrow{U}_w to \overleftarrow{u}_w .

Term-level orientation (Figure 3.3E). Importantly, since we've computed the forwards- and backwards-ranges $\vec{\sigma}_w$ and $\overleftarrow{\sigma}_w$ using distances in the same space \mathbb{T} , their values are comparable. We then compute the orientation of w as their difference: $\Omega_w = \overleftarrow{\sigma}_w - \vec{\sigma}_w$.

Utterance-level orientation. To compute the orientation of an utterance c_i , we

first compute the orientation of each sentence in c_i as the tf-idf weighted average Ω_w of its constituent terms.⁴ In manually inspecting the data, we observed that an utterance with multiple sentences can orient in *both* directions—e.g., a counselor could concatenate c_2 and c_1 from Figure 3.1 in a single utterance, addressing the texter’s previous utterance before moving ahead. To model this heterogeneity, we consider both the minimum and maximum sentence-orientations in an utterance: Ω^{\min} captures the extent to which the utterance looks backwards, while Ω^{\max} captures the extent to which it aims to advance forwards (see Section 3.6.5 for further discussion).

3.6 Application to counseling data

We apply our method to characterize messages from crisis counselors on the platform. We construct our training set from conversations involving a random sample of 20% of *counselors* in the data (whose conversations are omitted in subsequent analyses), resulting in a collection of 351,862 counselor messages and adjacent texter turns. We use dependency-parse arcs as counselor terms and unigrams as texter terms, reflecting the comparatively structured language of the counselors vs. the texters (counselors are instructed to write grammatically well-formed sentences); we used 25 SVD dimensions to induce \mathbb{T} . Further details about data and parameter choices are in the appendix (Section A.2).

Table 3.1 shows representative terms and sentences of different orientations.⁵

Around two-thirds of terms and sentences have $\Omega < 0$, echoing the importance

⁴Equivalently, we can take tf-idf weighted averages of $\vec{\sigma}_w$ and $\overleftarrow{\sigma}_w$, and then subtract the sentence-level ranges.

⁵Example sentences are derived from real sentences in the data, and modified for readability. The examples were chosen to reflect common situations in the data, and were vetted by the platform to ensure the privacy of counselors and texters.

Orientation	Example terms	Example sentences
Backwards-oriented (bottom 25%)	sounds frustrating, totally normal, great ways, on [your] plate, be overwhelming, sometimes feel frightening, on top [of] been struggling, feeling alone	You have a lot of things on your plate, between family and financial problems. [reflection] It's totally normal to feel lonely when you have no one to talk to. [reflection] Those are great ways to improve the relationship. [affirmation]
(middle 25%)	happened [to] make, mean [when you] say, is that, you recognized, source of the moment, are brave	Has anything happened to make you anxious? [exploration] It's good you recognized the need to reach out. [affirmation] Can you tell me what you mean when you say you're giving up? [risk assessment]
Forwards-oriented (top 25%)	plan for, confided [to] anyone, usually do, has helped, been talking, best support have considered, any activities	Can you think of anything that has helped when you've been stressed before? [problem solving] I want to be the best support for you today. [problem solving] We've been talking for a while now, how do you feel? [closing]

Table 3.1: Example terms and sentences with labeled strategies from crisis counselors' messages, at varying orientations: backwards-oriented (from the bottom 25% of Ω), middle, and forwards-oriented (from top 25%).

of addressing the texter's previous remarks.

In what follows, we examine counselor behaviours in terms of orientation, and illustrate how the measure can be used to analyze conversations. We start by validating our method via two complementary approaches. In a subset of sentences manually annotated with the counseling strategies they exhibit, we show that orientation meaningfully reflects these strategies (Section 3.6.1). At a larger scale, we show that the orientation of messages over the course of a conversation aligns with domain knowledge about counseling conversation structure (Section 3.6.2). We also find that other measures for characterizing messages are not as rich as orientation in capturing counseling strategies and conversation structure (Section 3.6.3). We then show that a counselor's orientation

reflection (113)
re-wording to show understanding and validate feelings <i>It can be overwhelming to go through that on your own.</i>
affirmation (60)
pointing out the texter’s positive qualities and actions <i>You showed a lot of strength in reaching out to us.</i>
exploration (44)
prompting texters to expand on their situation <i>Is this the first real fight you’ve had with your boyfriend?</i>
problem solving (110)
identifying the texter’s goals and potential coping skills <i>What do you usually do to help you feel calmer?</i>
closing (43)
reviewing the conversation and transitioning to a close <i>I think you have a good plan to get some rest tonight.</i>
risk assessment (19)
assessing suicidal ideation or risk of self-harm <i>Do you have access to the pills right now?</i>

Table 3.2: Counseling strategies and representative examples derived from the training material. The number of sentences (out of 400) assigned to each label is shown in parentheses (11 were not labeled as any action).

in a conversation is tied to indicators of their effectiveness in helping the texter (Section 3.6.4), before providing some nuance on whether orientation reflects an either-or decision to advance or address (Section 3.6.5).

3.6.1 Validation: Counseling strategies

Even though it is computed without the guidance of any annotations, we expect orientation to meaningfully reflect strategies for advancing or addressing that crisis counselors are taught. The author hand-labeled 400 randomly-selected sentences with a set of pre-defined strategies derived from techniques highlighted in the training material. Table 3.2 provides descriptions of these strategies; we also note examples in Table 3.1 that exemplify each strategy.

Figure 3.4A shows the distributions of orientations across each label. We

find that the relative orientations of different strategies corroborate their intent as described in the literature. Statements **reflecting** or **affirming** what the texter has said to check understanding or convey empathy (characterized by terms like *totally normal*) tend to be backwards-oriented; statements prompting the texter to advance towards **problem-solving** (e.g., *[what] has helped*) are more forwards-oriented. **Exploratory** queries for more information on what the texter has already said (e.g., *happened to make*) tend to have middling orientation (around 0). The standard deviation of orientations over messages within most of the labels is significantly lower than across labels (bootstrapped $p < .05$, solid circles), showing that orientation yields interpretable groupings of messages in terms of important counseling strategies.

The measure also offers complementary information. For instance, we find sentences that aren't accounted for by pre-defined labels, but still map to interpretable orientations, such as backwards-oriented examples assuaging texter concerns about the platform being a safe space to self-disclose.

3.6.2 Validation: Conversation structure

We also show that orientation tracks with the structure of crisis counseling conversations as described in the training material. Following Althoff et al. [2016], we divide conversations with at least ten counselor messages into five equal-sized segments and average Ω^{\max} and Ω^{\min} over messages in each segment. Figure 3.4B (black lines) shows that over the course of a conversation, messages tend to get more forwards-oriented (higher Ω^{\max} and Ω^{\min}). This matches a standard conversation structure taught in the training: addressing the texter's existing problems before advancing towards problem-solving. While this correspon-

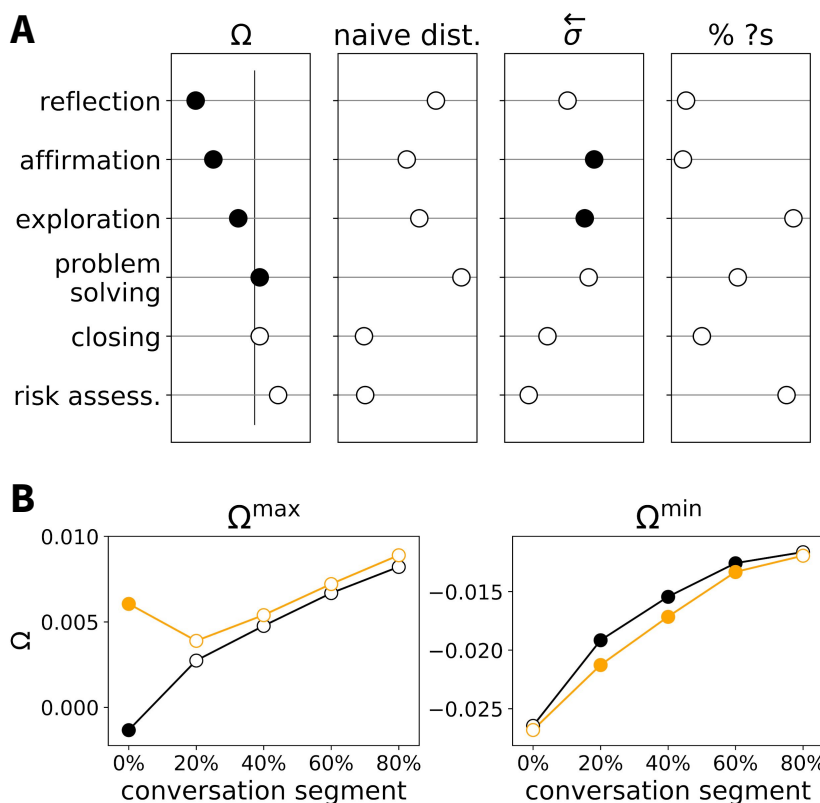


Figure 3.4: Validating the orientation measure and comparing to alternatives. **A**: Leftmost: Mean Ω per counseling strategy label (vertical line denotes $\Omega = 0$). Next three: same for other measures. **B**: Mean Ω^{\max} and Ω^{\min} per segment for risk-assessed (orange) and non-risk-assessed (black) conversations. Both: Solid circles indicate statistically significant differences (Wilcoxon $p < 0.01$, comparing within-counselor).

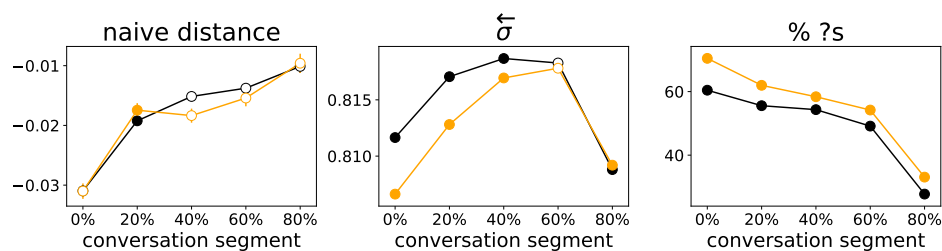


Figure 3.5: Mean naive distance, backwards-range ($\bar{\sigma}$), and % of messages with questions, per segment for risk-assessed (orange) and non-risk-assessed (black) conversations; solid circles indicate statistically significant differences (Wilcoxon $p < 0.01$, comparing conversation types within counselor).

dence holds in aggregate, orientation also captures complementary information to advancement through stages—e.g., while problem-solving, counselors may still address and affirm a texter’s ideas (Table 3.1, row 3).

We also consider a subset of conversations where we expect a different trajectory: for potentially suicidal texters, the training directs counselors to immediately start a process of **risk assessment** in which actively prompting the texter to disclose their level of suicidal ideation takes precedence over other objectives. As such, we expect more forwards-oriented messages at the starts of conversations involving such texters. Indeed, in the 30% of conversations which are risk-assessed, we find significantly larger Ω^{\max} in the first segment (Figure 3.4B, orange line; Wilcoxon $p < 0.01$ in the first stage, comparing within-counselor). Interestingly, Ω^{\min} is *smaller* in each subsequent stage, suggesting that counselors balance actively prompting these critical disclosures with addressing them.

3.6.3 Alternative measures

We compare orientation to other utterance-level measures:

Naive distance. We consider the naive direct-comparison approach mentioned in Section 3.5, taking a difference in cosine distances between tf-idf representations of a message and its reply, and a message and its predecessor.

Backwards-range. We consider just the message’s backwards-range. For each sentence we take tf-idf weighted averages of component $\vec{\sigma}_w$ and take the minimum value for each message.⁶

Question-asking. We consider whether the message has a question. This was used in Walker and Whittaker [1990] as a signal of taking control, which could

⁶We get qualitatively similar results with maximum $\vec{\sigma}$.

be construed as forwards-oriented; indeed, 61% of sentences with '?' have $\Omega > 0$, compared to 21% of sentences without.

Within-label standard deviations of each alternative measure are generally not significantly smaller than across-label (Figure 3.4A), indicating that these measures are poorer reflections of the counseling strategies. Label rankings under the measures are arguably less intuitive. For instance, reflection statements have relatively large (naive) cosine distance from their predecessors; indeed, the training encourages counselors to *process* rather than simply restate the texter's words. We also see that explicitly-marked questions are inexact proxies of forwards-oriented sentences—as in Table 3.1, questions can address a past remark by prompting clarifications, while counselors can also use non-questions to suggest an intent to advance stages (e.g., to transition to problem-solving).

These measures also track with the conversation's progress differently (Figure 3.5, plotting averages per conversation segment for each alternate measure). Notably, none of them clearly distinguish the initial dynamics of risk-assessed conversations as reflected in Ω^{\max} : for instance, simple counts of questions do not distinguish between questions geared towards risk-assessment versus more open-ended problem exploration.

3.6.4 Relation to conversation effectiveness

Past work on counseling has extensively discussed the virtues of addressing a client's situation [Rogers, 1957, Hill and Nakayama, 2000]. Some studies also suggest that accounting for *both* addressing and advancing is important [Mishara et al., 2007]—as such, we'd expect effective counselors to mix backwards- and forwards-oriented actions.

We use orientation to examine how these strategies are tied to conversational effectiveness in crisis counseling at a larger scale, using our framework to provide a unified view of advancing and addressing. To derive simple conversation-level measures, we average Ω^{\max} and Ω^{\min} over each counselor message in a conversation, acknowledging that future work could consider much more sophisticated ways to capture the dynamics across a conversation. We perform all subsequent analyses on a subset of 234,433 conversations, as detailed in the appendix.

Adjudicating counseling conversation quality is known to be difficult [Tracey et al., 2014]. As a starting point, we relate our conversation-level measures to two complementary indicators of a conversation’s effectiveness:

- **Perceived helpfulness.** We consider responses from a post-conversation survey asking the texter whether the conversation was helpful, following Althoff et al. [2016]. Out of the 26% of conversations with a response, 89% were rated as helpful.⁷
- **Conversation length.** We consider a conversation’s length as a simple indicator of the pace of its progress: short conversations may rush the texter, while prolonged conversations could suggest stalling and could even demoralize the counselor [Landphair and Preddy, 2012].⁸

Figure 3.6A compares Ω^{\min} and Ω^{\max} in conversations rated as helpful and unhelpful by texters. Both measures are significantly *smaller* in conversations

⁷We note that this indicator is limited by important factors such as the selection bias in respondents; see Chapter 6 for further discussion.

⁸As the training material notes, conversation length and texter perception may signal complementary or even conflicting information about a texter’s experience of a conversation and its effectiveness: “Some texters resist the end of the conversation. They ruminate [...] causing the conversation to drag on without any progress.”

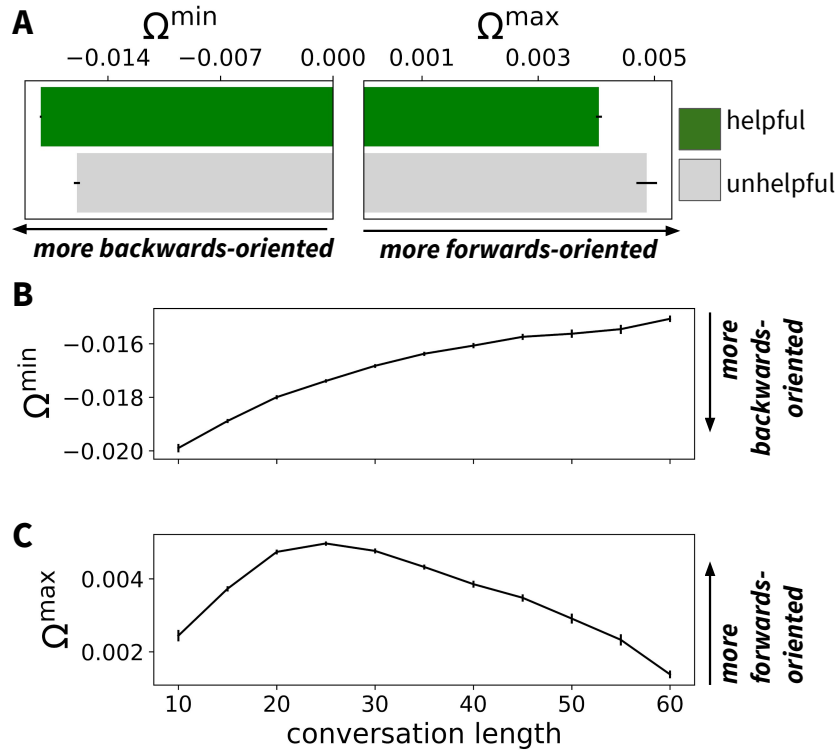


Figure 3.6: Relation between orientation and conversational effectiveness. **A:** Mean Ω^{\min} and Ω^{\max} in conversations rated as helpful (green) or unhelpful (grey) (macroaveraged per conversation). Differences in both measures are significant (Mann Whitney U test $p < 0.001$). **B, C:** Mean Ω^{\min} and Ω^{\max} of conversations with varying lengths (in # of messages). Both plots: Error bars show 95% bootstrapped confidence intervals.

perceived as helpful, suggesting that texters have a better impression of relatively *backwards-oriented* interactions where the counselor is inclined towards addressing their situation. As such, this result echoes past findings relating addressing to effectiveness.

Figure 3.6B compares Ω^{\min} in conversations of varying lengths, showing that Ω^{\min} increases with length, such that counselors exhibit less propensity for addressing in longer conversations. Anecdotal observations cited in interviews with the platform’s staff suggest one interpretation: conversations in which a

texter feels their concerns were not satisfactorily addressed may be prolonged when they circle back to revisit these concerns.

Figure 3.6C relates Ω^{\max} to conversation length. We find that Ω^{\max} is smaller in the lengthiest conversations, suggesting that such prolonged interactions may be stalled by a weaker impulse to advance forwards. Extremely short conversations have smaller Ω^{\max} as well, such that premature endings may also reflect issues in advancing. These findings echo the previously-positing benefits of mixing addressing and advancing: backwards-oriented actions may be helpful for the texter, but forwards-oriented actions are also tied to making progress.

Counselor-level analysis. These findings could reflect various confounds: for instance, texters with particularly difficult situations might affect a counselor's behaviour, but may also be more likely to give bad ratings, independent of how the counselor behaves. Alternatively, an overly long conversation could arise because the counselor is less forwards-oriented, or because the texter is reluctant to make progress from the outset, making it hard for the counselor to attempt to prompt them forwards.

To separate a counselor's decisions from these situational factors, we take a counselor-level perspective, drawing on an approach presented in Zhang et al. [2020] to address such confounding factors (see also Chapter 6.3). Our key intuition here is that counselors can exhibit cross-conversational inclinations for particular behaviours. We therefore relate these cross-conversational *tendencies* in orienting a conversation to a counselor's long-term propensity for receiving helpful ratings, or having long vs. short conversations.

We characterize a counselor's orienting tendency as the average Ω^{\max} and Ω^{\min} over their conversations; we likewise take the proportion of their (rated)

conversations which were perceived as helpful, or the average length of their conversations. We restrict our counselor level analyses to the 20th to 120th conversations of the 1,495 counselors with at least 120 conversations (ignoring their initial conversations when they are still acclimatizing to the platform).

To cleanly disentangle counselor tendency and conversational circumstance, we *split* each counselor’s conversations into two interleaved subsets (i.e., first, third, fifth . . . versus second, fourth . . . conversations), measuring orientation on one subset and computing a counselor’s propensity for helpful ratings, or their average conversation length, on the other. Here, we draw an analogy to the machine learning paradigm of taking a train-test split, “training” counselor tendencies on one subset and “testing” their relation to rating or length on the other subset. In general, the directions of the effects we observe hold with stronger effects if we do not take this split.

Echoing conversation-level effects, counselors that tend to be less forwards-oriented and more backwards-oriented (those in the bottom thirds of Ω^{\max} and Ω^{\min} respectively) are more likely to be perceived as helpful; this contrast is stronger in terms of Ω^{\min} (Cohen’s $d = 0.30$, $p < 0.001$) than Ω^{\max} ($d = 0.13$, $p < 0.05$), suggesting that a counselor’s tendency for advancing factors less into their perceived helpfulness than their tendency for addressing. Also in line with the conversation-level findings, counselors with smaller Ω^{\max} tend to have longer conversations ($d = 0.54$, $p < 0.001$), as do counselors with larger Ω^{\min} ($d = 0.17$)—here, a counselor’s tendency for advancing is more related to their propensity for shorter conversations than their tendency for addressing.⁹

⁹We note that counselors cannot selectively take conversations with certain texters; rather, the platform automatically assigns incoming texters to a counselor. As such, the counselor-level effects we observe cannot be explained by counselor self-selection for particular situations.

3.6.5 Relation to message construction

We note that in practice, the choice to orient forwards or backwards is not necessarily either-or. For instance, consider a message c_3 where c_2 and c_1 from Figure 3.1 are concatenated, such that the counselor empathetically responds to the texter *and* moves the conversation forwards.

To examine such potential heterogeneities, we analyze the orientations of each sentence in the 64% of counselor messages containing multiple sentences. We find that 52% of these messages have $\Omega^{\min} < 0$ and $\Omega^{\max} > 0$. We also find that in 74% of multi-sentence messages, the last sentence has a higher orientation than the first (as is reflected in our concatenated example, c_3): perhaps unsurprisingly, counselors construct these messages to first address what was said already, before trying to advance forwards.

Combining sentences with different properties is just one potential strategy for turn construction [Drew et al., 2011]; for instance, a counselor could suggest a way to move the conversation forwards that directly builds on what the texter has disclosed. Future work could fruitfully examine a broader range of strategies employed by counselors to manage conversational balances and tensions.

3.7 Discussion

In this chapter, we sought to examine a key balance in crisis counseling conversations between advancing forwards and addressing what has already been said. Realizing this balance is one of the many challenges that crisis counselors must manage, and modeling the actions they take in light of such challenges could point to policies to better support them. For instance, our method could

assist human supervisors in monitoring the progress of ongoing conversations to detect instances of rushing or stalling, or enable larger-scale analyses of conversational behaviours to inform how counselors are trained. The unsupervised approach we propose could circumvent difficulties in getting large-scale annotations of such sensitive content.

Future work could bolster the measure’s usefulness in several ways. Technical improvements like richer utterance representations could improve the measure’s fidelity; more sophisticated analyses could better capture the dynamic ways in which the balance of objectives is negotiated across many turns. The preliminary explorations in Section 3.6.4 could also be extended to gauge the causal effects of counselors’ behaviors [Kazdin, 2007, Zhang et al., 2020].

We expect balancing problems to recur in conversational settings beyond crisis counseling. In settings such as court proceedings, interviews, debates and other mental health contexts like longer-term therapy, individuals also make potentially consequential choices that span the backwards-forwards orientation axis, such as addressing previous arguments [Tan et al., 2016, Zhang et al., 2016], asking follow-up questions [Huang et al., 2017], or asking leading questions [Leech, 2002]. Our measure is designed to be broadly applicable, requiring no domain-specific annotations; we provide exploratory output on justice utterances from the US Supreme Court’s oral arguments, and on utterances from the Switchboard Dialog Act Corpus, in Chapter 5.6. However, the method’s efficacy in the present setting is likely boosted by the relative uniformity of crisis counseling conversations. Future work could develop approaches that better accommodate settings with less structure and more linguistic variability.

Part II

Framework

CHAPTER 4

THE EXPECTED CONVERSATIONAL CONTEXT FRAMEWORK

4.1 Overview

We've presented two approaches that characterize utterances based on:

- the response prompted by a question (Chapter 2);
- the replies and predecessors surrounding utterances in the midst of an interaction (Chapter 3).

In both cases, we've examined utterances in terms of how they relate to surrounding utterances in an interaction, showing that the resultant characterizations reflect other important aspects of the setting, such as partisan affiliation or conversation structure.

We now present a general framework to characterize utterances in terms of their *expected conversational context*. Concretely, the framework relates utterances to the surrounding turns in an interaction—in particular, the replies and predecessors. The framework yields methods to compute a variety of utterance characterizations—including the approaches presented in Chapters 2 and 3—given a collection of conversation transcripts. These characterizations can be interpreted as measures of an utterance's role in an interaction. Henceforth, we refer to this framework as the *Expected Conversational Context Framework*.

We start by conceptually outlining the Expected Conversational Context Framework (Section 4.2), and then relating it to other approaches to analyzing and modeling conversations from literature in sociology, computational linguistics and NLP (Section 4.3). We then detail a particular method for operational-

izing the framework that we use throughout this dissertation (Section 4.4), and suggest variants for future work.

4.2 Conceptual description

At a high level, the Expected Conversational Context Framework ties together four main ideas, that we've already made use of in Chapters 2 and 3:

1. It relates utterances to their *conversational context*;
2. It derives characterizations of utterances, and of their constituent terms, that quantify properties of their *expected* context;
3. It computes these characterizations given a *dataset* of conversations.
4. It performs such computations by embedding terms, utterances and contexts in a *shared* latent vector space.

In conjunction, these ideas result in a range of term- and utterance-level characterizations, which we outline in Table 4.1 (see Section 4.2.5), and which we empirically examine throughout the dissertation.

4.2.1 Conversational context

The core intuition underlying the Expected Conversational Context Framework is that the role of an utterance in an interaction is informed by the context in which it appears. In general, context encompasses an innumerable broad range of factors (see Chapter 6 for further discussion). Taking a necessarily narrower view, our framework relates utterances to a subset of its surrounding turns in a

conversation. Henceforth in this chapter, we will use “context” to refer to this limited conception of context, and will refer to an utterance’s surrounding turns as its *context-utterances*. As we discuss in Section 4.3, the framework’s focus on surrounding turns reflects the significance accorded to this aspect of context by existing approaches in sociology and computational linguistics, notably conversation analysis and centering theory.

Note that the framework’s output depends on our choice of what to consider as a context-utterance. In this dissertation, we focus on two such choices: we relate utterances to their immediate replies—the *forwards* context, or to their immediate predecessors—the *backwards* context, yielding *forwards* or *backwards* characterizations. Other choices of context-utterance are possible as well: later, we briefly explore an application of the framework that relates utterances to subsequent turns from the same speaker (skipping over the intervening reply).

4.2.2 Expected context

Our framework characterizes utterances in terms of their *expected* conversational context. Intuitively, given an utterance, we can imagine some replies as being more likely than others; likewise, we can imagine some predecessors that the utterance was more likely or less likely to follow. In other words, we conceive of a distribution over possible context-utterances; the characterizations yielded by the framework correspond to quantitative properties of this distribution. Roughly speaking, to computationally operationalize the framework, we formulate a method of inferring such distributions, given conversation data.

Alternative approach: direct comparison. We distinguish our approach from methods that might directly compare an utterance to the particular turns that

surround it in a conversation—i.e., the reply and predecessor that *actually* occurred, rather than what was *expected*. To restate the shortcomings of such a direct comparison approach, as described in Chapter 3.5.1:

- such an approach would struggle to draw connections between utterances and surrounding turns that are lexically distinct, such as when two interlocutors have different roles (like a counselor and a patient);
- the approach would fail to draw connections between utterances and replies that go beyond restating what came before (as when a responder somehow reframes a preceding statement);
- the approach would be unable to disentangle the utterance’s intended role, on the part of its speaker, from the potentially unexpected nature of its reply (as when a responder dodges a question) or predecessor (as when a speaker misunderstands the nature of the preceding turn);
- in a practical scenario where the framework is deployed to analyze conversations in real time, the approach would have to wait for an utterance’s reply to be observed before being able to characterize it, incurring a delay.

From terms to utterances. To set our expectations about the context of an utterance, we start by setting our expectations about the contexts of its constituent *terms* (e.g., n-grams, dependency-tree arcs). For a given term w , the framework derives a representation that reflects the distribution of context-utterances corresponding to utterances containing w . These term-level properties are informative in their own right; aggregating them across all of the terms in an utterance also enables us to characterize the utterance itself. Here, we make use of two assumptions that are common across NLP:

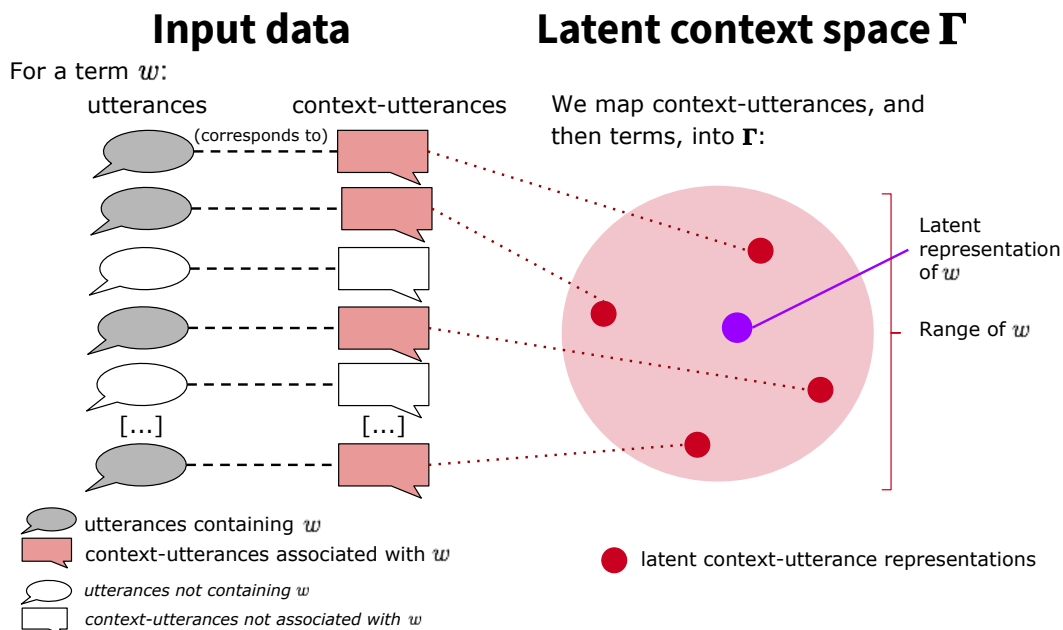


Figure 4.1: Sketch of procedure to compute latent term representations, as outlined in Sections 4.2.3 (left) and 4.2.4 (right).

- terms have stable characteristics across the multiple utterances and contexts in which they may appear;
- term characteristics can be somehow composed together to characterize an utterance.

4.2.3 Inference from conversation data

We compute our characterizations of the expected contexts of terms, and hence of utterances, given a collection of conversation transcripts. Our procedure is sketched in Figure 4.1. Concretely, we associate each term with the set of utterances containing that term, and with the set of context-utterances surrounding those utterances. For instance, if we consider replies as context, then a term is associated with the replies to the utterances that contain it (Figure 4.1, left).

For each term, we use its context-utterance set to model the distribution of its associated context-utterances.

Note that our characterizations are dependent on our choice of dataset. For instance, we do not expect a term to have similar characteristics when it occurs in parliamentary question periods and when it occurs in counseling conversations. As such, we derive measures that are tailored to a particular setting.

4.2.4 Embedding-based approach

To represent terms with respect to their conversational context, we draw on a common, distributional semantics paradigm in NLP, where linguistic objects like terms and utterances are represented as points in a vector space. In this space, two utterances are geometrically close together if they are similar in some way. Most commonly, such spaces are thought of as modeling semantic similarity; for our purposes, we will use the space to model similarity in associated or expected conversational contexts.

We start by deriving a *latent context space* that models semantic similarity among context-utterances, such that representations of context-utterances are geometrically close if the context-utterances themselves are similar. Consequently, each term is associated with a collection of vectors representing its corresponding set of context-utterances. Computing some property of this geometric collection then corresponds to deriving some characterization of the term with respect to its context.

In this way, we *map* terms from the input data into the latent space. In particular, we can represent each term in the latent space as a point in the centre of the region spanned by the embeddings of its associated context-utterances (Figure

4.1, right). By combining the latent representations for each term in an utterance (e.g., via taking an average), we can represent the utterance in the context vector space as well.

By embedding terms, utterances, and contexts in the *same* context vector space, we can meaningfully interpret distances between their representations. For instance, if we take replies to be conversational context, then:

- two terms, or two utterances, have latent representations that are close together if we expect them to be followed by similar replies;
- an utterance and a context-utterance representation are close together if we expect the context-utterance to be a likely response to the utterance.

4.2.5 Derived characterizations

Building off of the term, utterance and context embeddings in the shared latent space, we can derive a range of characterizations of terms and utterances. These characterizations are outlined in Table 4.1. In Chapters 2, 3 and 5, we empirically explore them and the analyses they enable.

Latent representations. The primary outputs of the framework are the latent representations of terms and utterances, mentioned above, that model their expected conversational context. In particular, we compute *forwards*-representations—modeling replies—and *backwards*-representations—modeling predecessors. In Chapter 2, we clustered forwards-representations of questions (with respect to answers) to arrive at a typology of questions based on the replies they aim to prompt.

	<i>Term</i>	<i>Utterance</i>
Vector representations		
forwards-representation (replies) <i>Chapters 2, 5.2.2, 5.6.1, 5.6.3</i>	$\vec{\phi}(w)$	$\vec{\Phi}(a)$
backwards-representation (predecessors) <i>Chapters 5.2.1, 5.2.2, 5.6.3</i>	$\overleftarrow{\phi}(w)$	$\overleftarrow{\Phi}(a)$
skip-representation (speakers' next turn) <i>Chapter 5.2.3</i>	$\phi'(w)$	$\Phi'(a)$
Expectation strengths		
forwards-range <i>Chapter 5.3.1</i>	$\vec{\sigma}_w$	$\vec{\Sigma}_a$
backwards-range <i>Chapter 5.3.1</i>	$\overleftarrow{\sigma}_w$	$\overleftarrow{\Sigma}_a$
Measures comparing forwards and backwards contexts		
orientation (compares forwards- and backwards-ranges) <i>Chapters 3, 5.6.2, 5.6.3</i>	Ω_w	Ω_a
shift (distance b/n forwards and backwards representations) <i>Chapters 5.3.2, 5.6.3</i>	δ_w	Δ_a
Other comparative measures		
skip-shift (skip vs. LSA representations) <i>Chapter 5.3.3</i>	δ'_w	Δ'_a
unexpectedness (forwards vs. reply representations) <i>Chapter 5.5</i>	n/a	$\mathcal{U}(a; r)$

Table 4.1: Overview of characterizations of terms w and utterances a derived via the Expected Conversational Context Framework, and empirically examined in the indicated dissertation chapters.

Expectation strengths. Intuitively, we may have stronger expectations of the replies prompted by some utterances versus others—for instance, a leading question might suggest a narrower range of answers than an open-ended one. Likewise we may have a stronger or weaker sense of what predecessor an utterance follows. To quantify the strengths of these expectations, we can measure the extent to which a distribution of context-utterances is concentrated or spread out—an approach we considered in Chapter 3. For a term, we measure the size of the region in the context latent space spanned by the representations of its associated context-utterances; a smaller measure indicates we have stronger expectations (since the context-utterances are more narrowly concentrated). By aggregating this measure across terms, we can characterize utterances as well. We refer to this quantity as the *range* of a term or utterance, distinguishing between *forwards-ranges*—modelling our expectations of replies—and *backwards-ranges*—modelling our expectations of predecessors.

We can draw a rough (if not statistically rigorous) analogy between forwards-representations and ranges, and the mean and variance of a distribution of replies, respectively. Likewise, we can think of backwards-representations and ranges as means and variances of distributions of predecessors. We show these two key properties, at the term level, in Figure 4.1 (right).¹

Comparing properties across contexts. By comparing the characterizations that the framework yields given different choices of context, we can derive further measures of how terms and utterances relate to the surrounding interaction. For this dissertation, we focus on making comparisons between forwards and backwards characterizations (i.e., those based on expected replies, versus expected

¹Strictly speaking, the characterizations derived under our present approach, as detailed in Section 4.4, do not have probabilistic interpretations: we do not specify a way to convert latent representations or the distances between them into probabilities. This gap could be fruitfully addressed in future work.

predecessors).² In particular, we explore the following measures:

- We can compare forwards- and backwards-ranges by taking a difference between the two. As such, we derive a measure, *orientation*, that contrasts the relative strengths of our expectations of what reply will follow an utterance, versus what precedes it. In Chapter 3, we interpreted orientation as capturing how an utterance directs the flow of the interaction.
- We can compare forwards- and backwards-representations by measuring the distance between them. The resultant measure, *shift*, models the extent to which an utterance is expected to move the conversation from one focus to another; larger differences correspond to utterances we expect to shift focus more drastically. We explore this property in Chapter 5.3.2.

Comparing expected and actual context-utterances. To what extent did we expect the reply that an utterance ultimately received in a conversation? Note that our framework provides one possible way to address this question: we derive forwards-representations of utterances that model our expectations of their replies, and that we can directly compare to vector representations of the replies in the context vector space. As such, if the reply vector is far in the space from the utterance vector, we'd interpret the reply as being more unexpected. In Chapter 5.5, we empirically explore this measure of unexpectedness.

4.3 Related work

As we've described, our framework operates on a key structural relationship in a conversation, between an utterance and its surrounding turns. Here, we look

²As we detail later, forwards and backwards characterizations are directly comparable if we use the same context vector space to model both choices of context.

to other literature that elaborates on how this relation is informative.

A central focus of conversation analysis is on making sense of utterances with respect to their surrounding turns [Schiffrin, 1994, Hoey and Kendrick, 2017]. In particular, a fundamental structural unit in conversation analysis is the *adjacency pair*—two utterances from two different speakers, in succession [Schegloff, 1968, Heritage, 1991]. By virtue of this structure, the second utterance is seen as *conditionally relevant* to the first: what we understand of the first utterance impacts how we interpret the second. This may mean that we have some expectation of what a reply will look like; it also means that the way we read *any* subsequent behaviour is contingent on the first utterance (for instance, a seemingly random remark, placed after a question, could be read as a dodge or a non-sequitur). In the middle of a conversation, these adjacency relations are significant in both directions. Per Heritage [1991], utterances are “doubly contextual;” they are “both context-shaped and context renewing” [Schiffrin, 1994]. In other words, our understanding of an utterance, as conversationalists and as analysts, depends both on how it relates to the prior context, and on what expectations it subsequently sets up. Our framework can be seen as a way of statistically modeling how utterances are shaped by, or how they renew conversational context.

Other work on discourse has also elaborated on the significance of such forwards and backwards relations [Webber, 2001]. In centering theory [Grosz et al., 1995], which proposes a model of discourse and its coherence, utterances are conceived of as having *backwards-* and *forwards-looking* centres: backwards-looking centres reference objects in the preceding discourse; forwards-looking centres serve as potential referents for the subsequent interaction. Other theoretical frameworks in that vein consider alternate formulations of backwards

and forwards-looking relations [Prince, 1981, Strube, 1998]. While past empirical considerations of these theories have sought to devise ways of identifying centres within utterances [Byron and Stent, 1998], our work more abstractly considers properties of our *expectations* of forwards and backwards relations. The characterizations we derive, as such, are not informative of particular referents so much as higher-level intentions or focal points in an interaction.

A range of computational work has also sought to model structural relations in conversations by way of graphical models. The starting assumption of such approaches is that discourse is comprised of latent types that may be sequentially related; an utterance’s membership in one or more types then generates its linguistic form. In Ritter et al. [2010], these types are interpreted as dialog acts in conversations on Twitter; in Althoff et al. [2016], they are interpreted as stages in a crisis counseling conversation (from the same setting that we consider). These methods proceed by algorithmically deriving latent types and then characterizing utterances in light of them; our framework can be seen as working in the reverse, deriving utterance characteristics and then inferring types as an additional interpretative step. We later comment more specifically on how our method’s output compares with that described in Althoff et al..

We use our framework to derive various properties pertaining to what utterances do in an interactions. Here, we draw a contrast to a common paradigm that aims at similar characterizations, by specifying ontologies. In early philosophical work, Searle [1976] proposes a taxonomy of five speech acts; more granular taxonomies, that have been taken up computationally, can be found in the DAMSL annotation scheme [Core and Allen, 1997] and in rhetorical structure theory [Mann and Thompson, 1988]. In particular domains, including those we consider in this dissertation, researchers and practitioners have developed

more specialized taxonomies that enumerate such categories as types of parliamentary question [Bates et al., 2014], types of discourse acts in online discussions [Zhang et al., 2017a], or types of actions or strategies in counseling or therapy interactions [Houck, 2008, Can et al., 2015, *inter alia*]. The computational task, given these taxonomies, is generally to derive linguistic signals that enable utterances to be automatically tagged according to them. Our work, by contrast, does not start by assuming an ontology: rather, we inductively infer similarities among groups of utterances based on linguistic and conversational patterns; if there are well-defined categories, we wish to discover rather than presuppose them. In our analyses, we compare our framework’s output with several of these ontologies, pointing out similarities, as well as distinctions or analogies that we draw and that are absent from them.

The technical operationalization of our framework, which we subsequently detail, draws heavily on distributional semantics [Firth, 1957, Landauer and Dumais, 1997]. The main ideas we take from this area are that we can make inferences about terms—and by extension, the utterances they comprise—based on the contexts in which terms occur in data, and that we model terms as points in a vector space, where distances correspond to semantic similarities. Our method likewise represents terms in a vector space; in particular, we build on one approach from this line of work, latent semantic analysis [Deerwester et al., 1990], to derive term vectors. The key distinction we make to conventional distributional approaches is that we explicitly consider conversation structure: we derive representations based on the *conversational* context, rather than context as defined by surrounding terms within a document or utterance. In Chapter 5.4, we illustrate some of the empirical consequences of this distinction, suggesting that we yield characterizations that more clearly reflect the interaction.

4.4 Methodology

We now detail our method to computationally operationalize the Expected Conversational Context Framework. We outline notation and key equations used in our method in Table 4.2. To focus our description, we consider replies or predecessors as the conversational context, noting that the method can be analogously applied for other choices of context as well.

4.4.1 Specifying the input data

The framework takes as input a collection of conversation transcripts, containing utterances $\{a_1, a_2, \dots\}$ and context-utterances $\{c_1, c_2, \dots\}$. We can think of this collection as our “training data.”³

Per our choice of conversational context, we associate each utterance a_i with its corresponding context-utterance c_i , e.g., its reply or its predecessor. Formally, we let $\vec{\gamma}(a_i)$ denote the reply of a_i (i.e., the *forwards*-context), and $\overleftarrow{\gamma}(a_i)$ denote its predecessor (the *backwards*-context). As noted above, the dissertation focuses on these two choices of context-utterance, but the subsequent description can be adapted for other choices of context-utterance.

Note that the sets of utterances and context-utterances may overlap, since both are comprised of utterances in the data. That said, we can make distinctions between them on the basis of domain knowledge. For instance, in this work, the conversational settings we examine involve speakers with different roles (e.g., question-askers versus answerers in Chapter 2; counselors versus texters in Chapter 3). As such, we distinguish between utterances and context-

³Note that our unsupervised method does not use any annotations of the data, so this collection is not training data as conventionally thought of in a supervised learning paradigm.

Input data

- a_i utterances (e.g., parliamentary questions, counselors' messages)
- w_j terms (e.g., dependency-parse arcs, unigrams)
- c_i context-utterances (e.g., ministers' answers, texters' messages)
- v_j context-terms
- $\vec{\gamma}(a_i)$ the reply of a_i in the data, i.e., the *forwards-context*
- $\overleftarrow{\gamma}(a_i)$ the predecessor of a_i in the data, i.e., the *backwards-context*

Input representations

- A utterance-term matrix, where row A_i represents a_i and column $A^{(j)}$ represents w_j (e.g., as tf-idf reweighted vectors)
- C context-utterance-term matrix, where row C_i represents c_i and column $C^{(j)}$ represents v_j
- \mathcal{C} input term space, where dimension i corresponds to context-utterance c_i . Note that $C^{(j)} \in \mathcal{C}$.

We can permute rows of A to align with rows of C :

- \vec{A} matrix where row \vec{A}_i represents a_i with $\vec{\gamma}(a_i) = c_i$, and column $\vec{A}^{(j)}$ represents w_j
- \overleftarrow{A} matrix where row \overleftarrow{A}_i represents a_i with $\overleftarrow{\gamma}(a_i) = c_i$, and column $\overleftarrow{A}^{(j)}$ represents w_j

Note that both $\vec{A}^{(j)}$ and $\overleftarrow{A}^{(j)}$ are in \mathcal{C} .

Latent context representations

- Γ latent context space, a d -dimensional vector space modeling semantic similarity among contexts
- U, s, V d -dimensional factors approximating \mathcal{C} , derived via SVD: $\mathcal{C} \approx UsV^T$. Equivalently, this step derives:
 - ϕ mapping from \mathcal{C} to Γ , where $\phi(C^{(j)}) = C^{(j)T}Us^{-1}$ is the representation of v_j in Γ

Latent term representations

- $\vec{\phi}(w_j)$ forwards-representation of w_j in Γ , where $\vec{\phi}(w_j) = \phi(\vec{A}^{(j)}) = \vec{A}^{(j)T}Us^{-1}$
- $\overleftarrow{\phi}(w_j)$ backwards-representation of w_j in Γ ; $\overleftarrow{\phi}(w_j) = \phi(\overleftarrow{A}^{(j)}) = \overleftarrow{A}^{(j)T}Us^{-1}$
- \vec{W} matrix where row $\vec{W}_j = \vec{\phi}(w_j)$
- \overleftarrow{W} matrix where row $\overleftarrow{W}_j = \overleftarrow{\phi}(w_j)$

Latent utterance representations

Let a be a vector representation of a new utterance a . Then:

- $\vec{\Phi}(a)$ forwards-representation of a in Γ , where $\vec{\Phi}(a) = a\vec{W}s^{-1}$
 - $\overleftarrow{\Phi}(a)$ backwards-representation of a in Γ , where $\overleftarrow{\Phi}(a) = a\overleftarrow{W}s^{-1}$
-

Table 4.2: Notation and key equations used in computationally operationalizing the Expected Conversational Context Framework.

utterances according to the roles played by their speakers: we take utterances to be questions asked by Members of Parliament or messages sent by counselors, and context-utterances to be responses given by government ministers or messages sent by texters.

Specifying terms. We break utterances into their constituent terms w_j , and context-utterances into constituent context-terms v_j .⁴ There are numerous ways in which terms and context-terms can be defined. Throughout this work, we consider unigrams or dependency-parse arcs. Note that the choice of terms can reflect particular intuitions about the relation between an utterance’s role in an interaction and its surface form: for instance, consider the particular choice of question terms detailed in Chapter 2, which we tailor for deriving rhetorical rather than topic-based representations.

Representing the input data. We represent the collection of utterances and context-utterances as utterance-term matrices (i.e., document-term matrices, where documents are utterances). We denote the matrix representing utterances as A , with rows A_i representing utterances and columns $A^{(j)}$ representing terms; we denote the matrix of context-utterances as C , with rows C_i and columns $C^{(j)}$ representing context-utterances and context-terms, respectively. We tf-idf reweight A and C , noting that future work could explore other reweighting schemes. We refer to A , C , and their rows and columns as *input representations*.

4.4.2 Deriving latent term representations

Following our high-level sketch, our method starts by representing terms with respect to context-utterances. In particular, we will embed terms and contexts

⁴We introduce separate notation to disambiguate between terms and context-terms, but note that they could be defined in the same way, and that the respective vocabularies may overlap.

in a shared latent context space. As such, we must derive this latent space, and then derive a mapping of the input representations of terms—i.e., columns $A^{(j)}$ of utterance-term matrix A —into the latent space. We can then aggregate latent representations of terms to derive utterance-level representations as well.

Our approach uses latent semantic analysis to derive the latent context space, and then to derive a *linear* mapping between the space of input representations and the latent context space. Throughout our description, we also suggest alternative methodological choices that future work could fruitfully explore.

Deriving the latent context space. We first derive a latent space Γ that models similarity among conversational contexts, where context-utterances c_i and c_j are represented as points that are geometrically close in Γ if c_i and c_j are similar.

Our approach for deriving Γ , and for embedding context-utterances in it, is to use latent semantic analysis (LSA) [Deerwester et al., 1990, Landauer and Dumais, 1997]. Per LSA, utterances that share similar terms, and terms that co-occur across many utterances, are considered to be semantically similar; these co-occurrences are modeled as linear relationships.

In particular, we use LSA to derive d -dimensional representations of context-utterances and context-terms, given C , the tf-idf reweighted context-utterance-term matrix representing the input data (we set d to be smaller than the number of context-terms). Via singular value decomposition (SVD), we approximate C as a product of d -dimensional factors: $C \approx USV^T$. Rows of U are then latent representations of context-utterances in Γ , while rows of V are representations of context-terms. As we later illustrate (Chapter 5.1.1), having both types of embedding enables us to easily interpret the framework’s outputs at the level of both utterances and terms.

Deriving a mapping into Γ . Ultimately, we wish to derive a mapping of the input representations of terms into Γ . Note that the SVD we performed to derive Γ gets us partway there, in computing latent representations of *context-terms*.

Formally, consider a vector space C , where the i th dimension corresponds to c_i , the i th context-utterance in the data. Note that columns of C , representing context-terms v_j , can be seen as elements of C : the i th entry of column $C^{(j)}$ denotes the weight of v_j in c_i , so $C^{(j)} \in C$. Then the SVD operation derives a mapping $\phi : C \rightarrow \Gamma$, where $\phi(C^{(j)}) = C^{(j)T} U S^{-1}$ (i.e., the j th row of V).

Embedding terms in context-space. We now describe how we embed terms in Γ , with respect to their associated contexts. Our procedure builds on the mapping ϕ we've just derived.

Note that ϕ is defined over vectors in C , i.e., representations of *context-terms* in columns $C^{(j)}$ of context-utterance-term matrix C . However, thus far, we've represented terms w_j in columns $A^{(j)}$ of utterance-term matrix A . In other words, our input term representations $A^{(j)}$ aren't in the correct domain C , so we can't directly apply ϕ to map them into Γ . Rather, we first need to convert $A^{(j)}$ to a vector in C . Here, we make use of the correspondence between utterances and context-utterances. Intuitively, we will "line up" the rows of A and C according to this correspondence.

Suppose we consider replies as our choice of context. Formally, we construct a new matrix \vec{A} by permuting rows of A , such that the i th row of \vec{A} , \vec{A}_i , represents an utterance a_i whose reply c_i is represented as the i th row of C , i.e., $\vec{\gamma}(a_i) = c_i$.⁵ Importantly, given the correspondence between a_i and c_i , we can

⁵If a context-utterance c_i does not have a corresponding utterance, then we set all entries of row \vec{A}_i as zero. This could be the case if we use a larger set of context-utterances than utterances as training data, e.g., when context-utterances can be replies or predecessors to the utterances in the data, as in Chapter 3. Equivalently, we can construct a new matrix comprised of a subset

view columns $\vec{A}^{(j)}$ of \vec{A} , representing terms w_j , as points in C as well: the i th entry of $\vec{A}^{(j)}$ denotes the weight of w_j in a_i , the utterance that c_i is replying to.

We can therefore use ϕ to map each term w_j , represented as $\vec{A}^{(j)}$, to its forwards-representation in Γ , as $\phi(\vec{A}^{(j)}) = \vec{A}^{(j)T} U s^{-1}$. As shorthand, we write the mapping as $\vec{\phi}(w_j) = \vec{A}^{(j)T} U s^{-1}$ (the upper arrow clarifies that the mapping pertains to the term's forwards context, i.e., replies).

We can follow an analogous procedure to represent w_j in terms of other choices of context. For instance, suppose we now take predecessors to be context-utterances. We construct matrix \overleftarrow{A} by permuting rows of A such that the i th row, \overleftarrow{A}_i , represents the utterance a_i whose predecessor is c_i , i.e., $\overleftarrow{\gamma}(a_i) = c_i$. Here, columns $\overleftarrow{A}^{(j)}$ can be seen as elements of C , again due to the correspondence between a_i and c_i . We can then apply ϕ to derive backwards-representations, which we write as $\overleftarrow{\phi}(w_j) = \overleftarrow{A}^{(j)T} U s^{-1}$.

Correspondence to formulation in Chapter 2. Let \vec{W} be the matrix whose j th row is $\vec{\phi}(w_j)$. Note that we can solve for \vec{W} in the equation $\vec{A} = U s \vec{W}^T$, corresponding to the method for deriving question term representations from Chapter 2. We can analogously define \overleftarrow{W} , a matrix whose j th row is $\overleftarrow{\phi}(w_j)$, and that satisfies the equation $\overleftarrow{A} = U s \overleftarrow{W}^T$.

Correspondence to formulation in Chapter 3. For a term w_j , let \vec{w}_{w_j} denote a vector containing the non-zero entries of $\vec{A}^{(j)}$, i.e., its weights in the utterances in which it occurs. Additionally, let \vec{U}_{w_j} denote a matrix containing rows of U representing the subset of replies to the utterances in which w_j occurs—i.e., the replies associated with w_j . Then forwards-representation $\vec{\phi}(w_j)$ can be written as $\vec{\phi}(w_j) = \vec{w}_{w_j}^T \vec{U}_{w_j} s^{-1}$. We therefore see that $\vec{\phi}(w_j)$ is equivalent to the central points \vec{u}_{w_j} derived in Chapter 3.

of rows in C for which there is a corresponding utterance.

Note that this formulation explicitly highlights that our approach associates terms with a subset of context-utterances and their corresponding latent representations in Γ : to compute term representations, we aggregate these context-utterance representations (as rows of \vec{U}_{w_j}) by taking a linear combination (with coefficients w_{w_j}), and scaling dimensions by s .

An analogous argument applies to other choices of context: we take a linear combination of the rows in U representing context-utterances associated with w_j . In particular, we note that backwards-representation $\overleftarrow{\phi}(w_j)$ is equivalent to central point \overleftarrow{u}_{w_j} from Chapter 3.

Computing expectation strengths. As in Chapter 3, to quantify the strengths of our expectations of term w_j 's replies, we measure how “spread out” are the latent representations of its corresponding context-utterances: we take the average cosine distance between each context-utterance representation and the latent representation of w_j . We refer to the resultant quantity as the *range*, distinguishing in particular between *forwards-range* $\vec{\sigma}_{w_j}$ —comparing w_j 's forwards-representation to the latent representations of its replies—and *backwards-range* $\overleftarrow{\sigma}_{w_j}$ —comparing w_j 's backwards-representation to its predecessors.

Comparing forwards and backwards contexts. Note that our procedure can yield forwards and backwards characterizations that come from the same latent space Γ : for either choice of context, we use the same set of context-utterances to derive Γ , but use different correspondences between terms and context-utterances (i.e., taking replies versus predecessors) to map terms into Γ . As such, the forwards and backwards characterizations are directly comparable. This enables us to compute the orientation measure introduced in Chapter 3—comparing forwards- and backwards-ranges, and the shift measure we examine Chapter 5.3.2—comparing forwards- and backwards-representations.

Alternative approaches. Our LSA-based method offers a procedure for embedding and characterizing terms that naturally derives from our procedure for embedding contexts, via SVD. Future work could consider other options for representing context-utterances, and then terms, in a shared latent space. As a starting point, approaches such as word2vec [Mikolov et al., 2013] and BERT [Devlin et al., 2019] model semantic similarity via more expressive, non-linear functions. These neural methods may therefore derive embeddings of context-utterances that better model semantic similarity. An operationalization of the framework that uses such embedding methods would need to specify a way to embed terms as well. Here, we could draw on neural approaches to modeling or generating discourse, which formulate ways to “predict” subsequent utterances [Serban et al., 2016].

As a further extension of the framework, we suggest ways of modeling context-utterances that aren’t directly premised on lexical characteristics. For instance, one could consider some measure of the emotional valence of a reply, resulting in a variant of the framework that characterizes utterances in terms of the sentiment elicited.

Future work could also consider other approaches to compute forwards- and backwards-ranges, that might better reflect the distribution of context-utterances for a term. For instance, while our present formulation compares replies or predecessors to a single point given by $\vec{\phi}$ or $\overleftarrow{\phi}$, other methods could account for the possibility of multiple, semantically distinct types of expected context-utterances (e.g., consider the potentially bifurcated space spanned by expected responses to yes-or-no questions).

Finally, while we draw a rough analogy between our latent representations and forwards/backward-ranges, and the mean and variance of a distribution of

context-utterances, future work could pursue approaches with more explicitly probabilistic interpretations. Such efforts could allow us to more rigorously address problems such as quantifying the extent to which an utterance’s reply is expected or unexpected, as we briefly explore in Chapter 5.5.

4.4.3 Aggregating from term to utterance-level representations

To represent an utterance with respect to its expected conversational context, we aggregate representations of its terms. We want the relative contribution of a term’s representation to be informed by its relative importance in the utterance. In our approach, we model term importance via tf-idf reweighting; given an utterance a_i , we denote its tf-idf reweighted vector representation as \mathbf{a}_i .

Utterance-level latent representations. Let $\vec{\Phi}(a_i)$ denote the forwards-representation of a_i , i.e., the point in Γ representing its expected replies. We take $\vec{\Phi}(a_i) = \mathbf{a}_i \vec{W} s^{-1}$. In words, we represent the utterance as a linear combination of forwards term representations (from \vec{W}), with tf-idf weights (from \mathbf{a}_i) as coefficients; we rescale dimensions of this weighted sum by s to properly map it into Γ . As such, this formulation corresponds to our derivation of question representations from Chapter 2. Likewise, we take backwards-representation $\overleftarrow{\Phi}(a_i)$, representing the expected predecessors of a_i , to be $\overleftarrow{\Phi}(a_i) = \mathbf{a}_i \overleftarrow{W} s^{-1}$.

We further motivate this formulation as follows. Suppose our conversation dataset consists of pairs of utterances and replies where the reply simply repeats the utterance, i.e., within each pair, utterance a_i and reply c_i are identical. As such, the input matrices we’ve constructed above, \vec{A} and C , are also identical to each other, as are the latent term and context-term representations given by \vec{W} and V . Accordingly, $\vec{A} = C = U s V^T = U s \vec{W}^T$. Since a_i and c_i are identical, we

would like for their vector representations to be identical as well, i.e., $\vec{\Phi}(a_i)$ is equal to the i th row of U . Solving for U in the above equation gives $U = \vec{A}Vs^{-1}$, whose i th row is equivalent to our above expression for $\vec{\Phi}(a_i)$. An analogous argument applies for $\overleftarrow{\Phi}(a_i)$.

Utterance-level expectation strengths. As in Chapter 3, we take the forwards-range of a_i , $\vec{\Sigma}_{a_i}$ to be a weighted average of term-level forwards-ranges $\vec{\sigma}_{w_j}$, using tf-idf weights given by entries of a_i . Likewise, we take backwards-range $\overleftarrow{\Sigma}_{a_i}$ to be a weighted average of term-level backwards-ranges.

Note that we can modify this formulation to reflect domain-specific intuitions about utterance form. For instance, in the counseling data from Chapter 3, we observed that counselors often use different sentences from the same utterance to address different aspects of the conversation (an idea we more extensively explore in the subsequent analyses). Accordingly, we computed sentence-level forwards and backwards ranges, and statistics derived from these, before taking aggregates across sentences like maximums and minimums.

Alternative approaches. Our approach relies on two main assumptions:

- that utterances can be modeled as terms that can be separately characterized and then combined together (per a standard bag-of-words assumption in NLP);
- that our weighting scheme, of taking tf-idf weights, adequately models the extent to which each term should inform the utterance’s characterization.

Future work could explore approaches that build on or relax these assumptions. For instance, via neural models, we could combine term characteristics in ways that more expressively reflect properties like term order and syntactic

structure. More sophisticated weighting schemes could better reflect the relative importance of terms; here, we could draw on approaches such as attention mechanisms [Bahdanau et al., 2014].

4.4.4 Particular implementation choices

When applying our method, we need to make several practical choices pertaining to the input data and representations, as well as the latent context vector space. In the appendix, we include further details about particular decisions we made in each setting we analyzed in the dissertation.

Here, we highlight two particular implementation choices we found to empirically produce better output, across all the settings we considered. At a high level, our framework builds off of term-level characterizations, and we would like to ensure that these characterizations are not skewed by the relative frequencies of terms. To this end:

- We scale *columns* of the matrices C , \vec{A} and \overleftarrow{A} , representing the occurrence of terms and context-terms in the training data, to unit ℓ_2 norm;
- We remove the first dimension from our latent term, utterance and context representations in Γ , since, across the datasets we considered, this dimension strongly corresponds to term frequency.

To encourage further experimentation, we release an implementation of the Expected Conversational Context Framework as part of the ConvoKit library [Chang et al., 2020],⁶ along with code demonstrating its use in various analyses presented throughout the dissertation.

⁶<https://convokit.cornell.edu/>

CHAPTER 5

EXPLORING FRAMEWORK OUTPUT

5.1 Overview

Having described the Expected Conversational Context Framework in general terms, we now revisit and elaborate on the work presented in Chapters 2 and 3. In particular, we can explicitly reframe the methods we've proposed as instantiations of the broader framework:

- In Chapter 2, we characterized questions asked by Members of Parliament (MPs) in terms of their expected replies. As such, the answers provided by government ministers served as our choice of conversational context; we accordingly derived forwards-representations of questions and question terms, clustering them to derive a typology of questions based on their rhetorical intent.
- In Chapter 3, we characterized messages sent by crisis counselors in terms of their expected replies and predecessors. As such, the messages sent by texters served as our choice of conversational context. We accordingly computed and compared forwards- and backwards-ranges of counselors' terms and utterances, to quantify how counselors used their messages to orient the flow of the interaction.

In the subsequent vignettes, we explore a range of extensions to these analyses that follow from our more general formulation, and that serve as seeds for future work. Through this exploration, we show that the framework is generative of a broad range of characterizations and lines of inquiry on utterances

and their roles in interactions. By more extensively examining the framework’s output, we also clarify some of its properties, and highlight technical and conceptual limitations.

To briefly summarize, in Sections 5.2 and 5.3, we explore utterance characterizations that extend those presented in the preceding chapters, and that come out of applying the framework with different combinations of conversational context. In Section 5.4, we empirically compare representations derived from the framework, and those derived under a standard distributional paradigm, in which context is considered to be surrounding terms in an utterance, rather than surrounding utterances in a conversation. In Section 5.5, we consider an additional line of inquiry that the framework enables: contrasting our expectations of an utterance’s reply with the reply it actually receives.

We perform these analyses on the parliament and counseling datasets, using the same implementation choices as in the preceding chapters (and as detailed in the appendix). We use the former setting to explore forwards-representations in a structured, socially rich question-answer scenario, and we use the latter to explore how the framework can be applied to an extended interaction where utterances both prompt and respond; we also use such structural differences to draw some informative contrasts. Finally, in Section 5.6, we explore the framework’s output on a selection of other interactional settings.

Note on source material. Section 5.6.1 references material originally found in Zhang et al. [2018], which used an earlier version of the framework; we report results on the datasets from that paper using an updated implementation. Section 5.6.2 contains updated results to those originally reported in Zhang and Danescu-Niculescu-Mizil [2020].

5.1.1 Examples of nearby vector representations

Before we explore the framework’s output more broadly, we use some hand-selected examples to concretely illustrate its core geometric idea: representing terms, utterances and contexts in the same latent vector space. Mapping all of these conversational objects to a shared space enables us to make well-defined comparisons between them. To reiterate Chapter 4.2.4:

- Two terms or utterances have geometrically close latent representations if we expect them to occur in similar contexts;
- Latent representations of terms and utterances are geometrically close to latent representations of their expected context-utterances.

Table 5.1 lists examples of terms w from questions asked in the parliamentary setting, along with other question terms and answer terms, whose latent representations are close to forwards-representation $\vec{\phi}(w)$ in cosine distance.¹ Consider term *can do* (e.g., *What **can** the Government **do** to help?*). Nearby answer terms suggest that questions with *can do* tend to be met with replies voicing that the government is, or is going to see to a matter (e.g., *I **am keen** to ..., I **had recently** met with...*); nearby question terms suggest that other questions that tend to prompt similar replies might ask what officials *are doing* or whether they will *work with* a relevant party.

Table 5.2 lists examples of terms w from counselors. Note that in the counseling setting, we can characterize counselor utterances and terms with respect to either replies or predecessors. As such, we list counselor and texter terms whose latent representations are close to $\vec{\phi}(w)$, pertaining to typical replies, and $\overleftarrow{\phi}(w)$,

¹Each example is taken from the twenty nearest question or answer terms.

<i>will [you] explain</i> Nearby question terms: <i>why, will [you] admit, cost</i> Nearby answer terms: <i>is simple, apologise for, as described</i>
<i>agree is</i> Nearby question terms: <i>agree are, agree further, agree need</i> Nearby answer terms: <i>agree strongly, agree with</i>
<i>can do</i> Nearby question terms: <i>are doing, work with, is taking</i> Nearby answer terms: <i>had recently, am keen, need [to] make</i>

Table 5.1: Example parliamentary question terms w , with question and answer terms whose latent representations are close to forwards-representation $\vec{\phi}(w)$.

<i>a lot</i> Forwards: Nearby counselor terms: <i>feel overwhelmed, are dealing, carrying</i> Nearby texter terms: <i>tired, sick, lost</i> Backwards: Nearby counselor terms: <i>dealing [with], an amount, understandable feeling</i> Nearby texter terms: <i>injury, loss, destroyed</i>
<i>how long</i> Forwards: Nearby counselor terms: <i>been feeling, for [some] time, ago</i> Nearby texter terms: <i>months, years, since</i> Backwards: Nearby counselor terms: <i>is causing, feeling like, [what] happened [to] make</i> Nearby texter terms: <i>horrible, depressed, trapped</i>
<i>anyone [to] talk [to]</i> Forwards: Nearby counselor terms: <i>to who, know anyone, you trust</i> Nearby texter terms: <i>talk, friends, family</i> Backwards: Nearby counselor terms: <i>be difficult, deserve support, able [to] share</i> Nearby texter terms: <i>depressed, disappoint, isolated</i>

Table 5.2: Example counselor terms w , with counselor and texter terms whose latent representations are close to forwards-representation $\vec{\phi}(w)$ or backwards-representation $\overleftarrow{\phi}(w)$.

pertaining to typical predecessors. Consider *how long* (e.g., ***How long*** have you been feeling this way?). Nearby texter terms to $\vec{\phi}(w)$ suggest that such messages (unsurprisingly) tend to be followed by responses indicating durations of time (*months*); nearby counselor terms suggest other messages that might get at similar comments on duration (e.g., have you ***been feeling this way for some time?***).

<p>Will the Minister explain why she is allowing companies that are making massive profits to be subsidized by the taxpayer?</p> <p>Nearby questions:</p> <p>Q: <i>Why does the Minister not obtain advice from all over Wales, rather than adopt such an incestuous relationship with failed Tory party candidates?</i></p> <p>Q: <i>Will they now admit that they should be doing more to help the Hutu refugees return to their country?</i></p> <p>Nearby answers:</p> <p>A: <i>We clearly cannot appoint a Regulator until the necessary legislation has passed through Parliament.</i></p> <p>A: <i>The simple reason is the existing problem we have with pollution, which has nothing to do with any future decision about the expansion of Heathrow.</i></p>
<p>Does the Minister agree that the best way to tackle low pay in Scotland is to create more job opportunities?</p> <p>Nearby questions:</p> <p>Q: <i>Does my hon Friend agree that it is by forging links with such groups that we will achieve our goals in Iraq?</i></p> <p>Q: <i>Does my right hon Friend agree that that is because under the previous Government school funding was allocated on the basis of party politics?</i></p> <p>Nearby answers:</p> <p>A: <i>I agree wholly with my hon Friend. It is important for jobs in industry[...]</i></p> <p>A: <i>Yes, I agree with a great deal of what my hon Friend has said. That is why we have such a large motorway programme.</i></p>
<p>What can the Government do to get solar energy on to the big roofs of warehouses?</p> <p>Nearby questions:</p> <p>Q: <i>What can the Government do alongside international partners to try to protect the humanitarian space?</i></p> <p>Q: <i>What can my right hon Friend do to end this school place lottery and get more good school places in my constituency?</i></p> <p>Nearby answers:</p> <p>A: <i>I will indeed meet the hon Gentleman to talk about that case.</i></p> <p>A: <i>My hon Friend raises a serious matter, and I shall refer it to the Attorney General for full consideration.</i></p>

Table 5.3: Example parliamentary questions a , with questions and answers whose latent representations are close to forwards-representation $\vec{\Phi}(a)$.

<p>You've had a lot on your plate and it's perfectly normal to feel down.</p> <p>Forwards:</p> <p>Nearby counselor messages:</p> <p><i>It's totally normal to feel overwhelmed when you're dealing with so much all at once.</i></p> <p><i>It makes sense to feel overwhelmed with so many difficult things at once.</i></p> <p>Nearby texter messages:</p> <p><i>I just get so tired of having to be strong.</i></p> <p><i>I've had it for so long and it makes life hard to go on.</i></p> <p>Backwards:</p> <p>Nearby counselor messages:</p> <p><i>It's understandable to be feeling overwhelmed with all these things at once.</i></p> <p><i>It makes sense that you're feeling stressed with all that's going on.</i></p> <p>Nearby texter messages:</p> <p><i>I feel like I'm losing my mind.</i></p> <p><i>I'm losing the strength to fight.</i></p>
<p>How long have you been feeling this way?</p> <p>Forwards:</p> <p>Nearby counselor messages:</p> <p><i>I'm wondering how long you've been feeling sad and depressed?</i></p> <p><i>How long have you felt like you can't make him happy?</i></p> <p>Nearby texter messages:</p> <p><i>I started feeling like this a few months ago.</i></p> <p><i>I've had thoughts like this for years.</i></p> <p>Backwards:</p> <p>Nearby counselor messages:</p> <p><i>Did something change today to make you feel this way?</i></p> <p><i>It's normal to feel confused or unsure in this situation.</i></p> <p>Nearby texter messages:</p> <p><i>I'm very depressed and stressed.</i></p> <p><i>My anxiety has been really bad lately.</i></p>
<p>I'm wondering if there's anyone else you can talk to.</p> <p>Forwards:</p> <p>Nearby counselor messages:</p> <p><i>Is there someone supportive you can talk to?</i></p> <p><i>Do you have anyone else in your life who you trust?</i></p> <p>Nearby texter messages:</p> <p><i>I get help from a therapist.</i></p> <p><i>A few of my best friends understand what's going on.</i></p> <p>Backwards:</p> <p>Nearby counselor messages:</p> <p><i>It must be difficult that your mom doesn't understand your situation.</i></p> <p><i>Have you told him how you need some support right now?</i></p> <p>Nearby texter messages:</p> <p><i>My boyfriend knows I'm depressed but we never talk about it.</i></p> <p><i>My brother calls me a liar all the time.</i></p>

Table 5.4: Example counselor messages a , with counselor and texter messages whose latent representations are close to forwards-representation $\vec{\Phi}(a)$ or backwards-representation $\overleftarrow{\Phi}(a)$.

Nearby texter terms to $\overleftarrow{\phi}(w)$ are different: the preceding messages seem to consist of texters talking about the state they're in (*horrible*). Accordingly, nearby counselor terms suggest other ways the counselors might respond to or reflect on such disclosures (e.g., *is something in particular **causing** this?*).

We can likewise inspect nearby representations of utterances and context-utterances, shown in Tables 5.3 and 5.4.² Crucially, we note that these replies and predecessors are different from the *actual* replies or predecessors surrounding the utterance in an interaction—rather, they represent context-utterances we'd *expect*, given our model of the utterance. We draw this point to examine the contrast between expected and actual contexts in Section 5.5.

5.2 Exploring latent representations

We now more broadly explore the latent representations derived via the framework, beyond the individual examples we've just discussed. To structure our exploration, we identify and analyze a few salient regions in the latent vector space Γ . In particular, using the K-Means algorithm, we cluster the forwards-representations of utterances in each dataset, $\overrightarrow{\Phi}(a)$, identifying k regions in which a set of utterance representations are close together; we likewise cluster the backwards-representations of utterances $\overleftarrow{\Phi}(a)$. We also assign terms and contexts to these inferred clusters based on their latent representations in Γ . We interpret each region as delineating a *type* of utterance or term, and refer to *forwards types* and *backwards types* to distinguish between typologies derived from clustering forwards- or backwards-representations, respectively. To ex-

²For privacy, we produced the examples we cite in this chapter from the counseling dataset as follows: we inspected examples from the data, and then wrote fictional pairs based on these examples and on examples found in the counselor training curriculum.

plore the derived representations, we qualitatively interpret these regions, and then quantitatively compare various term and utterance properties across them.

Note that this is essentially the procedure we described in Chapter 2 to derive a typology of questions in parliamentary discourse—there, we interpreted the resultant regions as different rhetorical roles, or “things that an asker is getting at.” While our aim in that work was to produce the typology, here our focus is broader: we use the types as a way of making sense of the latent space that the framework derives.³

5.2.1 Characterizing answers in parliamentary discourse

Thus far, in the parliamentary setting, we have derived-forwards representations of questions, $\vec{\Phi}$, given their expected answers. The 8 types of questions we’ve identified are outlined in Chapter 2.6.1. We now explore a natural extension: representing *answers* in terms of the expected or typical questions that prompt them. In particular, we take *questions* asked by MPs to comprise the conversational context, derive a latent space on the basis of the set of questions, and then derive backwards-representations of terms w and answers a , denoted as $\overleftarrow{\phi}(w)$ and $\overleftarrow{\Phi}(a)$, respectively. In this way, we structure the ways in which government ministers respond to the questions posed to them.

In manually inspecting the dataset, we find that the language used in answers tends to be less structured than that of the questions, potentially making answers more difficult to computationally model. Indeed, past work [Sacks, 1989a, Schiffrin, 1994] has suggested that while questions often contain distin-

³Note that focusing on these inferred typologies means that our analyses may overlook groups of terms and utterances that our particular application of the clustering algorithm fails to identify. In other words, this analysis approach enables us to systematically take a broad, but non-exhaustive view of the vector space.

guishing linguistic constructions, the primary commonality shared by answers might simply be that they follow questions. Accordingly, we'd find fewer linguistic regularities in answers than in questions—as is the case in our data. With this consideration in mind, we apply our framework; our particular methodological choices are listed in the appendix (Section A.1).

Types of answers. We identify five backwards-types of answers (choosing this number via manual inspection). Below, we name and interpret each type; in Table 5.5 we show example terms and utterances, and include further examples in the appendix (Table B.2).

0. Progress report. Reporting on, and generally committing to continue, a particular course of action.

1. Statement. Stating facts, often as a rebuttal to a question. Note that in addition to these prototypical rebuttal-like examples, this type also seems to encompass a much broader range of answers that is difficult to interpret.

2. Endorsement. Positively responding to, and voicing support for, the points raised in a question.

3. Comment. Acknowledging, commenting, and perhaps speculating on the points raised in a question. Note that this type is somewhat hard to interpret; we suggest that it captures lexical patterns indicative of polite, hedging responses to relatively tough but not outright hostile questions.

4. Commitment. Stating that the minister recognizes the importance of, and is seeing to, a matter raised in a question.

Compared to the question representations derived in Chapter 2, the answer representations cluster into less interpretable and evenly-sized groups. In particular, the **statement** type comprises almost 40% of answers, and seems to en-

<p>Progress report (4.2% answers, 18.4% terms) Answer terms: <i>looking at, will consider</i> Question terms: <i>report, will [you] follow</i> Q: Will the Home Secretary follow the advice of the head of counter-terrorism? A: We are looking at the best way to implement it [...]</p>
<p>Statement (39.9% answers, 14.9% terms) Answer terms: <i>is clear, have said</i> Question terms: <i>why do, is not</i> Q: Is not the truth that job losses are on the increase? A: It is clear that we are [actually] in balance [...]</p>
<p>Endorsement (12.5% answers, 14.5% terms) Answer terms: <i>am glad, support</i> Question terms: <i>agree is, will [you] welcome</i> Q: Does my hon. Friend agree that raising awareness is very important in preventing child abduction? A: I am extremely glad to join my hon Friend in paying tribute to the work of organisations [dealing with that matter].</p>
<p>Comment (17.6% answers, 20.1% terms) Answer terms: <i>am (not) sure, think</i> Question terms: <i>is important, not agree</i> Q: Does the Attorney-General not agree that [we] would be better served if the right for individuals to seek prosecutions were preserved rather than handed over? A: I think that [we] would be best served through proper and targeted work by the police and prosecutors [...]</p>
<p>Commitment (25.8% answers, 21.0% terms) Answer terms: <i>committed to, take seriously</i> Question terms: <i>will recognise, will ensure</i> Q: Will the Minister ensure that the newly unemployed receive the maximum possible assistance? A: Given the present downturn, I take [this point] seriously [...]</p>

Table 5.5: Representative examples of answer and question terms, and question-answer pairs, for each **answer** type inferred from the parliamentary question periods data.

compass a broad range of responses that do not follow the linguistic constructions typical to the other answer types identified. This suggests the following high-level picture: while some answers reflect recurring rhetorical tropes (e.g., **commitments**), a substantial proportion are linguistically more ad-hoc.

Relation to the preceding question. We expect that the representations we've derived of a question and the answer it receives will correspond, given the intuitive dependency of answers on questions, and given that our framework aims to model this dependency. Figure 5.1 shows positive pointwise mutual information statistics between the types of questions and answers in each exchange.⁴ Indeed, we find several unsurprising correspondences; for instance, **endorsements** are highly likely to follow **agreement** questions and unlikely to follow **demands for account**, relative to random chance. At the same time, we note that the type of a question does not entirely determine the answer it receives; for instance, while **commitments** are less likely to follow **demands for account**, 21% of such questions are still followed by such answers (e.g., a minister might address a criticism raised in a hostile question, rather than outright refute it).

Relation to party affiliation. We also explore the relation between the political affiliation of question-askers and the nature of the answers they receive.⁵ As in Chapter 2.6.3, to quantify the relative extent by which a particular answer type t_a is given to a government-affiliated, versus an opposition-affiliated asker, we compute the log-odds ratio of type t_a answers given to government versus opposition MPs. Figure 5.2 shows the resultant log-odds ratios for each answer type. We see that **commitments** and **endorsements** are significantly

⁴We use pointwise mutual information statistics to adjust for the relative frequencies of question and answer types. For the remainder of the chapter, the order in which answer types appear in figures corresponds to the relative extent to which a type responds to a question from a government vs. an opposition MP, as computed in this section

⁵In question periods, while the affiliation of question-askers varies, answers are almost always provided by government ministers. As such, we do not compare affiliations of answerers.

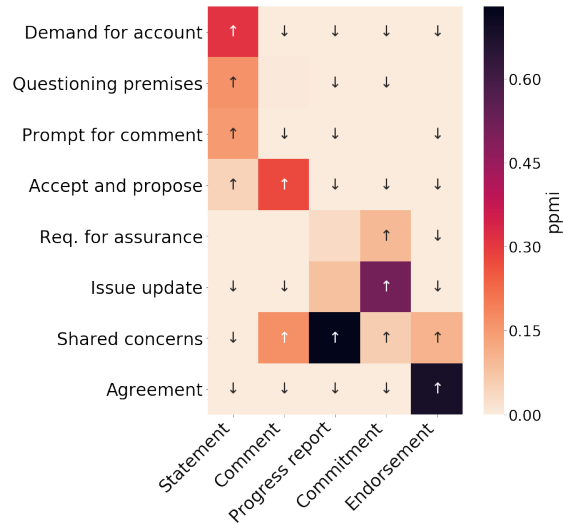


Figure 5.1: Positive pointwise mutual information statistics between question types (rows) and answer types (columns). Darker red denotes that an answer of that type is more likely to follow a question of that type, relative to random chance. \uparrow and \downarrow indicate significant differences in each direction (Fisher’s $p < 0.05$, Bonferroni-corrected in the number of possible type pairs).

more likely to serve as responses to government MPs (Fisher’s exact $p < 10^{-4}$ for each, Bonferroni-corrected in the number of answer types). Interestingly, we see that despite its somewhat nebulous nature, the **statement** answer type is significantly more likely to respond to opposition MPs ($p < 10^{-4}$). A possible explanation is that ministers tend to rely on rhetorical tropes—as encompassed by the other answer types—in responding to friendlier questions; when responding to more difficult or hostile questions from opposition MPs, they might depart from such linguistic regularities.

Since we’ve shown that the nature of an answer is related to its preceding question, we repeat this analysis per question type: for each question type t_q , we compute the log-odds ratio of type t_a answers given to government vs. opposition MPs, among the *subset* of answers to t_q type questions. We see that the

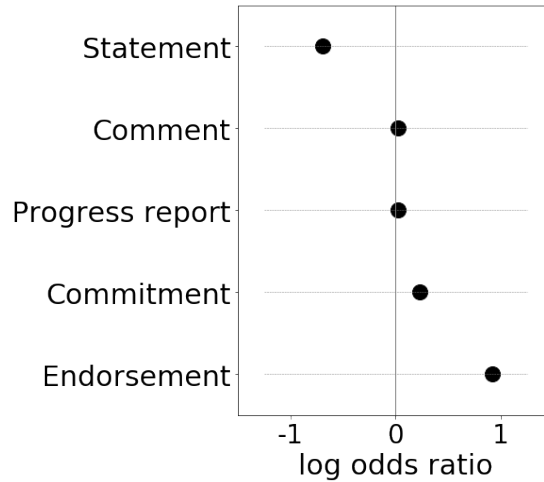


Figure 5.2: Log-odds ratios of answers of each type given to askers who are affiliated with the government, versus the opposition.

partisan differences shown in Figure 5.2 are largely replicated.⁶

Relation to labeled data. Finally, we compare the answer types to labels from the annotated dataset of Prime Minister’s Questions used in Chapter 2.6.2 [Bates et al., 2014]. In particular, question-answer pairs in that data are labeled as *answered*, *deferred* (i.e., the Prime Minister didn’t have the knowledge or capability to provide an answer) or *not answered*. We derive representations for each answer provided in this dataset, and assign these answers to our inferred types.

Correspondences between our inferred answer types and the annotated labels are visualized as pointwise mutual information statistics in Figure 5.3. We note clear correspondences between our answer types and these annotations: *answered* questions are highly associated with **endorsements** (constituting 59% of questions of that type, compared to 36% over the entire dataset, binomial test $p < 0.01$), while questions that are *not answered* are highly associated with

⁶As we later discuss in Section 5.5, this might reflect differences in how the answerer responds to the same type of question from MPs of different affiliations, or that our approach conflates different types of question asked by government vs. opposition MPs.

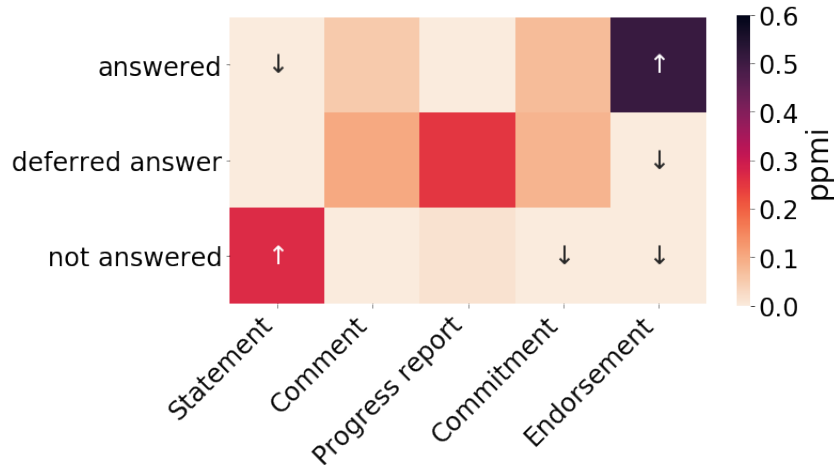


Figure 5.3: Positive pointwise mutual information statistics between answer types (columns) and annotated labels of answers (rows). Darker squares denote types and labels that are more associated with each other. \uparrow and \downarrow indicate statistical significance.

statements (33% in-type versus 25% over the entire dataset, $p < 0.01$).

These correspondences are somewhat surprising. At face value, the annotators for the dataset would have needed to use both the answer and the preceding question to determine whether or not the question was adequately answered. In contrast, our answer representations are based off of the terms contained in the answer, and do not account for the *particular* question that an answer follows. This raises a few possible interpretations. First, given the richness of the parliamentary institution, there might be routinized linguistic devices for answering or dodging questions. Second, since answers tend to reflect the nature of their preceding questions, the correspondence could simply reflect that some *questions* are more answerable than others.⁷

⁷The correspondences we observe are still statistically significant if we repeat the analysis over the subset of questions labeled as *standard* (other labels don't have enough representatives in the data to draw statistically meaningful conclusions). Per manual inspection of the labeled data, we suggest that this might reflect the diverse range of "standard" questions, i.e., stratifying our analyses by label does not properly control for question type.

5.2.2 Characterizing messages in counseling conversations

We now examine the latent representations our framework derives in the counseling conversation setting introduced in Chapter 3. Here, we continue our analyses of the counselors' messages.

In contrast to the parliamentary setting, which is largely constrained to questions and answers, counselors send messages in the context of extended conversations. As such, we can derive two representations for each message and term: in terms of expected replies, in the forwards direction (denoted $\vec{\phi}$ and $\vec{\Phi}$ for terms and messages, respectively), or in terms of expected predecessors, in the backwards direction (denoted $\overleftarrow{\phi}$ and $\overleftarrow{\Phi}$). As we proceed to show, the more actively conversational nature of this setting modulates how we interpret the representations and the inferred forwards and backwards types.

Note that we have already made use of term-level representations in this setting: in Chapter 3, we referred to $\overleftarrow{\phi}$ and $\vec{\phi}$ as "central points" and used them to compute the key measure presented in that chapter, orientation. Here, we more extensively examine these representations and their message-level counterparts. As in Chapter 3, to address heterogeneity within messages, we derive representations of *sentences* within messages, and cluster these sentence-level representations to derive forwards- and backwards-types.⁸

Counselor message types. We identify 8 forwards-types and 8 backwards-types of counselor messages. Interpretations of these types, along with example terms and sentences, are in Tables 5.6 and 5.7; further examples are provided in the appendix (Tables B.3 and B.4).

⁸While we perform the subsequent analyses sentence-by-sentence for consistency with Chapter 3, we note that the types of messages inferred from utterance-level representations are similar.

<p>Risk assessment (9.7% messages, 9.3% terms) Assessing risk of suicidal ideation or self-harm. C terms: <i>take [your] life, have [a] plan</i>; T terms: <i>pills, knife</i> C: When you say you want to end your life, do you have a plan for how you would do it? T: I have pills, but I'm so scared.</p>
<p>Service statement (13.8% messages, 13.9% terms) Telling the texter that they are here for them, often to start/end the conversation. C terms: <i>are here, talk tonight</i>; T terms: <i>thanks, bye</i> C: We are here to help: you're brave for reaching out tonight. T: Thanks, I really appreciate what you are doing.</p>
<p>Situation comment (12.7% messages, 14.2% terms) Talking about the difficult nature of the texter's situation. C terms: <i>be exhausting, feel overwhelmed</i>; T terms: <i>frustrating, everyday</i> C: It can be exhausting to deal with so much pain and sadness. T: It's frustrating to have all these problems.</p>
<p>Relationship comment (12.6% messages, 14.6% terms) Commenting on a situation involving a relationship difficulty. C terms: <i>communicate, you both</i>; T terms: <i>ignores, argues</i> C: It sounds like he didn't communicate this in an appropriate way. T: He argues with me every other day.</p>
<p>Coping mechanism (13.4% messages, 13.1% terms) Prompting the texter to think of ways to cope with their situation. C terms: <i>relieve stress, help relax</i>; T terms: <i>music, exercise</i> C: Can you think of things that help you relieve stress? T: Sometimes I listen to music.</p>
<p>Support system (12.3% messages, 9.9% terms) Prompting the texter to talk about support systems they could look to. C terms: <i>could reach, shared with</i>; T terms: <i>counselors, friend</i> C: Is there a friend you could reach out to instead? T: I have a close friend who doesn't ignore me...</p>
<p>Exploration (5.9% messages, 9.4% terms) Prompting the texter to say more about a situation. C terms: <i>share more, tell me</i>; T terms: <i>accused, attacked</i> C: Can you share more about how they reacted? T: They accused me of not trying hard enough.</p>
<p>Suggestion (19.6% messages, 15.5% terms) Offering to give the texter suggestions, and assurances of finding support. C terms: <i>at least, could help</i>; T terms: <i>suggestion, assistance</i> C: You might find it useful, and at least it's a good start. T: I guess, maybe I'll check out that suggestion.</p>

Table 5.6: Examples of counselor (C) and texter (T) terms, and message-reply pairs, for each inferred **forwards** type.

<p>Coping mechanism (13.2% messages, 15.3% terms) Commenting on or discussing coping mechanisms with the texter. C terms: <i>great way, activity</i>; T terms: <i>paint, cooking</i> T: I love to paint. C: Painting is a great way to keep your mind focused on something else.</p>
<p>Situation comment (13.9% messages, 10.8% terms) Commenting on the difficult nature of a texter’s situation. C terms: <i>incredibly difficult, can imagine</i>; T terms: <i>recently, suffered</i> T: My friend died recently and I feel horrible. C: That sounds incredibly difficult for anyone to endure.</p>
<p>Social comment (10.0% messages, 11.2% terms) Commenting on a situation involving social difficulties. Counselor terms: <i>closest to, no one</i>; T terms: <i>drama, friends</i> T: All of our friends pay more attention to her. C: It sounds isolating to be ignored by the people closest to you.</p>
<p>Feeling comment (14.9% messages, 10.3% terms) Responding to and prompting the texter to elaborate on feelings they’ve expressed. C terms: <i>how long, understandable [to] feel</i>; T terms: <i>worthless, hate</i> T: I feel worthless, and sad all the time. C: How long have you felt like that?</p>
<p>Suggestions (11.2% messages, 16.5% terms) Discussing potential sources of support with the texter. C terms: <i>can send, other options</i>; T terms: <i>services, local</i> T: I’m not sure what services would be available for me. C: I can send you some pointers to help you find some support.</p>
<p>Relationship comment (11.0% messages, 13.4% terms) Commenting on relationship difficulties (to contrast with social comments). C terms: <i>from someone, loves</i>; T terms: <i>cheating, blames</i> T: He blames me for messing us up. C: It must be heartbreaking to hear that from someone you care about.</p>
<p>Service statement (8.4% messages, 10.2% terms) Telling the texter about the service. C terms: <i>i am, are here</i>; T terms: <i>service, bot</i> T: Thanks, it was my first time using the service and I wasn’t sure what to expect. C: I am glad you texted in tonight.</p>
<p>Appreciation for disclosure (17.4% messages, 12.3% terms) Thanking the texter for sharing information, often for particularly difficult disclosures. C terms: <i>brave, willing [to] share</i>; T terms: <i>dying, cutting</i> T: I’ve been thinking about dying a lot lately. C: It was really brave of you to share that with me.</p>

Table 5.7: Examples of counselor (C) and texter (T) terms, and message-predecessor pairs, for each inferred **backwards** type.

We outline some high-level observations that we proceed to explore more systematically. First, many of these types seem to correspond to counseling strategies identified in the counselors’ training materials, and considered in

Chapter 3.6.1. Second, many of these types appear to reflect how typical counseling conversations are structured, as examined in Chapter 3.6.2. Finally, several of the forwards and backwards types seem to relate to each other. Henceforth, when we refer to a type that's surfaced in both the forwards and backwards direction, e.g., **coping mechanism**, we do not disambiguate between whether we are discussing the forwards or backwards type, unless the distinction is relevant.

Relation to labeled counseling strategies. Figure 5.4 visualizes correspondences between the forwards and backwards types, and the annotated strategy labels used in Chapter 3.6.1, depicting positive pointwise mutual information statistics between labels and types.⁹ Indeed, we see that many of the inferred types reflect these labeled strategies. For instance, we find that sentences labeled as *exploration* or *risk assessment* have forwards-representations that are clustered together; in both directions, we note similar correspondences between the *options* label and the **coping mechanisms** type.

Note that our latent representations also reflect finer distinctions: in both directions, they distinguish between **coping mechanism** and **suggestions** sentences among those labeled as *options*; in the backwards direction, sentences annotated as *reflection* are further delineated into the types of content being reflected on (e.g., **situation**, **feeling**). We note that these geometric distinctions are important in the domain: for instance, while counselors are encouraged to help texters brainstorm **coping mechanisms**, they are cautioned against immediately offering explicit and overly directive **suggestions** that cut short the problem-solving process. We note that the set of labels considered could be expanded to

⁹For the remainder of the chapter, the order in which forwards and backwards types are displayed in figures correspond to how early they occur in the conversation, in median % of elapsed messages, and as visualized in Figure 5.5.

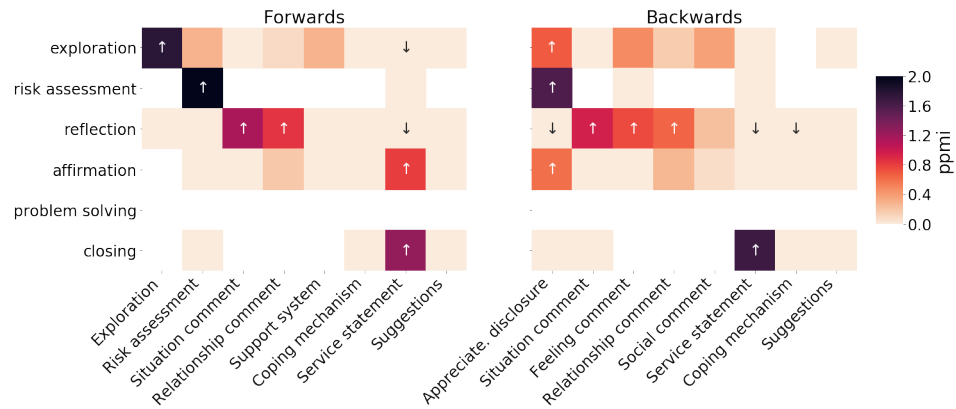


Figure 5.4: Positive pointwise mutual information statistics between annotated labels and (left) forwards types or (right) backwards types of sentences written by counselors. Darker red denotes types and labels that are more associated with each other. ↑ and ↓ indicate statistical significance; white squares indicate the type and label did not co-occur in the data.

account for these distinctions. That said, this analysis illustrates that these distinctions can be derived using our framework, from particular patterns in the data based on the conversational context—suggesting ways in which domain knowledge and inductive analyses of conversations can inform each other.

Relation to conversation structure. We next relate the forwards- and backwards-representations to the structure of counseling conversations. Following the approach in Chapter 3.6.2, we divide each conversation with more than ten counselor messages into five equally-sized segments. For each type, we consider the messages containing a sentence of that type, and examine the proportion of such messages in each segment. The distribution of messages across segments is visualized in Figure 5.5.¹⁰

These distributions reflect the typical progression of counseling conversa-

¹⁰Here, we do not double-count multiple of sentences of the same type within a message. While we use the segment-based approach for visual clarity in the figures, we note an alternative that arrives at similar conclusions: consider the percentage of the conversation elapsed at each message, and take the distribution of these percentages per type.

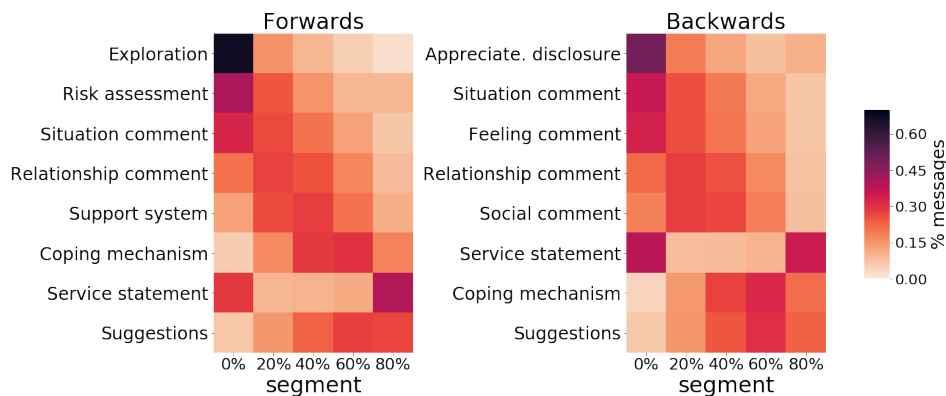


Figure 5.5: Proportion of sentences occurring in each segment of the conversation for (left) forwards types and (right) backwards types. Darker squares indicate segments of the conversation where each type occurs more often.

tions, as taught during training, in intuitive ways: explorations of the texter’s situation occur earlier in the conversation (indicated by dark squares towards the left of the figure for types such as **exploration** and **situation comment**), while talk of solutions occurs later on (dark squares towards the right of the figure for types such as **coping mechanism** and **suggestions**). Interestingly, we see that **relationship comments** and sentences pertaining to **social concerns** tend to occur slightly later in the conversation, compared to comments on **situations** and **feelings**. Manually inspecting examples of these types suggest a progression from broad discussion of feelings to more specific exploration of the texters’ social situations. For both forwards and backwards types, we see that **service statements** are concentrated both at the starts and at the ends of conversations, perhaps reflecting that at either end of the conversation, the counselor establishes and then reiterates what the service can offer for the texter.

In Althoff et al. [2016], which analyzes the same setting as us, the authors identify five stages via an unsupervised Markov model approach that explicitly models cohesive, temporally-ordered stages: *introductions*, *problem introduction*,

problem exploration, problem solving, and wrap up. Our framework yields types that reflect a similar progression, even though we do not directly account for high-level conversation structure. Here, we suggest that by modeling messages in terms of their replies and predecessors, we recover some low-level coherence (i.e., consecutive messages are likely focused on the same thing) that, at least in this setting, reflects how the conversation is organized more broadly. Of course, in contrast to the Markov model approach, our approach cannot explicitly model the high-level structure.

We also note contrasts between our low-level perspective and the stage-based model. We make finer distinctions within particular stages of the conversation; for instance, we identify **risk assessments** as a distinct (forwards) type from **exploration** statements. We also allow for different types to occur in different orders in a conversation. For instance, **suggestions**, on average, occur slightly later in a conversation than **coping mechanism** (shown in Figure 5.5 as **suggestions** occurring more in the last segment, in both directions). However, in around 20% of conversations, **coping mechanism** messages occur after **suggestions** (for both forwards and backwards types), corresponding to cases where counselors (continue to) discuss coping mechanisms after they've raised a particular suggestion.

Relation between forwards- and backwards-representations. We've noted that several of the forwards and backwards types identified seem to correspond to each other. Figure 5.6 visualizes these correspondences, depicting positive pointwise mutual information statistics between forwards and backwards types per sentence. Indeed, we see that several forwards types have backwards counterparts, shown as dark squares corresponding to types like **relationship comment**, **coping mechanism** and **service statement**.

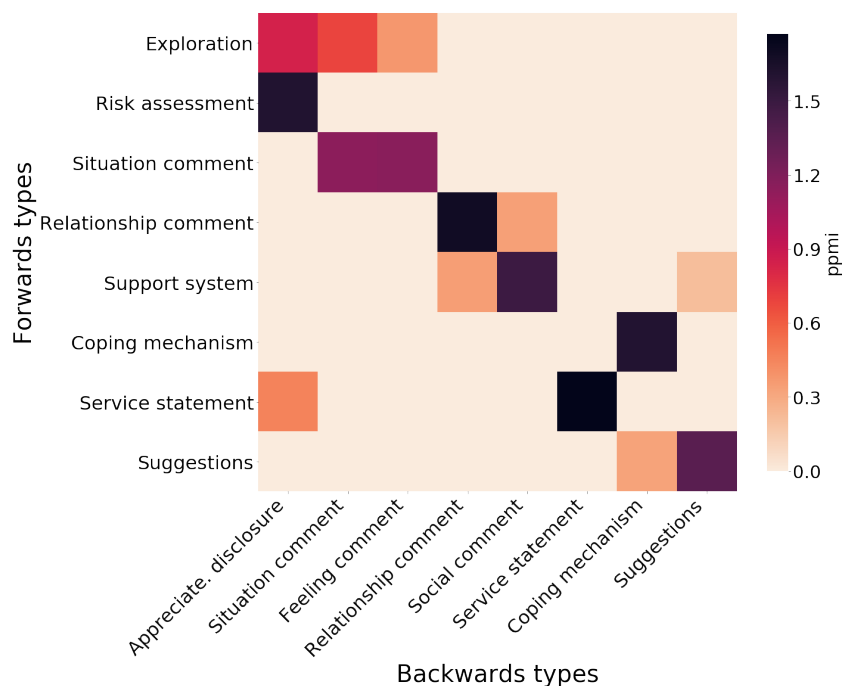


Figure 5.6: Positive pointwise mutual information statistics between forwards types (rows) and backwards types (columns) for sentences in counselor messages. Darker squares indicate that sentences are more likely than chance to be of a particular forwards type and backwards type.

We also note cases where distinctions between sentences are made under one choice of context but not the other. Among sentences of the **exploration** forwards type, we identify **appreciation for disclosure**, **situation comment** and **feeling comment** backward types. This suggests that while we expect texter responses to be relatively similar among **exploration** sentences (modeled as forwards-representations that are close together), counselors make systematic distinctions in responding to what texters say (corresponding to backwards-representations that cluster into distinct neighbourhoods). Likewise, among sentences of the **appreciation for disclosure** (backwards) type, we find **risk assessment** and **exploration** forwards types. We suggest that while counselors respond to a wide range of difficult disclosures in consistent ways (e.g., thanking

them for their bravery and honesty), they prompt the texter to different types of discussions depending on whether the texter expressed suicidal ideation (i.e., exploring a situation vs. engaging in a focused risk assessment).

Finally, we draw a high level contrast between the counseling and question periods settings. We interpreted the parliamentary question types as different rhetorical intentions: where does an asker aims to direct the subsequent answer? In the counseling setting, interpreting the forwards types as rhetorical prompts may be somewhat problematic. For some of the types (e.g., **situation/relationship comments**), the sentences assigned to that type seem related because they are responding to similar things that we expect the texter to continue discussing in their next turn. In other words, since these sentences arise in the *middle*, rather than at the start of an interaction, we conflate rhetorical intentions with conversational continuity. In Chapter 3 and in subsequent analyses, motivated by this observation, we explore ways of combining forwards and backwards characterizations.

5.2.3 Other choices of conversational context: next turn

Beyond replies and predecessors, we note that the framework can be applied with other choices of conversational context. For instance, in an extended conversation, we could use the framework to model a speaker’s utterance in terms of what the *same* speaker says in their next turn, skipping over intervening replies from other participants.

We explore this variant in the counseling conversation setting. Following the framework, we use *counselor* messages as conversational context, and derive a context vector space Γ_c . We derive ϕ' , a mapping of counselor terms w into

Γ_c , by modifying the derivation of forwards-representations: rather than using correspondences between terms and texter replies, we use correspondences between terms and the counselor’s next utterances, skipping over the texter’s reply. Analogous to the derivation of $\vec{\Phi}$, we then aggregate ϕ' over terms to derive representations of messages Φ' . We refer to the resultant latent representations in Γ_c as *skip-representations*; as such, two counselor messages a_i and a_j have similar skip-representations if we expect that the next message the counselor sends after a_i will be similar to the next counselor message after a_j . We include further methodological details in the appendix.

To explore the derived representations, we infer six types of messages, which we refer to as *skip types*. As above, we examine and cluster per-sentence representations. We outline and provide examples for these types in Table 5.8.

We offer a few explanations for why two terms or sentences could have similar skip-representations, resulting in the inferred types. First, counselors could be following certain scripts, such that different terms that occur at one step of the script lead to similar next steps. For instance, in the **hello** type, we find different ways that counselors greet texters (e.g., asking for their name, establishing that they’re here to help), which all lead to next turns in which the counselor asks an initial exploratory question (*what has happened recently?*). Additionally, in this setting, counselors are taught procedures to systematically risk-assess texters for suicidal ideation; accordingly, in the **risk assessment** type, we find language that addresses hints that the texter might be suicidal, and that lead to next turns in which the counselor tries to glean more concrete details.

In general, however, these conversations aren’t intended to be script-based: counselors are taught to be reactive to texters, rather than to prefigure their next turn. As such, another interpretation is that counselors likely talk about similar

<p>Hello statement (9.3% messages, 2.5% terms) Greeting the texter; establishing that the counselor is here to talk with them. C terms: <i>[I']m here, texting in</i>; C' terms: <i>you mentioned, happen recently</i> C: I'm here to listen to you. C': Did anything happen recently to make you feel that way?</p>
<p>Situation comment (38.4% messages, 39.6% terms) Commenting on the texter's situation. C terms: <i>can see, deal with</i>; C' terms: <i>decision, reasons</i> C: I can see that this is really troubling you. C': It sounds like a really difficult decision to have to make.</p>
<p>Wrap up (10.0% messages, 10.2% terms) Messages directed towards wrapping up the conversation. C terms: <i>[I']m glad, remember</i>; C' terms: <i>of course, stronger</i> C: I'm glad you are feeling at least a little better. C': I know you have a lot to think about and figure out, but you're stronger than you realize.</p>
<p>Risk assessment (6.7% messages, 6.8% terms) Performing a risk assessment for suicidal ideation. C terms: <i>your life, appreciate sharing</i>; C' terms: <i>plans, the medication</i> C: I'm wondering if you have thoughts of ending your life. C': I appreciate your honesty, have you thought of plans to do this?</p>
<p>Problem solving (18.8% messages, 30.4% terms) Discussing coping mechanisms as well as recommendations for other forms of support. C terms: <i>you think, what do</i>; C' terms: <i>is great, a solution</i> C: Do you think it would be helpful to put on some music you like? C': That is a great idea.</p>
<p>Exploration (16.7% messages, 10.6% terms) Messages sent in the course of exploring a texter's situation. C terms: <i>feeling overwhelmed, depressed</i>; C' terms: <i>afraid, a burden</i> C: It sounds like you're feeling overwhelmed right now. C': It makes sense that you are worried about being a burden.</p>

Table 5.8: Representative examples of counselor terms and terms in subsequent counselor messages (C and C', respectively), and pairs of successive counselor messages, for each **skip** type inferred from the counseling conversation data.

things from turn to turn—messages on **problem solving** tend to be followed by similar messages, since the problem solving process usually extends over multiple turns. Within a particular focus, however, the conversation could still evolve: for instance, in examining **exploration** messages, we suggest that counselors might try to move from a general comment on a situation (e.g., *feeling overwhelmed*) to a more specific detail in their next turn (e.g., *being a burden*). In Section 5.3.3, we quantitatively explore these various interpretations.

5.3 Exploring derived properties

As we described in Chapter 4, beyond yielding latent representations, the framework also outputs other characterizations that reflect other aspects of terms and utterances. In Chapter 3, we examined one such property, orientation. Here, we suggest a few additional properties and examine them in the parliament and counseling settings.

5.3.1 Expectation strengths

Recall that the forwards range of a term or utterance ($\vec{\sigma}$ and $\vec{\Sigma}$) quantifies the strength of our expectations of its potential replies, while the backwards range ($\overleftarrow{\sigma}$ and $\overleftarrow{\Sigma}$) quantifies the strengths of our expectations of potential predecessors. Here, we further explore and elaborate on these characteristics.

Forwards-ranges in parliamentary question periods. We explore the forwards-ranges $\vec{\Sigma}$ of parliamentary questions. Here, we revisit and elaborate on the intuition that higher and lower $\vec{\Sigma}$ should correspond to more or less open-ended questions. At the extreme, if we ask a leading question, we aim to prompt a

specific answer, corresponding to a lower $\vec{\Sigma}$. As in the preceding section, we structure our analyses around the inferred question types, and examine how the statistic varies across different types.

Figure 5.7 shows the distribution of $\vec{\Sigma}$ among questions in each inferred type.¹¹ In inspecting these distributions, along with prototypical examples per type, we suggest the following interpretation: questions with lower $\vec{\Sigma}$ point to somewhat formulaic answers (e.g., “Will the minister *meet* with me?” “I would *be delighted* [...]”). At the extreme, **agreement** questions tend to uniformly prompt statements of agreement. In contrast, questions with higher $\vec{\Sigma}$ prompt the answerer for responses that aren’t simply rhetorical—either the asker wants some new information (as is the case with **issue update** questions), or is **demanding an account** for a specific failure. Such questions therefore point to less routinized answers.

The question and answer terms we use are designed to omit more topical words. As such, it makes sense that $\vec{\Sigma}$ in this context captures the degree to which we expect rhetorically uniform answers. We note that there are other, perhaps more intuitive ways for questions to be open versus closed-ended. Concretely, we see that with **accept and propose** questions, the asker often wants to engage the asker to speculate on a proposed hypothetical; as such, the *content* of the answer might be open-ended. However, such questions have relatively low $\vec{\Sigma}$, perhaps because they prompt fairly regular lexical constructions (e.g., *I am certain*), even if the answerer uses these constructions to frame a more open-ended response. Future work could more carefully examine and operationalize

¹¹Throughout this section, when we relate each characterization to an utterance type, we consider the 50% of utterances in each type whose latent representations are closest to the corresponding cluster centroid for that type. We perform this subsetting to facilitate interpretation; the results we discuss are generally qualitatively similar if we consider all utterances, but come with larger error bars, since we include utterances that are less representative of each type.

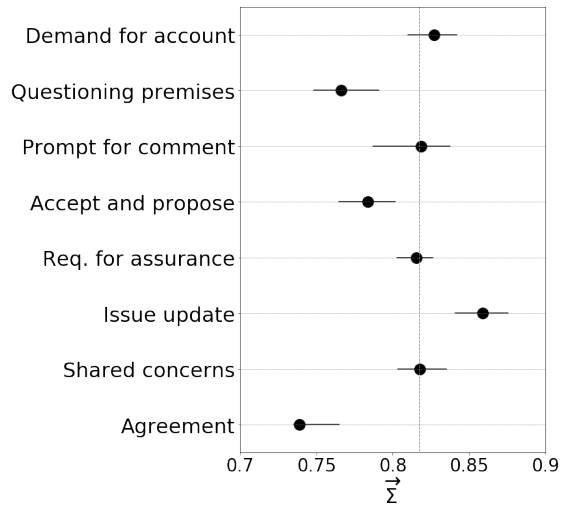


Figure 5.7: Distributions of $\vec{\Sigma}$ of questions for each parliamentary question type. Points correspond to median values, while error bars denote bottom and top quartiles. Median $\vec{\Sigma}$ over all questions is shown as the dotted line.

the relation between form and content.

Forwards- and backwards-ranges in counseling conversations. In Chapter 3, we combined forwards- and backwards-ranges into a single measure, orientation. Here, we briefly elaborate on the relation between these two measures.

To start, we note that these measures are highly correlated (Spearman's $\rho = 0.59$ at the term level and $\rho = 0.52$ at the sentence level). This perhaps reflects that conversations tend to exhibit some continuity from one turn to the next: for instance, in a relatively focused phase of an interaction, utterances that prompt well-defined responses (low forwards-range) are likely responding to well-defined predecessors (low backwards-range) as well.

Figure 5.8 visualizes the distribution of $\overleftarrow{\Sigma}$ and $\overrightarrow{\Sigma}$ per forwards and backwards type: each heatmap shows the proportion of sentences of that type with different $\overleftarrow{\Sigma}$ and $\overrightarrow{\Sigma}$, binned into tertiles. Heatmaps with darker cells in the top left and bottom right correspond to types that have low or high orientations,

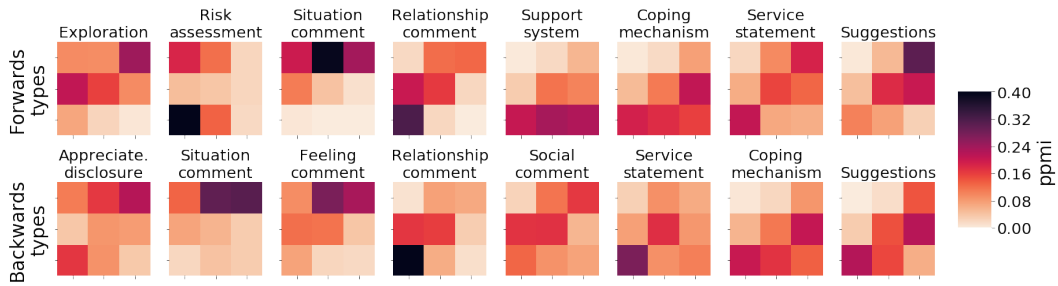


Figure 5.8: Distributions of $\vec{\Sigma}$ and $\overleftarrow{\Sigma}$ for counselor sentences per (top) forwards and (bottom) backwards type. Rows correspond to $\vec{\Sigma}$, binned into tertiles; columns correspond to $\overleftarrow{\Sigma}$. Bottom left corner corresponds to utterances with low forwards- and backwards-range; upper right corner corresponds to utterances with high forwards- and backwards-range.

respectively (e.g., **situation** messages have lower orientation relative to **coping mechanisms** messages). Heatmaps with darker cells in the bottom left correspond to types with low $\overleftarrow{\Sigma}$ and $\vec{\Sigma}$, as is the case for **relationship comments**; one interpretation is that such messages occur in stretches of the conversation where the counselor and texter are discussing relatively well-defined problems. Heatmaps with darker cells in the top right corner correspond to types with high $\overleftarrow{\Sigma}$ and $\vec{\Sigma}$; we find that such messages (corresponding to types such as **exploration** and **situation comment**) tend to occur in more exploratory parts of the interaction.¹²

5.3.2 Characterizing expected shifts

In addition to comparing forwards- and backwards-ranges, we can also compare the derived latent representations. Given a term w , if its forwards and back-

¹²We note that these qualitative groupings are not mutually exclusive. For instance, **relationship comments** tend to be low in both measures, and to also have low orientation; i.e., a counselor might focus on someone’s relationship problems over multiple turns, without raising forwards-oriented prompts for specific information.

wards representations $\overleftarrow{\phi}(w)$ and $\overrightarrow{\phi}(w)$ are far, then we expect its typical replies to differ from its typical predecessors: that is, we'd interpret such terms as those occurring at points where a conversation shifts focus. Conversely, if $\overleftarrow{\phi}(w)$ and $\overrightarrow{\phi}(w)$ are nearby, we'd expect replies and predecessors associated with w to be similar, and infer that the term occurs during parts of conversations where the focus is relatively stable across successive turns. We can similarly interpret comparisons made between utterance-level representations $\overleftarrow{\Phi}$ and $\overrightarrow{\Phi}$.

To formalize this idea, we propose a measure, *shift*, as the cosine distance between forwards- and backwards-representations of terms or of utterances; we denote term- and utterance-level shifts as δ and Δ , respectively. We expect terms and utterances with high values of δ and Δ to shift the focus of the conversation to a greater degree than those with lower values.

Table 5.9 shows examples of counselor terms w with low and high δ , and texter terms whose representations are close to $\overleftarrow{\phi}(w)$ or $\overrightarrow{\phi}(w)$. For high δ terms, we note clear contrasts in the two sets of texter terms displayed. For instance, *been coping* seems to prompt talk of coping mechanisms, but is preceded by mentions of a bad situation the texter is currently in.

We explore Δ across messages in counseling conversations by comparing its distribution across the inferred forwards and backwards types, visualized in Figure 5.9. We find that sentences aimed at exploring a situation (e.g., **exploration, situation/feeling comments**) tend to have higher Δ , perhaps reflecting how the counselor moves the discussion across different aspects of a texter's problem. Sentences on **risk assessment** and **coping mechanisms** tend to have lower Δ , suggesting that these are topics of discussion that counselors might try to sustain (i.e., until they glean adequate information about potential suicidal ideation, or until the texter has come up with a way to cope).

<p>Low δ</p> <p>your relationship (It seems like you're struggling with your relationship.) Nearby texter terms to $\overleftarrow{\phi}$ and $\overrightarrow{\phi}$: <i>he, arguing, cheated</i></p> <p>music (Listening to music is an excellent coping mechanism.) Nearby texter terms to $\overleftarrow{\phi}$ and $\overrightarrow{\phi}$: <i>draw, relaxing, listen</i></p> <p>high school (It sounds like your high school is causing a lot of stress for you.) Nearby texter terms to $\overleftarrow{\phi}$ and $\overrightarrow{\phi}$: <i>senior, classes, grades</i></p>
<p>High δ</p> <p>been coping (How have you been coping with all of this?) Nearby texter terms to $\overleftarrow{\phi}$: <i>dumped, cancer, recently</i> Nearby texter terms to $\overrightarrow{\phi}$: <i>usually, distractions, exercising</i></p> <p>told anyone (Have you told anyone about how you've been feeling?) Nearby texter terms to $\overleftarrow{\phi}$: <i>depressed, horrible, flashbacks</i> Nearby texter terms to $\overrightarrow{\phi}$: <i>talk, mum, friends</i></p> <p>you hope (What do you hope to gain from this conversation?) Nearby texter terms to $\overleftarrow{\phi}$: <i>ignoring, ruin, wants</i> Nearby texter terms to $\overrightarrow{\phi}$: <i>reassurance, continue, advice</i></p>

Table 5.9: Examples of counselor terms with low and high δ (in the bottom and top 25%). For examples w with low δ , we show texter terms whose representations are close to $\overleftarrow{\phi}(w)$ and $\overrightarrow{\phi}(w)$. For w with high δ , we show examples of texter terms which are close to $\overleftarrow{\phi}(w)$, contrasting them with examples that are close to $\overrightarrow{\phi}(w)$.

We also note that the shift measure makes some interesting distinctions within types. For instance, in the forwards **coping mechanisms** type, terms like *art, movie* and *music* have particularly low δ (in the bottom 5% of δ among terms); terms like *been coping* and *felt [has] helped* have particularly high δ (in the top 5%). While the low-shift terms point to discussions where the texter elaborates on specific coping mechanisms, the high-shift terms point to sentences where the counselor explicitly prompts the texter to *start* thinking of coping mechanisms, marking a break from the previous discussion.¹³

¹³Indeed, the high-shift terms are assigned to the **situation** and **feeling** backwards types.

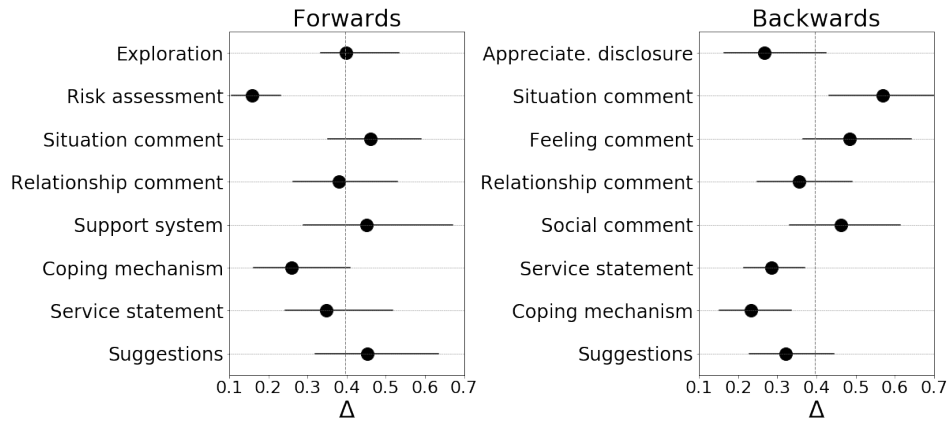


Figure 5.9: Distributions of Δ for sentences of each (left) forward type and (right) backward type in the counseling conversation setting, shown as box plots indicating quartiles. Median Δ over all sentences is shown as the dotted lines in each plot.

The relation between the various measures we've discussed would be interesting to explore in future work. Here, we make some high-level observations. We find that shift is correlated with forwards- and backwards-ranges (Spearman's $\rho = 0.49$ and 0.60 for forwards- and backwards-ranges at the term level, respectively; $\rho = 0.52$ and 0.57 at the sentence level). One interpretation follows from the intuition that conversations exhibit some continuity: having a strong expectation of what's being replied to or prompted (low $\overleftarrow{\Sigma}$ and $\overrightarrow{\Sigma}$) might suggest that the interaction is focused on a particular idea, which we do not expect will change (low Δ). Conversely a weaker expectation in either direction might correspond to moments in a conversation when the counselor is trying to shift focus to something new that is less predetermined by the existing conversation (high Δ). On the other hand, shift is only weakly, if at all, correlated with orientation ($\rho = 0.17$ and 0.00 at the term and sentence levels, respectively). Here, we point to comments that might move a conversation from one idea to the next (high Δ), but without widening or narrowing the range of possible things to discuss (Ω near zero): consider parts of a conversation that explore different

aspects of a **situation** or **feeling**.

5.3.3 Shifts to next turn

In Section 5.2.3, we applied the framework with the counselor’s next turn as the conversational context, deriving skip-representations Φ' of counselor sentences. There, we raised the question of whether our expectations of the counselor’s next turn reflect them advancing through well-defined scripts, or continuing to discuss similar things, or progressing in more nuanced ways. Here, we quantitatively explore these possibilities.

In particular, note that in the process of computing skip-representations, we actually derive two representations of a counselor’s sentence in the same latent space Γ_c : one in which the sentence plays the role of conversational context, and that’s derived from the LSA step used to induce Γ_c ;¹⁴ the other in which it is then mapped into Γ_c as Φ' . By comparing the skip-representation with the LSA-representation, we can quantify the extent to which the counselor’s next turn is expected to differ from their present turn.

Formally, we define the *skip-shift* Δ' of a sentence as the cosine distance between its skip-representation and LSA-representation. Figure 5.10 depicts the distribution of Δ' across the six inferred skip types. We see that the differences between types largely corroborates our intuition: **hello** sentences follow well-defined routines that quickly shift focus from greeting the texter to starting to explore their situation, reflected in high Δ' ; to a lesser degree, while **risk assessments** focus on suicidal ideation, they also follow consistent progressions, reflected in middling values of Δ' . We may infer that discussions of **situations**

¹⁴Technically, the LSA step yields latent representations of messages; we derive sentence-level representations following the procedure outlined in Section 5.4.

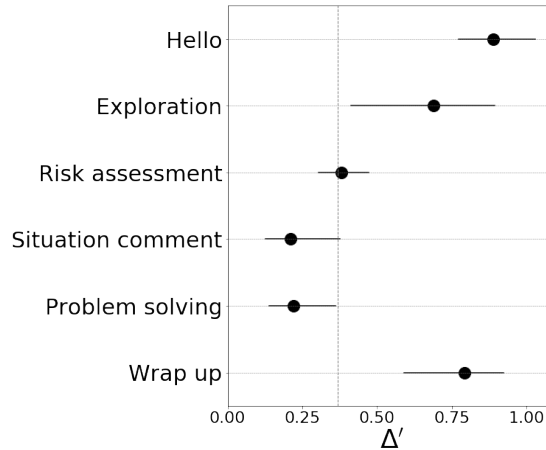


Figure 5.10: Distributions of Δ' for sentences of each skip type in the counseling conversation setting, shown as box plots. Median Δ' over all sentences is shown as the dotted line.

or **problem solving** are more static, as reflected in low Δ' : in the process of problem solving with a texter, we'd expect the counselor to spend an extended portion of the conversation talking to them about various options.

5.4 Utterance- vs. conversation-based representations

What do we gain from accounting for conversational context? Here, we revisit the parallel between our framework and distributional methods for representing words (Chapter 4.3). Per Firth [1957], “You shall know a word by the company it keeps.” Approaches that draw on this idea (such as LSA and word2vec) aim to derive representations of terms given the other surrounding terms in a document (or an utterance, in a conversational setting). Our framework aims to derive representations of terms given the surrounding *utterances* in a conversation. As such, we contrast the two notions of context that these approaches operationalize: utterance-based context, and conversational context.

Concretely, we compare the forwards- and backwards-representations from our framework with representations derived via LSA, demonstrating on the parliament and counseling datasets. While future work could make contrasts to other approaches, we focus on LSA because it's directly comparable to our approach, which uses it as an intermediate step. Another way to frame the following analysis is as follows: instead of first applying LSA to replies and predecessors to derive a latent space Γ , and then mapping terms and utterances into Γ , why not directly use LSA to derive these term and utterance representations?

Deriving LSA-based representations. Our method for computing LSA-based representations parallels the approach for computing context-utterance and context-term representations detailed in Chapter 4.4: we construct a tf-idf reweighted matrix A and use singular value decomposition to approximate it as a product of lower-dimensional factors $A \approx \bar{U}\bar{s}\bar{V}^T$, each with the same number of dimensions as our forwards- and backwards-representations. Rows of \bar{V} contain the resultant LSA representations of terms w , which we denote $\bar{\phi}(w)$. To compute the representation of an utterance a , we take its tf-idf reweighted vector representation a and compute $\bar{\Phi}(a) = a\bar{V}\bar{s}^{-1}$.

Comparing term representations. To compare the approaches based on utterance-level and on conversational context, we formulate a way to measure the differences between the representations they derive for each term. We describe how we compare forwards- and LSA-based representations; the approach for backwards-representations is analogous.

Concretely, for each term w , we wish to compare its representations $\vec{\phi}(w)$ and $\bar{\phi}(w)$. Note these representations aren't directly comparable— $\vec{\phi}(w)$ is defined in a vector space derived from the set of context-utterances (e.g., texters' messages, ministers' answers), while $\bar{\phi}(w)$ is derived from a potentially differ-

ent set of utterances (e.g., counselors’ messages, MPs’ questions). As such, the geometric distance between $\vec{\phi}(w)$ and $\bar{\phi}(w)$ is not well-defined.

Instead, we start from the following intuition: if $\vec{\phi}(w)$ and $\bar{\phi}(w)$ are similar, then other terms w' that have $\vec{\phi}(w')$ close to $\vec{\phi}(w)$ —i.e., are *similar to w under $\vec{\phi}$* , should be similar to w under $\bar{\phi}$ as well. In particular, consider the *neighbourhood* of the k terms closest to w under $\vec{\phi}$, which we denote $\mathcal{N}(w; \vec{\phi})$. If $\bar{\phi}(w)$ is similar to $\vec{\phi}(w)$, then this neighbourhood should be *retained* rather than *dispersed*: $\bar{\phi}(w)$ and $\bar{\phi}(w')$ should be close for $w' \in \mathcal{N}(w; \vec{\phi})$ and far for $w' \notin \mathcal{N}(w; \vec{\phi})$. By identifying terms whose neighbourhoods are retained versus dispersed from $\vec{\phi}$ to $\bar{\phi}$, we arrive at a comparison of these representation approaches.

Formally, we define a measure, *dispersion*, between $\vec{\phi}$ and $\bar{\phi}$. To compute the dispersion of a term w , which we denote $d(w)$,¹⁵ we start by identifying the k nearest terms to w under $\vec{\phi}$, $\mathcal{N}(w; \vec{\phi})$. Next, we compute the cosine distances under the *other* representation, between $\bar{\phi}(w)$ and $\bar{\phi}(w')$ for all other w' in our vocabulary. We take the percentile rank of each w' per this distance. Finally we compute $d(w)$ as the median percentile rank of each $w' \in \mathcal{N}(w; \vec{\phi})$.¹⁶

If $d(w)$ is low, then terms close to w under $\vec{\phi}$ are also relatively close to w under $\bar{\phi}$, so $\bar{\phi}$ retains the neighbourhood around w . At the extreme, suppose that the k words in $\mathcal{N}(w; \vec{\phi})$ are also the k closest words under $\bar{\phi}$. Then, letting N denote the vocabulary size, we have $d(w) = (k/2)/N$. If $d(w)$ is high, then $\bar{\phi}$ disperses the neighbourhood. As another reference point, if under $\bar{\phi}$ distances between $\vec{\phi}(w)$ and $\vec{\phi}(w')$ are randomly permuted, then $\mathbb{E}[d(w)] = 1/2$.¹⁷

¹⁵Technically we’d specify which representations the measure is comparing, but we omit this for notational clarity.

¹⁶Note that dispersion is asymmetric between $\vec{\phi}$ and $\bar{\phi}$. Results when computing the measure in the other direction suggest similar interpretations to the ones we subsequently provide, and we do not report them here.

¹⁷We also considered an alternate measure that directly compares the sets of k nearest words under $\vec{\phi}$ and $\bar{\phi}$, via Jaccard similarity. Various properties of this measure make interpretation

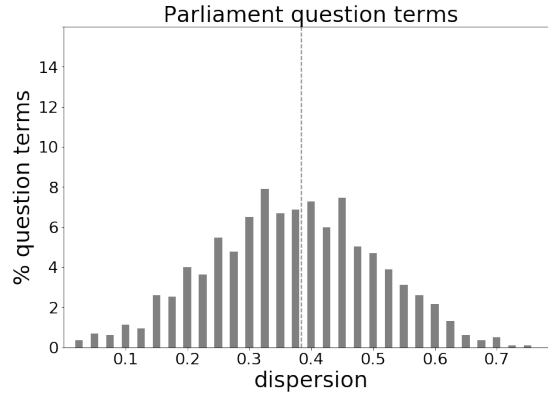


Figure 5.11: Histogram of d for each question term in the parliament setting. Dotted grey line indicates median d over all terms.

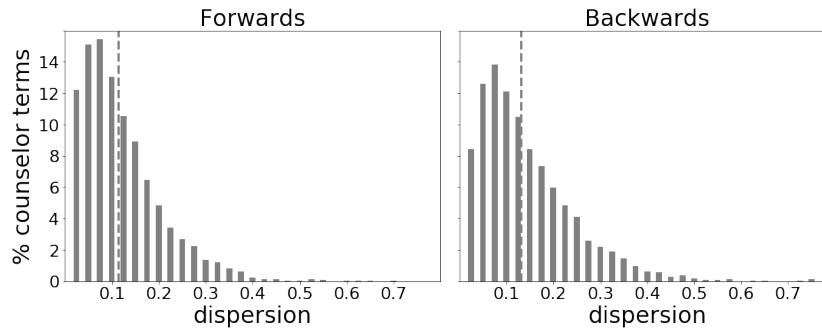


Figure 5.12: Histogram of d for each counselor term, comparing LSA representations with (left) forwards- and (right) backwards-representations of terms. Dotted grey lines indicates median d over all terms, for each comparison.

Analysis of terms. We compute dispersions for parliamentary question terms and counselor terms. In the parliamentary setting we compare $\vec{\phi}$ with $\bar{\phi}$; in the counseling setting we make comparisons with both $\vec{\phi}$ and $\overleftarrow{\phi}$. In each comparison, we take k , the size of the neighbourhood, to be 5% of the total vocabulary size (such that the lower bound for dispersion is 0.025).

In Figures 5.11 and 5.12 we show distributions of d for each representation we compare to $\bar{\phi}$. Comparing with the aforementioned extremal reference points, we see that most terms exhibit some degree of dispersion, though not more difficult: many terms have equivalent dispersions; a large portion have a dispersion of 0.

at the level of random permutation. We also note that dispersions are higher in the parliament than in the counseling setting.

To further interpret these results, we inspect examples of terms with low and high dispersions for each comparison, listed in Tables 5.10, 5.11 and 5.12. We suggest the following high-level intuition: terms with low dispersion tend to be those that occur in relatively formulaic utterances and in routinized parts of the interaction. In such cases, the low value of the measure would reflect that what's talked about within an utterance is highly related to what's subsequently talked about in a reply, or what's being responded to from a predecessor. We see this intuition reflected in the low-dispersion examples (e.g., *agree* is in the parliamentary setting; *considered talking* and *a website* in the counseling setting, which correspond to standard actions asking about support systems, or offering to share a helpful resource). More broadly, we speculate that routines and formulaic utterances—reflected in lower dispersions—are more prevalent in the counseling setting than the parliament setting, given the fact that counselors receive the same training that might result in relatively standardized language. In the counseling setting, we also see examples that tend to occur in parts of the conversation focused on a particular topic (e.g., *dad*, *relaxing*). We suggest this is because what's discussed in one turn is similar to what's discussed in surrounding terms; as such, each of the representations reflects similar information.

Inspecting high-dispersion examples offers suggestions about what additional information is captured when accounting for conversational, beyond utterance-level context. We see that our framework more clearly reflects the structure of a conversation, drawing analogies between terms on the basis of the moments in a conversation when they'd appear. For instance, in the parliamentary setting, though *will visit* and *will come* might be semantically close,

<p>Low d</p> <p>agree is Nearby terms to $\vec{\phi}$ and $\bar{\phi}$: <i>agree are, agree be, agree have</i></p> <p>can tell Nearby terms to $\vec{\phi}$ and $\bar{\phi}$: <i>will tell, could tell, please tell</i></p> <p>consider Nearby terms to $\vec{\phi}$ and $\bar{\phi}$: <i>will consider, consider making, also consider</i></p>
<p>High d</p> <p>is taking Nearby terms to $\vec{\phi}$: <i>doing ensure, work with, being done</i> Nearby terms to $\bar{\phi}$: <i>is given, is what, is there</i></p> <p>will visit Nearby terms to $\vec{\phi}$: <i>draw to, consider is, does know</i> Nearby terms to $\bar{\phi}$: <i>will come, come to, come clean</i></p> <p>happened to Nearby terms to $\vec{\phi}$: <i>take seriously, tell why, does regret</i> Nearby terms to $\bar{\phi}$: <i>whatever, what has, got</i></p>

Table 5.10: Examples of parliamentary question terms with low and high d (in the bottom and top 25%), comparing **forwards**- and LSA representations. For w with low d , we show other question terms whose representations are close to $\vec{\phi}(w)$ and $\bar{\phi}(w)$. For w with high d , we show examples of terms which are close to $\vec{\phi}(w)$, contrasting with examples close to $\bar{\phi}(w)$.

in practice, *will [the Minister] visit* is used to voice **shared concerns**, as is *[may I] draw to [your attention]*; in contrast, *will come* is more likely used in a more aggressive demand for the minister to *come clean* about a wrongdoing. In the counseling setting (comparing backwards-representations), while *inspired* and *perseverance* are both used to affirm the texter’s strength, *inspired* tends to be used to more specifically praise the texter for taking steps *in [the right] direction*.

Analysis of utterances. To see how the contrasts between representation methods play out across a dataset, we contrast the representations of *utterances* de-

<p>Low d</p> <p>what helps Nearby terms to $\vec{\phi}$ and $\bar{\phi}$: <i>distract yourself, some things, helps feel</i></p> <p>considered talking Nearby terms to $\vec{\phi}$ and $\bar{\phi}$: <i>to counselor, a therapist, a professional</i></p> <p>dad Nearby terms to $\vec{\phi}$ and $\bar{\phi}$: <i>your mother, your father, parents</i></p>
<p>High d</p> <p>what say Nearby terms to $\vec{\phi}$: <i>explaining, reaction, believe you</i> Nearby terms to $\bar{\phi}$: <i>they say, im sorry, say that</i></p> <p>been coping Nearby terms to $\vec{\phi}$: <i>with feelings, what helped, in past</i> Nearby terms to $\bar{\phi}$: <i>been handling, stress of, the loss</i></p> <p>open with Nearby terms to $\vec{\phi}$: <i>have plan, kill yourself, how end</i> Nearby terms to $\bar{\phi}$: <i>opening up, really appreciate, sharing this</i></p>

Table 5.11: Examples of counselor terms with low and high d (in the bottom and top 25%), comparing **forwards**- and LSA representations. For w with low d , we show other counselor terms whose representations are close to both $\vec{\phi}(w)$ and $\bar{\phi}(w)$. For w with high d , we show examples of terms that are close to $\vec{\phi}(w)$, contrasting with examples close to $\bar{\phi}(w)$.

rived via our framework, $\vec{\Phi}$ and $\overleftarrow{\Phi}$, and $\bar{\Phi}$, derived via LSA. All of these approaches derive utterance representations that somehow reflect the characteristics of the terms that the utterances contain; $\bar{\Phi}$ stems from characterizations of terms that are based on within-utterance context, while $\vec{\Phi}$ and $\overleftarrow{\Phi}$ reflect term characterizations based on conversational context. As before, we detail our comparison approach for $\vec{\Phi}$ and note that the approach for $\overleftarrow{\Phi}$ is analogous.

We adapt the term-level dispersion measure to define an utterance-level dispersion $d(a)$ for each utterance a , between $\vec{\Phi}$ and $\bar{\Phi}$: we take the k nearest utter-

<p>Low d</p> <p><i>relaxing</i> Nearby terms to $\overleftarrow{\phi}$ and $\bar{\phi}$: <i>a book, walking, activity</i></p> <p><i>listen to</i> Nearby terms to $\overleftarrow{\phi}$ and $\bar{\phi}$: <i>am here, about anything, whatever</i></p> <p><i>a website</i> Nearby terms to $\overleftarrow{\phi}$ and $\bar{\phi}$: <i>can find, a resource, links</i></p>
<p>High d</p> <p><i>inspired</i> Nearby terms to $\overleftarrow{\phi}$: <i>in direction, track, working on</i> Nearby terms to $\bar{\phi}$: <i>dealing with, such situation, perseverance</i></p> <p><i>feeling is</i> Nearby terms to $\overleftarrow{\phi}$: <i>someone love, helpless, are coping</i> Nearby terms to $\bar{\phi}$: <i>feeling upset, feeling frustrated, feeling sad</i></p> <p><i>anxious about</i> Nearby terms to $\overleftarrow{\phi}$: <i>nervous, test, college</i> Nearby terms to $\bar{\phi}$: <i>frustrated, depressed, isolated</i></p>

Table 5.12: Examples of counselor terms with low and high d (in the bottom and top 25%), comparing **backwards**- and LSA representations. For w with low d , we show other counselor terms whose representations are close to both $\overleftarrow{\phi}(w)$ and $\bar{\phi}(w)$. For w with high d , we show examples of terms that are close to $\overleftarrow{\phi}(w)$, contrasting with examples close to $\bar{\phi}(w)$.

ances to a under $\vec{\Phi}$, $\mathcal{N}(a; \vec{\Phi})$, and compute the median percentile rank of each $a' \in \mathcal{N}(a; \vec{\Phi})$, given distances between the LSA-based representations $\vec{\Phi}(a)$ and $\vec{\Phi}(a')$. To match our other analyses, in the counseling setting, we compute dispersions for each *sentence* within a counselor’s message.

Note that computing distances between every pair of utterances in our datasets would be too computationally intensive. As such, we consider a sampled version of the measure: we take a random subset of utterances, S . To compute $d(a)$ we take the k nearest neighbours of a in S and only consider ut-

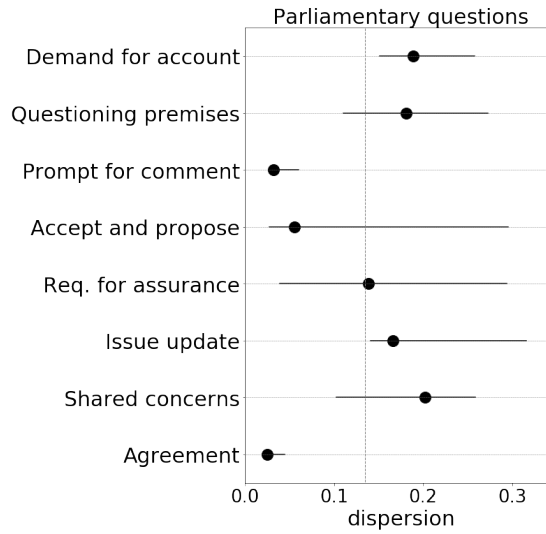


Figure 5.13: Distributions of d for parliamentary questions of different types, shown as box plots. Median d over all questions is indicated by the dotted line.

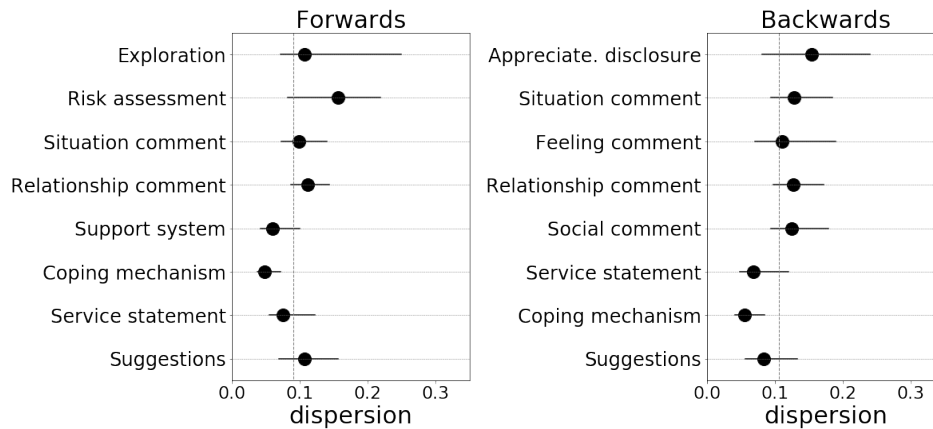


Figure 5.14: Distributions of d for counselor sentences of different (left) forwards and (right) backwards types, shown as box plots. Median d over all sentences is indicated by the dotted lines.

terances in S when computing distances under $\bar{\Phi}$ and percentile ranks. For both the parliament and counseling settings, we take $|S|$, the size of the subset, to be 40,000, and k , the size of the neighbourhood, to be 2,000, or 5% of $|S|$.¹⁸

¹⁸We find that the computed measures are quite consistent from sample to sample, and qualitatively similar under minor changes to these parameters.

To structure the subsequent analyses, we compare dispersions across utterances of different forwards and backwards types. Figure 5.13 visualizes the distribution over each question type in the parliament setting; Figure 5.14 visualizes these distributions per forwards and backwards type in the counseling setting. We find that our term-level intuitions largely play out: utterances that reflect routine interactional procedures as well as formulaic language, such as **agreement** and **prompt for comment** questions in the parliament setting, tend to have smaller dispersions. In the counseling setting, the utterances with the lowest dispersions likewise suggest routine ways to perform actions like asking about **coping mechanisms** or **support systems**; indeed, counselors are explicitly taught these strategies when they're trained. In contrast, when exploring aspects of the texter's situation (e.g., **exploration, relationship comments**), counselors might exhibit more linguistic flexibility in reacting to particular disclosures; additionally, these explorations may be less reflective of preexisting routines. As such, these types of messages tend to have higher dispersion.

Counterintuitively, in the counseling setting, **risk assessment** statements have high dispersion, even though risk assessing for suicidal ideation is a very well-defined procedure. In inspecting examples of corresponding terms with high dispersion, we suggest that while the particular questions that counselors ask might follow a routine, they also say things that are less bound to the procedure for gauging suicidal intent, and are instead aimed at empathetically responding to the wide range of harrowing disclosures that come up in this routine, e.g., *Thanks for being **open with me**.*

5.5 Comparing expected and actual replies

We now turn to the following question: what is the relationship between the reply we expect and the reply we actually get? In any interaction, interlocutors might dodge or misconstrue what was previously said, potentially signaling or leading to conflict. We point to a range of scenarios where divergences between expected and actual replies could be sociologically meaningful. Receiving an unexpected answer to a question could relate to failures to hold people to account via question-asking, as suggested in studies of question dodging or deferral in interviews [Rogers and Norton, 2011] and legislative proceedings [Bates et al., 2014]. In a counseling setting, unexpected replies could signal a reluctance to self-disclose following a counselor’s question, or a refusal to move to another part of the conversation per the counselor’s suggestion. Such conversational troubles are highlighted as particular sources of difficulty in the training manual used by the crisis counseling service we examined.

In this section, we use the Expected Conversational Context Framework to conduct exploratory analyses of expected versus unexpected replies. In particular, we suggest a method to quantify the extent to which a reply is expected given its predecessor. In general, such a method requires a way to model what’s expected given an utterance, and a way to compare the actual reply with our expectations. Note that our framework offers approaches to address both criteria. Recall that the framework derives representations of utterances and replies in a shared context space Γ . The forwards-representation of an utterance a , $\vec{\Phi}(a)$, models our expectations of what reply a will get, in the sense that $\vec{\Phi}(a)$ is close to latent representations $\Phi(r)$ of replies r that are more expected, and far from representations $\Phi(r)$ of r that are unexpected.

Formally, we quantify the *unexpectedness* of a reply r given an utterance a as the cosine distance between $\vec{\Phi}(a)$ and $\Phi(r)$; we denote this measure as $\mathcal{U}(r; a)$. Larger $\mathcal{U}(r; a)$ mean that r is more unexpected, and smaller $\mathcal{U}(r; a)$ mean that r is more expected.¹⁹

Application to data. We apply this method to the parliament and counseling settings. In the parliament setting, we compare MPs’ questions with the corresponding answers provided by ministers. We also contrast unexpectedness in replies to questions from government versus opposition MPs. We additionally compare the measure with the labels of answer types provided in the annotated dataset examined throughout the preceding analyses [Bates et al., 2014].

In the counseling setting, we compare counselors’ messages to texters’ replies, and relate unexpectedness to the conversation’s progress. As with the preceding analyses, we take a sentence-by-sentence approach. For each sentence s in a counselor’s message a that receives reply r , we compute $\mathcal{U}(r; s)$. We consider the sentence with the minimum \mathcal{U} to be what the texter is most likely replying to, and take this minimum value as the utterance’s unexpectedness. To mitigate noise, we restrict our analyses to counselor sentences with at least 5 counselor terms, and texter messages with at least 10 texter terms. Note that our method’s inability to model extremely short utterances constrains our analyses in a systematic way: we cannot account for extremely terse responses from the texter, which could also be seen as unexpected.

¹⁹For future work, we note that an equivalent idea could be formulated in the backwards direction: comparing backwards representations and preceding utterances could be used to examine misunderstandings, or to formalize the idea of “[answering] the question you wished had been asked of you” (Robert McNamara, quoted in Weissman [2012]). Indeed, this is a dodging strategy explored in Rogers and Norton.

5.5.1 Properties of the unexpectedness measure

Tables 5.13 and 5.14 list examples of utterance-reply pairs with low or high unexpectedness (bottom and top 25%) in the parliamentary and counseling settings. The examples we include are selected as follows: we randomly sample 20 low-unexpectedness pairs and 20 high-unexpectedness pairs; among these examples, we then select particular pairs to highlight in the discussion. In the remainder of the section, we will more substantively discuss these examples.

Comparison to answer labels. In the parliamentary setting, we compare the method’s output to labels provided in Bates et al. [2014] on the nature of answers provided by Prime Ministers during questions period. In Section 5.2.1 we related these labels with backwards types of answers; here, the unexpectedness measure explicitly relates answers to their antecedent questions.

Concretely, we compare the unexpectedness of question-answer pairs labeled as *answered* with question-answer pairs labeled as either *unanswered* or *deferred*—henceforth, we collectively refer to the latter two types as *unanswered*.²⁰ If our measure meaningfully models the extent to which a reply is expected, then the unexpectedness of *answered* pairs should be smaller than the unexpectedness of *unanswered* pairs, in which the point the asker was getting at was somehow evaded. Indeed, we see that this is the case: unexpectedness for *answered* pairs are statistically significantly smaller (Mann Whitney U test $p < 0.05$, Cohen’s delta $d = 0.16$).

Comparison to other approaches. The assumption underlying our approach for modeling unexpectedness is that a reply r to an utterance a is expected if similar utterances from the data, containing similar terms to a , received replies

²⁰We group these types together due to data size; none of the methods we consider in this section are able to distinguish between unanswered and deferred pairs.

<p>Low unexpectedness</p> <p>[PL1] Q: What can the Minister do to make sure there are no further unnecessary repossessions? A: We are working with the Ministry of Justice to improve provision of advice at the courts.</p> <p>[PL2] Q: What is the Minister doing to make sure that young people have valuable activities all year round? A: We are supporting local authorities through programmes such as the Centre for Youth Impact.</p> <p>[PL3] Q: What can [the PM] do to alleviate the difficulties on the train line? A: I visited the station recently, and I hope that Network Rail can sort something out.</p> <p>[PL4] Q: Does my hon. Friend agree that it is only by cooperating with our European partners that we can tackle organized crime? A: I agree very much that this is an important issue for joint work.</p> <p>[PL5] Q: Does the PM agree that this is exactly the kind of business-led course that the nation needs? A: I absolutely agree, and commend the college for the steps it is taking to work with businesses.</p>
<p>High unexpectedness</p> <p>[PH1] Q: What can the government do to ensure that steel is included in such contracts? A: [Purchasing steel] is an important movement in the right place.</p> <p>[PH2] Q: What can Ministers do to better protect parents of [these students]? A: I am afraid to say that this is about the hon. Gentleman yet again putting more barriers in the way of that school improving.</p> <p>[PH3] Q: What can the Minister to do improve the quality of management at the Post Office? A: We have appointed a new finance director [...] I am pleased to announce [other] new personnel [who will] strengthen the management.</p> <p>[PH4] Q: Does she agree that the new fire station is a splendid example of a station that will serve the people? A: I was pleased to be part of that wonderful community event . It is a fantastic new facility...</p> <p>[PH5] Q: Does the Minister agree that the action taken by the Government on teacher training is inadequate? A: It will not surprise you to hear that I disagree [...] we have invested a great deal specifically in this.</p>

Table 5.13: Examples of question-answer pairs in the parliament setting with low or high unexpectedness.

<p>Low unexpectedness</p> <p>[CL1] C: Do you have a plan for how you would do it? T: Yes, I will buy a gun.</p> <p>[CL2] C: Do you have a specific plan for this? T: I'm going to overdose.</p> <p>[CL3] C: What are some ways you have coped in the past? T: Taking a walk, listening to music...</p> <p>[CL4] C: What are some things you have tried to make yourself feel better? T: I've done writing and coloring.</p> <p>[CL5] C: It's normal to feel overwhelmed about this. T: I'm more than overwhelmed, I'm going crazy...</p> <p>[CL6] C: It's difficult to feel like you're being judged for being sad T: It's like we're not supposed to feel anything at all.</p>
<p>High unexpectedness</p> <p>[CH1] C: Do you have a plan for how you would end your life? T: I would have done it this morning but my friend called me...</p> <p>[CH2] C: Do you know how you would kill yourself? T: That's not what I meant when I said that...</p> <p>[CH3] C: Are there things that help you relax when you're feeling this way? T: No...I feel like there is no hope</p> <p>[CH4] C: Can you think of things that make you feel a bit better? T: That's hard. My girlfriend is still mad at me.</p> <p>[CH5] C: Are there other activities you can try to help you relax? T: I go on tumblr sometimes, people there are really understanding.</p> <p>[CH6] C: That's a lot to have going on at once. T: Plus my relationship is not going well.</p> <p>[CH7] C: That can seriously be overwhelming. T: I'm so stressed I laid down for an hour instead of working</p> <p>[CH7] C: That can seriously be overwhelming. T: I'm so stressed I laid down for an hour instead of working</p> <p>[CH8] C: It can be disheartening to feel that no one appreciates your effort. T: I was suicidal a week ago, not sure if you got that message.</p>

Table 5.14: Examples of counselor message-texter reply pairs in the counseling setting with low or high unexpectedness.

that are similar to r . Here, we compare to approaches that implement a simpler assumption: r is expected given a if r shares similar terms to a .

Concretely, we define a family of measures that make such direct comparisons. We derive tf-idf representations of a and r , and take the cosine distance of these representations; we refer to the resultant measure as the *tf-idf-unexpectedness*. To address linguistic noisiness, we also consider a variant where we take cosine distance between representations of a and r derived via LSA, which we refer to as *LSA-unexpectedness*.²¹ To make these measures as comparable as possible to the unexpectedness measure derived from our framework, we derive these representations from the same data that was used by our framework, and use the same number of dimensions as our forwards-representations for the LSA-unexpectedness approach.

First, we find that these direct-comparison measures do not reflect the distinction between answered and unanswered pairs in the labeled data. The difference in LSA-unexpectedness between these two classes is not statistically significant ($p = 0.21$, $d = 0.09$). While the difference in tfidf-unexpectedness is statistically significant ($p < 0.05$, $d = 0.06$), we note some problems with the distribution of the measure, resulting in the small effect size: for 70% of the labeled pairs, the median value of the measure (i.e., the cosine distance between representations) is 1, indicating that there was no overlap between question and answer terms. This relates to a conceptual problem with the direct-comparison approach: if an utterance and its reply are linguistically different, then the assumption that expected replies have similar terms as their preceding utterance doesn't hold. In the case of the tf-idf-unexpectedness measure, the statistical

²¹To ensure that these representations are comparable, we construct a tf-idf reweighted term-document matrix encompassing utterances and replies and use singular value decomposition to decompose this larger matrix.

significance probably reflects that when utterances and replies actually share terms, this overlap is still informative.

Admittedly, the comparison between measures is somewhat disingenuous: we designed our question and answer terms in a restrictive way (as detailed in Chapter 2, we remove nouns and only take arcs from the roots of the dependency parses), in order to reflect functional rather than topical information. However, the annotators providing these labels would have also evaluated the extent to which the subject matter itself is adequately addressed. If we compare the similarity in tf-idf representations of questions and answers that use bigrams, rather than question and answer terms, we find a larger difference between unanswered and answered pairs ($p < 0.001$, $d = 0.32$). This suggests that our choice of terms may not be the most appropriate for examining unexpectedness, but also points to an ambiguity in defining what counts as an “expected” reply; we later consider the distinction between echoing topical content (as would be captured by the bigram-based measure) and picking up on the rhetorical gist of a question (as would be captured by our measure).²²

While we do not have labels to compare with in the counseling setting, we find problems with the direct-comparison approaches that are arguably more straightforward to interpret (especially since we do not take such a restrictive definition of counselor and texter terms). The median tf-idf-unexpectedness is 0.98, indicating that counselor and texter messages are fairly linguistically dissimilar (unsurprisingly so, given their vastly contrasting roles). Inspecting examples of message-reply pairs with high tf-idf- or svd-unexpectedness (in

²²In fact, the two measures aren’t correlated (Spearman’s $\rho = 0.05$). As a back of the envelope experiment, we train logistic regression classifiers to distinguish between answered vs. unanswered pairs. A classifier that combines the framework-derived and bigram-based measures attains a higher accuracy than classifiers using just one of the features; however, these differences are not statistically significant, perhaps owing to the small data size.

the top 25%) further underlines that we shouldn't assume that more expected replies share terms with their corresponding utterances. For instance, in the following exchange, while it's clear from the data that remarks about *music* tend to follow questions about *activities*, we wouldn't expect the counselor to ask the question and talk about music in the same message (in fact, doing so might be seen as overly prescriptive):

Counselor: *What are some activities that could help you get your mind off things and relax?*

Texter: *I love music [...]*

We notice a similar issue appearing in the following example; here we point to the relation between *anyone you can trust* and *my mom*:

Counselor: *Is there anyone you can trust to talk about all of this?*

Texter: *I was thinking about telling my mom [...]*

Relation to utterance type. Intuitively, unexpectedness depends on the type of utterance. Indeed, if our expectations of a reply aren't particularly strong to start out with (as would be the case in an open-ended question or remark), then we'd expect a wider range of replies.

Numerically, we see that unexpectedness is positively correlated with forwards-range, meaning that unexpectedness tends to be larger for utterance-reply pairs where we have weaker expectations of the reply (Spearman's $\rho = 0.19$ in the parliamentary setting, $\rho = 0.25$ in the counseling setting). We also see that utterances of types with smaller forwards-range $\vec{\Sigma}$ (e.g., **questioning premises** in the parliament setting, **risk assessment** and **coping mechanism** in the counseling setting) tend to have lower unexpectedness, while types with larger $\vec{\Sigma}$ (e.g., **issue update** and **demand for account** in the parliamentary set-

ting, **situation comment** in the counseling setting) have higher unexpectedness.

We suggest that the notion of unexpectedness is harder to define and model if we don't really have expectations of the reply we'll get, and return to this idea later in the discussion.

5.5.2 Analysis of parliamentary question periods

In the parliamentary setting, we examine how unexpectedness relates to the party affiliation of a question-asker: are the responses provided by a minister more or less unexpected when the asker is a government vs. an opposition MP?

As shown in Bates et al. [2014], Prime Ministers are more likely to defer on, or not answer, questions asked by an opposition MP. Building on this finding, we form a hypothesis about ministers more broadly: that the answers they provide are less likely to match the question when the asker is in the opposition party. We'd see this reflected in lower \mathcal{U} for question-answer pairs where the question is asked by a government MP versus an opposition MP. Indeed, we find lower \mathcal{U} in the pairs involving government versus in opposition askers, though the difference is very slight (Mann Whitney U $p < 0.01$, $d = 0.047$).

More substantial differences emerge when we make this comparison per question type. For each type, we take the 25% of questions of that type that are closest to the corresponding cluster centroid, to ensure that we examine government and opposition-asked questions that are rhetorically comparable. We find statistically significant differences in two types: for **agreement** and **issue update** questions, replies are less unexpected when the asker is a government versus an opposition MP ($p < 0.01$ for each type; $d = 0.16$ for **agreement** questions and $d = 0.13$ for **issue update** questions). This corroborates our hypoth-

esis in these two cases. To interpret these findings, we more closely examine question-answer pairs with low or high unexpectedness (in the bottom and top 25%, respectively) for questions of the **issue update** and **agreement** types.

Among **issue update** pairs with low unexpectedness, we see that ministers' responses includes commitments that they are seeing to a particular course of action to address the issue raised by the asker (examples PL1 and PL2 from Table 5.13). Examples with high unexpectedness reflect a variety of reasons for why the value might be large. Indeed, recall that **issue update** questions tend to have high forwards-range—i.e., our expectations of what answers they get are less well-defined to begin with. The minister might broadly acknowledge the issue without explicitly committing to anything (example PH1); they may also refute the premise of what the asker has brought up (example PH2). Here, we revisit the ambiguity we noted earlier, of whether we should consider a reply to be expected on the basis of its topical or its rhetorical nature. Arguably, in both examples, the minister has acknowledged the content of the question, though, at least given what we've seen of other **issue update** questions in the data, we might have expected them to supplement their acknowledgement with an explicit course of action.

Which view of unexpectedness is more meaningful, in the broader project of holding governments to account? A minister might check the rhetorical boxes ("we are seeing to it") without actually answering to the issue raised (example PL3). Alternatively, they might wax poetic on the issue without concretely stating what they'll do about it. It would be fruitful for future work to examine up these different possibilities; a starting point could be to consider richer ways of defining question and answer terms that include topical information.

Other examples more unambiguously point to modeling errors. This is ex-

emplified in PH3, where the Minister seems to adequately address the question by highlighting very specific courses of action—so specific, in fact, that terms like *appointed* and *announced* are either not in our vocabulary, or are not conventionally used to address **issue update** questions. Here, we highlight a key limitation: our framework derives representations from patterns in how replies *tend to follow* terms in the data; as such, these representations statistically reflect *conventions* that don't necessarily hold in a *particular* instance of a question-answer pair (a point we return to in Chapter 6). In fact, this example suggests that our measure systematically fails to give credit for MPs who go beyond routines to provide answers that are actually tailored to the particular question.

Among **agreement** pairs with low unexpectedness, we find the routinized exchanges we've already noted, of MPs and ministers collectively bolstering the government's position on a matter (examples PL4 and PL5). Among pairs with high unexpectedness, we see cases where answers express assent but without the typical language associated with answers of this type (example PH4); we also see cases where *questions* of the **agreement** type use the construction in an unusually combative way (example PH5). Here, our model again leads us astray because it only accounts for conventions: the "agreement" exchange is so routinized that an exchange that deviates from it in form but not in rhetorical function would be considered unexpected.

In short, our model seems to measure deviation from conventional question-answer patterns, rather than the extent to which an answer addresses the rhetorical point of the question. If so, the differences in the unexpectedness measure between questions asked by government or opposition MPs might reflect, in a roundabout way, the following: askers and answerers are more likely to use rhetorical tropes in standardized ways when talking within, versus across party

lines. This hypothesis would be interesting to rigorously explore in future work, though we acknowledge that it reflects a restrictive and arguably unsatisfying conceptualization of what counts as an unexpected answer.

5.5.3 Analysis of counseling conversations

We now explore how unexpectedness relates to *when* a message occurs in a counseling conversation. Given the structured nature of these interactions, messages that are somehow out of place could signal a conversational difficulty. For instance, a counselor that asks about coping strategies before the texter has adequately explained their problem might be rushing the conversation; a counselor that risk-assesses too late might have had difficulty inferring that the texter was potentially thinking of suicide. We hypothesize that these out-of-place actions are associated with more unexpected replies: for instance, a texter may not be willing to respond to a particular message that comes too abruptly.

To start, we consider the relation between unexpectedness and timing. Formally, we define the *index* of a message as the number of messages that were sent prior to it in the conversation.²³ We compare the indexes of counselor messages that receive more unexpected, or more expected replies. To ensure that these two classes of message are comparable, we perform this analysis stratified by forwards type. For each conversation and type, we take the *first* message in which that type occurs in the conversation, and ignore all future messages of that type. Among these first occurrences, we take the eighth of messages in each type that are closest to the cluster centroid corresponding to that type. Within the resultant subset of messages, we take the more unexpected set to be com-

²³An alternate analysis, which we leave to future work, would consider the clock time elapsed rather than the number of messages sent.

prised of message-reply pairs whose unexpectedness is in the top third, and the less unexpected set to be comprised of message-reply pairs whose unexpectedness is in the bottom third.²⁴

We find significant differences for the **risk assessment** type: low-unexpectedness pairs occur earlier in the conversation than high-unexpectedness pairs (mean index= 5.9 for low- \mathcal{U} and 8.2 for high- \mathcal{U} , Mann Whitney U $p < 0.01$, $d = 0.30$). As shown in Section 5.2.2, risk assessment messages tend to occur very early on in the conversation, while later risk-assessments are more unusual; as such, this finding corroborates our hypothesis above.

Next, we examine what happens after a message is replied to in an expected or unexpected way. In particular, for how much longer does the conversation continue afterwards? On the one hand, an unexpected reply could point to a reluctance on the texter's part to move on per the counselor's suggestion, resulting in a prolonged interaction. Alternatively, if the unexpected reply signals some point of tension, then the texter might cut short the conversation.

Concretely, we compare the number of subsequent messages between a message and the end of a conversation, for messages that get expected or unexpected replies. As established in the preceding analyses, we expect the significance of an unexpected reply to vary by message type; we also expect that the amount of time left in a conversation will be related to the time that's already passed. As such, we conduct this analysis in a paired fashion. Within each forwards type, we match message-reply pairs with low unexpectedness and pairs with high unexpectedness. We enforce that the indexes of each message within a matching are the same, and as before, only consider the first instance of each type in a conversation. Within each matching, we compare the number of mes-

²⁴The reported results are similar under minor modifications of these parameters.

sages left in the interaction.

We find that **coping mechanism** and **suggestions** messages with more expected replies occur closer to the end of the conversation than messages with more unexpected replies (in 57% of **coping mechanism** pairs and 67% of **suggestions** pairs, low-unexpectedness pairs are closer to the end, Wilcoxon $p < 0.01$). In contrast, **risk assessment** messages with *less* expected replies occur closer to the end the conversation (in 55% of pairs, $p < 0.01$).

To interpret these findings, we examine examples of message-reply pairs with low or high unexpectedness, for different forwards-types. Among **risk-assessment** pairs with low unexpectedness, we see the texter providing information about their ideation that the counselor asked for (Table 5.14, examples CL1 and CL2). For examples with high unexpectedness, we see indications that the counselor has somehow misread the situation, prompting the texter to clarify that they don't actually have thoughts of suicide (examples CH1 and CH2; alternatively, the texter is deflecting the question by suggesting the counselor misunderstood). Inspecting the subsequent conversation suggests that in the latter case, a texter might not think the service is going to be appropriate for them, and may therefore end the conversation earlier.

Among **coping mechanism** pairs with low unexpectedness, we see the texter describing coping mechanisms when the counselor asks about them (examples CL3 and CL4). Among pairs with high unexpectedness, we see instances of the texter declining to answer or stating that nothing works (example CH3), suggesting that the counselor might have to spend more time in the conversation building up to a point where the texter feels more ready to start problem solving. We also see cases where the texter instead revisits a point raised earlier in the conversation (example CH4), suggesting that maybe the texter feels like they

haven't completely gotten everything off their chest, and doesn't want to move on as a result. More erroneously, we find examples where the texter responds with a less common coping strategy (example CH5), underlining our method's reliance on conventional replies.

We note that this analysis does not lead to causal interpretations: responding in unexpected ways doesn't necessarily *cause* the conversation to go on for longer afterwards. Indeed, in the above examples, we note that the prolonged interaction could reflect the difficulty of the texter's situation to start out with, rather than a particular event that occurred during the conversation. We elaborate on the difficulty of drawing causal inferences about conversations in the next chapter.

We've speculated that when our expectations of a reply aren't particularly strong, indicated in high forwards-range, unexpectedness may not be such a well-defined concept. Here, we revisit this idea by examining examples of such instances. In particular we consider low- and high-unexpectedness pairs with **situation comment** messages, where the counselor reflects on the texter's situation. Inspecting the low-unexpectedness examples, we see the texter echoing a sentiment the counselor raises (examples CL5 and CL6). Inspecting the high-unexpectedness examples, we see cases where the texter raises new information (example CH6) or continues on the same topic without necessarily echoing anything the counselor said (example CH7).

Looking further backwards in a conversation could also add information that our method is missing, as in cases where a texter responds to a counselor's remark about a particular detail by recalling a different detail they mentioned earlier on, that they might feel is more important (example CH8). As such, we underline that our framework, and the unexpectedness measure we derive

from the framework, only account for a limited snapshot of the interaction. For a very well-defined question, as in the **risk assessment** and **coping mechanism** cases, the information found in a message-reply pair might already be meaningful. However, while the counselor is still exploring the texter’s problems—and as they seek to respond to and understand many different aspects of a texter’s situation—the dependencies across messages—and our inability to model them—becomes more problematic.

Finally, we draw a distinction between our statistical measure of unexpectedness and the degree to which, in a psychological sense, the counselor finds something unexpected. Of course, we have no way of directly accessing what the counselor thinks at a particular moment. However, across the high-unexpectedness examples, we can speculate on ways in which a reply might actually be *expected*. From the preceding interaction, the counselor might already anticipate that a particular texter will be reluctant to progress through a conversation; they’d hence expect “I have no coping mechanisms” to be a valid answer, even if this is not captured in the model and data. In sum, our discussion underlines that unexpectedness is multifaceted: our data-driven operationalization of the idea necessarily leaves out many nuances.²⁵

²⁵Indeed, stepping through the technical details of our approach, we provide some back-of-the-envelope intuition for why negative responses to questions about coping mechanisms tend to be modeled as unexpected. Such responses tend not to share terms with positive responses (e.g., that actually mention a coping mechanism like music or art), and might be more linguistically similar to negative responses to other types of prompt (e.g., “I have no intention of suicide”, “I have no one else to talk to”). This suggests conceptual problems in representing utterances as a *single* vector modeling its expected replies, which would be worth revisiting in future work.

<p>casual</p> <p>Terms: <i>guess, yeah, lol, oh</i></p> <p>Lol, sorry about that man. Yeah, I saw those glitches.</p>
<p>coordination</p> <p>Terms: <i>appreciate, help, let [me] know, [i]’ll try</i></p> <p>Let me know if you need help with those articles. Your suggestions are a great help, I’ll try my best.</p>
<p>procedures</p> <p>Terms: <i>restored, please remove, be deleted, was reverted</i></p> <p>Those images should be deleted. I restored the image on Commons.</p>
<p>contention</p> <p>Terms: <i>is not, why, does [that] make, understand</i></p> <p>What part of that is not a source? Is there any reason why you removed that from the infobox?</p>
<p>editing</p> <p>Terms: <i>should start, thinking about, would prefer, be added</i></p> <p>I was thinking about adding an external link here. Would you prefer making the text smaller or the box wider?</p>
<p>moderation</p> <p>Terms: <i>explain, warned, block, report</i></p> <p>I will report you for violating this policy. You should warn him about his disruptive editing.</p>

Table 5.15: Examples of terms and comments, for each comment type inferred from the Wikipedia talk page discussions data.

5.6 Application to other datasets

We provide a brief overview of how we’ve applied our framework to other settings. We note that these additional datasets raise complexities that may have been less salient in the parliament and counseling conversation settings; as such, we also highlight some further challenges for the framework.

5.6.1 Forwards representations in Wikipedia discussions

In Zhang et al. [2018], we applied the framework to infer types of comments at the starts of online discussions between Wikipedia editors. Such discussions concern various matters surrounding the editing process of Wikipedia articles.

On a dataset of discussions, detailed in Zhang et al. [2018], we derive forward-representations of comments, and use K-Means clustering to infer six types of comment; methodological details are listed in the appendix (Section A.3). As with the parliamentary setting from Chapter 2, we interpret these types as different *rhetorical intentions*, distinguishing between what the comment-writer intends to drive the conversation towards. Table 5.15 lists these types, along with representative examples of terms and comments.

Wikipedia discussions span an interesting mix of structured and unstructured talk: freeform exchanges take place alongside more procedural interactions related to article-editing. Our typology reflects this range: we find types corresponding to formal procedures (**moderation**, **procedures**), types reflecting routine activities like **coordination** of work or **editing** decisions, and types reflecting **contentious** or **casual** conversation. We note that it is particularly difficult for our method to expressively model how comments are responded to in less routinized discussions; as such, the latter comment types mentioned are harder to interpret.

We applied the inferred typology in a study of conversations that derail into overt hostility and personal attacks—a problem that’s especially disruptive in such collaborative interactions. Concretely, our goal was to detect early warning signs of eventual derailment, when a conversation still appears civil; we compiled a labeled dataset of derailed and on-track discussions, further detailed in Zhang et al. [2018]. We find differences in the distributions of comment types at the starts of eventually-derailed versus on-track conversations, shown in Figure 5.15 as log-odds ratios: types that might signal some existing or impending confrontation (**contention**, **moderation**) are more likely to occur at the starts of conversations that go awry; types that reflect discussions about collaborating

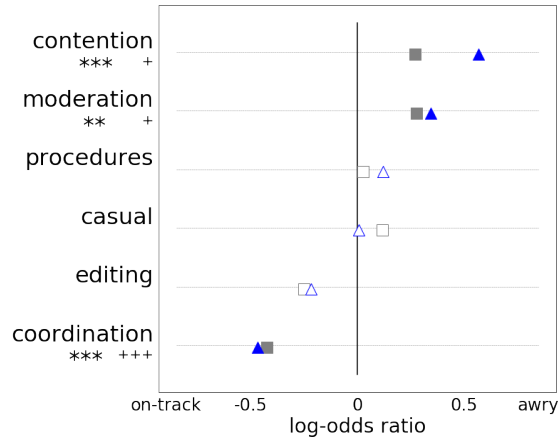


Figure 5.15: Log-odds ratios of comment types exhibited in the first and second comments of conversations that turn awry, versus those that stay on track. \triangle and \square denote log-odds ratios in the first and second comments, respectively; points are solid if they reflect significant (two-tailed binomial test $p < 0.05$) differences. * denotes statistical significance at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) levels for the first comment; + denotes corresponding p -values for second comment.

and editing (**coordination**) are more likely to occur in conversations that stay on track. In Zhang et al. [2018], we showed that features derived from this typology have predictive power in forecasting eventual personal attacks.

As with many other online discussion settings, a single comment in this data can receive multiple replies. Our framework does not account for this additional structure; we included only the earliest reply to each comment and ignore the rest. For future work, it would be interesting to adapt our approach to more expressively model discussions that diverge into multiple branches.

5.6.2 Orientation in US Supreme Court oral arguments

We present an exploratory study of how our approach for measuring orientation, as introduced in Chapter 3, could be adapted to analyze domains beyond

Orientation	Example terms	Example sentences
More backwards-oriented (bottom 25%)	does[n't] mean so you reasonable and you entitled particular	Well, does that not mean they are available in public? (Burger) So you have really three kinds of prices. (Rehnquist) And you really think you can prove beyond a reasonable doubt which one was the aggressor? (Scalia) And you dont think anybody is entitled to try to find out? (Marshall)
More forwards-oriented (top 25%)	suppose would [you] say difference [between] [do] you think your position agree	Suppose that were shown, that scene from the opera. (Ginsburg) Do you agree that Mr. Byrd could have excluded a carjacker? (Gorsuch) Is there any difference between its power as a contractor and its power as an owner? (Stevens) Is it your position that we should focus entirely on voting age population figures? (O'Connor)

Table 5.16: Example terms and sentences from utterances of Supreme Court justices which are more backwards- or more forwards-oriented (bottom and top 25% of Ω).

crisis counseling conversations. We apply the method to oral arguments in the US Supreme Court, where justices engage in exchanges with lawyers. Here, we aim to characterize the justices' utterances, using the lawyers' utterances as the conversational context.

We scrape transcripts of 6,733 cases from the Oyez project.²⁶ We used a training set of 91,924 justice and 372,268 lawyer utterances, that were sufficiently long and that had context-utterances that were also sufficiently long (see the appendix for further details). Both lawyer and justice utterances were represented as dependency-parse arcs.

Table 5.16 shows representative terms and (paraphrased) sentences with dif-

²⁶<https://www.oyez.org/>; the dataset can be found here: <https://convokit.cornell.edu/documentation/supreme.html>. Note that a similar analysis of the Supreme Court setting is presented in Zhang and Danescu-Niculescu-Mizil [2020] on a smaller set of transcripts; the dataset we use in this dissertation was collected independently of the smaller dataset used in the paper, which was originally collected by Ana Smith.

ferent orientations. We find that highly forwards-oriented terms and utterances tend to reflect justices pressing the lawyers to address a point in a particular way (e.g., [do you] *agree*, [what’s the] *difference between*); the least forwards-oriented terms involve the justice rehashing and reframing (not always in a complimentary way) a lawyer’s prior utterances (e.g., *so you [...], does [that] mean*).

Oral arguments contain more linguistic and topical heterogeneity than counseling conversations, since they cover a wide variety of cases, and because the language used by each justice is more differentiated. In addition, the dataset is much smaller. As such, our framework is sensitive to the particularities of each justice utterance, and produces messier output.

We note that some of the decreased interpretability might also come from the particular interactional dynamics of the institution. In particular, justices are tasked with scrutinizing the arguments made by lawyers; accordingly, 70% of terms and 87% of sentences have $\Omega > 0$ in this setting—a stark contrast to the counseling setting, where orientations tended to be negative. If justices aren’t strongly required to reflect on the lawyers’ statements, then it may be unclear what balance of objectives orientation is modeling, or what being backwards-oriented even means; many of the terms we found with negative orientation could be backwards-oriented not because they respond to a well-defined class of lawyer utterances, but simply because they aren’t forwards-oriented.

5.6.3 Exploration of the Switchboard Dialog Act Corpus

As a further exploration of our framework’s adaptability, we apply it to examine the Switchboard Dialog Act Corpus [Stolcke et al., 2000]. This dataset consists of transcripts of telephone conversations where interlocutors discuss a pre-set

topic. We use the framework to characterize utterances said by either interlocutor in each conversation. As such, we take *each* utterance in the data to be the conversational context (used to derive the context vector space Γ), and we compute characterizations of each utterance as well. We focus on examining utterance types inferred from forwards- and backwards-representations, as well as the orientation and shift measures. We compare these characterizations to the extensive tagset that the data is annotated with.

Data description. Several aspects of the dataset present challenges for our framework. First, in contrast to other settings we’ve examined, there are no institutionally-set practices; as such, the particularities of what each interlocutor says, rather than recurring conventions, are especially salient. Second, the corpus consists of transcribed speech. Individual utterances contain various disfluencies and self-corrections; there are numerous short interjections or backchannels (19% of utterances are labeled as such [Jurafsky et al., 1997]). In relating utterances and context-utterances, our framework assumes that a conversation consists of clearly-delineated turns; in a telephone conversation, this assumption becomes fraught.²⁷

In contrast to other settings, the Switchboard conversations are also relatively symmetric: interlocutors in a conversation do not have clearly differentiated roles.²⁸ Additionally, these interactions aren’t goal-directed: the interlocutors discuss a particular topic, but don’t need to achieve anything beyond that. Per manual inspection of the dataset, these features give the Switchboard

²⁷Note that the Supreme Court Oral Argument transcripts also contain similar challenges relating to transcribed speech, but the turn-taking dynamics between justices and lawyers are more structured. The parliament data is also transcribed, though disfluencies and background interruptions were not recorded, and turn-taking is strictly moderated.

²⁸We suggest that by contrast, in Wikipedia discussions, the person initiating the interaction tends to play a distinct role from the person responding to the first-commenter, since they take the initiative to navigate to the page to make a request or comment.

conversations a somewhat static quality: interlocutors share remarks without driving towards a particular point.

We skirt around most of these challenging features in our exploratory analyses. We process the data to remove disfluencies and backchannels. To avoid capturing topic-specific information, and to minimize the noise incurred from characterizing rare terms, we curate a vocabulary of 381 unigrams that occur in at least half of the conversation topics and in at least 200 conversations. Our preprocessing steps result in a collection of 34,562 utterances, spanning 1,155 conversations and 440 speakers; further details are included in the appendix.

Utterance tags. Utterances in Switchboard are annotated with several labels, under the SWBD-DAMSL annotation scheme. These tags reflect many different types of properties, spanning sociolinguistic indicators, discourse relations and form-based labels. Interestingly, many of the tags are grouped into *forwards* and *backwards* communicative functions. Forwards tags pertain to the type of speech act the utterance constitutes [Searle, 1976]; backwards tags denote what type of response the utterance is (an utterance can be labeled with both types of tag). This conceptualization of forwards and backwards is different from ours: for instance, our framework might characterize different speech acts (all corresponding to different forwards tags) as more forwards- or backwards-oriented.

In the subsequent analysis, we compare the framework’s output with the nine tags that occur in at least 1000 utterances:

Forwards: *statement-non-opinion, statement-opinion, yes-no question, wh-question*

Backwards: *acknowledgement, accept, appreciation, yes answer*

Other: *hedge*

Comparison to latent representations. We derive forwards- and backwards-representations of utterances, and then infer forwards and backwards types. For this analysis, we derive two types in either direction, as clusterings involving more than two types were hard to interpret. Among both forwards- and backwards-types, we identify the following:

- **Personal:** Utterances recounting personal experiences (e.g., “*Um, I haven’t gotten too terribly much into my major yet*”; “*And of course you need a baby sitter for that, but I’d really like to get out to the movies more often*”), comprising 58% and 54% of utterances for forwards and backwards types, respectively.
- **Commentary:** Utterances providing commentary, generally about the assigned topic rather than about personal matters (e.g., “*I’d say that the role of the teacher has gotten lower and lower [in] society*”, “*One of the long-term solutions would be to have some sort of solar power satellites up*”).

To measure how much an utterance type is associated with a particular tag, we compute log-odds ratios between type and tag.²⁹ We see that the **commentary** type is highly associated with *opinion statements* and *hedges* (log odds of 1.2 and 0.4 respectively, for both forwards and backwards types). The **personal** type is highly associated with *non-opinion statements*—such as those recounting personal anecdotes—and *yes answers*—which often seem to accompany personal anecdotes (log odds of 0.38 and 0.48, respectively, for the forwards type; 0.42 and 0.56 for the backwards type).

Note that we do not infer such a typology if we clustered LSA-based representations, instead of those derived by our framework: in that case, we output types distinguishing between **questions** (37% of utterances) and **statements**

²⁹Since there are only two types, the log-odds values are symmetric between the types.

(63%). Log-odds ratios between *opinion* or *non-opinion statements*, and the **statement** type, are 0.39 and 0.82 respectively; ratios between *wh-* or *yes-no questions* and the **question** type are 0.85 and 0.60. Notably, while the forwards and backwards types distinguish between opinion and non-opinion statements, the LSA-based typology does not.

We suggest the following explanation for the difference in typologies, which future work could more rigorously evaluate. Structurally, different stages of a conversation seem to be more focused on trading personal anecdotes, or on trading impersonal commentary, with relatively little intermixing between the two. Characterizing an utterance based on the nature of the adjacent utterances, as per our framework, recovers this structural coherence. Lexically, questions and statements are quite distinct from one another (for instance, questions appear to contain more discourse particles like *um* and *uh*), so it makes sense that this distinction is reflected in the LSA-based representations. However, questions and statements can occur both when the conversation is focused on personal anecdotes, and when it is focused on commentary, so such a distinction is not captured in the forwards- and backwards-representations.

Comparison to orientation. 81% of utterances have negative orientation. We suggest that this skew reflects the symmetric, goal-less nature of the conversations: people mostly respond to each other, and there are very few attempts to direct the conversation in a more systematic way.

For each tag, we compare the orientation of utterances with that tag to utterances without. We find that *opinion statements* and *appreciations* tend to have higher orientation (Mann Whitney U $p < 0.001$ for each, comparing utterances with or without the tag; Cohen's $d = 0.14$ and 0.33 respectively), while *wh-questions* have lower orientation ($p < 0.001$, $d = 0.46$). These findings underline

the differences between what the tagset conceives of as forwards or backwards, and what we consider to be forwards- or backwards-oriented. In posing a *wh-question* (a forwards communicative function, per the tagset), an interlocutor opens up a range of possible responses, rather than leading to a specific response (e.g., “*What do you think can be done about that?*”); as such, under our framework, we’d reasonably consider the forwards expectations to be fairly weak (modeled as low orientation). It’s less clear why *appreciations* (a backwards tag) have relatively high orientation: this could reflect localized interaction patterns where voicing appreciation for a statement (e.g., *I can imagine [...]*) is often followed by a similarly positive remark (e.g., *Yeah, completely [...]*).

Comparison to shift. For each tag, we also compare the shifts of utterances with or without that tag. We find that *wh-questions* tend to have relatively high shifts ($p < 0.001$, $d = 0.49$ and 0.20), while *opinion* and *non-opinion statements* have low shifts ($p < 0.001$, $d = 0.25$ and 0.51), as do *hedges* ($p < 0.001$, $d = 0.46$). We suggest that in asking questions, an interlocutor is inviting new information into the conversation and potentially departing from what was previously being discussed (e.g., “*How did they end up so far away?*”, “*what kind of business is it?*”). In simply offering statements, an interlocutor may add onto the existing thread of discussion without prompting such a departure.³⁰

5.7 Discussion

Throughout this section, we’ve explored a range of characterizations that can be derived from the Expected Conversational Context Framework, demonstrating

³⁰Admittedly, it seems that most of these high-shift utterances have a high value of the statistic by virtue of the word “how”; it’s worth further investigating the extent to which the framework returns interpretable characterizations for a wider range of terms.

how it's generative of a range of ways to describe and analyze conversations. In particular, we note that by accounting for an utterance's expected context, we are able to compute representations of utterances that provide some indication of how they fit into the broader structure of a conversation, or how they might focus or shift the interaction around them. We are also able to address questions that are inherently interactional, such as whether our expectations of a reply are realized. It would be fruitful for future work to continue these explorations on several fronts: by more rigorously interpreting and validating the characterizations we've presented, by refining our methods for deriving them, by applying them to address substantive questions in a domain, or by considering other properties.

We underline that the framework's output varies by setting, as does our interpretations of this output. Notably, while we could interpret forwards-representations of utterances at the starts of interactions (as in the parliament setting) as reflecting some form of intent, in the midst of back-and-forth exchanges (as in the counseling setting), the backwards dependencies make this interpretation less plausible. Our outputs and interpretations are also tied to the varied goals and incentives of conversationalists—we contrast, for instance, the distributions of orientations in the counseling, Supreme Court, and Switchboard datasets. An interesting line of future work could more substantively consider such a comparative approach, mapping out the dependencies between features of a conversation and features of the setting in which it occurs. The statistics that our framework computes could serve as quantitative bases for making these comparisons. The range, orientation and shift measures could be used to draw meaningful analogies or distinctions across settings even when the particular terms and utterances in each setting are very different; for instance,

such measures could enable us to perform a cross-domain analysis of leading or open-ended questions.

Throughout our analyses, we've also noted various ways in which the framework's output doesn't adequately reflect the phenomenon being modeled. We saw clear examples of faults and ambiguities as we examined the utterances and replies surfaced in our analysis of unexpectedness, and we acknowledge that the findings we report may be more informative of these methodological limitations than of the actual phenomenon of replying in unexpected ways. We also found that the framework was sensitive to the lack of routinized language in some of the settings we considered, and that it cannot address significant structural phenomena like branching discussions, or backchannels in the middle of turns. Some of these issues might be technical; more expressive ways of modeling utterances, such as those mentioned throughout Chapter 4, could mitigate them. However, as we suggest in the next chapter, some of these issues may reflect more fundamental complexities in conversations.

Part III

Action?

CHAPTER 6

TOWARDS ACTIONABLE UNDERSTANDINGS?

6.1 Overview

How do we get from analyses of conversations to actionable understandings?
What challenges do we encounter along the way?

To frame the subsequent discussion, we consider the types of actions we might ultimately wish to enact, using the crisis counseling setting as a demonstration. A better understanding of what's difficult for counselors, and what's effective at helping people in crisis, could point to ways of assisting counselors within and across conversations, and could lead to improvements in how they're trained. Imagine the clarity we could offer by informing counselors of the optimal moments in a conversation to move from exploring a problem to collaboratively problem-solving—a type of conclusion that our method from Chapter 3 potentially lays the groundwork for.

In this chapter, we consider such interventions in light of the nuances that might give us pause. In the preceding analyses, as we've explored our framework's descriptive capabilities, we've also come across various shortcomings that point to complexities outside its scope. More intuitively, our experiences as human beings having conversations (in the author's case, briefly being a crisis counselor) may make us suspicious of easy, prescriptive conclusions.

We start by highlighting some key qualities of conversations that inform and complicate our descriptive analyses (Section 6.2). We draw on existing sociological work to suggest ways in which conversationalists' actions are embedded in their particular conversational and situational contexts, and discuss how this

is reflected both in our methodological choices and our empirical shortcomings. We then draw on the causal inference literature to mathematically illustrate how these contextual factors have concrete implications for efforts towards rigorously establishing prescriptive insights (Section 6.3). We end with a discussion of the additional challenges that the formal analysis leaves open (Section 6.4).

Note on source material. Portions of this chapter consist of excerpts from Zhang et al. [2020], with Sendhil Mullainathan and Cristian Danescu-Niculescu-Mizil. We’ve modified some of the terminology used in that work to align more closely with the language used in the rest of the dissertation.

6.2 Conversational complexities

Conversations and utterances are embedded in the contexts in which they arise. We’ve drawn on this idea in developing and applying the Expected Conversational Context Framework; we’ve also encountered it when we observed ways in which the framework seems to fall short. In fact, critiques—originating in linguistics, sociology and anthropology—of computational approaches like ours have specifically highlighted inadequacies in how such approaches account for and conceive of context [Suchman, 1987, Schiffrin, 1994, Clark, 1996]. To discuss the limitations of our framework and the relevance of those critiques, we consider two interlocking sources of contextual information: the other utterances, and the broader situation.¹

¹We use context in a broad sense here, to denote any factor that could inform how we as conversationalists or as analysts interpret utterances in a conversation [Schiffrin, 1994]. We admit that the term is somewhat overloaded. In computer science circles, it seems that context could refer to various forms of “world knowledge”, or anything not present in linguistic data. As the Suchman quote we later cite suggests, the “et cetera” quality of context is no accident—it simply speaks to the wide range of things that can be meaningful in a conversation.

6.2.1 Conversational context

Conversations are dense in potentially significant information: many aspects of the interaction govern how utterances are constructed by one conversationalist and construed by another. This density of signal is central in sociological approaches like conversation analysis: per Heritage [1989], “no order of detail can be dismissed a-priori as disorderly, accidental, or irrelevant.”

Our framework underlines and operationalizes one key aspect of conversational context: the surrounding utterances. We addressed this aspect by deriving latent representations that model the language used in utterances and in collections of associated replies and predecessors. Indeed, our analyses suggest that this contextual information is important: accounting for it enables us to yield characterizations that, broadly speaking, meaningfully reflect how utterances fit into the interaction.

What “orders of detail” have we missed? As noted in Chapter 5.5, our approach mischaracterizes replies as unexpected when we do not account for key aspects of an utterance, such as its topical content or the particular way in which it’s constructed. Since we also don’t account for the full sequence of utterances leading to the present one, a reply that brings up something mentioned earlier in the conversation could be seen by our approach as completely out of place. In addition, our approach relies solely on linguistic data, ignoring other sources of signal whose significance has been documented in past work [Clark, 1996, Gumperz, 1982]—prosody, backchanneling, timing [Erickson, 2012], to name a few. In settings like the Switchboard corpus, we filtered out such signals, even though they are highly informative of uptake and understanding. In the parliament setting, by focusing on the text alone, we overlook the distinctly per-

formative nature of the exchanges, which likely has some bearing on how we interpret the rhetorical role of questions.²

Even in text-only scenarios like the crisis counseling setting, we leave key details unaddressed. For instance, we do not account for the timing of messages. However, from personal experience, from guidance written into the training manual, and from the broader literature on social interactions [Erickson, 2012], we know that timing and silence are strongly felt in conversations—no less in those that occur during time-sensitive crisis situations. A message that arrives after a delay could be read as disengaged; a message that arrives surprisingly quickly could be read as flippant.

In short, conversational context encompasses much more than the words used in nearby utterances. Additionally, it is continually reshaped: every subsequent utterance can modify how the interaction is understood by its participants [Heritage, 1989, Schiffrin, 1994]. In contrast, we note that our framework considers a static representation of conversational context.

The density and complexity of potentially relevant conversational signals has methodological implications. Conversation analysts, like us, collect and study many instances of a phenomenon. However, they tend to identify these instances one-by-one, revising their descriptions of the phenomenon along the way to constantly accommodate further nuances [Schiffrin, 1994, Hoey and Kendrick, 2017]. We contrast this qualitative approach with a computational one like ours, which automatically identifies large collections of utterances

²We admit that our argument conflates the details in an utterance’s context with details in the utterance, i.e., both utterances and context-utterances contain information we don’t model. Interestingly, one critique of conversation analysis [Schiffrin, 1994] is that in its emphasis on accounting for information in the surrounding conversation, it may miss pertinent details in the linguistic construction of the utterance itself. For now, we leave open the question of whether a theoretical treatment of context encompasses an account of utterances as well.

given pre-set criteria—e.g., utterances with large forwards-range, which we compute on the basis of a particular vocabulary of terms and a particular algorithm. Our framework may provide an informative birds-eye view, but is brittle to the long tail of relevant details.

6.2.2 Situational context

Across the settings we've analyzed, conversations arise in situations of social significance. Importantly, many aspects of the broader situation continually inform what happens in a conversation. In the counseling setting, for instance, past work has shown that counselors have more trouble helping texters of different racial backgrounds and gender identities [Helms and Cook, 1999, Fischer, 2021]; other accounts have also suggested that some people might be especially skeptical of mental health and medical services [Swami, 2012], leading to difficulties in building trust with the counselor. In other settings, we've drawn connections between conversational dynamics and an MP's career trajectory (Chapter 2), or the health of an online collaboration (Chapter 5.6.1).

What role do conversations play in these overarching contexts? As our efforts suggest, addressing this question is challenging. For instance, consider the problem of relating a measure like orientation to a counseling conversation's effectiveness (Chapter 3). We note that the indicator we use—texter rating—is riddled with caveats: there is a gulf of missing information between a texter saying they found the conversation helpful via text-message survey, and the texter actually experiencing some sort of positive improvement in a broader, longitudinal sense. In fact, we inherit such measurement problems from the broader domain: consider the numerous challenges of quantifying a complicated public

health outcome [McGlynn, 1997, Derose et al., 2002].

None of these extenuating factors are directly recorded in the linguistic transcripts we used as input to our method. Rather, we relied on intuition, expertise, and experience in the norms and practices of a domain to conceive of and to address them. Here, we again suggest that our approach is at odds with the density of the domain to which it's applied. Computational frameworks like ours are built on modeling a finite set of variables (e.g., the conversational context), using datasets that capture particular aspects of the broader situation (e.g., the words used). As such, we can only account for additional situational aspects in a piecemeal fashion—by incrementally adding new dependent or independent variables to our analyses, as they come to mind.

6.2.3 Implications for descriptive accounts

One takeaway from the above discussion is that in order to arrive at better descriptive accounts of conversations, we need richer data and more expressive models. We may juxtapose our limited view of conversations, as sequences of text, with the transcription practices found in conversation analysis, in which details like pauses and backchannels are accounted for [Jefferson, 1984, Hepburn and Bolden, 2013], giving the resultant transcripts some resemblance to musical scores. We may more extensively engage with domain experts who could point us to more meaningful social attributes in the setting. We may also look to work in NLP that aims to more systematically describe the range of possible contextual factors [Bisk et al., 2020, Hovy and Yang, 2021], laying the groundwork to more holistically account for them.

To somewhat qualify such recommendations, we note that we can only ac-

count for the contextual factors we can think of—let alone are able to systematically capture via our methods. Tellingly, in a survey of six different approaches to analyzing discourse, spanning a variety of academic disciplines, Schiffrin [1994] observes that all of them have their own theory of what counts as context, and how to account for it. As Suchman [1987] notes, in attempting to enumerate and account for contextual factors, we're faced with their inherent innumerability: "Every utterance's situation comprises an indefinite range of possibly relevant features, [...] as if we always included in our utterance an implicit *ceteris paribus* clause and closed with an implicit *et cetera* clause."

A corollary of this idea—that conversations and utterances are dependent on uncountably many contextual factors—is that they're dependent on the *particular* contexts in which they arise. As conversationalists, we may have some intuitive appreciation for *the* right thing said at *the* right moment; as analysts, we may be all too familiar with the challenges of accounting for an ever-growing list of social variables, resulting in objects of study that look singular in such a high-dimensional space. We may wonder if approaches like ours—that aim to derive abstract representations of conversational phenomena—are actually able to account for this essential particularity.

For instance, by design, our framework computes representations of utterances that represent our expectations of what contexts they arise in, without reference to the actual contexts in which they appear. We note that this opens several descriptive opportunities; notably, by distinguishing between the expected and actual replies of an utterance, we enable a range of analyses on this front. Even so, we find instances where our model of what's expected leads us astray because there was something particular to the utterance or its predecessors that we didn't account for. As Clark [1996] would put it, our framework

arrives at *conventional*, rather than *contextual* understandings, that “specify only the *potential* uses of a word or construction—and only some of these; they never specify the *actual* uses.”³

When we started on our study of parliamentary questions, what led us to appreciate the richness of the setting was such rhetorically interesting examples as that quoted in Chapter 2: “*The Prime Minister is rightly shocked by the revelations that many food products contain 100% horse. Does he share my concern that many of his answers may contain 100% bull?*” Our framework doesn’t quite live up to our initial excitement yet: it would locate this example in the neighbourhood of questions that might unironically voice a shared concern, overlooking its singularly ironic nature.

In the counseling setting, we were motivated by the prospect of accounting for how counselors exhibit conversational ingenuity in the face of difficult situations. Would a computational approach like ours be able to give counselors full credit for being responsive to the particular details of a conversation? Even in such a routinized setting as counseling, we suspect that such approaches might overlook such ingenuities. Our oversights might even be systematic: in Zhang et al. [2019], we demonstrate that, with experience, counselors become more linguistically distinctive relative to each other, which we suggest is a sign of “finding their own voice.” This evolution would literally take their language use and conversational behaviours outside the realm of the conventional.

³We suggest, perhaps as a subtle point worth expanding on, that the alternate measures of unexpectedness we considered also fall short of accounting for the particular. Consider the operation of comparing vector representations of an utterance and the actual reply it receives: neither the vectors, nor the comparison, are tailored in any way to the particular pair being analyzed. As such, short of retroactively adding special cases to our method, we wouldn’t be able to accommodate such extenuating knowledge as “in this particular set of circumstances, a question-asker is really interested in receiving a response on this particular thing.”

6.3 Challenges for deriving prescriptive conclusions

How does the contextual and particular nature of conversations impact our ability to arrive at actionable conclusions about them? We've already seen one implication of these complexities: they constrain our ability to produce rich, descriptive accounts. Now, we more explicitly show how these key properties also raise challenges when we attempt to translate our analyses of the data to prescriptive insights.

Suppose that in our data of counseling conversations, we observe that counselors who have more effective conversations also tend to use more positive language. Does this imply that we should somehow encourage counselors to behave more positively in conversations? We note that this question is inherently *causal*: addressing it requires us to establish that behaving more positively *causes* better outcomes.

Rigorously addressing such causal questions requires addressing the types of conversational complexities we've discussed. To precisely illustrate why this is the case, we draw on the causal inference literature and show, via a theoretical analysis, that these complexities are central to the inference problem. We then empirically demonstrate their practical implications on the counseling dataset.

Scope of the causal analysis. For the following analyses, we focus on a category of settings that we term *goal-oriented asymmetric conversational platforms*. Consider a platform that maintains a roster of *agents* who are expected to interact with incoming *clients*. First, the platform is *goal-oriented*: it has an overall objective that it seeks to use its agents to maximize. Second, it is *conversational* in the sense that agents work towards this objective by having conversations with clients; as such, their behaviours within these conversations are consequential.

Finally, the platform has an *asymmetric* degree of leverage: it can implement policies that affect its agents, but is unable to control its clients' characteristics. We note that the counseling service is a prototypical example of such a platform, with counselors and texters playing the role of agents and clients, respectively. The paradigm recurs across many other domains like customer service [Packard et al., 2018, Hu et al., 2019]—where sales representatives interact with customers, interviews—where interviewers interact with interviewees, and education—where teachers interact with students [Graesser et al., 1995].⁴

To clearly lay out our argument, we focus on a specific policy that such a platform can enact: deciding how to *allocate* its agents. The platform may allocate more conversations to agents it identifies as being more effective; as such, it may seek data-driven guidance on how to best select these effective conversationalists. Given the inherently conversational setting, we consider allocation policies where agents are allocated on the basis of behaviours they exhibit over past conversations they've taken, which we refer to as behavioural *tendencies*.⁵ As such, we analyze how these aggregate tendencies—e.g., an inclination to use more positive language or to write longer messages—can be used by the platform to identify and hence allocate conversations to more effective agents. Intuitively, observing that certain behavioural signals are correlated with desired conversational outcomes would suggest that the platform should allocate agents on the basis of tendencies inferred from these signals. The subsequent analyses more rigorously examines this intuition.

We note that helping the platform make allocation decisions is arguably

⁴A clear negative example, that's neither goal-oriented nor asymmetric, would be the Switchboard setting.

⁵As an alternative, the platform could make allocation decisions without directly accounting for agents' behaviours, relying instead on measures of past performance, or on demographic and personality attributes. See Zhang et al. [2020] for a more extensive comparison to these approaches.

much less satisfying than the prospect raised at the start of this chapter: helping the *agents* make decisions *during* a conversation through training and other forms of support. Here, our choice of what to focus on motivated by descriptive tractability: both types of approach are subject to the inference challenges we proceed to discuss, but the allocation policy is easier to theoretically analyze. In Section 6.4, we briefly discuss additional challenges that must be addressed with the training approach.

Overview of inference challenges. Before following through with an allocation policy, the platform needs to ensure that the policy would actually have a desired effect. Concretely, we must consider a counterfactual question: if the platform had allocated another agent with a different tendency to a conversation, would the conversation have had a better outcome? While this question could in principle be addressed via randomized experiments, an experimental approach is often infeasible given the sensitivity of a conversational setting like counseling, and the difficulty of specifying treatments involving complex interactional signals [Egami et al., 2018, Wang and Culotta, 2019]. Addressing such inherently counterfactual questions with observational data has been a core focus of causal inference (for surveys, see Angrist and Pischke [2008], Rosenbaum [2010] and Hernán and Robins [2020]). Such literature, however, has not dealt with the setting of conversations, or with the complexities we’ve just discussed.

To outline the difficulties of the inference task in a conversational domain, consider a naive approach for relating conversational behaviours and outcomes: if we observe that good outcomes follow conversations where agents exhibit a certain behaviour, we may naively infer that this behaviour is a useful signal of effectiveness. For instance, suppose we find that client mood tends to improve after conversations involving agents who use language with a greater degree of

positive sentiment. Such a finding could motivate us to allocate more positive agents to more future conversations.

At a high level, this initial approach suffers from a crucial pitfall that we've already anticipated: while an outcome may indeed arise as a result of an agent's behaviours, many *contextual* factors could also influence both the outcome and the behaviours that the agent exhibits. For instance, an agent may say more positive things in a conversation involving a congenial client who might also be more easily satisfied. However, in a situation involving a client with genuinely difficult concerns, a tendency for positivity may not even be appropriate, let alone effective. As such, a naive correlational approach cannot answer the counterfactual question posed above—it cannot inform us on how more positive agents would fare in conversations with less congenial clients. Crucially, it would be impossible for the platform or the agent to somehow influence the difficulty of the concerns that the client comes in with. This means that the correlation we observed isn't prescriptive: we can't establish if making agents more positive has a desired effect, but that's all the leverage the platform has.

6.3.1 Analysis of the causal inference task

We now proceed to more rigorously examine the entanglement between behaviour, outcome and context, focusing on the policy of allocating agents given their conversational tendencies. In particular, the allocation policy takes an aggregated view: the platform makes allocations based on how agents tend to behave over their past conversations. Intuitively, taking agent-level aggregates decouples our analyses from the contextual particularities that might constrain an agent's behaviour in a single conversation: over many conversations, their

personal inclinations may materialize as conversational tendencies. Likewise, an agent may exhibit a systematic *propensity* to elicit certain outcomes, even if the outcome of a single interaction is contingent on the context.

In what follows, we draw on the causal inference literature to formally examine the inference task underlying the allocation policy [Angrist and Pischke, 2008, Rosenbaum, 2010, Hernán and Robins, 2020]. First, we define this task in terms of the causal effect of allocation that we wish to estimate. We then discuss the challenges we face in quantifying this effect. We decompose these challenges into two key difficulties: the *observational* nature of our data entangles our correlational findings with *situational* aspects of the conversation; the *interactional* nature of our setting induces further entanglements with the *conversational* context. We analyze each of these challenges by concretely identifying statistical biases that arise in naive estimators of the effect of allocation; we also discuss particular assumptions we'd need to make to address these biases.

Formalizing the inference task

Our goal is to evaluate the potential effectiveness of a policy that allocates agents to conversations, given their conversational tendencies. We now discuss the central measurement in this task, which corresponds to the counterfactual question introduced in the preceding section: given two agents J and K , who have different tendencies with respect to some behavioural signal (e.g., J tends to use more positive language than K), what is the effect of allocating one agent to a conversation versus the other, on a given outcome? We henceforth refer to this quantity as the *allocation effect*.

Under our observational approach, we wish to estimate the allocation ef-

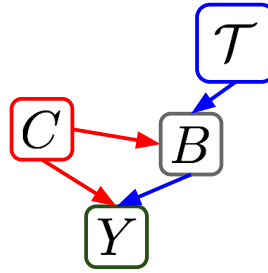


Figure 6.1: Graphical representations of the key dependencies underlying the inference task, between tendency \mathcal{T} , outcome Y , behaviour B and context C . Our goal is to estimate the effect of tendency on outcomes (blue path), however the contexts in which the behaviours and outcomes are observed confound this estimation (red arrows).

fect from data on conversations that J and K have already taken. As such, we must use the data in two ways. First, we must use past observations to estimate the propensity of each agent to get a desired outcome (e.g., proportion of their clients who improved their mood). Second, we must estimate each agent’s behavioural tendencies from their past conversations.⁶

In order for our estimate of the allocation effect to have a causal interpretation, we must ensure it can be ascribed to differences in the tendencies of J and K , rather than to differences in the contexts in which J and K ’s outcomes and behaviours were observed. As noted in the preceding discussion, these contextual factors can shape both the outcome of a conversation and an agent’s behaviour within the conversation, which thus become entangled.

These problematic dependencies are summarized in the graphical representation [Pearl, 1995] depicted in Figure 6.1. We would like to estimate the effect of (allocating) tendencies \mathcal{T} on outcomes Y (blue path); to this end, we must

⁶Indeed, conversational datasets, such as the ones we’ve examined throughout this dissertation, seldom comes with a priori labels of how agents tend to act. We may contrast this data-driven approach with self-reported indicators.

use behaviours B and outcomes Y observed under particular contexts C . Context shapes both behaviours and outcomes (red paths); our challenge is thus to somehow disentangle the effects of context and tendency.

Potential outcomes formulation. To formally highlight the biases that are incurred as a result of this entanglement, we mathematically express the allocation effect in terms of the potential outcomes framework [Angrist and Pischke, 2008, Rosenbaum, 2010]. Let \mathcal{T} be a random variable denoting a conversational tendency of agents, and suppose that agents J and K have different tendencies τ^J and τ^K . Let Y be a random variable denoting a conversational outcome. The allocation effect is then the expected difference in outcome if J , rather than K , is allocated to a conversation:

$$\mathcal{D}(\tau^J, \tau^K) = \mathbb{E}[Y | \mathcal{T} = \tau^J] - \mathbb{E}[Y | \mathcal{T} = \tau^K] \quad (6.1)$$

Let $\mathbb{D}(\tau^J, \tau^K)$ denote an estimate of $\mathcal{D}(\tau^J, \tau^K)$ from the data. Formally, this estimate has a causal interpretation if it is unbiased, i.e., $\mathbb{E}[\mathbb{D}(\tau^J, \tau^K)] = \mathcal{D}(\tau^J, \tau^K)$. Conversely, the estimate is misleading if it is contingent on the contexts C in which the observed conversational behaviours occurred.⁷

As we have intuitively noted and as shown in Figure 6.1, such dependencies on C arise when we estimate Y with observed outcomes, and \mathcal{T} with observed behaviours. We now articulate the challenges that are incurred from these dependencies. For each challenge, we provide an intuitive description supplemented with a graphical representation of the relationships between the vari-

⁷Throughout, the notation we use adopts the following convention: uppercase denotes random variables (e.g., Y , \mathcal{T} and \mathcal{D} are random variables for conversational outcome and tendency, respectively), lowercase denotes realizations of these variables (e.g., τ^J is an observed value of \mathcal{T}), and empirical estimators are listed in blackboard bold (e.g., \mathbb{D} is an empirical estimate of \mathcal{D} based on the observed data).

ables involved [Pearl, 1995], before drawing on potential outcomes arguments to formally express the biases incurred in the estimates. Our formal descriptions also point to particular assumptions under which we could mitigate these biases, as well as solutions that are premised on these assumptions.

Estimating outcomes: observed assignment and situational context

We first address the difficulties stemming from estimating agents' propensities for outcomes Y using our observations of their past conversations. To simplify the discussion, we provisionally suppose that we are given explicit labels of the agents' tendencies, returning to this point in discussing the second challenge.

At a high level, our estimates are subject to a problem that pervades observational studies: we can only observe outcomes in conversations that were actually *assigned* to agents exhibiting these tendencies. Here, we describe the implications of this problem in conversational settings.

Let A denote the observed assignment—i.e., the matching between each agent J and their conversations in the data. The assignment mechanism potentially exposes different agents to contrasting *situations*: for example, agent J may be assigned to more challenging clients than K . As such, these assignment-induced differences in situation, rather than differences in the agents' tendencies, could drive observed differences in outcome. In this way, A skews our estimation of the allocation effect.

The graphical model depicted in Figure 6.2 highlights the problematic dependencies between tendency \mathcal{T} and outcome Y , as indicated by the red edges: assignment A determines both \mathcal{T} and C , which in turn determine Y . As such, we cannot discount the effect of differences in assignment (red), beyond differences

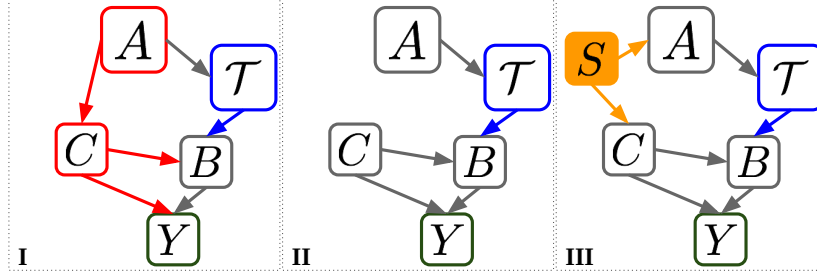


Figure 6.2: Graphical representations of the dependence between assignment A and outcome Y through behaviour B and context C that result from the observational nature of our analyses, giving rise to the selection bias exposed in (6.3). **I**: the problematic pathways from A to Y ; **II**: an idealized setting where conversations are randomly assigned to agents, in which the dependency is trivially broken; **III**: a scenario where assignment is governed by a set of observable *selection variables* S .

in tendency (blue), on the observed outcome.

Potential outcomes formulation. To surface the bias incurred from assignment, we formally examine the estimation of Y . As a first attempt, we can estimate the propensity of an agent J to get an outcome using the average outcome over their past conversations, denoted \mathbb{Y}^J . As such, we would measure $\mathcal{D}(\tau^J, \tau^K)$ as $\mathbb{D}(\tau^J, \tau^K) = \mathbb{Y}^J - \mathbb{Y}^K$. We note that our empirical estimators are contingent on A , i.e., we can only observe J in the conversations in which they actually participated. As such,

$$\mathbb{E}[\mathbb{Y}^J] = \mathbb{E}[Y | \mathcal{T} = \tau^J, A = J]$$

Substituting this expression into the above equation for $\mathbb{D}(\tau^J, \tau^K)$, we see that our estimator of the allocation effect is biased:⁸

⁸In the last derivation we subtract and re-add the second term.

$$\begin{aligned}
\mathbb{E}[\mathbb{D}(\tau^J, \tau^K)] &= \mathbb{E}[Y^J - Y^K] \\
&= \mathbb{E}[Y | \mathcal{T} = \tau^J, A = J] - \mathbb{E}[Y | \mathcal{T} = \tau^K, A = K] \\
&= \mathbb{E}[Y | \mathcal{T} = \tau^J, A = J] - \mathbb{E}[Y | \mathcal{T} = \tau^K, A = J] & (6.2) \\
&+ \mathbb{E}[Y | \mathcal{T} = \tau^K, A = J] - \mathbb{E}[Y | \mathcal{T} = \tau^K, A = K] & (6.3)
\end{aligned}$$

The equations highlight that our observed difference could have two sources. The first (6.2) corresponds to the effect of varying the tendencies over a shared set of situations (i.e., that were assigned to J). This is the value we need to estimate in order to answer the counterfactual question: what outcomes would have been attained had the conversations that were assigned to J been instead handled by an agent with a different tendency τ^K ? The second (6.3) reflects the *selection bias* that arises because J and K were actually exposed to different situations via assignment, as illustrated in Figure 6.2I.

An idealized setting: random assignment. As with many causal inference questions, selection bias would be eliminated if agents were *randomly assigned* to conversations, and are hence exposed to the same distributions of situations. As such, observed differences in outcome could no longer be ascribed to assignment-induced differences in situation. Formally, random assignment makes assignment and outcome independent for each agent (Figure 6.2II), such that the problematic term (6.3) trivially cancels out.

However, this selection bias remains in more realistic conversational settings, where assignment mechanisms are seldom random. At the extreme, if an agent *selects* their conversations, a record of positive conversational outcomes could be ascribed to picking clients who are easier to help, rather than having some replicable conversational proficiency. The problem persists beyond self-

selection—e.g., agents who work during the day may encounter more congenial clients than those who work at night.

A limited solution: controlling for situation. We may try to mitigate selection biases by controlling for situation C , for instance by comparing \mathbb{Y}^J and \mathbb{Y}^K only over conversations that match on attributes of the situation, e.g., are about the same issue. Indeed, many prior studies of conversations have employed such techniques [Jaech et al., 2015, Pavalanathan and Eisenstein, 2015, Tan et al., 2016, Choudhury and Kiciman, 2017, Zhang et al., 2018, Sridhar and Getoor, 2019, Saha and Sharma, 2020]. *Completely* controlling for situation would certainly break the problematic pathway from A to Y : Figure 6.2I shows that the two variables are conditionally independent given C and \mathcal{T} (formally written as $Y \perp\!\!\!\perp A \mid \{C, \mathcal{T}\}$).⁹

However, this approach is fundamentally limited: we can only control for the situational attributes that we can imagine, observe, and systematically measure. This leaves other important but inaccessible aspects (e.g., the client’s mental state) unaccounted for.

Enabling assumption: observed selection variables. We now describe an assumption under which this bias can be mitigated: at a high level, the challenge is addressable if we know that the assignment mechanism operates on particular aspects of the situation in a structured way. In particular, suppose that assignment is random up to a set of completely observable *assignment selection variables* S (Figure 6.2III, orange edges). As a natural example, consider conversational platforms where agents work during different *shifts*, and clients are randomly assigned to agents within each shift time. While different agents and

⁹Conditional independence corresponds to the criterion of d-separation in the graphical representation [Pearl, 1995].

clients may select different shifts, within a single shift these factors play no role in who gets assigned to whom. Furthermore, for each conversation, the platform knows the shift in which it took place. Beyond shift times, other examples of selection variables include geographic location and organizational divisions like departments of a store.¹⁰

Importantly, conditioning on S breaks the pathway between A and Y ; that is, Y and A are conditionally independent given S and \mathcal{T} ($Y \perp\!\!\!\perp A \mid \{S, \mathcal{T}\}$). Controlling for selection variables can be seen as a special case of controlling for observable situational attributes, where we know how these attributes S are related to the assignment mechanism. *Within each value of the selection variable*, our observations of agents' conversational outcomes are hence decoupled from situational differences due to assignment. As such, we modify our estimator to first measure the allocation effect for a particular selection variable (e.g., within a shift), comparing outcomes attained by agents with tendencies τ^J versus τ^K only for conversations with that selection variable.

Formally, for a given selection variable s , denote the corresponding estimator of the allocation effect as $\mathbb{D}(\tau^J, \tau^K \mid S = s)$. By conditional independence, we have that:

$$\begin{aligned} & \mathbb{E}[Y \mid \mathcal{T} = \tau^J, A = J, S = s] \\ &= \mathbb{E}[Y \mid \mathcal{T} = \tau^J, A = K, S = s] \\ &= \mathbb{E}[Y \mid \mathcal{T} = \tau^J, S = s] \end{aligned}$$

Thus, after conditioning on S , the bias (6.3) cancels out. That is, among conversations with the same S , empirical differences in outcome are entirely driven by

¹⁰We are effectively using the assignment of agents as valid instrument, conditional on shift, for the kinds of conversation the client is exposed to [Angrist and Pischke, 2010, Brito and Pearl, 2012, Pearl, 2013].

tendency:

$$\begin{aligned}\mathbb{E}[\mathbb{D}(\tau^J, \tau^K | S = s)] &= \\ &= \mathbb{E}[Y | \mathcal{T} = \tau^J, S = s] - \mathbb{E}[Y | \mathcal{T} = \tau^K, S = s]\end{aligned}\tag{6.4}$$

Repeating this matching process across all S then yields an aggregate measurement of outcome differences arising from varied tendencies, rather than from differences in the situations that agents are assigned to.

Estimating tendencies: interactional effects and conversational context

We now address the difficulties stemming from estimating agents' tendencies \mathcal{T} using our observations of their past behaviours. To simplify the discussion, we suppose that the difficulty we've just described, in estimating outcomes, has been fully addressed.

At a high level, the problem we face stems from the interactional nature of conversations: as we've discussed, the behaviour of an agent both shapes, and is constantly shaped by the behaviour of the other participant. As such, our measurement of an agent's tendencies, and hence our inferences about the relation between tendency and outcome, is skewed by the conversational context that agents inevitably react to within an interaction. At an extreme, we may observe that agents say "you're welcome" precisely after clients thank them. This does not necessarily mean that saying "you're welcome" is a behavioural inclination some agents have, beyond a reaction to the preceding interaction; it certainly does not follow that we should encourage more frequently saying "you're welcome."

An interactional problem. As a thought experiment, consider a *non-interactional* scenario where an agent's behaviour can affect an outcome without any inter-

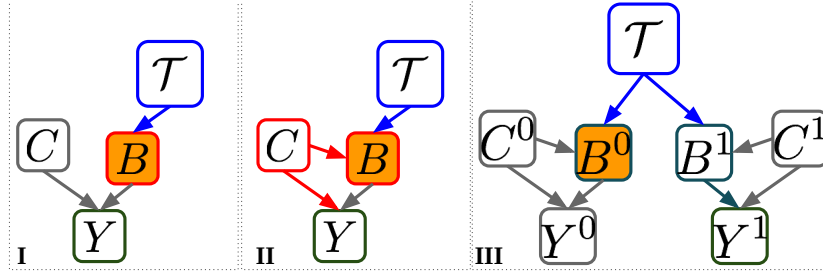


Figure 6.3: Graphical representations of the entanglement between conversational context C , behaviours B and outcomes Y , giving rise to the bias in (6.6). **I**: dependencies in a non-interactive setting; **II**: problematic dependencies when B interacts with contexts C that also shape Y ; **III**: our approach, observing behaviours and outcomes on different splits of data.

action with the client: a “secret santa” paradigm where an agent, the gift-giver, has no back and forth with their recipient (Figure 6.3I). In this case an agent’s behaviour is purely a reflection of the agent’s tendencies (e.g., an inclination for cheap gifts); and an empirical mismatch between \mathcal{T} and B simply reflects the noise with which a tendency gives rise to a behaviour. As we accrue more observations of the agent, we would expect such mismatches to diminish.

In contrast, in an interactive setting, these factors are problematically entangled (Figure 6.3II, red path): since the agent inevitably reacts to the client’s behaviour, B reflects C as well as \mathcal{T} . Furthermore, C can impact outcomes Y . An agent’s observed behaviour hence constrains the distribution of C that could have yielded our observed outcomes. As such, as with nonrandom assignment, differences in observed outcomes once again could reflect differences in context as well as in tendency—only this time, the problem comes from contextual factors within the conversation.

Potential outcomes formulation. Formally, let B be a random variable denoting observed agent behaviours. We use an aggregate of J ’s past behaviours,

denoted \mathbb{B}^J , to measure τ^J . Our empirical estimators are hence contingent on these observed behaviours:

$$\mathbb{E}[Y^J] = \mathbb{E}[Y|\mathcal{T} = \tau^J, B = \mathbb{B}^J]$$

Again, we highlight the bias in estimator $\mathbb{D}(\tau^J, \tau^K)$:

$$\begin{aligned} \mathbb{E}[Y^J - Y^K] &= \mathbb{E}[Y|\mathcal{T} = \tau^J, B = \mathbb{B}^J] - \mathbb{E}[Y|\mathcal{T} = \tau^K, B = \mathbb{B}^K] \\ &= \mathbb{E}[Y|\mathcal{T} = \tau^J, B = \mathbb{B}^J] - \mathbb{E}[Y|\mathcal{T} = \tau^K, B = \mathbb{B}^J] \end{aligned} \tag{6.5}$$

$$+ \mathbb{E}[Y|\mathcal{T} = \tau^K, B = \mathbb{B}^J] - \mathbb{E}[Y|\mathcal{T} = \tau^K, B = \mathbb{B}^K] \tag{6.6}$$

As before, two factors contribute to the observed difference in outcome. The first (6.5) arises from a difference in tendencies. The second (6.6), as we've described above and as depicted in Figure 6.3II, reflects a difference in conversational context, and is inherent to the interactional nature of conversations.

A limited solution: ignoring the interaction. The factor in (6.6) intuitively compounds as the conversation progresses and an agent's behaviour becomes increasingly contingent on the circumstances. As such, we may seek to address this bias by only considering behaviours from the start of the conversation, before behaviour and circumstance become tightly coupled. Indeed, prior work has taken such a limited view of conversations [Althoff et al., 2016, Zhang et al., 2018] with this confound in mind. However, insofar as this approach does not fully address the bias incurred by interaction, it also constrains the scope of the conversational tendencies we can study.

Enabling assumption: separable sets of conversations. To factor out this in-

teractional bias, we must decouple our observations of agent behaviours and outcomes from the conversational contexts they are both tied to. Consider a conversational platform where agents take many conversations, and where different subsets of these conversations are separable from each other. With such assumptions, we consider a simple fix: for each agent, we measure their behaviours over a *subset* of the conversations they’ve taken, and use a *separate* set of conversations to measure the outcomes they elicit.¹¹

Formally, suppose we split each of B , Y , C into two random variables, one for each subset. As shown in Figure 6.3III, the only pathway connecting an agent’s behaviours and outcomes *across* these splits is via their conversational tendencies. That is, B^0 and Y^1 are conditionally independent given \mathcal{T} , so

$$\mathbb{E}[Y^1 | \mathcal{T} = \tau^J, B^0 = \mathbb{B}^{J,0}] = \mathbb{E}[Y^1 | \mathcal{T} = \tau^J]$$

and the bias term (6.6) cancels out.

6.3.2 Empirical demonstration

Via our theoretical analysis, we’ve seen that unless we address the influence of situational and conversational context, we cannot rigorously establish whether an allocation policy leads to better outcomes. We now illustrate this empirically, instantiating our theoretical formulation in the counseling conversation dataset. Here, counselors play the role of agents, who have the goal of helping texters—

¹¹While we solve a different problem, our solution is analogous to separating train and test sets to mitigate overfitting—here we “train” our measurements of tendencies and “test” their effects on separate data splits. Note that throughout, we use “subset” to refer to a collection of conversations, not to a subset of messages within a single conversation.

the clients—to a calmer mental state.¹²

For the purposes of our present demonstration, we consider a small set of conversational behaviours, which we select as simple representatives of a broad range considered in past work in the counseling and mental health domain. These behaviours, along with studies that have demonstrated their correlations with mental health-related outcomes, are listed in Figure 6.4, along with studies that have demonstrated their correlations with mental health-related outcomes. In particular, conversation length, response length and response speed relate to the *fluency and pace* of the conversation [Althoff et al., 2016, Pérez-Rosas et al., 2018, Chikersal et al., 2020]; sentiment is a frequently-cited attribute of the *style or tone* of an utterance [Althoff et al., 2016, Pérez-Rosas et al., 2018, Chikersal et al., 2020, inter alia]; lexical similarity between utterances and linguistic coordination are often used to characterize *interactional* behaviours [Althoff et al., 2016, Sharma and De Choudhury, 2018] like adapting to a client’s language or reflecting their concerns.¹³

We relate these behavioural signals to two complementary indicators of a conversation’s outcome. First, we consider the **rating** provided by texters after conversations, introduced in Chapter 3. We also consider whether the conversation was properly **closed**—i.e., the counselor wraps up the interaction at a moment that feels appropriate for all participants—or **disengaged**—i.e., a counselor ends a conversation after a texter is unresponsive for a long period of time.

¹²We focus on analyzing the subpopulation of 4,861 counselors who take at least 80 conversations, reporting all statistics over the first 80 conversations taken by each of these sufficiently prolific counselors.

¹³We measure a counselor’s speed in a conversation as the number of words they write, per minute taken to reply to a texter. Following Althoff et al., we measure sentiment as the VADER compound score of each message [Hutto and Gilbert, 2014] and similarity as cosine similarity between a counselor’s message and the texter’s preceding message; we obtain conversation-level measures of response length, sentiment and similarity by averaging over the counselor’s messages in a conversation. As in Althoff et al., we use the approach from Danescu-Niculescu-Mizil et al. [2012] to measure coordination.

In our data, 72% of conversations are properly closed.

To illustrate the empirical consequences of the conversational complexities and inference challenges we've discussed, we compare various methods of estimating the relation between tendencies and outcomes: we consider estimates that address these challenges, as well as naive estimates that do not. Discrepancies between these estimates would therefore point to ways in which failing to account for the challenges leads to misleading conclusions.

Naive formulation: conversation-level effects. We first compare counselor behaviours in conversations that are rated positively versus in those rated negatively, as well as in conversations that are closed versus in those where the texter disengages. For each conversation-level behaviour and outcome, these comparisons yield statistically significant differences (Mann-Whitney U test $p < 0.01$), echoing several correlations reported in prior work between behaviours and outcomes in individual conversations. As we have argued, the usefulness of these relationships in guiding policies is unclear, since they could reflect contextual factors that the platform cannot influence. For instance, the sentiment of counselor messages is significantly more positive in positively- versus negatively-rated conversations; this could reflect the benefits of an upbeat tone, or that distressed texters who are harder to help also tend to discuss less positive things. At the extreme, closed conversations are much longer than disengaged ones (28.4 vs. 20.0 messages per conversation on average), perhaps tautologically: disengaged conversations end prematurely by definition.

Counselor-level correlations (Δ). To build up to a counselor-level approach that addresses the influence of context, we first consider correlations between counselor-level aggregates of behaviour \mathbb{B} ,¹⁴ and of outcome \mathbb{Y} (computed as a

¹⁴With the exception of coordination, which is already a counselor-level property, we derive

counselor’s proportion of positively-rated or closed conversations). This view corresponds to the counselor-level approach taken in Althoff et al. [2016].¹⁵ To quantify the extent to which an aggregated behaviour \mathbb{B} relates to an outcome propensity \mathbb{Y} , we compute Kendall’s tau correlations between \mathbb{B} and \mathbb{Y} , depicted in Figure 6.4 as Δ . At a high level, Kendall’s tau compares the rankings of counselors according to \mathbb{B} and according to \mathbb{Y} by capturing the extent to which, within each pair of counselors, differences in \mathbb{B} are in the same direction as differences in \mathbb{Y} . This mirrors our formulation of the allocation effect from Equation 6.1, which is likewise defined over pairs of counselors; here, however, we implicitly and naively assume that \mathbb{B} and \mathbb{Y} correspond to estimates of counselor tendency and outcome that can be meaningfully related (i.e., we ignore the two inference challenges).

Addressing bias from interactional effects (\square). As we’ve mathematically shown, both \mathbb{B} and \mathbb{Y} are entangled with the conversational context, by virtue of the interaction between counselors and texters. We note that in this setting, counselors have many conversations with different texters; further, given the service’s focus on providing support in acute crises, texters generally do not contact the service repeatedly, and the service does not deliberately assign repeat texters to the same counselor (contrasting, for instance, a therapy-oriented service). As such, we assume there are no dependencies between different conversations taken by a counselor, allowing us to address this entanglement. Concretely, we divide each counselors’ conversations into two splits, such that split 0 consists of their even-indexed conversations (i.e., the second, fourth, sixth, ...) and split 1 consists of their odd-indexed conversations. Using Kendall’s tau,

counselor-level aggregates by averaging a counselor’s per-conversation behaviours, e.g., their sentiment, across all of their conversations.

¹⁵Note that Althoff et al. [2016] only consider the top and bottom 40 counselors in terms of \mathbb{Y} , while we consider all counselors.

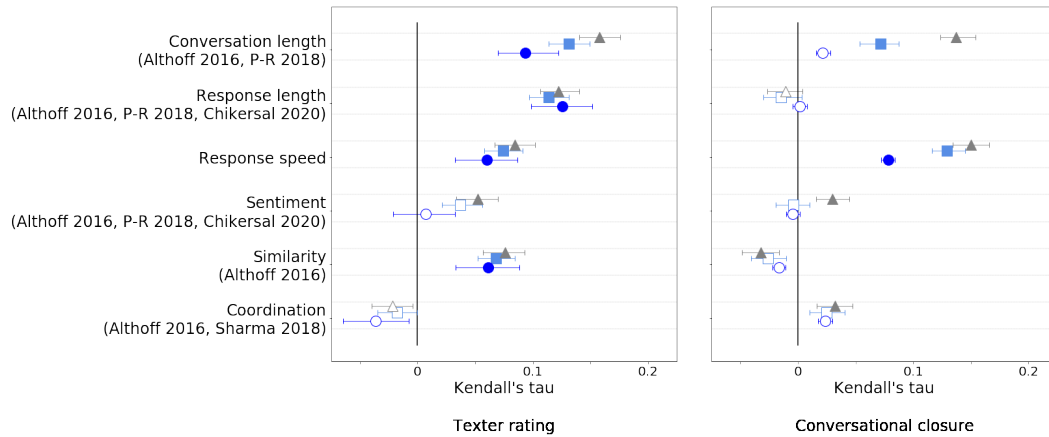


Figure 6.4: Relation between counselor-level behavioural tendencies and outcomes, measured as Kendall’s tau correlations, for different estimation approaches: \triangle correlates counselor behaviour and outcome propensity; \square computes this correlation across temporally-interleaved splits of conversations; \circ further controls for shift time, thus reflecting the allocation effect formulated in Equation 6.1 while accounting for the inference challenges we described. Error bars show bootstrapped 95% confidence intervals; shapes are filled for bootstrapped and Bonferroni-corrected $p < 0.01$. Abbreviated citations indicate studies that have demonstrated correlational relationships between the respective behaviours and outcomes.

we compare the ranking of counselors according to their average behaviour \mathbb{B}^0 over split 0 with their ranking based on their outcome propensity \mathbb{Y}^1 over split 1, depicted as \square in Figure 6.4. As such, \mathbb{B}^0 and \mathbb{Y}^1 correspond to estimates of counselor tendency and outcome which address this source of bias.

Addressing bias from observed assignment (\circ). The relation between \mathbb{B}^0 and \mathbb{Y}^1 is still subject to biases incurred by assigning counselors to different texter situations. As we’ve discussed, such a problem could be addressed given an observed selection variable. In this setting, we assume that the assignment of conversations to counselors is random up to the *shift times* the counselors sign up to take; indeed, while counselors can choose which shifts to sign up for, the

platform assigns counselors to conversations randomly within-shift. Accordingly, we control for shift time, as follows. For each counselor J and shift s , we compute a shift-specific outcome propensity $\mathbb{Y}_s^{J,1}$ (again over split 1). For counselors J and K , we then compute the difference in outcome propensity over each shift they coincide on, $\mathbb{D}^1(\tau^J, \tau^K | S = s) = \mathbb{Y}_s^{J,1} - \mathbb{Y}_s^{K,1}$. To aggregate across shifts, we take $\mathbb{D}^1(\tau^J, \tau^K)$ as the average of $\mathbb{D}^1(\tau^J, \tau^K | S = s)$, weighted by the number of conversations taken by the least-active of the two counselors within each shift. Finally, we compute Kendall’s tau between outcome differences from split 1 and behavioural patterns from split 0, shown in Figure 6.4 as \bigcirc .

Results. In comparing different counselor-level approaches, we highlight the two inference challenges that are in play, and show how addressing them moderates our understanding of how different tendencies and outcomes are related. This is depicted in Figure 6.4 as \bigcirc s which are hollow—indicating no statistical significance—or closer to the vertical line (at Kendall’s tau = 0) than corresponding \triangle s or \square s—indicating that the latter two approaches overestimated the effect size. For example, the large counselor-level effects of conversation length on closure (\triangle) diminish drastically after addressing interactional bias (\square), showing that length tautologically reflects closure. Further addressing the temporally-mediated assignment bias (\bigcirc) shows that this tendency does not have a significant effect on closure; the decreased effect size also suggests that the previously-observed relation may also have been contingent on shift time.

6.4 Discussion

Do our adjusted estimates, represented as \bigcirc in Figure 6.4, constitute prescriptive understandings? So long as we are confident of the assumptions about the

counseling service that we relied on above, we can interpret those estimates as quantifying a causal relation: the extent to which assigning a counselor to more conversations based on a particular behavioural tendency will improve a particular outcome. In Zhang et al. [2020], we built on this analysis to estimate the effects of an allocation policy, suggesting via a back-of-the-envelope simulation that such a policy could actually increase the share of positive ratings received by the service. Some natural next steps include validating the assumptions we've made and substantiating these estimates with actual randomized trials; we could also expand the range of behavioural signals considered.

In practice, many of the assumptions we've made might give us pause. As we've already discussed, the outcome measures we consider reduce the complex space of mental health outcomes to a few narrow points. Additionally, we note that our analysis of the allocation policy elides a key logistical question: can counselors actually be assigned to more conversations? Realistically, counselors may not be willing or able to take on additional load, and taking more conversations may adversely affect their ability to attain good outcomes. Such a policy also seems inhumane, especially in light of accounts demonstrating high levels of stress and burnout among people working in the area of mental health [Ferguson, 2016, Fischer, 2021]. The aspect of situational context we've ignored, so to speak, is that counselors are people in a difficult profession, and the strains they experience in one conversation can carry over into the next.

Of course, we made these assumptions for argument's sake: by taking them for granted, we could clearly show that situational and conversational context had direct bearing on our ability to make causal inferences. However, we stumble over such factors as soon as we wish to realize the real-world applications that the analysis is ostensibly driving towards.

Informing decisions within a conversation? Recall the types of policies we were initially motivated by: helping counselors by telling them what conversational behaviours are effective. We've instead examined a coarser-grained policy that helps the counseling service while giving the counselors very little agency. Formally, the causal claims we've considered are of the form, "X is a signal on the basis of which a platform can assign agents to future conversations, to influence outcome Y." To more directly help the agents, we'd need to address a more direct question: "does doing X in a conversation cause Y?"

We note that the theoretical formulation we've built up in the preceding section actually led us away from making such conclusions. The core idea—behind viewing the allocation policy as easier to analyze, and behind the enabling assumptions and solutions we proposed—was to abstract away from examining behaviours within conversations and take statistical aggregates. We interpreted the resultant quantities as behavioural tendencies, but we emphasize that what we've estimated is, by design, decoupled from the situational and conversational contexts that govern agent behaviour in a conversation.

How do we move towards the causal claims we want to make? We leave this as an open question, and suggest some further challenges that conversational settings raise for causal inference methods. Fundamentally, causal inference is unable to make conclusions about *individual* effects [Hernán and Robins, 2020]—i.e., we cannot use the data to show that in a particular conversation, a particular action was somehow consequential. Rather, we can only say that across the data, we can contrast actions taken in some conversations with alternate, counterfactual actions taken in other, comparable conversations, arriving at aggregated estimates of the actions' effects. Even if we are happy with such aggregates, we run into problems. A key condition that must be met in order to

estimate causal effects is *positivity*: that the data contains enough observations of the various counterfactual scenarios we wish to examine.

The *particular* nature of conversations, as discussed earlier, seem at odds with these statistical limitations. Our ability to do action X in a conversation, and whether that's effective, is highly dependent on the context. For instance, Figure 6.4 suggests that writing long messages might be causally related to good outcomes. In trying to translate this tendency-level result to a behaviour-level one, we're faced with a range of contextual contingencies. Consider an exchange with a texter who might be mistrustful of the service, who might be losing their patience, who might be in a life-threatening situation, who might have asked such a straightforward question that a long-winded answer seems like a dodge—such particularities constrain the counselor's ability to craft a long message, and whether that message is read positively or negatively by the texter, or not read at all. Translating a causal finding to a recommendation also raises issues. Training counselors to use more words seems problematic if verbosity is simply a signal of something less readily measurable, like eloquence, fluency in a language, experience with the texter's situation, or not being burnt out. Unilaterally increasing wordcount would superficially modify the messages that are produced without impacting the underlying action and its effects.

A standard way to address these issues in the computational and statistical literature [Angrist and Pischke, 2008, Hernán and Robins, 2020] is to control for aspects of the context: “in situation C, if you're a counselor of type T, then doing X will cause Y.” Sociologically, we may wonder how finely to specify the situation; statistically, as we more finely specify what C is, we approach the mathematically intractable task of reasoning about the particular.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this dissertation, we considered how computational approaches could arrive at actionable understandings of conversations. Our starting premises are that conversations are complex, and that they play crucial roles in broader tasks; analyses that can address these complexities could therefore inform ways to support these tasks and the conversationalists carrying them out. Building on studies of particular conversational phenomena (Chapters 2 and 3), we developed a computational framework for characterizing utterances and their roles in an interaction (Chapter 4). We demonstrated how the general approach enables a variety of analyses across a range of conversational settings (Chapter 5). We then critically appraised this framework, in terms of whether it could yield meaningfully rich and actionable accounts of conversations (Chapter 6).

7.1 Epistemological tensions

In the preceding chapter, we juxtaposed our descriptively generative, yet theoretically narrow, treatment of context with the dense ways in which utterances are informed by the contexts in which they arise. We discussed various ways in which this disconnect between statistical method and sociological complexity raises problems in translating analyses into actionable understandings. We find that computational approaches like ours yield exciting descriptive possibilities, as well as unsatisfying—and prescriptively consequential—gaps. The following quote, from Wittgenstein, aptly captures the disconnect we’ve arrived at:

“We have got on to slippery ice where there is no friction and so in a certain sense the conditions are ideal, but also, just because of that, we are unable to walk. We want to

walk, so we need friction. Back to the rough ground!" [Wittgenstein, 1953]

We believe that future computational work needs to square with this key tension, if we wish to better describe conversations, and to better inform how people have them.

As a starting point, we're reminded of a distinction proposed by Ryle, between someone navigating a village they live in, and then being tasked to draw a map of this village, as a cartographer:

"In the morning, he can walk from the church to the railway station without ever losing his way. But now, in the afternoon, he has to put down with compass bearings and distances in kilometres and metres the church, the railway station, and the paths and roads between [...] He has, so to speak, to translate and therefore to re-think his local topographical knowledge into universal cartographical terms." [Ryle, 1962]

Ryle's objective, in making this distinction, was to set straight confusions he felt were plaguing philosophers about the descriptive scope of their work; in effect, he wanted to clarify what exactly philosophers are doing. His point is that there is a difference between problems people are concerned with in living their life versus in doing philosophy; the latter type of problem concerns not just "new questions of an old sort," but "questions of a new sort." Analogously, our computational methods seem to lend us the descriptive power of "universal cartographical terms" but are unable to account for "local topographical knowledge." Drawing on this idea, we suggest that it's productive to recognize that our understanding of conversations as analysts, and as conversationalists, are distinct, and to clarify the ways in which these two views differ.

As seeds of a theory that elaborates on this distinction, we consider some case examples. We might use a computational approach to establish qualified

expectations about what happens next in a conversation (Chapter 5.5), recognizing that our model and the conversationalists have access to different sets of information (an entire dataset of past exchanges, versus the particularities of this one exchange). We might highlight the importance of a contextual factor by empirically demonstrating the consequences of failing to account for it (e.g., conflating correlational and causal effects, in Chapter 6.3), acknowledging that surfacing the problem is much easier than modeling the factor. More directly analogous to Ryles' own conclusion, a computational perspective could allow us to map out seeming contradictions that come up in an activity—like addressing backwards while also advancing forwards in an interaction (Chapter 3)—“[stating] their directions, their limits, and their interlockings.” In sum, the limitations we've noted, while serious, don't preclude that our methods are descriptively compelling. Per Sacks [1989b], if we could come up with systematic accounts of “abstract [conversational] objects which get used on singular occasions, then that's something which is exceedingly non-trivial to know”—even if accounting for singular occasions continues to elude us.

Importantly, the distinction that Ryle calls out has implications for how our approaches could inform actions: if our descriptions reflect a view of conversations that's fundamentally different from how conversationalists experience conversations, then who are we to try to intervene on what conversationalists do? Here, we point to work that aims to clarify how computational efforts, given their limitations, should be oriented with respect to social impact, and that calls out the harms that result when models that see the world in a limited way nonetheless are applied to shape it [Abebe et al., 2020, Green and Viljoen, 2020, Alkhatib, 2021]. For instance, drawing a parallel to similar epistemic tensions in legal scholarship, Green and Viljoen [2020] argue that computational

practitioners should adopt a stance of “algorithmic realism” that’s cognizant of the social and contextual factors that are important but left unaddressed by their methods. Together, this body of work suggests ways to scope and circumscribe the roles played by approaches like ours, to “leverage [their] particular strengths” [Abebe et al., 2020] without taking crucial real-world complexities for granted.

7.2 Future directions

To conclude, we outline some particular directions that future research could explore, in concert with these epistemic questions.

7.2.1 Conversations as processes

Thus far, our analyses have largely focused on characterizing individual utterances in an interaction. However, conversations—and crucial conversational phenomena—develop over many turns. For instance, the process of fostering trust is instrumental in settings like crisis counseling, as counselors build connections to total strangers experiencing turmoil. Many computational studies of conversational processes have largely focused on the task of forecasting their endpoints: over the course of an interaction, was someone persuaded [Tan et al., 2016], was a problem solved [Niculae and Danescu-Niculescu-Mizil, 2016], did a conflict bubble up [Niculae et al., 2015a, Zhang et al., 2018, Chang and Danescu-Niculescu-Mizil, 2019], was some sort of rapport established [Goldberg et al., 2020, Bao et al., 2021]? Beyond extracting predictive signals, it would be fruitful to more richly account for what happens in the middle of these processes.

For instance, was a conversational development gradual or sudden? Was there a pivotal moment, after which the outcome was all but inevitable, or are there always possible ways out?

A crucial quality that such an account must address is that conversations are *emergent*—they arise out of local decisions made by conversationalists that are contingent on, and continually renegotiated according to what the other participants do [Suchman, 1987, Clark, 1996]. We see this contingent quality in an account of how people end conversations from Schegloff and Sacks [1973]. Their analyses demonstrate that closing a conversation is not facilitated by pre-planned routines so much as continual motions towards opening up closings that may be taken up or rebuffed; we are reminded of meetings that are inordinately prolonged by one participant bringing up “just one more thing.”

How can a computational approach infer such a process from the data? Here, Clark offers a word of caution:

“Transcripts are like footprints in the sand. They are merely the inert traces of the activities that produced them, and impoverished traces at that. The structure we find in a transcript only hints at how a conversation emerged [...] the trace tells us nothing about the choices the speaker did not make; lulls us into assuming the choices were there from the start.” [Clark, 1996]

In other words, the goals, plans and routines we might infer in a post-hoc analysis of the data reflect not the intermediate dynamics of a conversational process, but rather the fact that we can only observe one among many counterfactual paths the conversation could have taken along the way. A challenge for future work is to develop methods that, as Schegloff [1982] puts it, are able to “retain a sense of the actual as an achievement among possibilities.”

7.2.2 Conversations as parts of broader tasks

Throughout this work, we've emphasized how conversations arise in the context of broader endeavours. We've also highlighted the limited extent to which our present methods can situate conversations in these endeavours, to arrive at rich understandings about the roles that conversations play. There are a variety of ways in which future work could address such a limitation: improving causal inference methods, conducting randomized trials, or pursuing ethnographic work to more directly engage with practitioners and policymakers.

A key challenge for these methodological extensions is holistically accounting for the broader task. In real-world conversational settings, conversations are often used in concert with other actions, such that the challenges faced by conversationalists aren't strictly conversational. For instance, consider another goal-oriented asymmetric setting: contact tracing. Contact tracers talk to community members to glean information about the spread of a disease. The conversations they conduct are important, but are only effective in conjunction with a host of other public health measures such as quarantining affected people, developing treatments and vaccines, informing the public, and offering economic and mental health support to populations affected by the disease. Recent accounts of contact tracers' experiences [Akam, 2020, Becker, 2020] point to the range of conversational challenges they face—jogging someone's memory, building trust, overcoming hostility—as well as logistical ones, like getting people to pick up their phone or answer their door. Quantifying the impact of a particular conversational behaviour, in light of the multitude of extenuating factors, raises statistical—but also sociological and epistemological—challenges that future work must address.

7.2.3 Conversational professions

In many settings, including several we've mentioned throughout the dissertation, conversationalists have conversations over and over again as part of their jobs. This is true of counselors, but also of educators, lawyers, journalists, physicians, contact tracers, Ph.D. advisors, and so on. In Zhang et al. [2019], we consider the question of whether crisis counselors learn new skills, or become better at their jobs, with experience. We show via a longitudinal analysis that counselors become more linguistically diverse as they take more conversations, but leave open further exploration of the mechanics behind this evolution, as well as a rigorous clarification of what constitutes "learning conversational skills." This longitudinal perspective—of having conversations as a profession—raises other important questions. For instance, in the counseling setting, how often do counselors get burnt out, how does this show through in the conversations they take, and how could we best support them?

Viewed as parts of overarching projects and careers, the boundaries between individual conversations become porous. As such, there are exciting opportunities for future work to move from analyzing individual utterances, to sequences of utterances, to utterances and conversations as situated in a broader interactional ecosystem.

APPENDIX A

FURTHER METHODOLOGICAL DETAILS

We elaborate on the datasets considered in this dissertation, as well as on particular methodological choices made in applying the Expected Conversational Context Framework to each dataset. Code implementing our framework, and demonstrating these choices on the datasets we publicly release, can be found at <https://convokit.cornell.edu>.

General. Across all of the settings we considered, we made the following methodological choices, detailed in Chapter 4.4.4: as input to the Expected Conversational Context Framework, we represent utterances as column-normalized tf-idf reweighted matrices; we also remove the first dimension of the derived latent representations.

A.1 UK parliamentary question periods

We provide further details on the UK parliamentary questions periods data introduced in Chapter 2. The full dataset consists of 216,894 question-answer pairs, covering 6 prime-ministerships (from Thatcher to Cameron), 1,975 different askers, and 1,066 different answerers, and can be accessed at <https://convokit.cornell.edu/documentation/parliament.html>.

Selection of terms. We extract terms from questions and answers as dependency-parse arcs, following the procedure described in Section 2.5. We consider a question-term vocabulary consisting of the 1,152 terms that occur in at least 100 and at most 10% of questions, and an answer-term vocabulary with the same filtering parameters, consisting of 2,706 terms. In particular, we set the maximum document frequency to be fairly low, since we found that including

extremely common terms results in less interpretable typologies.

Framework inputs and parameters. As input to the Expected Conversational Context Framework, we consider the subset of 183,084 questions and 192,192 answers that contain at least one term in the question or answer-term vocabulary, respectively. To characterize questions (in Chapter 2) we derive a latent context space of 25 dimensions, noting that since we remove the first dimension, we effectively work with 24-dimensional representations. To characterize answers (in Chapter 5.2.1) we derive a latent space of 15 dimensions (resulting in 14-dimensional representations).

Data subset used in partisanship, department and tenure analyses. In Chapter 2, when analyzing the relation between question types and various institutional attributes of the askers and answerers, we restrict our analyses to the subset of data for which information on these attributes is known. In particular, we consider the questions and answers that occur after 1997 (the start of the Blair government), since we were not able to consistently infer asker and answerer affiliations in the earlier subset of the data. Additionally, we consider only questions asked by MPs affiliated with the government party, or with the official opposition (i.e., the largest opposition party). Finally, we only consider questions which have at least one question-term in the vocabulary. In sum, these decisions result in a subset of 80,907 questions.

We use the same subsetting decisions when analyzing the relation between answer types and party affiliation (Chapter 5.2.1), resulting in a subset of 84,823 answers. In our analysis of expected versus actual replies (Chapter 5.5), we consider the question-answer pairs for which both the question and answer in each pair meet these filtering criteria, resulting in 80,461 pairs.

Labeled dataset of Prime Ministers’ Questions. Here, we provide further details on the labeled dataset of questions asked to Prime Ministers from Bates et al. [2014] used to validate and interpret the framework output. The data consists of 1,413 question-answer pairs (that we could successfully clean and extract from the corpus that the authors shared with us). Questions in the dataset are labeled as *standard* (1,003 questions), *helpful* (205 questions) and *unanswerable* (161 questions). Answers are labeled as *answered* (498 answers), *deferred* (550 answers) and *not answered* (353 answers).

A.2 Crisis counseling conversations

We provide further details on the crisis counseling conversation dataset from Crisis Text Line introduced in Chapter 3. The full dataset consists of over 1.5 million conversations, and was accessed via a fellowship program. Crisis Text Line’s present data access policy is detailed at <https://www.crisistextline.org/data-philosophy/>.

Selection of terms. We use dependency-parse arcs as counselor terms and unigrams as texter terms. For each role, we consider the 5,000 most frequent terms that occurred in at most 50% of counselor or texter messages, respectively.

Framework inputs and parameters. As training data for deriving forwards and backwards characterizations (Chapter 5) and term-level orientation (Chapter 3), we randomly sampled 20% of *counselors* in the data (whose conversations are omitted in subsequent analyses). From the conversations that this subset of counselors is involved in, we consider the texter messages with between 15 and 45 words, and the counselor messages with between 20 and 40 words; we further filtered the subset of counselor messages to include only those that occur

between the texter messages that meet our wordcount cutoff. This results in a collection of 351,862 counselor messages and 599,884 texter messages. We derive a latent context space of 25 dimensions (resulting in 24-dimensional representations, since we remove the first dimension).

To derive skip-representations (Chapter 5.2.3), we considered the same 20% of counselors used derive forwards- and backwards characterizations. Here, we use the counselor messages with between 20 and 40 words, and whose subsequent counselor message also had between 20 and 40 words. This results in a collection of 417,259 messages.

When inferring forwards, backwards and skip types, we consider a subset of the training data consisting of counselor sentences with at least 10 terms. This results in 580,060 sentences used to infer forwards and backwards types, and 696,098 used to infer skip-types.

Data subset used in analyses of conversation structure. To analyze the relation between orientation and conversation structure in Chapter 3, we randomly sample 20% of counselors (different from those included in the training data), and consider the 179,148 conversations that this subset of counselors is involved in, that have at least ten counselor messages. We also use this subset to analyze the relation between forwards/backwards-types and conversation structure in Chapter 5.2.2. Finally, we use this subset in our examination of unexpectedness in Chapter 5.5; here, we consider the 1,258,331 counselor message and texter response pairs, where the texter's message contains at least 10 terms, and where the counselor message contains at least one sentence with at least 5 terms.

Data subset used in analyses of conversation effectiveness. In analyzing the relation between orientation and indicators of conversation effectiveness (Chap-

ter 3.6.4), our aim was to explore conversational behaviour in relatively “typical” conversations, rather than in exceptional cases or those that reflected earlier versions of the training curriculum. As such, we only consider the 234,433 conversations that had at least five counselor messages, were not risk-assessed or prematurely disconnected before being closed by the counselor, and were taken by counselors who joined the platform after January 2017.

A.3 Other datasets

A.3.1 Wikipedia talk page discussions

We provide further details on the Wikipedia talk page discussions datasets examined in Chapter 5.6.1. The training data we used to infer comment types can be found at <https://convokit.cornell.edu/documentation/wiki.html>, and was introduced in Danescu-Niculescu-Mizil et al. [2012]; the data used to analyze awry versus on-track conversations can be found at <https://convokit.cornell.edu/documentation/awry.html>, and is further detailed in Zhang et al. [2018].

Details on deriving comment types. We use two overlapping sets of comments as utterances and as context-utterances: the set of 214,919 comments that receive at least one reply, and the set of 240,436 comments that are replies to a preceding comment. We found that it was preferable to derive separate input representations, and hence separate vocabularies of terms and context-terms, for these two sets. In particular, comments that initiate further discussion in this context often tend to be requests, and exhibit noticeable linguistic differences from comments that respond to these requests.

For both comment sets, we take terms to be dependency-parse arcs with nouns removed (following the same procedure to extract parliamentary question terms from Chapter 2.5); the choice to omit nouns reflects that we want to derive rhetorical information that is agnostic to the topic being discussed. We consider all such terms that occur in at least 50 comments, resulting in vocabularies of 5,022 terms and 5,233 context-terms.

We derive 25-dimensional latent representations using our framework, and infer 6 comment types. We found that clustering *term*-level representations, rather than utterance-level representations (as in the other settings we considered), produced more interpretable output. We suggest this reflects that Wikipedia comments vary greatly in length and structure, such that term-level representations smooth out much of this variation.

A.3.2 US Supreme Court oral arguments

We provide further details on the US Supreme Court Oral Arguments dataset explored in Chapter 5.6.2. The data consists of transcripts originally provided by the Oyez project (<https://www.oyez.org/>) and can be found at <https://convokit.cornell.edu/documentation/supreme.html>.

In this setting, we applied our framework to measure the orientation of justice utterances, using the surrounding lawyer utterances as conversational context. We consider the subset of justice utterances with between 10 and 50 words, and lawyer utterances between 10 and 75 words. We further filter the justice utterance set to contain only utterances whose replies and predecessors met our length cutoffs. This results in a dataset of 91,924 justice and 372,268 lawyer utterances. Both justice and lawyer utterances are represented as dependency-parse

arcs; we consider a vocabulary of 1,240 justice terms and 2,000 lawyer terms that occurred in at least 250 utterances. Finally, we derive 15-dimensional latent representations.

A.3.3 Switchboard Dialog Act Corpus

We provide further details on the Switchboard Dialog Act Corpus, which we explored in Chapter 5.6.3. The dataset was originally presented in Stolcke et al. [2000]. To sidestep some of the challenges of working with transcribed speech data, we consider a processed version of the data, found at <https://convokit.cornell.edu/documentation/switchboard.html>. In particular, we process the data to remove disfluences in utterances, via regular expressions. We also use a rough heuristic to remove backchannels: we ignore utterances shorter than 5 words, and merge utterances by the same speaker that would otherwise be separated by these interjections; we also merge the set of tags that each constituent utterance is labeled with.

To avoid capturing topic-specific information, and to minimize the noise incurred from characterizing rare terms, we curate a vocabulary of 381 unigrams that occur in at least 33 (50%) of the conversation topics and in at least 200 conversations. As input to the framework, we consider the subset of 34,562 utterances with at least 5 terms in the vocabulary, using this set both as utterances and as context-utterances. We derive 15-dimensional latent representations.

APPENDIX B

FURTHER EXAMPLES

We include further examples representative of utterance types inferred by applying the Expected Conversational Context Framework to the parliamentary question periods data, and to the crisis counseling data. In Tables B.1 and B.2, we include further examples of question and answer types in the parliamentary setting. In Tables B.3 and B.4, we include further examples of forwards and backwards types from the counseling setting. As in other examples from the counseling dataset shown in Chapter 5, to preserve the privacy of the conversation participants, we wrote fictional messages based on actual messages in the data, and on examples found in the counselor training curriculum.

<p>Demand for account</p> <p>Question terms: <i>not realize, how [can you] justify, will [you] stop</i></p> <p>Q: Why does the Leader of the House deny the publication of information on [MP spending]?</p> <p>Q: Will the Minister explain why he waited five days after having been informed on the presence of lead-contaminated feedstock before imposing restrictions?</p> <p>Q: Can the Minister explain why she thinks that regional assemblies are better placed to tackle social exclusion than elected and accountable authorities?</p>
<p>Shared concerns</p> <p>Question terms: <i>consider making, draw [attention] to, will [you] undertake</i></p> <p>Q: Will she undertake to see whether the contract can be put forward sooner?</p> <p>Q: Will [the PM] ensure that chatlines are properly regulated?</p> <p>Q: May I draw my hon. Friend's attention to the audit report?</p>
<p>Agreement</p> <p>Question terms: <i>do [you] share, agree with, agree [that we] need</i></p> <p>Q: Does [the Minister] agree that one of the best ways to improve the trade balance is to continue the Governments strong economic policies?</p> <p>Q: Is it not important that the Department continues its excellent work [on] flood defences?</p> <p>Q: Does he agree that taxpayers need to be considered when it comes to aid spending?</p>
<p>Issue update</p> <p>Question terms: <i>update [us] on, what specific, doing [to] help</i></p> <p>Q: What more can the Government do to help [...] disabled people in the work force?</p> <p>Q: What specific action is [the PM] taking to defend the British fishermen in the negotiations?</p> <p>Q: What can she do to ensure that local authorities [work] in partnership with landlords?</p>
<p>Questioning premises</p> <p>Question terms: <i>does [the Minister] believe, is not, will [you] concede</i></p> <p>Q: Is not the Minister aware that Norwich has suffered under this Government?</p> <p>Q: Does the Minister believe that the Film Council's [selection process] is entirely objective?</p> <p>Q: Does he not think that it would be timely for the Modernisation Committee to consider the laptop situation?</p>
<p>Request for assurance</p> <p>Question terms: <i>can [you] reassure, will [you] discuss, will [you] give</i></p> <p>Q: Will she give an assurance that the toll booths will be manned on all occasions?</p> <p>Q: Can he assure those explorations will be subject to rigorous environmental control?</p> <p>Q: Will [the PM] discuss with the Department of Trade whether any further help could be given to our own industry to remain competitive in world markets?</p>
<p>Prompt for comment</p> <p>Question terms: <i>say is, can [you] tell, can [you] say</i></p> <p>Q: Will he tell us who has been appointed to be responsible for green economic growth?</p> <p>Q: Can the Foreign Secretary tell the House how much of the increase is due to enlargement?</p> <p>Q: Can he say when the review of the motorway is likely to be completed?</p>
<p>Accept and propose</p> <p>Question terms: <i>does [the Minister] recognize, be better, is [it] possible</i></p> <p>Q: Does the Chancellor recognise that the debt problem [justifies] a larger rate of increase than the Government are willing to support?</p> <p>Q: Would it not be better to have direct answerability of the authority [via] direct election?</p> <p>Q: Does [the PM] accept that the failure of the last round of trade negotiations was due to the fact that the Japanese could get a much better bargain from each member state separately?</p>

Table B.1: Representative examples of question terms and questions, for each parliamentary question type.

<p>Progress report Answer terms: <i>are reviewing, am prepared, are examining</i> A: We are examining the timetabling of the additional paternity leave [...] A: As I have indicated, we are reviewing the detained fast track.</p>
<p>Statement Answer terms: <i>is not, am sorry, knows</i> A: That is what underpinned my approach to the pre-Budget report, and I am sorry that the Conservatives do not share that view. A: As I have said, we have an independent central Bank and I propose to keep it that way.</p>
<p>Endorsement Answer terms: <i>is right, is important, be interested</i> A: I certainly support the concept of extending unified grading. A: It is much more important to understand what is going on at a local level, as my hon. Friend has done.</p>
<p>Comment Answer terms: <i>believe, referred to, understand</i> A: I am not sure that I go as far as that, but it is important that they be sold correctly. A: I do believe that it would be a good idea to recognise it in the tax system.</p>
<p>Commitment Answer terms: <i>know, will give, continue</i> A: We are committed to seeking an efficient, not-for-profit operator. A: We will continue to support British firms and workers.</p>

Table B.2: Representative examples of answer terms and answers, for each parliamentary answer type.

<p>Risk assessment Terms: <i>would kill, thank [you] for, how [would you] take</i> C: Thank you for sharing that with me, do you have a means for doing it? C: You mentioned you want to take your life, and I want to make sure you're safe. C: Thank you for being honest, I'm wondering if you have a plan for when you would kill yourself?</p>
<p>Service statement Terms: <i>end conversation, if [you] need, i hope</i> C: I am more than glad to talk with you tonight. C: If you're feeling okay tonight, I'm going to end the conversation. C: If you need support again, we are here to help.</p>
<p>Situation comment Terms: <i>everything, a toll, are trying</i> C: It's understandable to feel overwhelmed when everything is hitting you at once. C: It must take a toll to deal with that pressure. C: It sounds like you are trying to deal with a lot that's been out of your control.</p>
<p>Relationship comment Terms: <i>the person, space, with someone</i> C: I'd imagine it's upsetting to hear that she needs space. C: It sounds like a very difficult situation to be in with someone. C: It sounds like you feel awful about arguing with the person you love.</p>
<p>Coping mechanism Terms: <i>some ways, that helps, get [your] mind [off of]</i> C: What sorts of things does your friend do to help you get your mind off of this? C: What are some ways that help you relax? C: Distractions can be great ways to help get your mind off of the sadness.</p>
<p>Support system Terms: <i>talk with, to anyone, get support</i> C: Is there another relative you could talk to about those frustrations? C: Have you shared any of your concerns with your counselor? C: Do you feel like being able to talk with someone would be helpful?</p>
<p>Exploration Terms: <i>more about, missed sorry, can tell</i> C: Do you want to tell me a bit more about your situation? C: I'm sorry, I missed your last message. C: Can you tell me more about what is causing this anger? Note: This type also includes phrases telling a texter a message was mis-sent, perhaps because such comments also tend to occur near the starts of conversations.</p>
<p>Suggestion Terms: <i>a try, option, good idea</i> C: I think this practice could help you feel better. C: It might be worth giving this a try. C: That could be a good option to start taking care of the larger issue.</p>

Table B.3: Representative examples of counselor (C) terms and messages, for each **forwards** type.

<p>Coping mechanism Terms: <i>something else, other things, distract from</i> C: What are other things you could do in your spare time? C: Would you be open to trying something else to relieve the stress? C: Those sound like good activities to help you relax.</p>
<p>Situation comment Terms: <i>through [a] lot, handle [as one] person, extremely</i> C: I'm hearing that you have been through a lot. C: That sounds like a lot for one person to handle. C: It can be extremely frustrating to deal with all of that.</p>
<p>Social comment Terms: <i>by people, supported, who are</i> C: It's disheartening to feel like no one supports you. C: It seems like you do not feel supported by those friends. C: It's tough when the people who are around you don't seem to understand.</p>
<p>Feeling comment Terms: <i>get sense, [i]'m hearing, is normal</i> C: It's understandable to feel that way in your circumstance. C: It is normal to be overwhelmed in a time like this. C: I'm hearing that you feel angry about the situation.</p>
<p>Suggestions Terms: <i>work with, recommend, are interested</i> C: If you are hoping to work on this, I can recommend some other options. C: I can send you a website that could help with this. C: Do you think it's something you are interested in trying?</p>
<p>Relationship comment Terms: <i>you care [about], boyfriend, girlfriend</i> C: I wonder if your boyfriend can see your perspective and the consequences of his actions? C: It sounds like you think she was overreacting about that situation. C: It sounds like you really care about your girlfriend, even though you've been arguing a lot.</p>
<p>Service statement Terms: <i>texted in, ctl, we</i> C: I can chat with you when you're in a crisis. C: We are here for you and we want to help. C: I'm so glad that we were able to help.</p>
<p>Appreciation for disclosure Terms: <i>telling me, suicidal, sharing with</i> C: I want to make sure you're safe - have you been having suicidal thoughts lately? C: Thank you for telling me about that. C: Thanks for being willing to share all of that with me.</p>

Table B.4: Representative examples of counselor (C) terms and messages, for each **backwards** type.

BIBLIOGRAPHY

- Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. Roles for Computing in Social Change. In *Proceedings of FAccT*, 2020.
- Simon Akam. On the hunt with Yorkshire’s virus-detectives. *The Economist*, 2020.
- Ali Alkhatib. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of CHI*, 2021.
- James Allen. *Natural Language Understanding*. Benjamin/Cummings Publishing Company, 1995.
- Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. In *Proceedings of ICWSM*, 2014.
- Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *TACL*, 4, 2016.
- Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, 2008.
- Joshua D. Angrist and Jörn-Steffen Pischke. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2), 2010.
- David C. Atkins, Mark Steyvers, Zac E. Imel, and Padhraic Smyth. Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1), 2014.
- J. Maxwell Atkinson and Paul Drew. *Order in Court*. Springer, 1979.
- Malika Aubakirova and Mohit Bansal. Interpreting Neural Networks to Improve Politeness Comprehension. In *Proceedings of EMNLP*, 2016.
- John L. Austin. *How to Do Things with Words*. Harvard University Press, 1962.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, 2014.
- Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of WWW*, 2021.

- Stephen R. Bates, Peter Kerr, Christopher Byrne, and Liam Stanley. Questions to the Prime Minister: A Comparative Study of PMQs from Thatcher to Cameron. *Parliamentary Affairs*, 67(2), 2014.
- Jo Becker. This Contact Tracer Is Fighting Two Contagions: The Virus and Fear. *The New York Times*, 2020.
- Howard B. Beckman and Richard M. Frankel. The Effect of Physician Behavior on the Collection of Data. *Annals of Internal Medicine*, 101(5), 1984.
- Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of ACL*, 2020.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of FAccT*, 2021.
- Giacomo Benedetto and Simon Hix. The Rejected, the Ejected, and the Dejected: Explaining Government Rebels in the 2001-2005 British House of Commons. *Comparative Political Studies*, 40(7), 2007.
- Shohini Bhattachali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. Automatic Identification of Rhetorical Questions. In *Proceedings of ACL*, 2015.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In *Proceeding of EMNLP*, 2020.
- Graham D. Bodie, Andrea J. Vickery, Kaitlin Cannava, and Susanne M. Jones. The Role of “Active Listening” in Informal Helping Conversations: Impact on Perceptions of Listener Helpfulness, Sensitivity, and Supportiveness and Discloser Emotional Improvement. *Western Journal of Communication*, 79(2), 2015.
- Amber E. Boydston, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. Tracking the development of media frames within and across policy issues. In *Proceedings of the APSA*, 2014.
- David Bracewell, Marc Tomlinson, and Hui Wang. Identification of Social Acts in Dialogue. In *Proceedings of COLING*, 2012.
- Carlos Brito and Judea Pearl. Generalized Instrumental Variables. In *Proceedings of UAI*, 2012.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prfulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, and Tom

- Henighan. Language Models are Few-Shot Learners. In *Proceedings of NeurIPS*, 2020.
- Peter Bull and Pam Wells. Adversarial Discourse in Prime Minister’s Questions. *Journal of Language and Social Psychology*, 31(1), 2012.
- Donna Byron and Amanda Stent. A Preliminary Model of Centering in Dialog. In *Proceedings of ACL*, 1998.
- Dogan Can, David C. Atkins, and Shrikanth S. Narayanan. A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *Proceedings of INTERSPEECH*, 2015.
- Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N’godjigui Junior Diarrassouba, Mark E. Whiting, and Michael S. Bernstein. My Team Will Go On: Differentiating High and Low Viability Teams through Team Interaction. *Proceedings of the ACM on Human-Computer Interaction*, 4(3), 2020.
- Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Sriku-mar. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of ACL*, 2019.
- Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW*, 2010.
- Dallas Card, Justin Gross, Amber Boydston, and Noah Smith. Analyzing Framing through the Casts of Characters in the News. In *Proceedings of EMNLP*, 2016.
- Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proceedings of EMNLP*, 2019.
- Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of SIGDIAL*, 2020.
- Daniel N. Chester and Nona Bowring. *Questions in Parliament*. Cambridge University Press, 1962.
- Perna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. In *Proceedings of CHI*, 2020.

- Munmun De Choudhury and Emre Kiciman. The Language of Social Support in Social Media and its Effect on Suicidal Ideation Risk. In *Proceedings of ICWSM*, 2017.
- Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.
- Mark G Core and James F Allen. Coding Dialogs with the DAMSL Annotation Scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.
- Philip Cowley. *The Rebels: How Blair Mislaid His Majority*. Politico's, 2005.
- Philip Cowley. Arise, Novice Leader! The Continuing Rise of the Career Politician in Britain. *Politics*, 32(1), 2012.
- Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of WWW*, 2012.
- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*, 2013a.
- Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of WWW*, 2013b.
- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 1990.
- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, H. Hill, Dan Jurafsky, and Tatsunori Hashimoto. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. In *Proceedings of ACL*, 2021.
- Stephen F. Derose, Mark A. Schuster, Jonathan E. Fielding, and Steven M. Asch. Public Health Quality Measurement: Concepts and Challenges. *Annual Review of Public Health*, 23(1), 2002.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, 2019.
- Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. Meeting Recorder Project: Dialog Act Labeling Guide. Technical report, 2004.
- Paul Drew, Traci Walker, and Richard Ogden. Self-repair and action construction. In *Conversational Repair and Human Understanding*. 2011.

- Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. How to Make Causal Inferences Using Texts. Working Paper, 2018.
- Andrew C. Eggers and Arthur Spirling. Ministerial Responsiveness in Westminster Systems: Institutional Choices and House of Commons Debate, 1832–1915. *American Journal of Political Science*, 58(4), 2014.
- Charles R. Epp, Steven Maynard-Moody, and Donald P. Haider-Markel. *Pulled Over: How Police Stops Define Race and Citizenship*. University of Chicago Press, 2014.
- Frederick Erickson. Rhythm in Discourse. In *The Encyclopedia of Applied Linguistics*. 2012.
- Cat Ferguson. With therapy app Talkspace, patient anonymity and safety collide. <https://www.theverge.com/2016/12/19/14004442/talkspace-therapy-app-reviews-patient-safety-privacy-liability-online>, 2016.
- John Rupert Firth. *Papers in Linguistics, 1934-1951*. 1957.
- Molly Fischer. The Mental Health Therapy-App Fantasy. *The Cut*, 2021.
- Sarah Ford, Lesley Fallowfield, and Shôn Lewis. Doctor-patient interactions in oncology. *Social Science & Medicine*, 42(11), 1996.
- Jacques Gaume, Nicolas Bertholet, Mohamed Faouzi, Gerhard Gmel, and Jean-Bernard Daeppen. Counselor motivational interviewing skills and young adult change talk articulation during brief motivational interventions. *Journal of Substance Abuse Treatment*, 2010.
- Eric Gilbert. Phrases That Signal Workplace Hierarchy. In *Proceeding of CSCW*, 2012.
- Erving Goffman. On Face-Work: An Analysis of Ritual Elements in Social Interaction. *Psychiatry*, 18(3), 1955.
- Erving Goffman. Replies and responses. *Language in society*, 5(3), 1976.
- Simon B. Goldberg, Nikolaos Flemotomos, Victor R. Martinez, Michael J. Tanana, Patty B. Kuo, Brian T. Pace, Jennifer L. Villatte, Panayiotis G. Georgiou, Jake Van Epps, Zac E. Imel, Shrikanth S. Narayanan, and David C. Atkins. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*, 67(4), 2020.
- Sandra Gonzalez-Bailon, Andreas Kaltenbrunner, and Rafael E Banchs. The structure of political discussion networks: A model for the analysis of online deliberation. *Journal of Information Technology*, 25(2), 2010.

- Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9(6), 1995.
- Ben Green and Salomé Viljoen. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of FAccT*, 2020.
- Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 2013.
- Justin Grimmer, Solomon Messing, and Sean J. Westwood. How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation. *The American Political Science Review*, 106(4), 2012.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2), 1995.
- John J. Gumperz. *Discourse Strategies*. Cambridge University Press, 1982.
- Poonam Gupta and Vishal Gupta. A Survey of Text Question Answering Techniques. *International Journal of Computer Applications*, 53(4), 2012.
- William L. Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Loyalty in Online Communities. In *Proceedings of ICWSM*, 2017.
- Sanda M. Harabagiu. Questions and Intentions. In *Advances in Open Domain Question Answering*. Springer Netherlands, 2008.
- Sanda M. Harabagiu, Steven J. Maierano, and Marius A. Pasca. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(3), 2003.
- Janet E. Helms and Donelda Ann Cook. *Using Race and Culture in Counseling and Psychotherapy: Theory and Process*. Allyn & Bacon, 1999.
- Alexa Hepburn and Galina B Bolden. The conversation analytic approach to transcription. In *The Handbook of Conversation Analysis*. Wiley Online Library, 2013.
- John Heritage. Current developments in conversation analysis. In *Conversation: An Interdisciplinary Perspective*. 1989.
- John Heritage. *Garfinkel and Ethnomethodology*. Wiley, 1991.

- John Heritage and Steven Clayman. *Talk in Action: Interactions, Identities, and Institutions*. John Wiley & Sons, 2011.
- Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
- Clara E. Hill and Emilie Y. Nakayama. Client-centered therapy: Where has it been and where is it going? A comment on Hathaway (1948). *Journal of Clinical Psychology*, 2000.
- Elliott M. Hoey and Robin H. Kendrick. Conversation analysis. In *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*. 2017.
- Jon Houck. Motivational Interviewing Skill Code (MISC) 2.1. 2008.
- Dirk Hovy and Diyi Yang. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of NAACL*, 2021.
- Yuheng Hu, Ali Tafti, and David Gal. Read This, Please? The Role of Politeness in Customer Service Engagement on Social Media. In *Proceedings of HICSS*, 2019.
- Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. It doesn't hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology*, 113(3), 2017.
- Clayton J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of ICWSM*, 2014.
- Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political Ideology Detection Using Recursive Neural Networks. In *Proceedings of ACL*, 2014.
- Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. Talking to the crowd: What do people react to in online discussions? In *Proceedings of EMNLP*, 2015.
- Gail Jefferson. Transcript notation. In *Structures of Social Action: Studies in Conversation Analysis*. 1984.
- Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding similar questions in large question and answer archives. In *Proceedings of CIKM*, 2005.
- Neil P. Jones, Alison A. Papadakis, Caitlin M. Hogan, and Timothy J. Strauman. Over and over again: Rumination, reflection, and promotion goal failure and their interactive effects on depressive symptoms. *Behaviour Research and Therapy*, 47(3), 2009.

- Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard DAMSL Coders Manual, 1997.
- Christopher Kam. *Party Discipline and Parliamentary Politics*. Cambridge University Press, 2009.
- Alan E. Kazdin. Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, 3, 2007.
- Greg P Kearsley. Questions and question asking in verbal discourse: A cross-disciplinary review. *Journal of Psycholinguistic Research*, 5(4), 1976.
- Thomas K Landauer and Susan T Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2), 1997.
- Juliette Landphair and Teri Preddy. More than talk: Co-Rumination among college students. *About Campus*, 17(3), 2012.
- Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, 2019.
- Beth L. Leech. Asking Questions: Techniques for Semistructured Interviews. *Political Science & Politics*, 35(4), 2002.
- Wendy G. Lehnert. *The Process of Question Answering*. PhD thesis, 1977.
- Tom Louwrese. Mechanisms of Issue Congruence: The Democratic Party Mandate. *West European Politics*, 35(6), 2012.
- Steven Lytinen and Noriko Tomuro. The use of question types to match questions in FAQFinder. In *AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002.
- James Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.
- William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Interdisciplinary Journal for the Study of Discourse*, 8(3), 1988.
- Elizabeth A. McGlynn. Six Challenges In Measuring The Quality Of Health Care. *Health Affairs*, 16(3), 1997.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of ICLR*, 2013.

- Brian L. Mishara, François Chagnon, Marc S. Daigle, Bogdan Balan, Sylvaine Raymond, Isabelle Marcoux, Cécile Bardon, Julie K. Campbell, and Alan D. Berman. Which helper behaviors and intervention styles are related to better short-term outcomes in telephone crisis intervention? Results from a Silent Monitoring Study of Calls to the U.S. 1-800-SUICIDE Network. *Suicide & Life-Threatening Behavior*, 37(3), 2007.
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4), 2008.
- Johanna D. Moore and Cécile Paris. Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, 19(4), 1993.
- Dong Nguyen, A. Seza Dođruöz, Carolyn P. Rosé, and Franciska de Jong. Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3), 2015.
- Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3), 2014.
- Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Conversational Markers of Constructive Discussions. In *Proceedings of NAACL*, 2016.
- Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game. In *Proceedings of ACL*, 2015a.
- Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. In *Proceedings of WWW*, 2015b.
- Susan Nolen-Hoeksema, Blair E. Wisco, and Sonja Lyubomirsky. Rethinking Rumination. *Perspectives on Psychological Science*, 3(5), 2008.
- Martin Nystrand, Lawrence L. Wu, Adam Gamoran, Susie Zeiser, and Daniel A. Long. Questions in Time: Investigating the Structure and Dynamics of Unfolding Classroom Discourse. *Discourse Processes*, 35(2), 2003.
- Myle Ott, Claire Cardie, and Jeff Hancock. Estimating the Prevalence of Deception in Online Review Communities. In *Proceedings of WWW*, 2012.
- Grant Packard, Sarah G. Moore, and Brent McFerran. (I'm) Happy to Help (You): The Impact of Personal Pronoun Use in Customer–Firm Interactions. *Journal of Marketing Research*, 55(4), 2018.

- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, 2002.
- Sungjoon Park, Donghyun Kim, and Alice Oh. Conversation Model Fine-Tuning for Classifying Client Utterances in Counseling Dialogues. In *Proceedings of NAACL*, 2019.
- Umashanthi Pavalanathan and Jacob Eisenstein. Emoticons vs. Emojis on Twitter: A Causal Inference Approach. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*, 2015.
- Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4), 1995.
- Judea Pearl. On the Testability of Causal Models with Latent and Instrumental Variables. In *Proceedings of UAI*, 2013.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. Understanding and Predicting Empathic Behavior in Counseling Therapy. In *Proceedings of ACL*, 2017.
- Verónica Pérez-Rosas, Xuotong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. Analyzing the Quality of Counseling Conversations: The Tell-Tale Signs of High-quality Counseling. In *Proceedings of LREC*, 2018.
- Anthony R. Pisani, Nitya Kanuri, Bob Filbin, Carlos Gallo, Madelyn Gould, Lisa S. Lehmann, Robert Levine, John E. Marcotte, Brian Pascal, David Rousseau, Shairi Turner, Shirley Yen, and Megan L. Ranney. Protecting User Privacy and Rights in Academic Data-Sharing Partnerships: Principles From a Pilot Program at Crisis Text Line. *Journal of Medical Internet Research*, 21(1), 2019.
- Hanna F. Pitkin. *The Concept of Representation*. University of California Press, 1967.
- Vinodkumar Prabhakaran and Owen Rambow. Written Dialog and Social Power: Manifestations of Different Types of Power in Dialog Behavior. In *Proceedings of IJCNLP*, 2013.
- Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, Prateek Verma, Nelson Morgan, Jennifer L. Eberhardt, and Dan Jurafsky. Detecting Institutional Dialog Acts in Police Traffic Stops. *Transactions of the Association for Computational Linguistics*, 6, 2018.
- Ellen F. Prince. Toward a taxonomy of given-new information. In *Radical Pragmatics*. Academic Press, 1981.

- Sven-Oliver Proksch and Jonathan B. Slapin. Parliamentary questions and oversight in the European Union. *European Journal of Political Research*, 50(1), 2011.
- Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. Identifying Rhetorical Questions in Social Media. In *Proceedings of AAIL*, 2016.
- Sujith Ravi, Bo Pang, Vibhor Rastogi, and Ravi Kumar. Great Question! Question Quality in Community Q&A. In *ICWSM*, volume 14, 2014.
- Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLOS One*, 11, 2016.
- Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised Modeling of Twitter Conversations. In *Proceedings of NAACL*, 2010.
- Carl R. Rogers. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2), 1957.
- Todd Rogers and Michael I. Norton. The artful dodger: Answering the wrong question the right way. *Journal of Experimental Psychology*, 17(2), 2011.
- Stephen Rollnick and William R. Miller. What is Motivational Interviewing? *Behavioural and Cognitive Psychotherapy*, 23(4), 1995.
- Amanda J. Rose, Wendy Carlson, and Erika M. Waller. Prospective Associations of Co-Rumination with Friendship and Emotional Adjustment: Considering the Socioemotional Trade-Offs of Co-Rumination. *Dev Psychol*, 43(4), 2007.
- Paul R. Rosenbaum. *Design of Observational Studies*. Springer, 2010.
- Sara Rosenthal and Kathleen McKeown. I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions. In *Proceedings of SIGDIAL*, 2015.
- Gilbert Ryle. Abstractions. *Dialogue (Canadian Philosophical Review)*, 1(1), 1962.
- Niharika Sachdeva and Ponnurangam Kumaraguru. Call for Service: Characterizing and Modeling Police Response to Serviceable Requests on Facebook. In *Proceedings of CSCW*. ACM, 2017.
- Harvey Sacks. Lecture Seven: On Questions. *Human Studies*, 12(3/4), 1989a.
- Harvey Sacks. Lecture One: Rules of Conversational Sequence. *Human Studies*, 12(3/4), 1989b.
- Harvey Sacks. *Lectures on Conversation*. Blackwell, 1992.

- Koustuv Saha and Amit Sharma. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *Proceeding of ICWSM*, 2020.
- Jonathan Sandoval, Amy Nicole Scott, and Irene Padilla. Crisis counseling: An overview. *Psychology in the Schools*, 46(3), 2009.
- Emanuel Schegloff. Discourse as an interactional achievement: Some uses of ‘uh huh’ and other things that come between sentences. In *Analyzing Discourse: Text and Talk, Georgetown University Roundtable on Languages and Linguistics*. 1982.
- Emanuel A. Schegloff. Sequencing in Conversational Openings. *American Anthropologist*, 70(6), 1968.
- Emanuel A. Schegloff. Some sources of misunderstanding in talk-in-interaction. *Linguistics*, 25(287), 1987.
- Emanuel A. Schegloff and Harvey Sacks. Opening up Closings. *Semiotica*, 8(4), 1973.
- Deborah Schiffrin. *Approaches to Discourse: Language as Social Interaction*. Wiley, 1994.
- John R. Searle. A classification of illocutionary acts. *Language in Society*, 5(1), 1976.
- Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of AAAI*, 2016.
- Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceeding of EMNLP*, 2020.
- Eva Sharma and Munmun De Choudhury. Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. In *Proceedings of CHI*, 2018.
- Anna Shtok, Gideon Dror, Yoelle Maarek, and Idan Szpektor. Learning from the Past: Answering New Questions with Past Answers. In *Proceedings of WWW*, 2012.
- Arthur Spirling and Iain McLean. UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons. *Political Analysis*, 15(1), 2007.
- Dhanya Sridhar and Lise Getoor. Estimating Causal Effects of Tone in Online Debates. In *Proceedings of IJCAI*, 2019.

- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3), 2000.
- Michael Strube. Never Look Back: An Alternative to Centering. In *Proceedings of ACL*, 1998.
- Lucy A. Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, 1987.
- Viren Swami. Mental Health Literacy of Depression: Gender Differences and Attitudinal Antecedents in a Representative British Sample. *PLOS ONE*, 7(11), 2012.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu, and Lillian Lee. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*, 2016.
- Michael Tanana, Kevin A. Hallgren, Zac E. Imel, David C. Atkins, and Vivek Srikumar. A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment*, 65, 2016.
- Yla R Tausczik and James W Pennebaker. Improving Teamwork Using Real-Time Language Feedback. In *Proceedings of CHI*, 2013.
- Jenny Thomas. Cross-Cultural Pragmatic Failure. *Applied Linguistics*, 4(2), 1983.
- Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, 2006.
- Terence J. G. Tracey, Bruce E. Wampold, James W. Lichtenberg, and Rodney K. Goodyear. Expertise in psychotherapy: An elusive goal? *The American Psychologist*, 69(3), 2014.
- Christoph Treude, Ohad Barzilay, and Margaret-Anne Storey. How do programmers ask and answer questions on the web? In *Proceedings of ICSE*, 2011.
- Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect. *PNAS*, 114(25), 2017.
- Marilyn Walker and Steve Whittaker. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In *Proceedings of ACL*, 1990.

- Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of ACL*, 2019.
- Zhao Wang and Aron Culotta. When do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception using Individual Treatment Effect Estimation. In *Proceedings of AACL*, 2019.
- Paul Webb and David M. Farrell. Party members and ideological change. In *Critical Elections: British Parties and Voters in Long-Term Perspective*. Sage Publications, 1999.
- Bonnie Lynn Webber. Computational Perspectives on Discourse and Dialog. In *The Handbook of Discourse Analysis*. Wiley, 2001.
- Harry Weger, Gina R. Castle, and Melissa C. Emmett. Active Listening in Peer Interviews: The Influence of Message Paraphrasing on Perceptions of Listening Skill. *International Journal of Listening*, 24(1), 2010.
- Jerry Weissman. When Someone Asks You a Question, Respond. *Harvard Business Review*, 2012.
- Joseph Weizenbaum. ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 26(1), 1983.
- Terry Winograd. What Does It Mean to Understand Language? *Cognitive Science*, 4(3), 1980.
- Ludwig Wittgenstein. *Philosophical Investigations*. Wiley, 1953.
- Diyi Yang, Miaomiao Wen, and Carolyn Penstein Rosé. Weakly Supervised Role Identification in Teamwork Interactions. In *Proceedings of ACL*, 2015.
- Diyi Yang, Robert Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. Seekers, Providers, Welcomers, and Storytellers: Modeling Social Roles in Online Health Communities. In *Proceedings of CHI*, 2019.
- Amy Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of ICWSM*, 2017a.
- Amy X. Zhang. Building Systems to Improve Online Discussion. In *CSCW '18 Companion*, 2018.
- Justine Zhang and Cristian Danescu-Niculescu-Mizil. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proceedings of ACL*, 2020.

- Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational Flow in Oxford-style Debates. In *Proceedings of NAACL*, 2016.
- Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. Asking too Much? The Rhetorical Role of Questions in Political Discourse. In *Proceedings of EMNLP*, 2017b.
- Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of ACL*, 2018.
- Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. Finding Your Voice: The Linguistic Development of Mental Health Counselors. In *Proceedings of ACL*, 2019.
- Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. Quantifying the Causal Effects of Conversational Tendencies. *Proceedings of the ACM on Human-Computer Interaction*, 4(2), 2020.
- Wei E. Zhang, Quan Z. Sheng, Jey Han Lau, and Ermyas Abebe. Detecting Duplicate Posts in Programming QA Communities via Latent Semantics and Association Rules. In *Proceedings of WWW*, 2017c.