PREDICTING SOCIAL DYNAMICS IN INTERACTIONS USING KEYSTROKE PATTERNS

ADAM GOODKIND

DEPARTMENT OF COMMUNICATION STUDIES

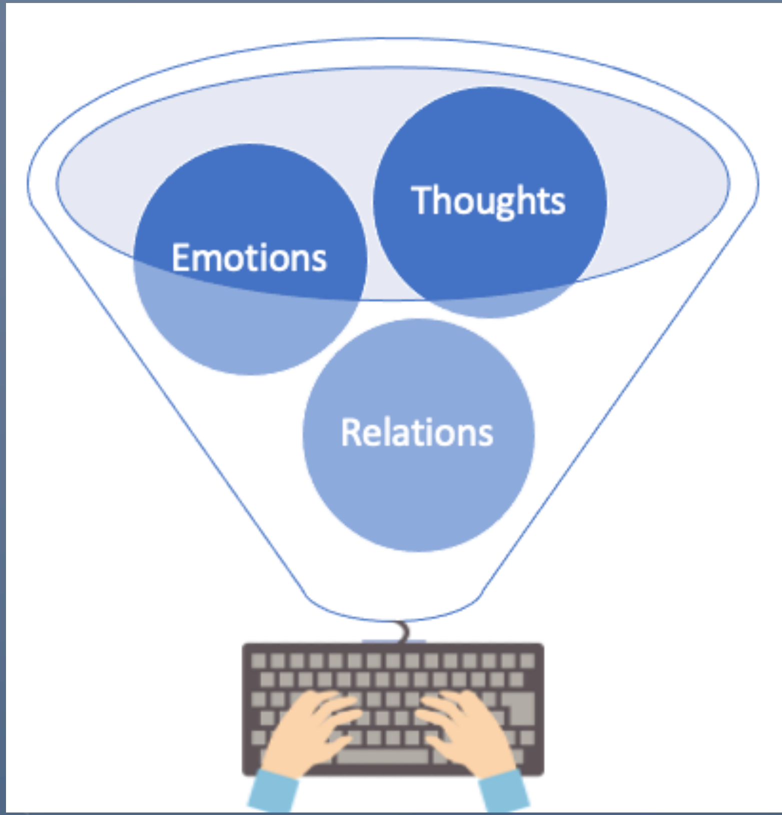
PHD PROGRAM IN MEDIA, TECHNOLOGY, AND SOCIETY

COMMITTEE: PROF. DARREN GERGLE (CHAIR), PROF. ANNE MARIE PIPER, PROF. DAVID-GUY BRIZAN

OUTLINE

1. Motivation
2. Research Questions
3. Data Collection
4. Background Work
5. Studies 1, 2, and 3
6. Future Directions and Possibilities

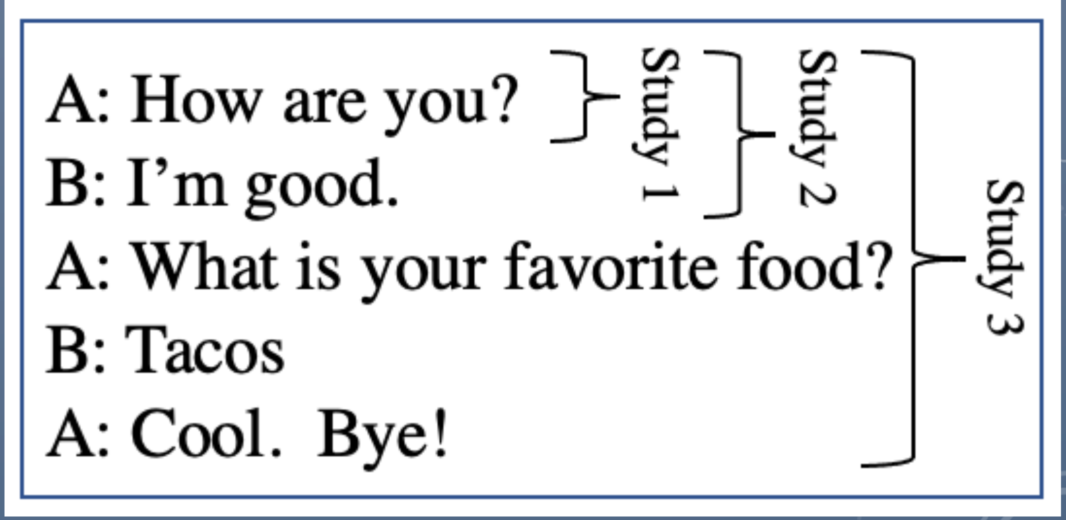
OVERALL – MOTIVATION



- Advance *affective computing* by understanding not just the literal words of a user, but their emotional content as well (Picard, 2000)
- Make text-based conversations more multi-dimensional
- Improve experiences like virtual healthcare (telehealth) and remote work

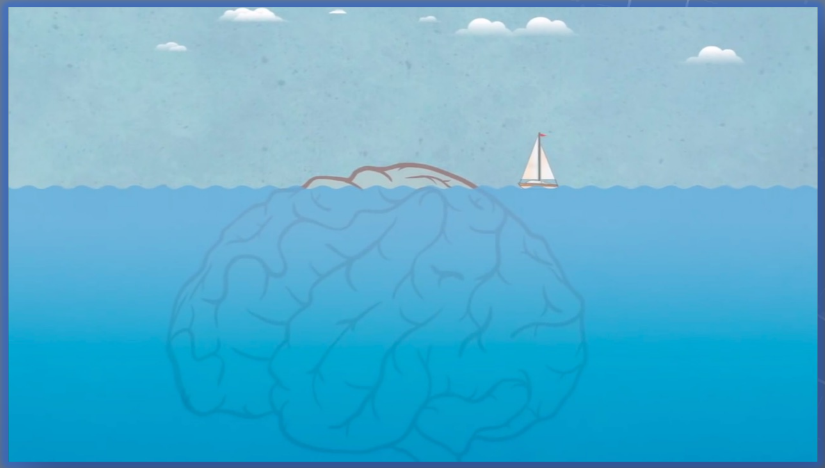
OVERALL – INTRODUCTION

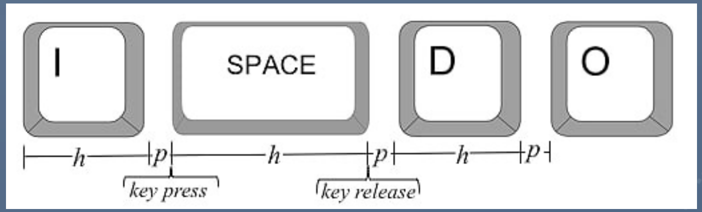
- Conversation analysis at 3 levels
  - Study 1 – Individual utterances
  - Study 2 – Adjacency pairs
  - Study 3 – Entire conversation
- Model the relationship between underlying intentions and keystroke timing

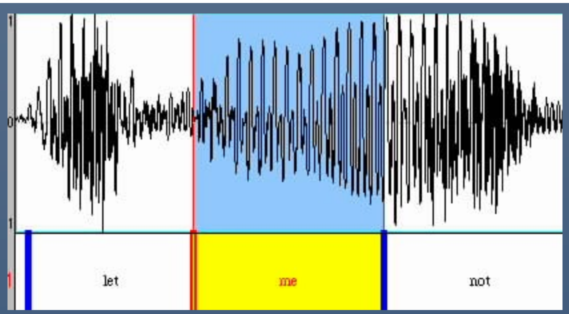


KEYSTROKE DYNAMICS

- *Keystroke dynamics* - detailed timing information about typing, when every key was pressed and released, to understand the manner and rhythm of keystroke production
- Why is it interesting?
    - Language production is a window onto the mind
    - Typing is precise and relatively easy to measure as compared to speech







OVERALL – RESEARCH QUESTIONS

Study 1	Can keystrokes detect the function of an utterance, e.g., whether it's functioning to clarify previous context or advance the conversation?

OVERALL – RESEARCH QUESTIONS

Study 1	Can keystrokes detect the function of an utterance, e.g., whether it's functioning to clarify previous context or advance the conversation?
Study 2	Can keystrokes detect sentiment changes between messages? Are keystrokes sensitive to the sentiment of a specific utterance and the overall opinions?

OVERALL – RESEARCH QUESTIONS

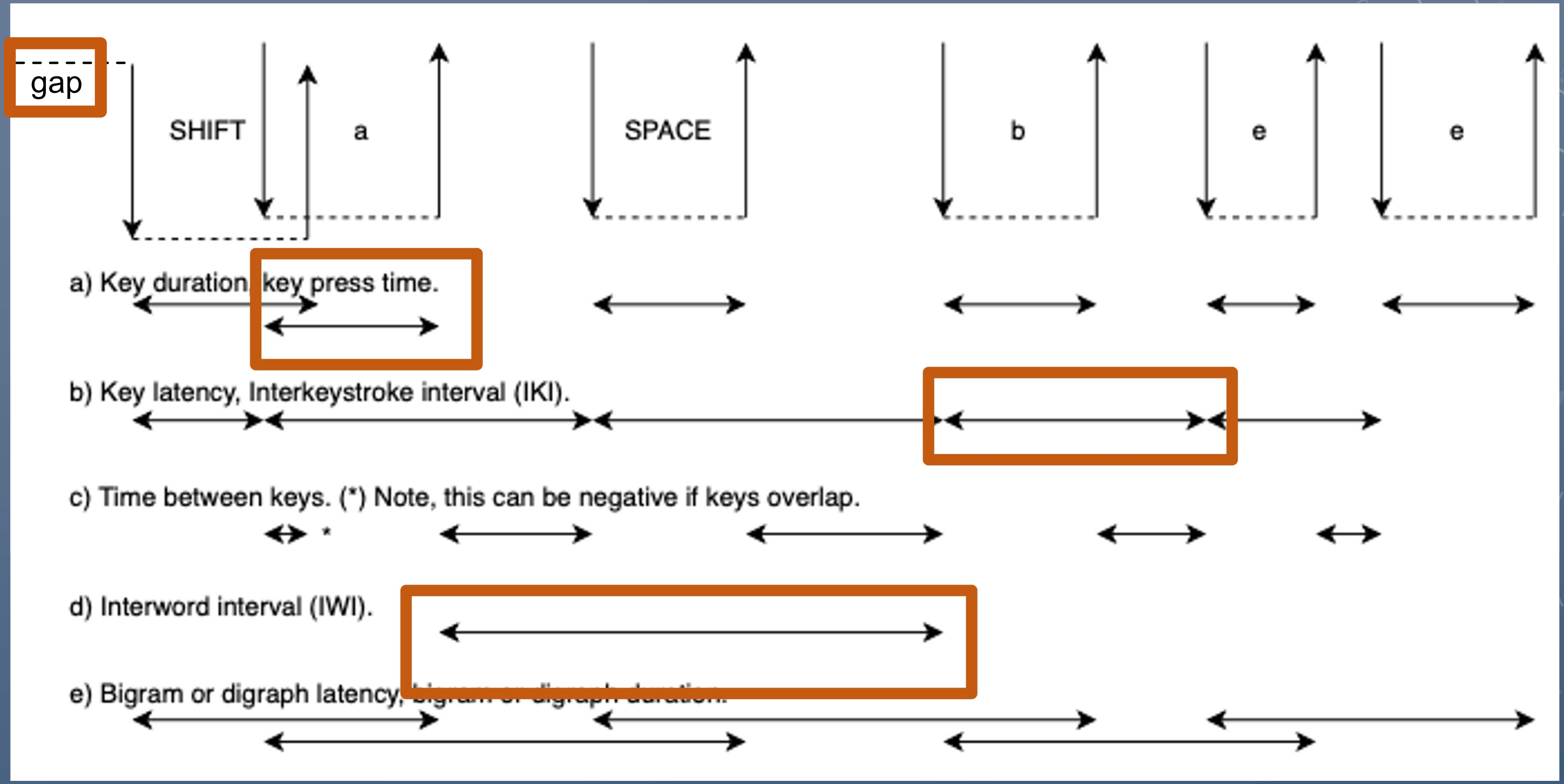
Study 1	Can keystrokes detect the function of an utterance, e.g., whether it's functioning to clarify previous context or advance the conversation?
Study 2	Can keystrokes detect sentiment changes between messages? Are keystrokes sensitive to the sentiment of a specific utterance and the overall opinions?
Study 3	Can keystrokes predict when users feel a low level of rapport with their partner?

KEYSTROKE FEATURES

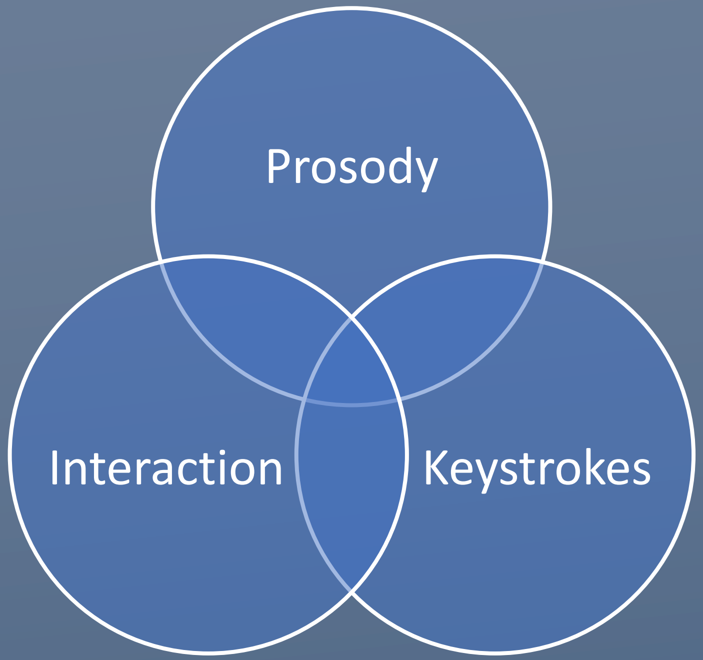
“A bee”

KEYSTROKE FEATURES

“A bee”



BACKGROUND WORK



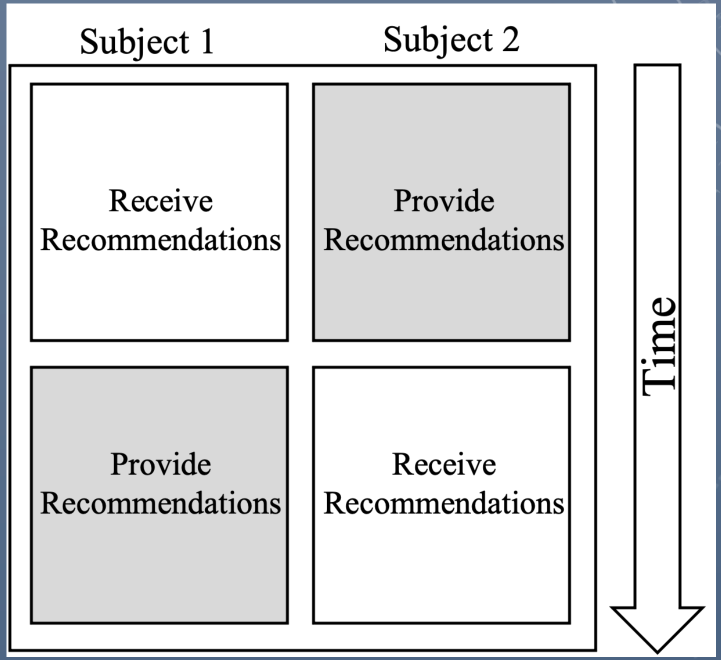
- *Speech prosody* - the patterns of stress and intonation in a language
- Prosody is determined by a number of social factors (Pierrehumbert & Hirschberg, 1990)
- The vast majority of prosody-related work studies *explicit* prosody
- Study typing using *implicit* or *silent* prosody (Fodor, 2002)
- Keystroke timing has been shown correspond to speech timing at both the syllable level and syntactic unit level (Ballier, et al., 2019; Goodkind & Rosenberg, 2015; Plank, 2016)
- My thesis looks at keystrokes as an element of an interaction, and how this reflects not only the user themselves, but the relationship between partners

DATA COLLECTION

Goal: Elicit strong opinions in a conversation

Procedure:

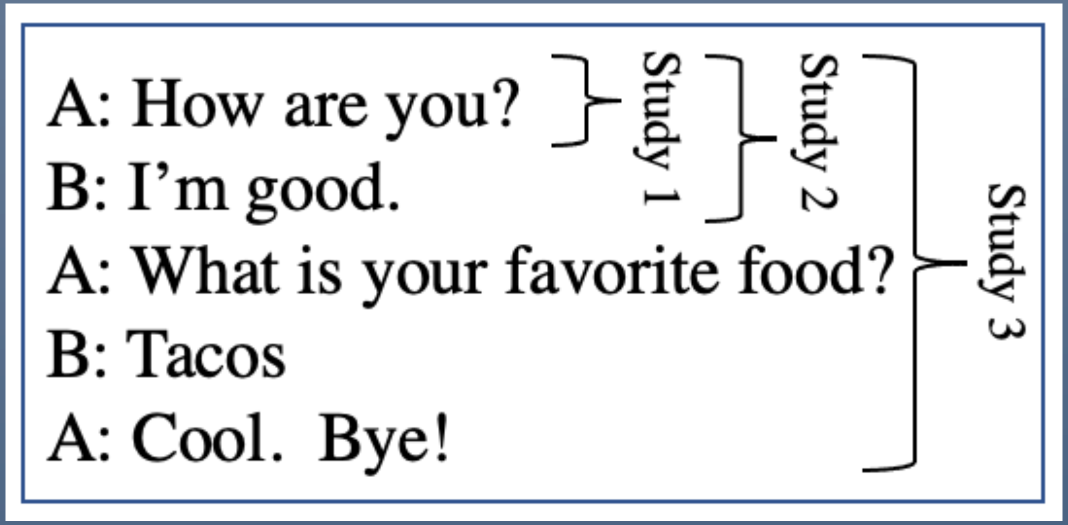
- Discussed movie and TV show recommendations for 16 minutes
  - 1st half: Subject 1 received recommendations from Subject 2
  - 2nd half: Switched roles, prompted to discuss different genre
- Followed by questionnaire asking participant to rate aspects of their partner as well as the overall conversation



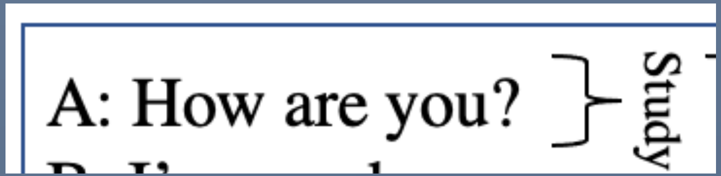
Dataset

- 102 conversations
- ~4,800 messages
- ~327,000 keystrokes

STUDY 1 – DIALOGUE ACTS



STUDY 1 – DIALOGUE ACTS



STUDY 1 – DIALOGUE ACTS BACKGROUND

- Models the conversational function an utterance can perform (Ivanovic, 2005)

  Albert: *She works at Apple.*

Backward	**Forward**
Beth: *Who works at Apple?*	Beth: *And she enjoys kayaking.*

- Different dialogue acts have different amounts of cognitive complexity (Gnjatović, 2013)
- Better dialogue act classification can lead to better human-computer interactions, such as improved experiences with chatbots (Bawden et al., 2016)

STUDY 1 – DIALOGUE ACTS METHODOLOGY

- Dialogue act classification performed in 2 ways
  - Automatically classified
    - I used the `DialogTag` library
  - Manually coded (considered “gold standard”)
    - Performed by a research assistant and me
- Approximately 15% of labels were different

STUDY 1 – DIALOGUE ACTS
EXP. 1A – DIFFERENTIATING DIALOGUE ACTS

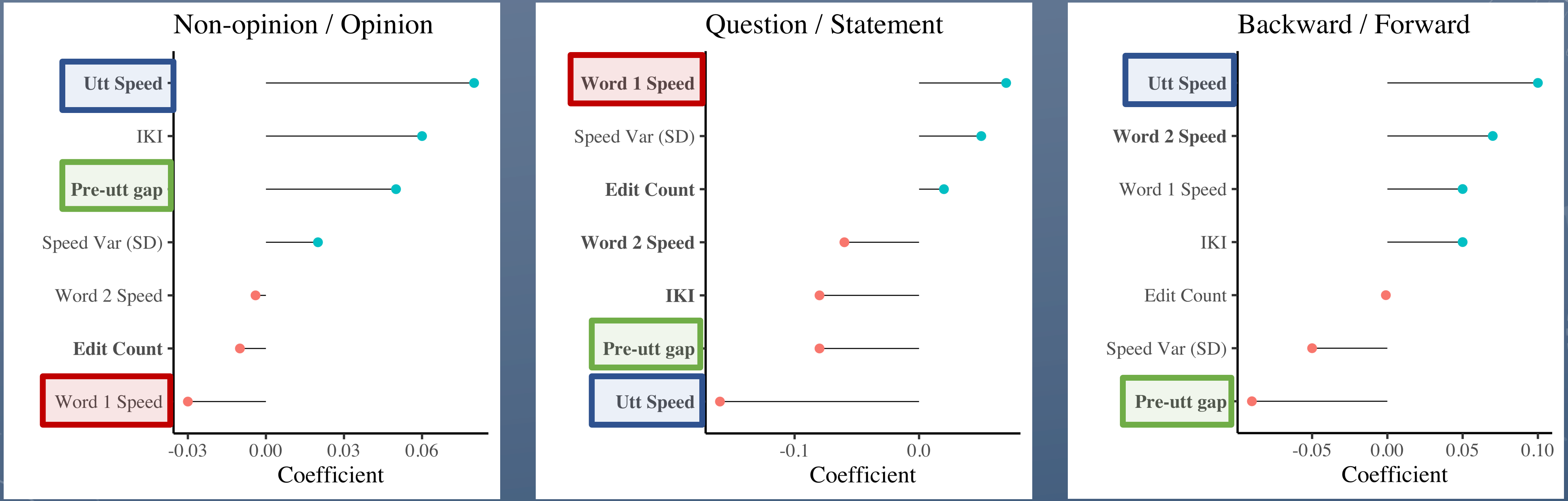
RQ 1a. Can typing patterns predict differences in pairs of dialogue acts, where each member of the pair would require a very different response?

- Binary classifications
  - Non-opinion/Opinion
  - Question/Statement
  - Backward/Forward

- Keystroke metrics
  - Pre-utterance gap
  - Overall mean typing speed
  - Overall typing speed variability (SD)
  - Edit count
  - Word 1 and 2 typing speeds
    - Make early predictions?
  - Various interactions

$$dialogue\_act\_binary \sim keystroke\_metric_1 + \ldots + keystroke\_metric_n + (1 \mid subject)$$

STUDY 1 – DIALOGUE ACTS
EXP. 1A – RESULTS



STUDY 1 – DIALOGUE ACTS
EXP. 1B – CONSISTENCY OF TYPING METRICS WITHIN DA

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

- Unlike in Exp. 1a, typing metrics do not need to be unique (just consistent within a DA)
- Used same features as Exp. 1a, and all DAs
- But flipped dependent and independent variables

$$keystroke\_metric \sim dialogue\_act_1^n + (1 \mid word\_count)$$

STUDY 1B – DIALOGUE ACTS
EXP 1B - RESULTS

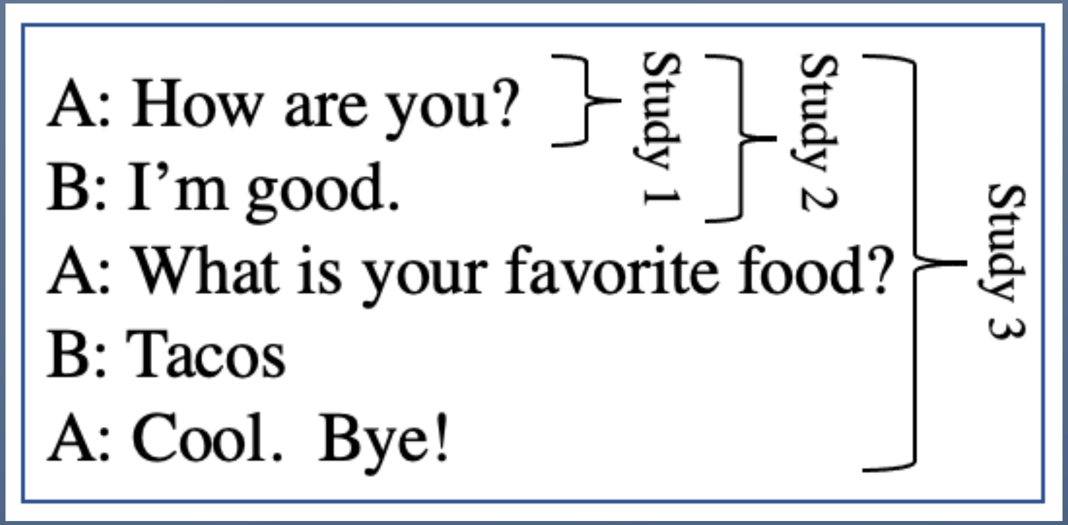
Dependent Variable	Dialogue act
Word count	19.57****
Utterance speed	9.55****
Edit count	6.29****
Speed variability	5.09****
Pre-utterance gap	3.89****
Word 1 - word 2 gap	1.87+
Word 1 speed	2.53*
Word 2 speed	4.45****

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

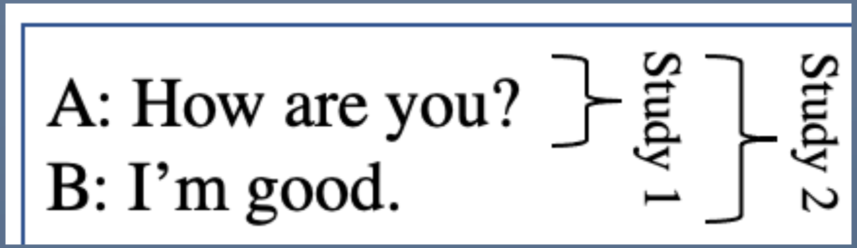
STUDY 1 – DIALOGUE ACTS
RESEARCH QUESTIONS REVISITED

- RQ 1a. Can typing patterns predict differences in pairs of dialogue acts, where each member of the pair would require a very different response?
  - Yes
  - Differentiation of opinions and non-opinions is especially useful
- RQ 1b. Does each dialogue act have a consistent set of typing patterns associated with it?
  - Maybe
  - Supports the notion that DAs differ in cognitive complexity

STUDY 2 – SENTIMENT AND OPINIONS

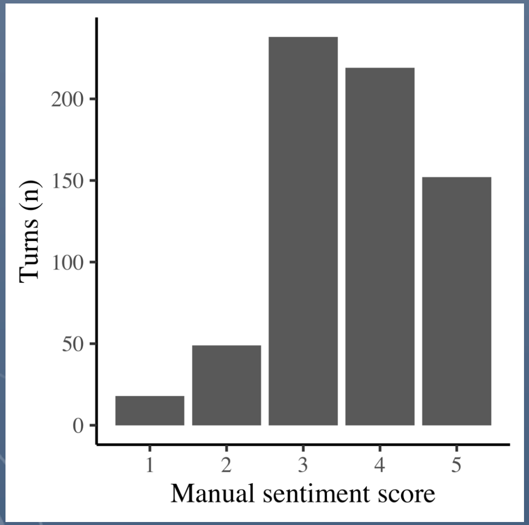


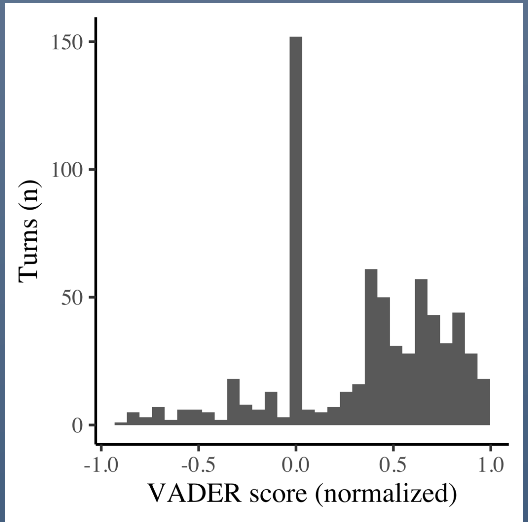
STUDY 2 – SENTIMENT AND OPINIONS


A: How are you?
B: I'm good.
Study 1
Study 2

SENTIMENT IN THE DATA

- 2 ways of labeling sentiment
  - Manually with human annotators (“gold standard”)
  - Algorithmically (used **VADER**)





	Following Utt		
Current Utt	Negative	Neutral	Positive
Negative	**65%**	9%	26%
Neutral	12%	**68%**	20%
Positive	10%	7%	**83%**

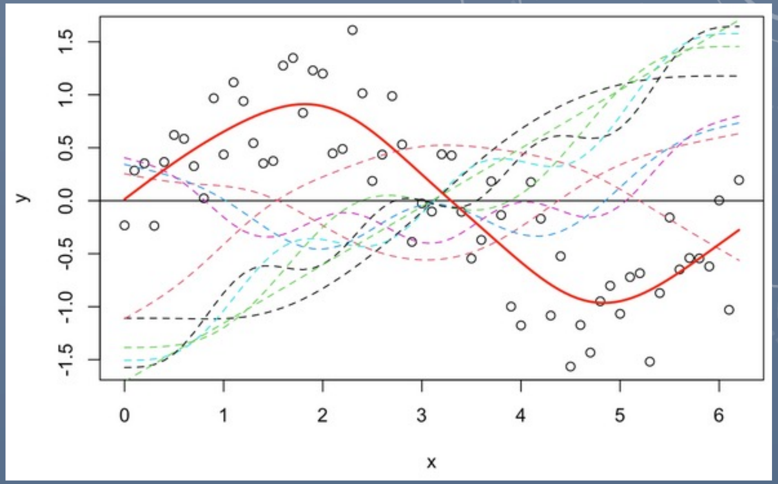
- Utterance sentiment in conversations is not independent, but is simultaneously sensitive to individual-, group-, and network-level properties (Gergle, 2017; Kenny et al., 2020)

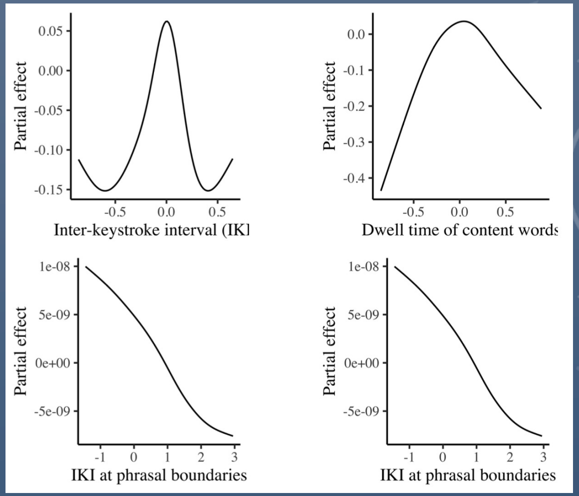
GENERALIZED ADDITIVE MODELS: GAMs

- Generalized Additive Models (GAMs) have been used for complex sentiment detection from scant data (Qi & Li, 2014)
- Linear models ($y \sim x\beta$), but with functions instead of coefficients

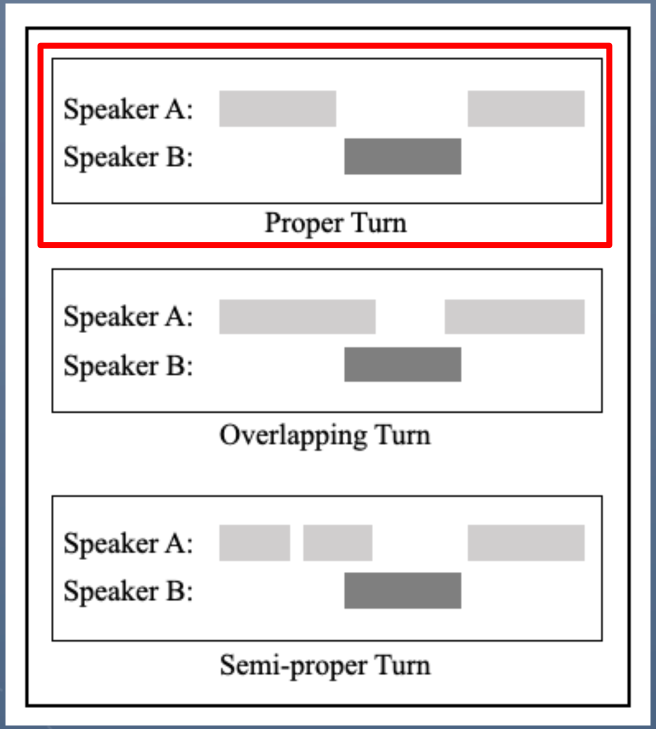
$$y = \beta_0 + x_1\beta_1 + \varepsilon, \Rightarrow g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2)$$

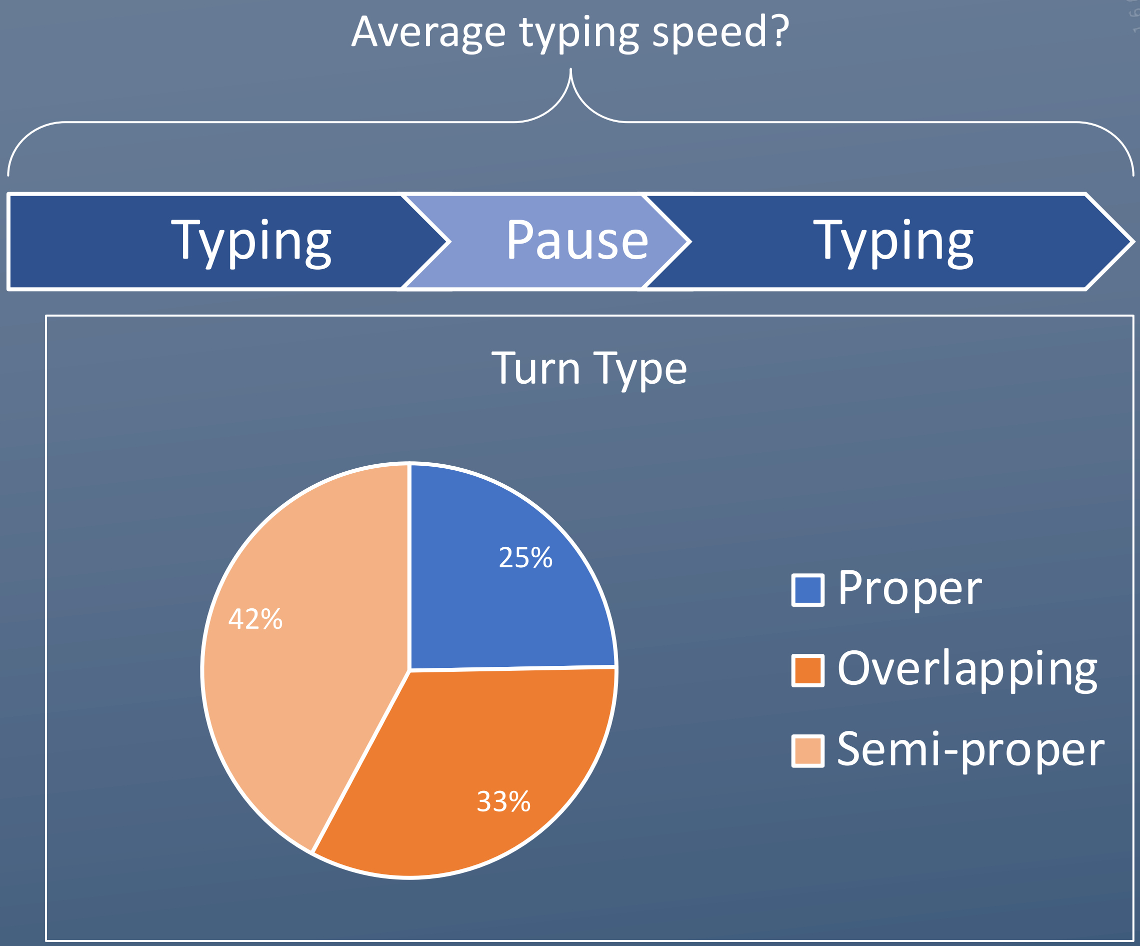
- Advantage: Can fit non-linear effects
- Disadvantage: Direction and magnitude of effect aren’t straightforward





STUDY 2 – SENTIMENT IN DIALOGUE
TURN TYPES





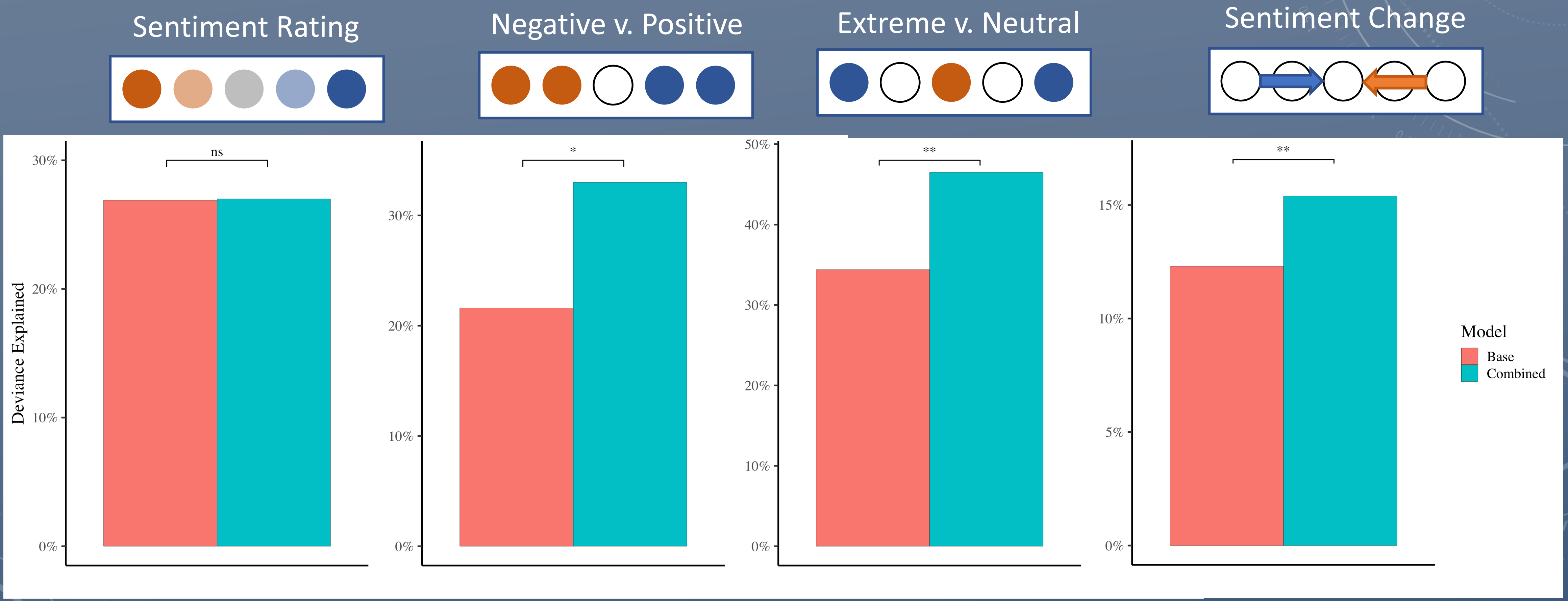
STUDY 2A – SENTIMENT IN DIALOGUE METHODOLOGY

RQ 1a. Does keystroke information provide additional information about sentiment and sentiment change above lexical information?

Base: $g(E(\textit{gold standard}) \sim f(\textit{VADER prediction})$

Combined: $g(E(\textit{gold standard}) \sim f(\textit{VADER prediction}) + f(\textit{keystroke features})$

STUDY 2 – SENTIMENT IN DIALOGUE
EXP 2A – RESULTS (4 TASKS)



STUDY 2 – SENTIMENT AND OPINION IN DIALOGUE
EXP 2B - METHODOLOGY

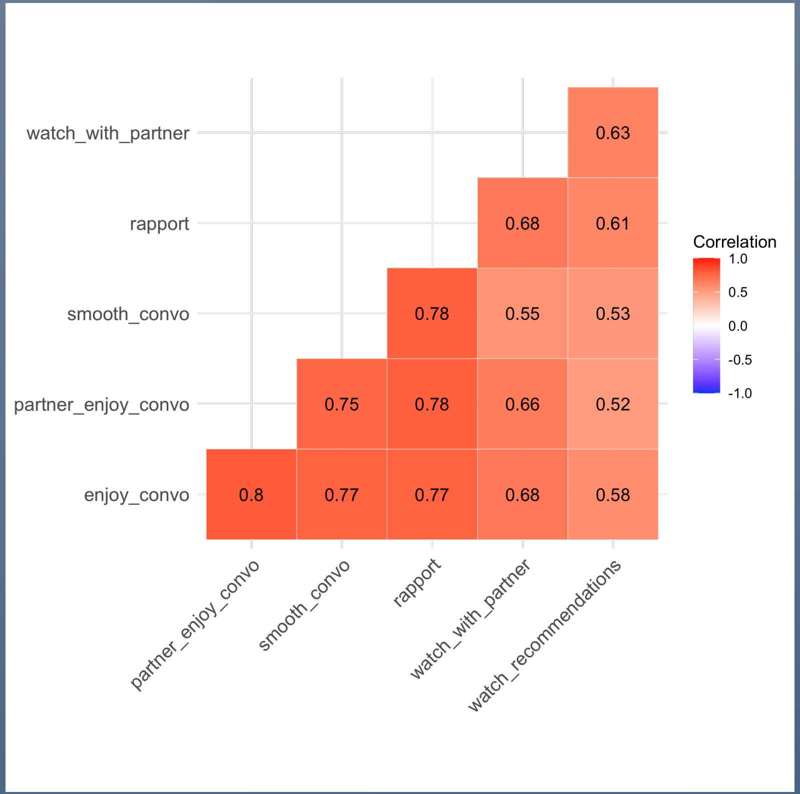
RQ 2b. Are typing patterns independently sensitive to both a user's overall opinion of their partner and the sentiment of a specific utterance?

Base: $g(E(\textit{keystroke feature}) \sim f(\textit{gold standard sentiment})$

Combined: $g(E(\textit{keystroke feature}) \sim f(\textit{gold standard sentiment}) + f(\textit{opinion})$

- Flipped independent and dependent variables
- Keystroke features were the same as the predictors in Exp. 2a
- Example opinion questions (from post-conversation questionnaire):
  - *How likely are you to watch a recommendation?*
  - *How smooth do you feel the conversation was?*

STUDY 2B – SENTIMENT IN DIALOGUE RESULTS



Keystroke Metric	Significance of opinion rating
Pre-turn pause	$p < .0001$ ***
IKI	$p < .0001$ ***
Dwell time	$p = .08^{+}$
Edit count	$p = .09^{+}$
Pause before send	$p = .09^{+}$
Phrase boundary pause	$p = .14$
Pre-word pause	$p = .28$

STUDY 2 – SENTIMENT IN DIALOGUE RESEARCH QUESTIONS REVISITED

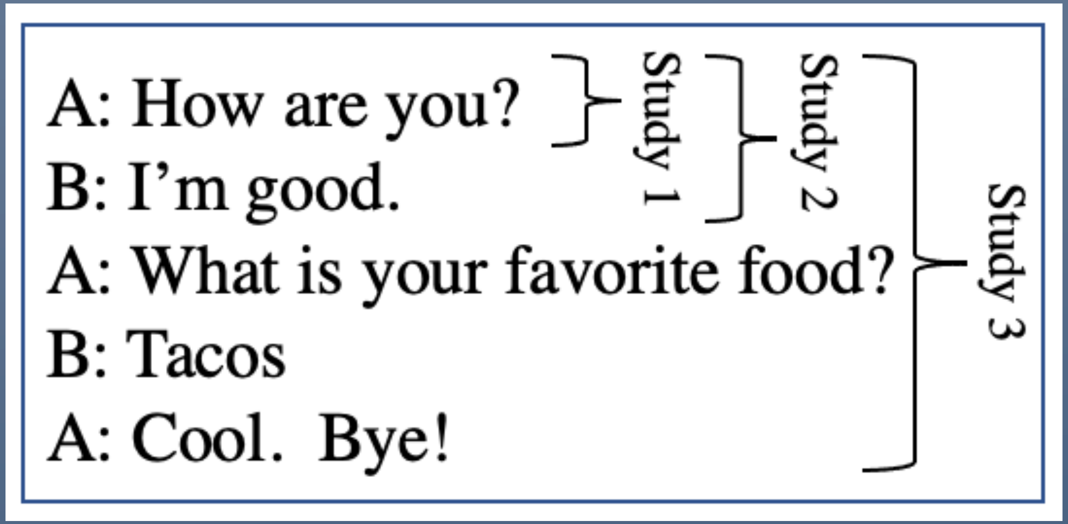
RQ 2a. Does keystroke information provide additional information about user sentiment and sentiment change, above lexical information?

- Yes

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

- Somewhat

STUDY 3 – LOW RAPPORT



STUDY 3 – LOW RAPPORT

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

Study 1
Study 2
Study 3

STUDY 3 – RAPPORT IN DIALOGUE RESEARCH QUESTIONS

a. Can typing patterns over an entire conversation be used to predict low levels of rapport between partners in an interaction?
b. How do subsets of keystroke data compare at predicting low rapport?

STUDY 3 – RAPPORT IN DIALOGUE BACKGROUND

- Rapport is tough to define succinctly:

  "...an individual's experience of harmonious interaction with another person, often described as 'clicking' or 'having chemistry'"

  Tickle-Degen & Rosenthal (1990)

- Rapport is critical for improved cognitive function (Barnett et al., 2020)
- Rapport can be detected from very thin slices of an interaction (Carney et al., 2007)

STUDY 3 – PREDICTING RAPPORT LEVELS
CLUSTERING THE PARTICIPANTS

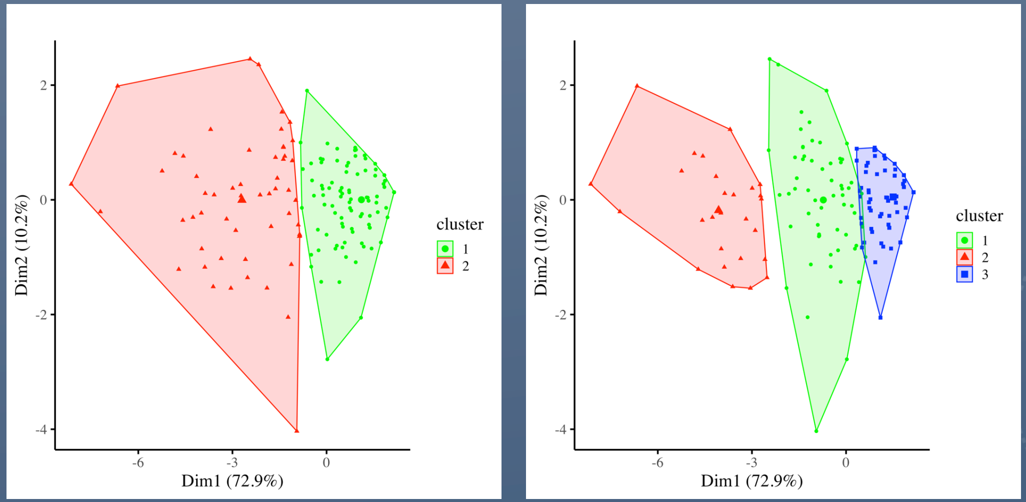
- Created a 6-dimensional vector from questionnaire ratings

$$\overrightarrow{rapport} = [5, 6, 7, 5, 4, 5]$$

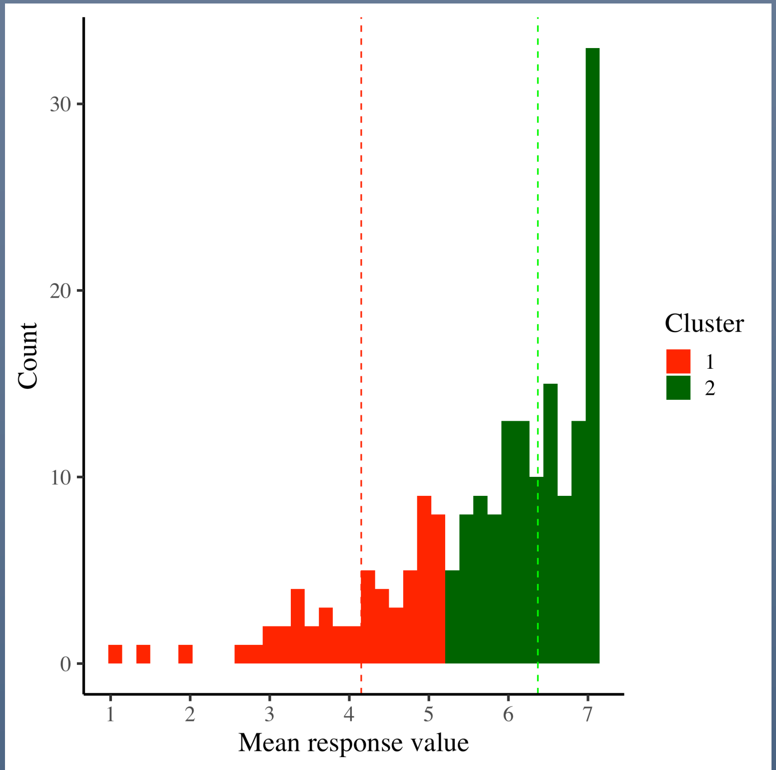
Watch?

Enjoy?

- An ensemble of distance metrics recommended 2 clusters



STUDY 3 – PREDICTING RAPPORT LEVELS CLUSTERING THE PARTICIPANTS

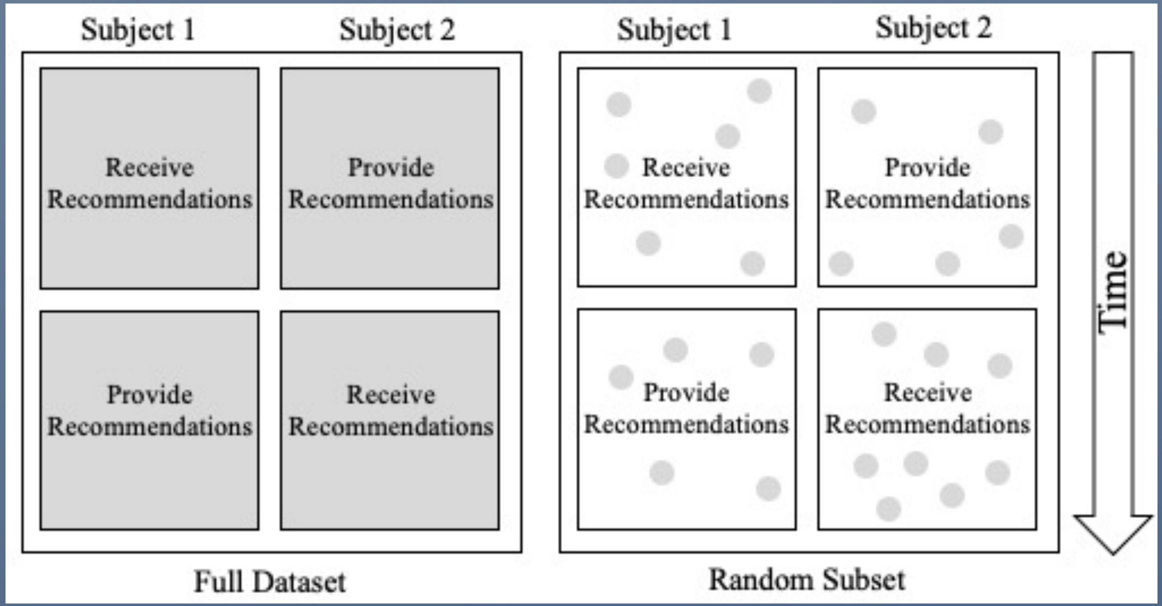


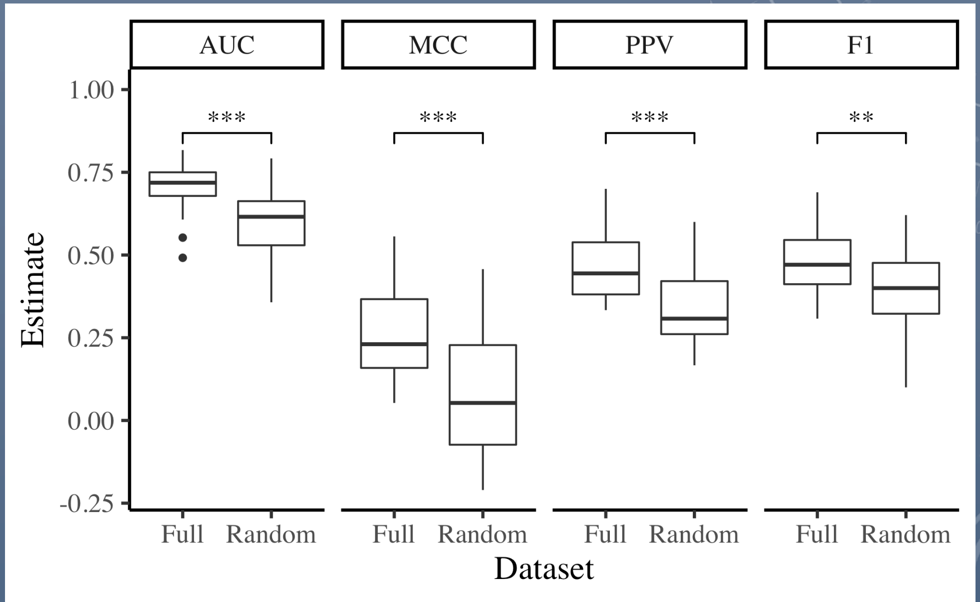
- Cluster characteristics
  - Low-mid rapport
    - 56 subjects (of 192)
  - High rapport
    - 136 subjects (of 192)
    - Mean questionnaire rating: 6.38 (of 7)

STUDY 3 – RAPPORT IN DIALOGUE METHODOLOGY – MODEL AND METRICS

- Tested a random forest, boosted tree, and neural network
  - A multilayer perceptron, with 10 hidden units performed best on a validation set
- Metrics were selected for their sensitivity to correct predictions of the minority class (low rapport)
  - ~~Accuracy~~ – Would be dominated by the majority class
  - Area under the ROC curve (AUC)
  - Matthews Correlation Coefficient (MCC)
  - Positive Predictive Value (PPV)
  - F1 Score

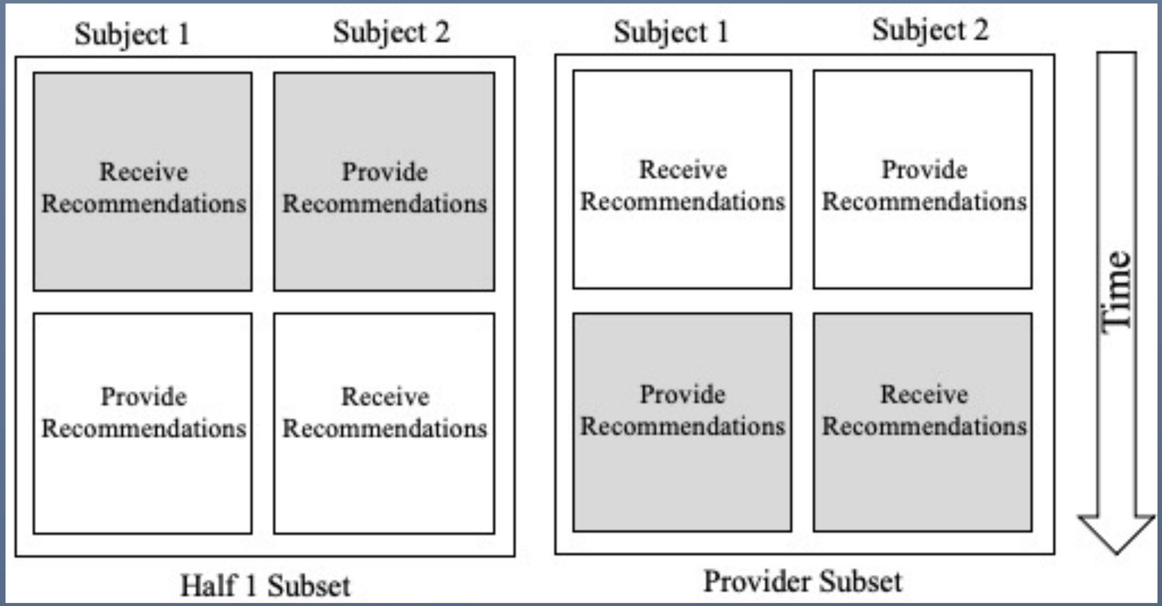
STUDY 3 – PREDICTING RAPPORT FULL DATASET VS RANDOM SUBSET

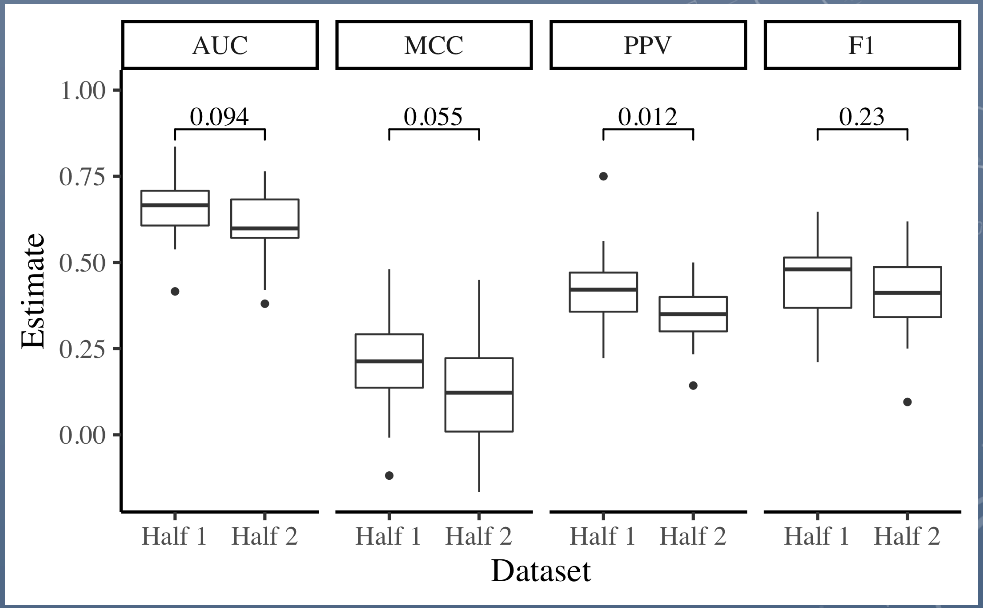




- Randomization needs to be redone using repeated subsampling

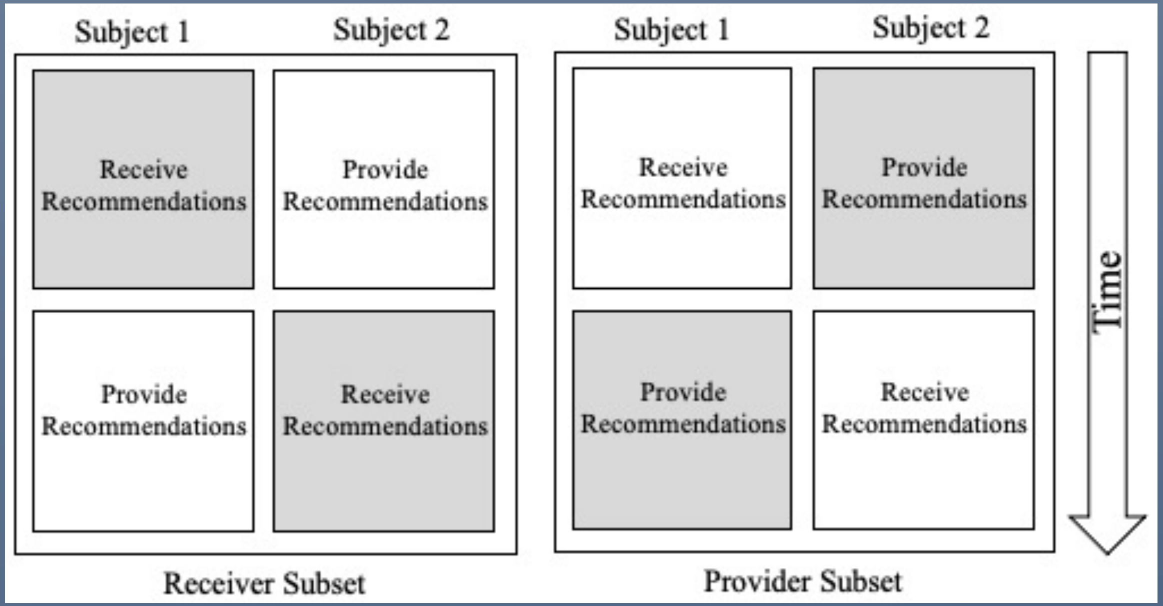
STUDY 3 – PREDICTING RAPPORT 1ST HALF VS 2ND HALF SUBSET

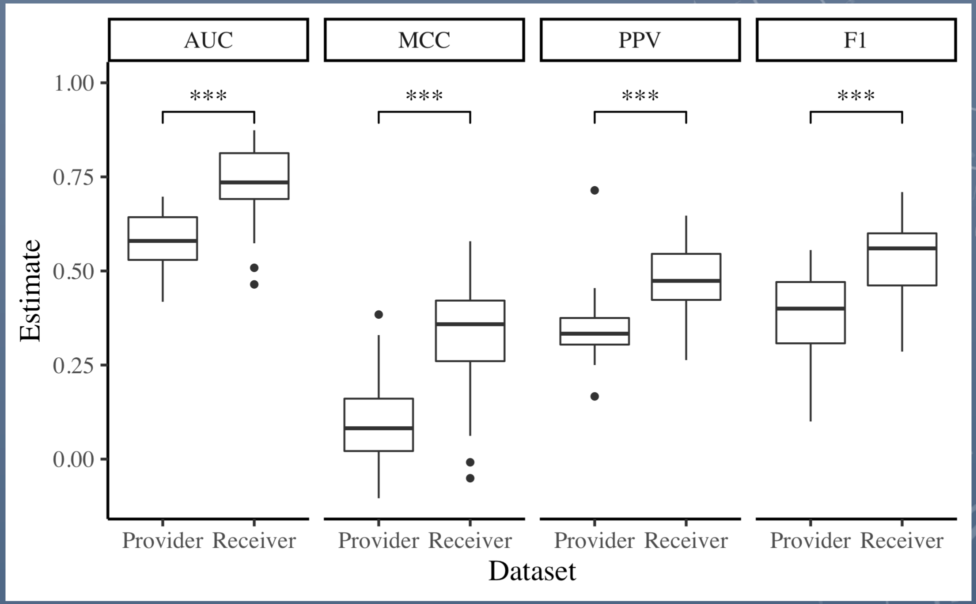




- Temporal halves not significantly different
- But first impressions matter (Tolmeijer et al., 2021)

STUDY 3 – PREDICTING RAPPORT PROVIDER VS RECEIVER SUBSET



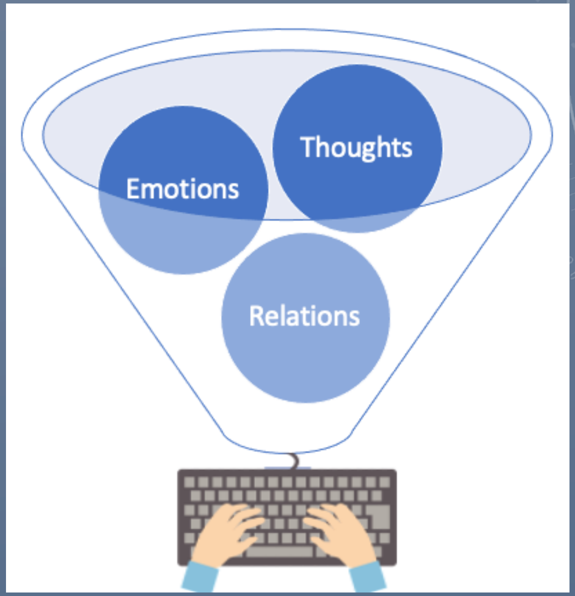


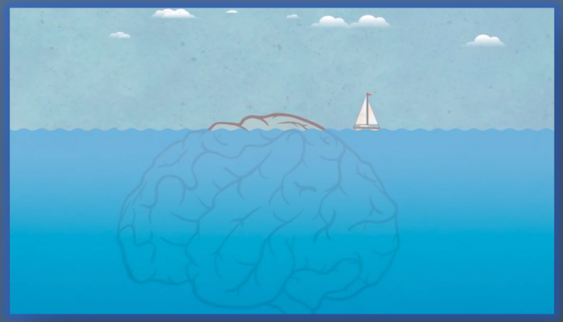
- Intriguing how much more useful receiver is versus provider
- Also extremely useful for the larger aim of my thesis

OVERALL

OVERALL TAKEAWAYS

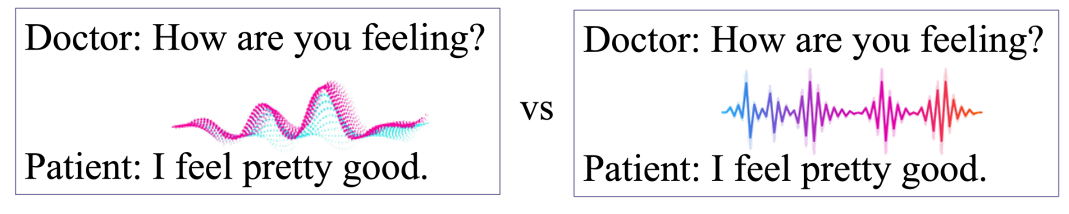
- Keystroke patterns are:
  - Complex
  - Associated with different underlying intentions, where those intentions may not be evident from word choice alone.
- Evidence that prosody is also realized implicitly, not just for a partner to hear
- Combining keystrokes and HCI holds a lot of possibilities



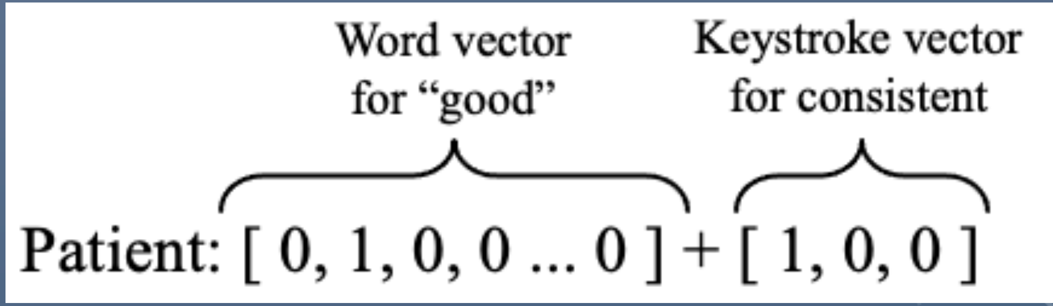


FUTURE DIRECTIONS AND POSSIBILITIES

- Human-to-Human
  - Visualizing typing to make it useful



- Human-to-Computer
  - Augment lexical information for computer agents (chatbots)

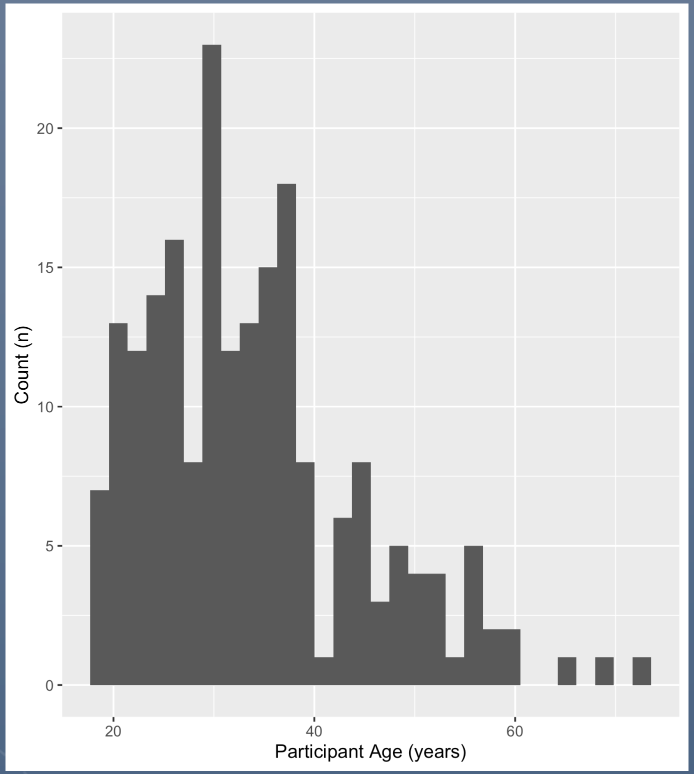


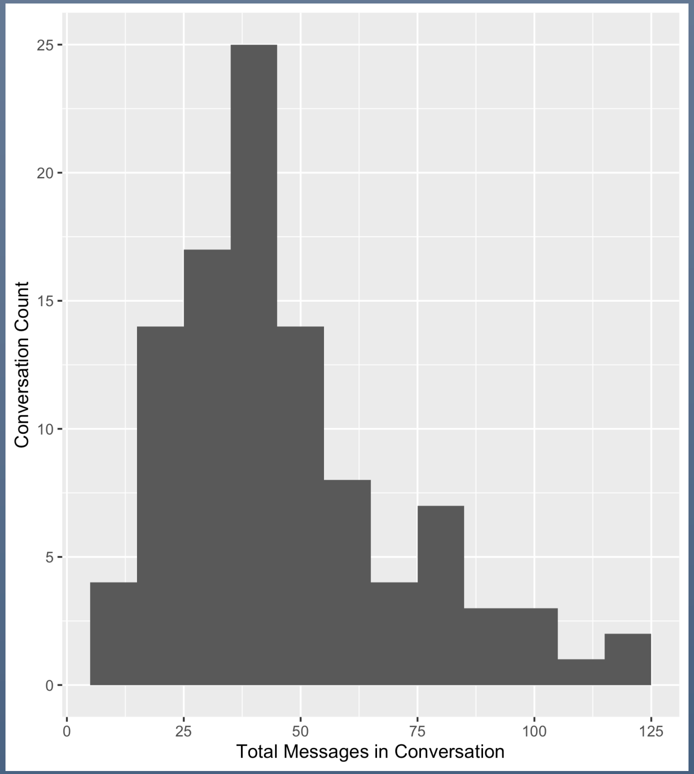
- Ethical implications must be accounted for when using keystroke data

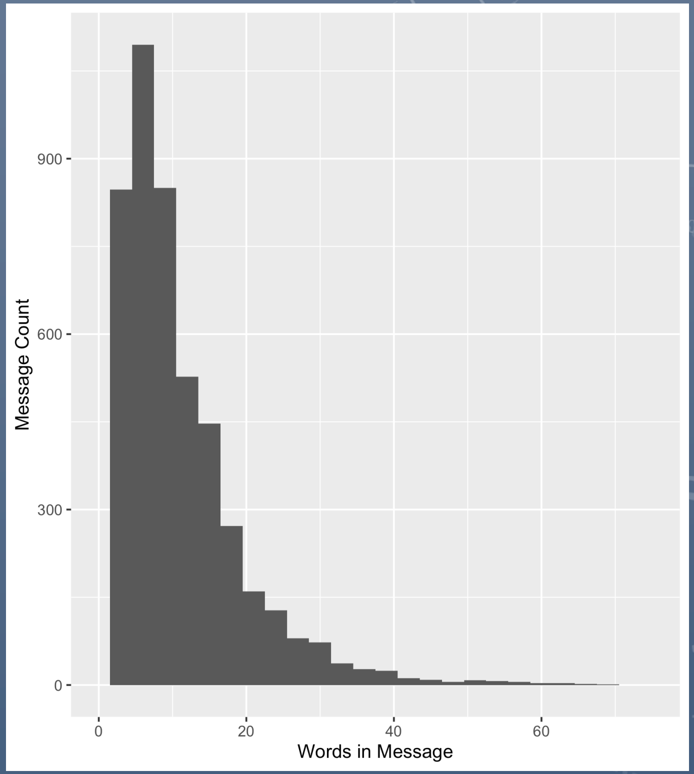
Thank you!

- Research assistant: Elana Laski
- Committee: Darren Gergle, Anne Marie Piper, David-Guy Brizan
- Members of the CollabLab and Language & Computation Lab

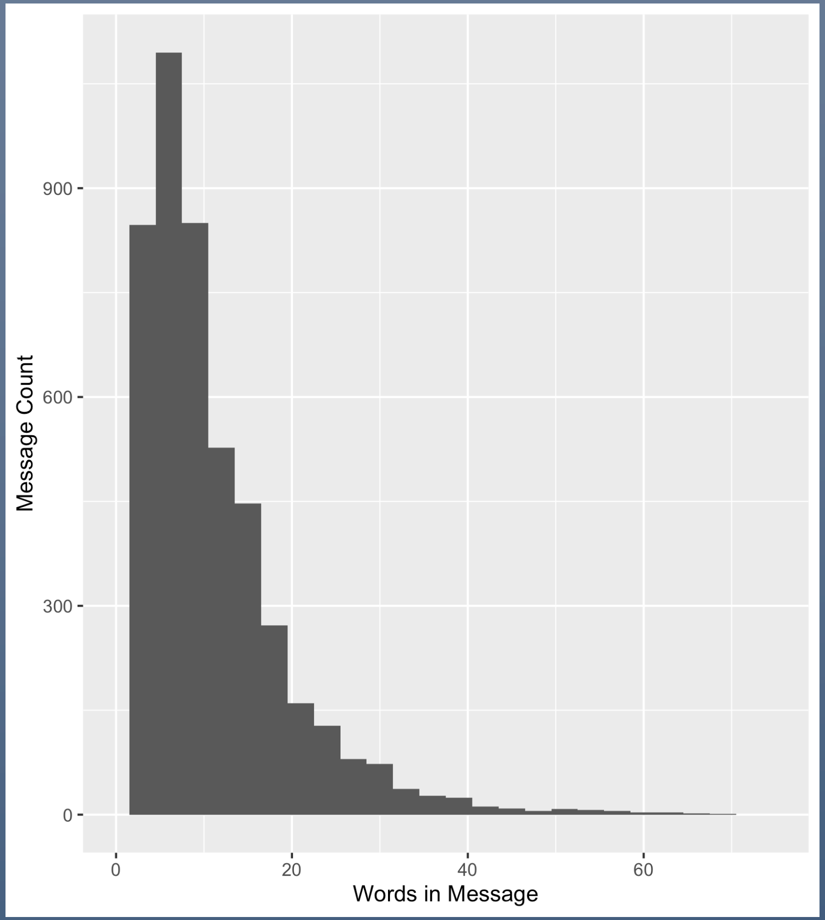
OVERALL – DATA COLLECTION



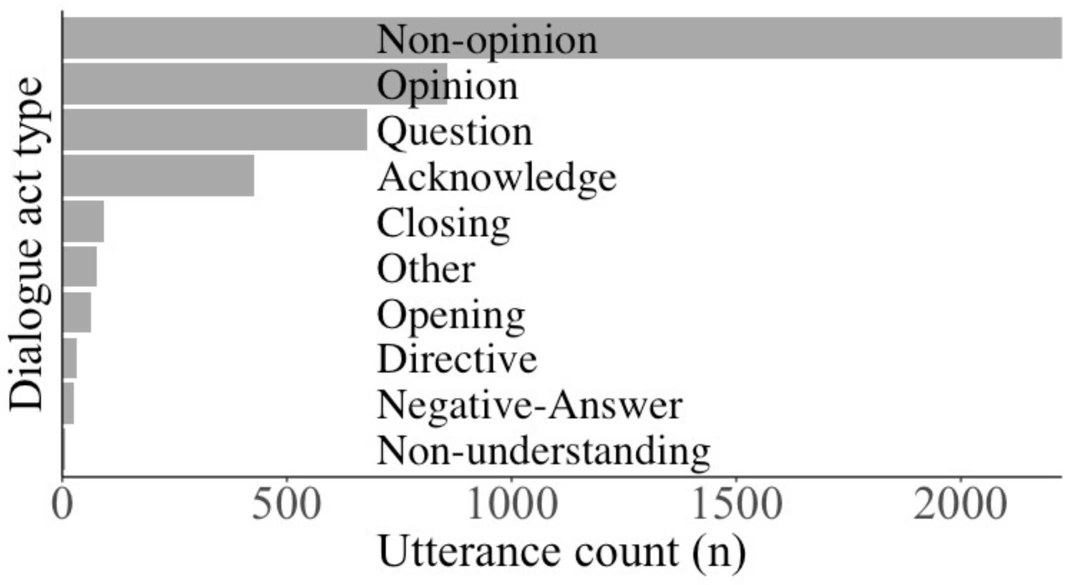




OVERALL - UTTERANCE LENGTHS

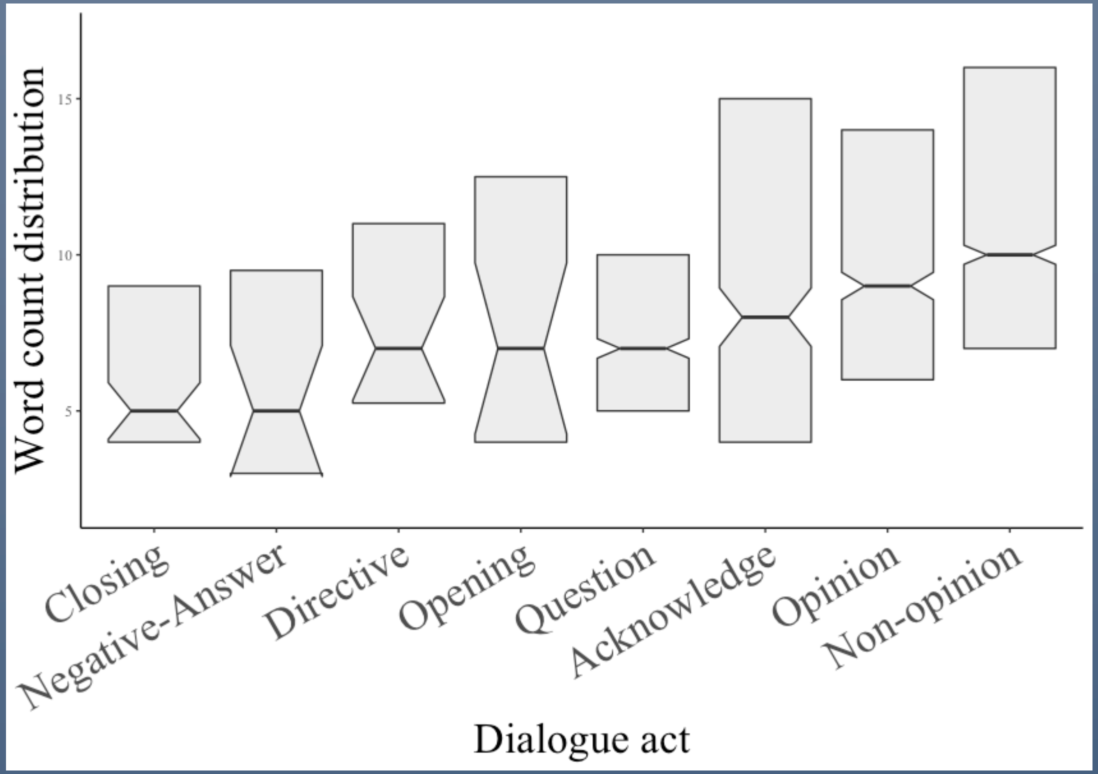


STUDY 1 – DIALOGUE ACTS DISTRIBUTION AND EXAMPLES

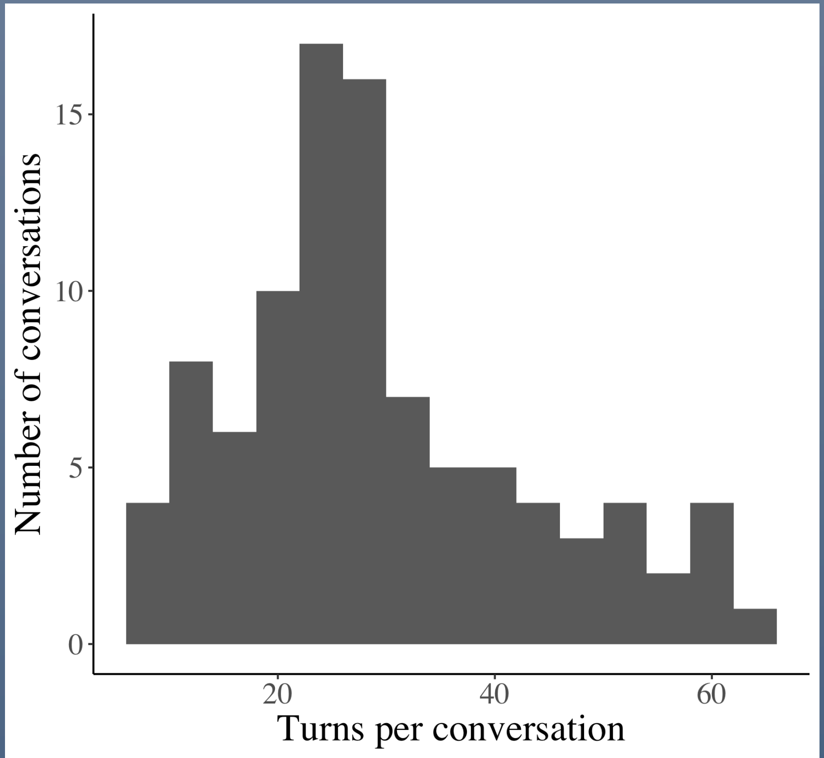


Dialogue Act	Example
Non-opinion	*It's on Netflix*
Opinion	*The whole premise is so good!*
Acknowledge	*Oh definitely.*
Directive	*Check out the trailer*
Negative-Answer	*No, not really*
Non-understanding	*Who?*

DIALOGUE ACT WORD COUNTS

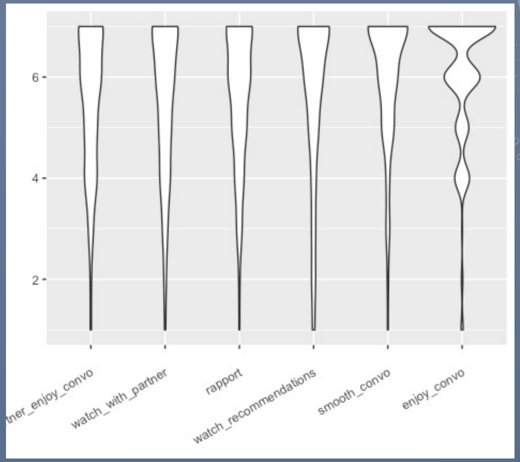


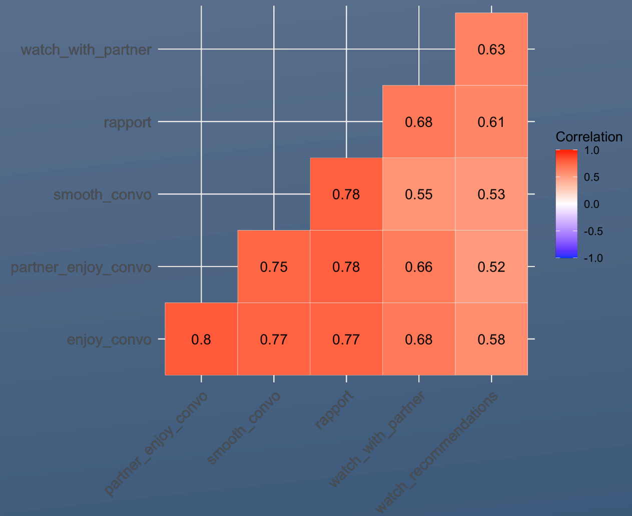
STUDY 2 – TURNS PER CONVERSATION



STUDY 2: OPINIONS

- Opinion questions
  - **To what degree did you enjoy the conversation?**
  - To what degree did the conversation go smoothly?
  - Hypothetically, how much do you think you'd enjoy watching a movie with your partner?
  - How would you rate the level of rapport established between you and your partner?
  - How likely do you think it is that you'll end up watching one of the movies your partner recommended?
  - To what degree do you think your partner enjoyed chatting with you? **(self-awareness)**



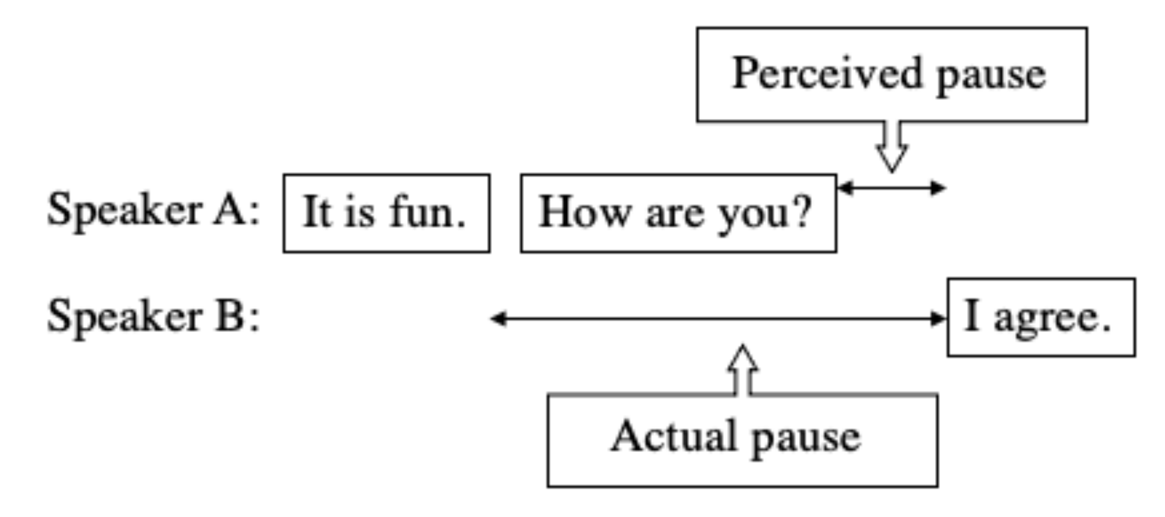


STUDY 2B – SENTIMENT IN DIALOGUE RESULTS

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

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

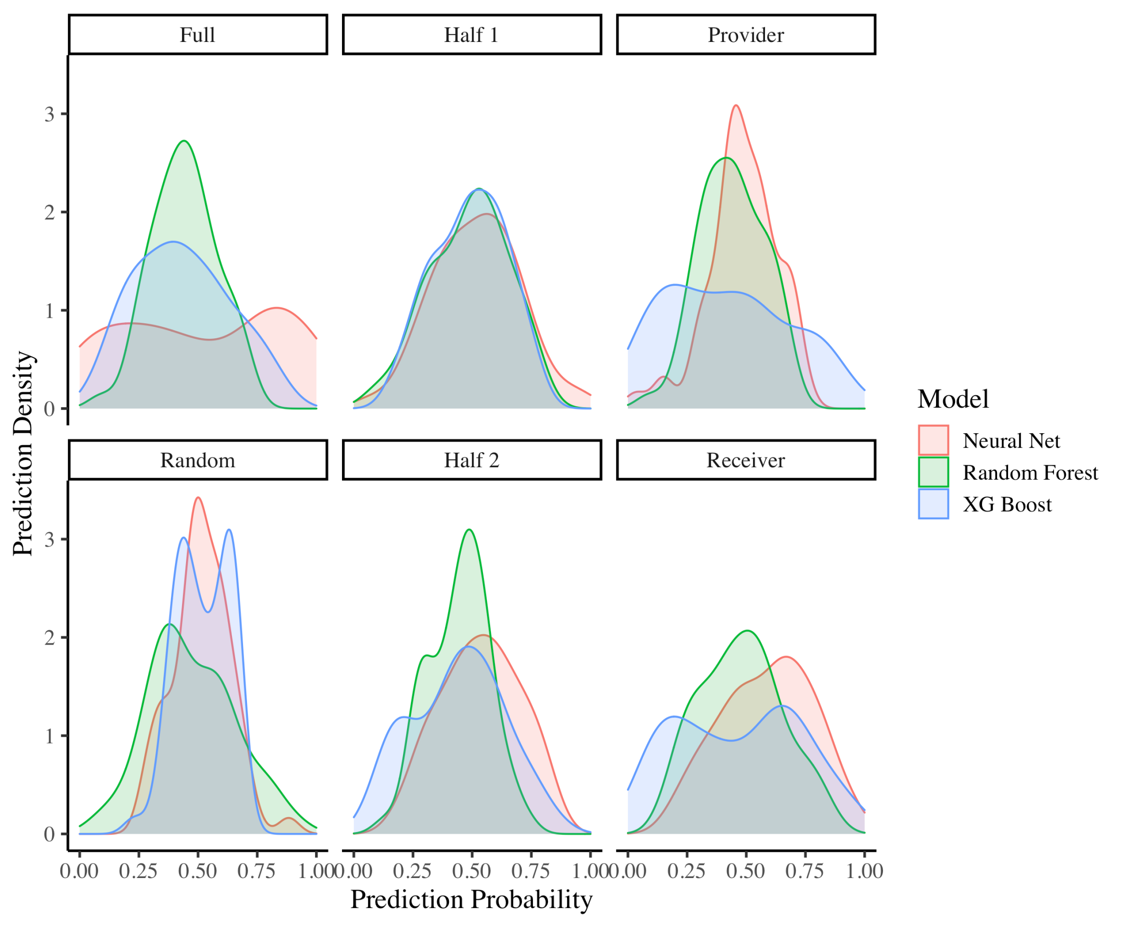
STUDY 2 – SENTIMENT IN DIALOGUE



STUDY 3 – RAPPORT IN DIALOGUE METHODOLOGY – METRICS

- Area under the ROC curve (AUC) – The proportion of true-positives to false-positives
- Matthews Correlation Coefficient (MCC) – A numeric representation of an entire confusion matrix: all 4 quadrants need to be accurate
- Positive Predictive Value (PPV) – The proportion of positive cases (*actual* low rapport) against the predicted class members, but accounting for *prevalence*, which is the proportion of the class of interest within the entire dataset
- F1 Score – The harmonic mean of precision and recall; NOT ACCURACY

STUDY 3: MODEL COMPARISONS



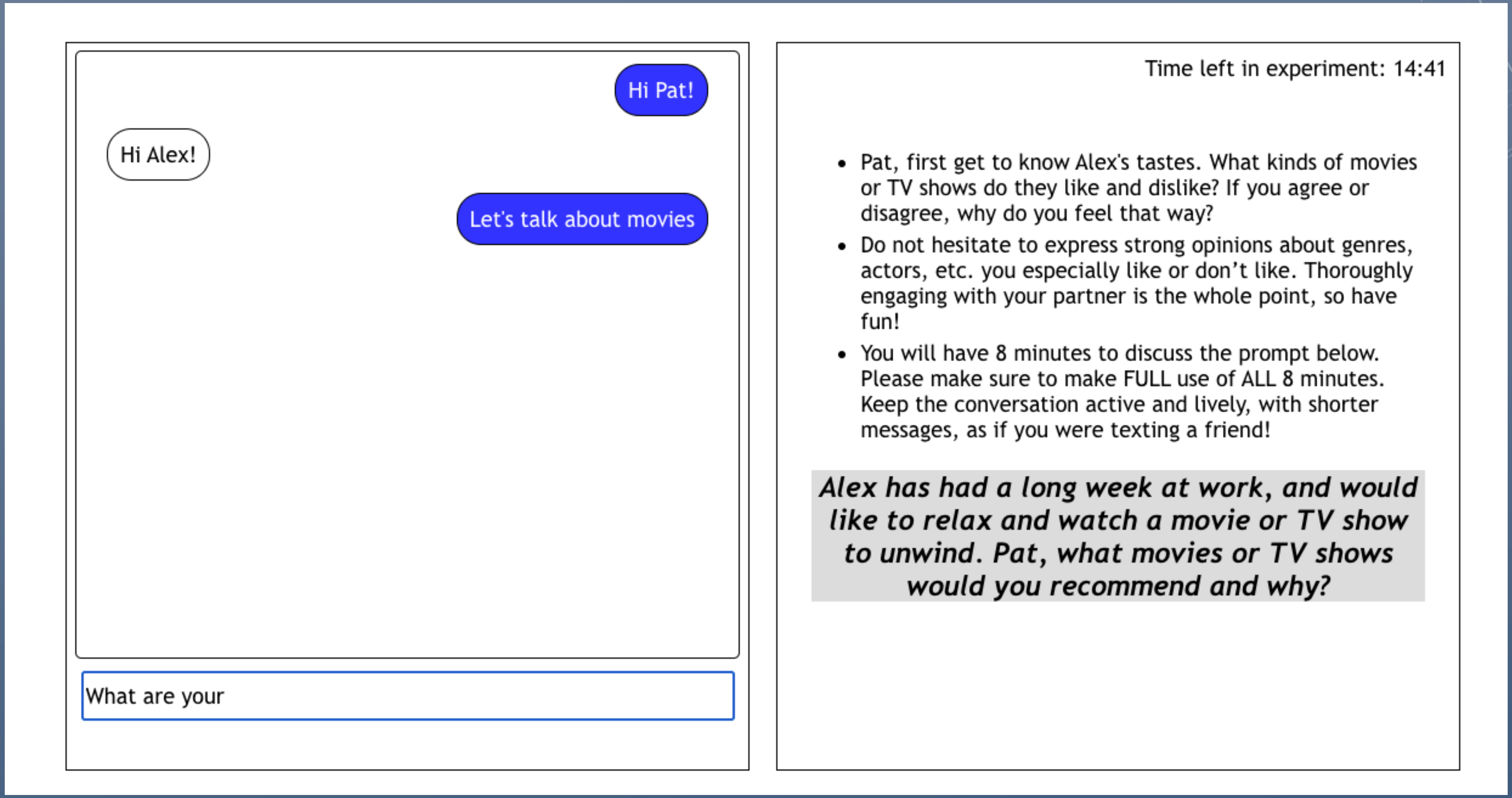
STUDY 3: MINORITY CLASS PREDICTIONS

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

ETHICAL ISSUES

- Every major browser allows you to write an extension that logs keystrokes
- Keystrokes can predict demographics
  - Age
  - Gender
  - Education level
  - Personal identity
- BUT keystrokes can be anonymized and still be helpful

EXPERIMENTAL APPARATUS



EXPERIMENT PIPELINE