Appendix A

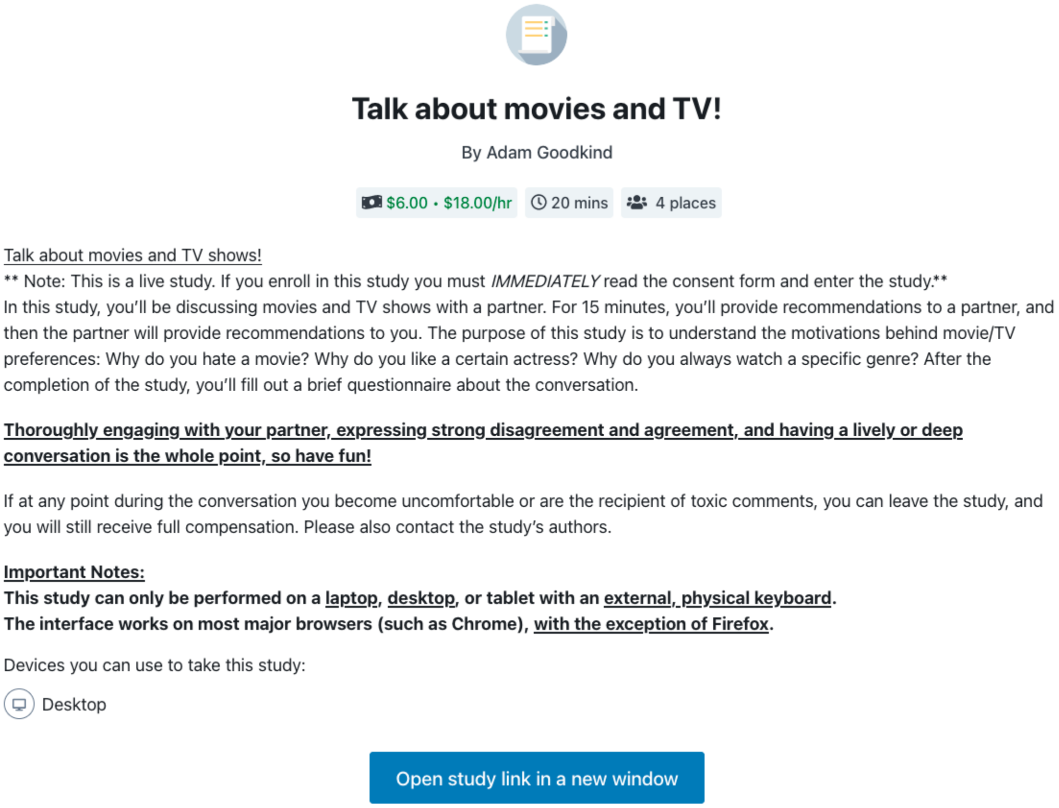
IRB Detail

Prior to IRB approval, a pilot study consisting of ten 10-minute dialogues was collected. This was used for testing purposes to ensure that the experimental apparatus was accurately collecting all of the data necessary for my studies. Further, I solicited comments from pilot participants in order to improve the experimental setup. Because the pilot study was conducted prior to IRB approval and used a different setup, all of the data was discarded after testing, and is not included in any of my thesis studies.

The IRB required consent before participants began the experiment. In a pre-experiment consent form, participant were told what the experiment would entail, and the approximate amount of time it would take. Although the subject matter they would be discussing was relatively innocuous (movie and TV preferences), participants were informed that if they felt uncomfortable with the conversation or it turned toxic, they could leave the experiment and still receive compensation.[1] Only after this initial consent was submitted could a participant enter the experiment apparatus.

In a post-experiment consent form, full disclosure was provided as to the true objective of the experiment, and exactly what data was collected and why. It was explained that keystrokes are considered a soft biometric, and can be personally revealing, akin to an iris scan (Banerjee and Woodard, 2012). However participants were also reminded that their conversation would be

[1] In my full data collection, consisting of over 200 participants, only one participant left the experiment for this reason. However, it should be noted that their specific objections were unclear, and I am unsure if they were mentally stable or intoxicated. In any case, the data from this experiment was thrown out and not included.



Talk about movies and TV!

By Adam Goodkind

Talk about movies and TV shows!
** Note: This is a live study. If you enroll in this study you must *IMMEDIATELY* read the consent form and enter the study.**
In this study, you'll be discussing movies and TV shows with a partner. For 15 minutes, you'll provide recommendations to a partner, and then the partner will provide recommendations to you. The purpose of this study is to understand the motivations behind movie/TV preferences: Why do you hate a movie? Why do you like a certain actress? Why do you always watch a specific genre? After the completion of the study, you'll fill out a brief questionnaire about the conversation.

Thoroughly engaging with your partner, expressing strong disagreement and agreement, and having a lively or deep conversation is the whole point, so have fun!

If at any point during the conversation you become uncomfortable or are the recipient of toxic comments, you can leave the study, and you will still receive full compensation. Please also contact the study's authors.

Important Notes:
This study can only be performed on a laptop, desktop, or tablet with an external, physical keyboard.
The interface works on most major browsers (such as Chrome), with the exception of Firefox.

Devices you can use to take this study:

Desktop

Figure A.1
The advertisement for the experiment, sent out by Prolific to potential participants.

manually anonymized, and that only the minimal demographics provided by Prolific would be used. Upon reading this, participants had the option to contact the researchers and ask to opt-out (while still receiving compensation). In the actual data collection, none of the participants chose to exercise this option after the experiment was complete, though.

Appendix B

Prolific Advantages and Disadvantages

Prolific provides a number of key benefits over Amazon's Mechancial Turk:

1. Prolific performs more stringent screening of participants, so researchers are less likely to encounter scammers or bots, and more likely to receive trustworthy and engaged participants.
2. Prolific encourages fair compensation for participants, encouraging a healthier environment for research (e.g. Hara et al., 2019). Before running an experiment a researcher must declare an hourly rate, and the number of participants required. Given this, they then deposit the necessary funds to pay every participant their quoted rate. Upon completion of the experiment, Prolific pays all participants based on the *actual* average time taken to complete the experiment. For example, if a researcher estimates that their experiment will only take five minutes, but in reality most participants take ten minutes, then all participants will be paid for ten minutes of work.
3. Prolific takes privacy very seriously: from initial recruitment through final payment a research only ever sees a participant's ID number and limited demographics, rather than names, credit card numbers, etc.
4. Pre-screening is free and easy on Prolific, which was essential for ensuring that all participants currently lived in the United States (per the IRB). We also pre-screened for other criteria, which will be enumerated in Section 4.1. On the other hand, MTurk charges for pre-screening.

One downside to Prolific is that their pool of participants is smaller than that of MTurk's. That being said, even after applying my pre-screening criteria for my data collection, I still had a pool of approximately 28,000 participants that were eligible to receive notification of my experiment. Of these, approximately 20,000 identified as female, and 8,000 identified as male. Because of this imbalance, Prolific also provides the option to distribute an experiment to a gender-balanced sample. I selected this option for a few batches, but it did not make much difference, i.e. my population was usually well-balanced anyway.

Prolific has also made their peak times for experiments available on their website. Because I needed to quickly pair participants for my dialogues, I chose to do small batches on a near-daily basis at the peak times (usually between 10 am and 12 pm Central Standard Time). Over the course of one month, I ran 13 batches of data collection, collecting 10 dialogues in each batch. Usually all 10 dialogues were collected within 45 minutes, since not all participants would join the moment a study was sent out.

Appendix C

Prosodic Parallels in Speech and Typing

Tables C.1 and C.2 below outline prosodic features of spoken language and then describes speech prosody's analogs in keyboard typing.

Feature in speech	Manifestation in speech	Manifestation in typing
Pauses	Pauses are usually only measured between words. Pauses between phonemes within a word are often difficult to impossible to measure. In fact, some phoneme transitions do not have any delimiting characteristics; rather, a speaker produces them in contiguous succession.	Pauses are relatively trivial to measure. Pauses can be measured in many ways, as seen in Figure 2.1. Pauses between intra-word keystrokes are typically measured in the same way as pauses in between-word keystrokes.
Energy/ intensity/ loudness	Speakers consciously and unconsciously choose how much energy to use in producing speech. These choices are usually directly perceivable by the listener.	A typist can choose to alter the visual attributes of their message, such as all capital letters or bold font. Evidence shows that if a typist is (silently) producing more intense language, this is manifested as increased typos, revisions, or longer keypresses (dwell time) (Lee et al., 2015a).
Length of sound/ duration	Syllable lengthening in speech is learned early in development and implemented for many different reasons (Snow, 1994). Measuring or at least comparing syllable duration is relatively robust in speech science.	Repeated letters are employed frequently in typing. Kalman and Gergle (2009) finds evidence for many uses of letter repetition, which parallel uses in spoken prosody.

Table C.1

A comparison of parallel features in spoken prosody and keystroke dynamics (continued in Table C.2)

Feature in speech	Manifestation in speech	Manifestation in typing
Speech rate	Speakers speed up and slow down for a large variety of reasons. The production rate of language, per se, can encode significant information about the intended message.	If a message is only transmitted upon completion, then the typing rate within that message is not necessarily known. If a number of messages are transmitted rapidly, it can be inferred by the receiver that the language is being produced rapidly. A real-time typing environment, which is less common today, would also facilitate awareness of production rate.
Pitch/ fundamental frequency	Humans continuously alter the pitch of their voice, e.g. high tones and low tones. These alterations can convey significant amounts of information about the affective or emotional properties of the speaker.	Aside from inferences drawn by the receiver from altered language production, pitch cannot be conveyed in typing production.
Timbre	Timbre is difficult to define succinctly, but it represents the quality of sound that makes a particular voice have a different sound from another, even when producing the same phoneme.	In CMC, the messaging medium outputs uniform text styling. Hand-written communication, such as shaky or sloppy text, could possibly be considered a parallel for voice timbre.

Table C.2

A comparison of parallel features in spoken prosody and keystroke dynamics (continued from Table C.1)

Appendix D

Experimental iterations

Minor elements of the experiment were improved upon. Because the IRB was ruled exempt from further review (see Section 4.4.1), these minor changes did not require IRB approval. In addition, the changes did not alter the nature of the experiment in any significant way. The following subsections (within Section 4.4.2) are included in order to explain the rationale for certain features of the experiment.

Timer and Timing

One of the most significant changes I made was actually adding a countdown timer to the experiment. Although I had included a "warning" during the experiment saying "One Minute Left," many participants felt this was not sufficient and that they easily lost track of time during the experiment. The pilot study was also only 10 minutes long, and many participants asked for additional time. In a way, I took both of these as positive signs that my experimental prompts were engaging and that a naturalistic conversation was taking place, which fully occupied the attention of the participants.

Adding a timer had the additional benefit of allowing the conversations to follow a more natural trajectory, especially towards the end of the experiment. Since participants could see the amount of time remaining, they could more properly "wrap up" their conversations, rather than being caught off guard when the experiment ended.

Since Study 1 (Chapter 5) looks at dialogue acts, which also facilitate the function of changing the direction of the conversation, it was important to also have data where utterances were intended to conclude a conversation, reflecting on what had taken place, rather than launching in to a new topic.

Age Constraints

A significant change that I made to the experiment setup was removing age constraints after the second batch of data collection. Initially I had limited the experiment to participants between ages 25-40, encapsulating the “Millennial” generation. My initial intention was that by constraining the age range, participants would be more likely to have cultural awareness of the same movies in general, even if they did not share the same preferences.

However, the downside to an age constraint was that it seemed to foster *too much* agreement. In other words, while my experiment was intended to evoke both agreement and disagreement along with positive and negative sentiment, implementing an age constraint led to very little disagreement and negative sentiment. This takeaway was gleaned from participants’ comments on my experiment after the pilot study and first two batches of data collection. Almost without exception, every participant said only positive comments about the conversation itself as well as what they thought of their partner. For example, one participant in the pilot study said "[My partner was] very informative and thoughtful. I will absolutely be looking into one of the movies they recommended."

In this situation, the changes I made to my experiment did not come from explicit feedback. Rather, participant feedback was qualitatively analyzed and adjustments were made based on this analysis.

Removing the age constraint did not materially effect my experiment. While in the first two batches, participants could not be more than fifteen years apart in age, throughout the entire data collection process a large number of pairs of participants were within fifteen years of each other. As an illustration of how removing age constraints did not affect the experiment, see Figure 4.3.

However, after collecting batches three and four (with age constraints removed), I tested whether larger age differences had a significant effect on participants' subjective enjoyment of the conversation. If age difference did significantly change the nature of the experiment, then I would not be able to include batches one and two with the rest of the data. As can be seen in Figure 4.3 though, larger age differences did not substantially increase or decrease enjoyment of the conversation.

Explicit Instructions and Prompt Wording Modification

The final change I made came from an online discussion with participants who frequently use Prolific. My initial experimental prompts and instructions were based on previous experiments that were run in-person primarily with undergraduate students enrolled in relevant courses, and participating in experiments as a course requirement. In my case, though, experiments were being run online with a large and diverse population, where participants are primarily financially motivated.

Given the heterogeneity of my participant population, it was also necessary to more explicitly write out my experiment instructions and prompts. As seen in the instructions below, I learned that it was necessary to instruct participants to enter the experiment immediately, since they needed to be paired with a partner at the same time. This is distinct from most online experiments which are asynchronous and involve stimuli much as surveys that can be taken at any time. Most likely this instruction limited who participated in my experiments, since they needed to have 15-20 minutes available at the moment the experiment was sent out.

In addition, many online participants take part in experiments as an additional source of income, and therefore want to complete as many experiments as possible as quickly as possible (Chandler and Kapelner, 2013). On the other hand, an undergraduate coming into a lab is prepared to begin an experiment immediately, and is not trying to maximize their earnings by taking part in multiple experiments. For this reason I added into both the initial instructions below as well as in the conversational prompts that participants should have fun, and that disagreements are completely acceptable.

I also observed that the pauses between utterances in my pilot study were unusually long in some places. It is also possible that online participants will be undertaking multiple experiments at one time on other platforms, and thereby take advantage of the conversational nature of my experiment by taking long pauses to either rest or work on other studies. For this reason I also added to my prompts, "Please make sure to make FULL use of ALL 8 minutes. Keep the conversation active and lively..." a a passive-aggressive way to discourage slow responses.

Appendix E

Study 1 details

Study 1a details

Model building

As a base model, I looked at the typing patterns of the entire utterance, where the model was calculated to predict whether the dialogue act was a Non-opinion or Opinion statement. The base predictors were chosen because they seemed more fundamental to the typing process, and some also had more obvious spoken prosody parallels. The base predictors were:

1. Utterance typing speed (keystrokes/utterance duration)
2. Utterance average inter-keystroke interval (IKI)
3. The interaction of speed and IKI
4. The pause between the previous message being sent and the beginning of the current utterance (pre-utterance gap)
5. The edit count (pressing BACKSPACE or DELETE)

The reason that the interaction of typing speed and IKI is included is because speed is more sensitive to word length, whereas IKI is not. While in future work I will find a more precise way to measure typing rate, it seems that the interaction of the two terms, and the variance absorbed by this interaction, improve the predictive power of the other predictors.

In model testing (see Table E.1), the pre-utterance pause, utterance speed, and edit count predictors were found to significantly predict whether the utterance was a Non-opinion or Opinion ($p < 0.05$). Because standardization was performed by-subject, the model coefficients are less meaningful and do not directly represent any definite "unit" of measurement. They are similar to a z-score, though. The model showed that the pre-utterance gap was shorter for Non-opinions, typing speed was faster when typing Opinions, and that opinion utterances contained significantly fewer edits. Using the variance inflation factor (VIF) test, none of the predictors exhibited collinearity.

The model was then extended to see if in addition to the average speed of typing, the variability of typing speed was a significant factor. To accomplish this, the standard deviation of typing speed was added as a predictor. Importantly, I only looked at the standard deviation of the IKI at the beginning of the word. This was because that gap reflects lexical recall, which could be expected to vary between different types of dialogue acts. On the other hand, intra-word intervals reflect motor skills, which are expected to be more consistent across a typing session (Logan and Crump, 2011).

While the additional variability predictor was not significant in and of itself ($p = 0.45$), adding variability increased the significance of the other predictors. In addition, the log-likelihood and BIC of the model also improved. The coefficient showed that variability is greater when typing an opinion utterance than when typing a statement.

In addition to typing patterns spanning the entire utterance, an additional goal was to see if typing patterns in the initial part of an utterance are also sensitive to dialogue act type. A new model was tested that incrementally added: the typing speed of the first word, the typing speed of the second word, and the duration of the gap between the first two words. Without any utterance-spanning predictors, none of these factors improved the model's accuracy or fit above a baseline model. However, the interaction of the gap before and after the first word was significant ($p = 0.05$), and the addition of the terms improved the fit of the model as well as the BIC. The rate at which the first and second words were typed did not improve the predictive power of the model.

One of the most surprising findings (not reported) was that adding a crossed or nested random effect of last sender did not improve the model. I had assumed that an utterance, especially a pre-utterance pause, would be effected by whether the message was a response to the partner, or a continuation of a turn by the same sender. The fact that conversation position explained so much variance is one reason for to the necessity of Study 2, which will look at the trends of dyadic pairs during the progression of a conversation.

Study 1b details

See the tables below for details of each model.

	Dependent variable: Non-opinion vs Opinion				
	Model				
Covariate	Base	Base	+ Typing	+ Gaps	+ Word Speeds
	(Fixed)	(Mixed)	Variability		+ intx
(Intercept)	-0.54^{***}	-0.54^{***}	-0.54^{***}	-0.54^{***}	-0.54^{***}
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
Pre-utterance gap	0.05^{**}	0.05^{**}	0.05^{**}	0.05^{**}	0.05^{**}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Interkey-interval (IKI)	0.07	0.08^{*}	0.07	0.07	0.06
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Utterance speed	0.07^{**}	0.08^{**}	0.08^{**}	0.08^{**}	0.08^{**}
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
IKI:speed	0.02	0.02	0.02	0.02	0.02
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Edit Count	-0.01^{**}	-0.01^{**}	-0.01^{**}	-0.01^{**}	-0.01^{**}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Speed variability (sd)			0.02	0.02	0.02
			(0.03)	(0.03)	(0.03)
Word 1-2 gap				0.002	0.001
				(0.03)	(0.03)
Pre-utt-gap:word 1-2 gap				-0.05^{*}	-0.05^{*}
				(0.03)	(0.03)
Word 1 speed					-0.03
					(0.03)
Word 2 speed					-0.004
					(0.03)
Word 1:word 2 speed					-0.04
					(0.03)
Observations	2,965	2,965	2,965	2,965	2,965
Log Likelihood	-1,744.21	-1,742.88	-1,742.66	-1,740.75	-1,739.17
AIC	3,500.42	3,499.76	3,501.32	3,501.50	3,504.33
BIC		3,541.73	3,549.27	3,561.44	3,582.26
Note:					*p<0.1; **p<0.05; ***p<0.01

Table E.1
Effects of adding covariates to a model predicting non-opinion vs opinion binary dialogue acts

	Dependent variable:
	Utterance speed
(Intercept)	−0.07
	(0.05)
Acknowledge	-0.15^{*}
	(0.06)
Closing	0.10
	(0.10)
Directive	0.39^{*}
	(0.16)
Negative-Answer	-0.30^{+}
	(0.17)
Non-opinion	0.07
	(0.05)
Opening	-0.53^{**}
	(0.17)
Opinion	0.15^{**}
	(0.05)
Question	0.26^{***}
	(0.05)
Word count	−0.002
	(0.002)
Observations	4,108
R^2	0.02
Adjusted R^2	0.01
Residual Std. Error	0.95 (df = 4099)
F Statistic	8.55^{***} (df = 8; 4099)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.2

Results of whether dialogue acts controlling for word count can predict the speed at which the utterance was produced.

	Dependent variable:
Variable	Edit count
(Intercept)	0.64^{*}
	(0.27)
Acknowledge	0.37
	(0.36)
Closing	−0.46
	(0.60)
Directive	−0.84
	(0.90)
Negative-Answer	−0.05
	(0.99)
Non-opinion	−0.45
	(0.27)
Opening	2.14^{*}
	(0.99)
Opinion	−0.48
	(0.30)
Question	−0.22
	(0.31)
Word count	0.34^{***}
	(0.01)
Observations	4,108
R^2	0.24
Adjusted R^2	0.24
Residual Std. Error	5.44 (df = 4099)
F Statistic	164.62^{***} (df = 8; 4099)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.3
Results of whether dialogue acts can predict edit counts, controlling for word count.

	Dependent variable:
Variable	Typing speed variability (sd)
(Intercept)	-0.12^{**}
	(0.05)
Acknowledge	0.17^{**}
	(0.06)
Closing	-0.16
	(0.11)
Directive	-0.24
	(0.16)
Negative-Answer	0.16
	(0.18)
Non-opinion	-0.03
	(0.05)
Opening	0.27
	(0.18)
Opinion	-0.03
	(0.05)
Question	-0.13^{*}
	(0.06*Z*)
Word count	0.01^{***}
	(0.002)
Observations	4,108
R^2	0.02
Adjusted R^2	0.02
Residual Std. Error	0.96 (df = 4099)
F Statistic	12.37^{***} (df = 8; 4099)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table E.4

Results of whether dialogue acts predict variation in typing speed, when controlling for word count.

	Dependent variable:	
	Pre-utterance gap	
Variables	Base model	+ last sender
(Intercept)	0.09^{*}	-0.11^{*}
	(0.05)	(0.05)
Acknowledge	-0.01	0.09
	(0.06)	(0.06)
Closing	0.02	0.04
	(0.10)	(0.10)
Directive	-0.08	-0.18
	(0.16)	(0.15)
Negative-Answer	-0.05	0.04
	(0.17)	(0.17)
Non-opinion	-0.05	-0.08^{+}
	(0.05)	(0.05)
Opinion	0.04	-0.003
	(0.05)	(0.05)
Question	0.12^{*}	0.09^{+}
	(0.05)	(0.05)
Word count	-0.01^{***}	-0.003
	(0.002)	(0.002)
Last sender		0.49^{***}
		(0.03)
Observations	4,069	4,069
R^2	0.01	0.06
Adjusted R^2	0.01	0.06
Residual Std. Error	0.97 (df = 4061)	0.95 (df = 4060)
F Statistic	6.02^{***} (df = 7; 4061)	34.64^{***} (df = 8; 4060)
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001	

Table E.5

Results of whether dialogue acts can predict pre-utterance gaps, when controlling for word count, and word count + who the last sender was.

	Dependent variable		
Variables	Word 1 speed	Word 2 speed	Word 1-2 gap
	(1)	(2)	(3)
(Intercept)	−0.05	0.03	−0.04
	(0.05)	(0.05)	(0.05)
DA:Acknowledge	−0.05	-0.17^{**}	0.13^{*}
	(0.06)	(0.06)	(0.06)
DA:Closing	−0.04	0.20^{+}	-0.19^{+}
	(0.11)	(0.11)	(0.11)
DA:Directive	−0.05	0.10	−0.01
	(0.16)	(0.16)	(0.16)
DA:Negative-Answer	−0.12	0.14	0.02
	(0.18)	(0.18)	(0.18)
DA:Non-opinion	0.14^{**}	0.01	−0.01
	(0.05)	(0.05)	(0.05)
DA:Opening	−0.01	-0.44^{*}	0.11
	(0.18)	(0.18)	(0.18)
DA:Opinion	0.10^{+}	0.02	0.002
	(0.05)	(0.05)	(0.05)
DA:Question	0.02	0.13^{*}	−0.05
	(0.06)	(0.06)	(0.06)
Word count	−0.002	−0.0004	0.004^{*}
	(0.002)	(0.002)	(0.002)
Observations	4,108	4,108	4,108
R^2	0.005	0.01	0.005
Adjusted R^2	0.003	0.01	0.003
Residual Std. Error (df = 4099)	0.98	0.97	0.96
F Statistic (df = 8; 4099)	2.43^{*}	3.90^{***}	2.36^{*}
Note:	+ p<0.1; * p<0.05; ** p<0.01; *** p<0.001		

Table E.6

Results of whether dialogue acts can predict the typing speed of word 1, word 2, and the length of time between words 1 and 2.

Appendix F

Manual Sentiment Analysis Guidelines

These guidelines were compiled by my research assistant, based on the guidelines set forth in Mohammad (2016).

1 - Negative

- There is an explicit or implicit clue suggesting that the speakers is conveying a negative opinion or is in a negative state
- Displeasure, anger, frustration, irritation, dislike, etc.

2 - Somewhat Negative

- There is evidence to suggest that the speaker is conveying a slightly negative opinion or is in a negative state, but is not conveying strong negativity
- On the cusp of negative feelings listed above
- Not a neutral statement due to some suggestion of negativity, but not outright negative in sentiment

3 - Neutral

- The statement is neither explicitly positive or negative. No emotional state or opinion is conveyed through the statement

4 - Somewhat Positive]

- There is evidence to suggest that the speaker is conveying a slightly positive opinion or is in a positive state, but is not conveying strong positivity
- On the cusp of positive feelings listed below
- Not a neutral statement due to some suggestion of positivity, but not outright positive in sentiment

5 - Positive

- There is an explicit or implicit clue suggesting that the speakers is conveying a positive opinion or is in a positive state
- Pleasure, optimism, joy, relaxed, admiring, like/interest, excitement, etc.