Chapter 5

Study 1: Keystroke patterns in dialogue acts

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

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

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

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

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

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

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

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

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

5.1 Methodology

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

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

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

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

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

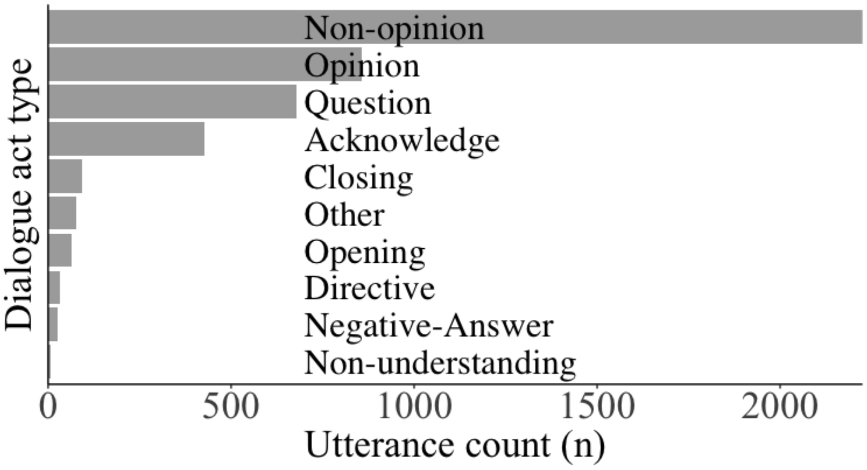

Figure 5.1
The distribution of high-level dialogue act categories within my collected dataset. These categories were mapped from the original 27 dialogue acts (see Table 5.1 for details).

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

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

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

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

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

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

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

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

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

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

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

Table 5.1

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

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

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

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

5.2 Results

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

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

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

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

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

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

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

5.2.1 Experiment 1a: Timing patterns predicting dialogue acts

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

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

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

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

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

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

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

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

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

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

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

5.2.2 Experiment 1b: Dialogue acts predicting timing patterns

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

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

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

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

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

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

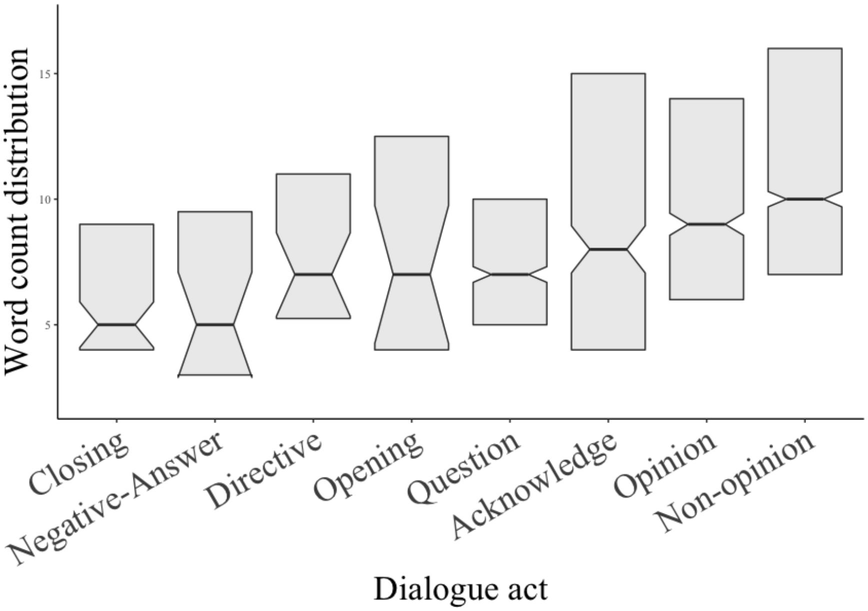

Figure 5.2
The distribution of word counts within each dialogue act, ordered from shortest word count to longest word count. Closing DAs have the shortest average word count will non-opinion statements have the longest average word count. This result makes sense since closings are more formalized and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

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

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

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

Table 5.4

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

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

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

Table 5.5

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.3 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.4 Future work

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

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

5.4.1 Limitations of future work

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

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

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

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

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

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

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

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

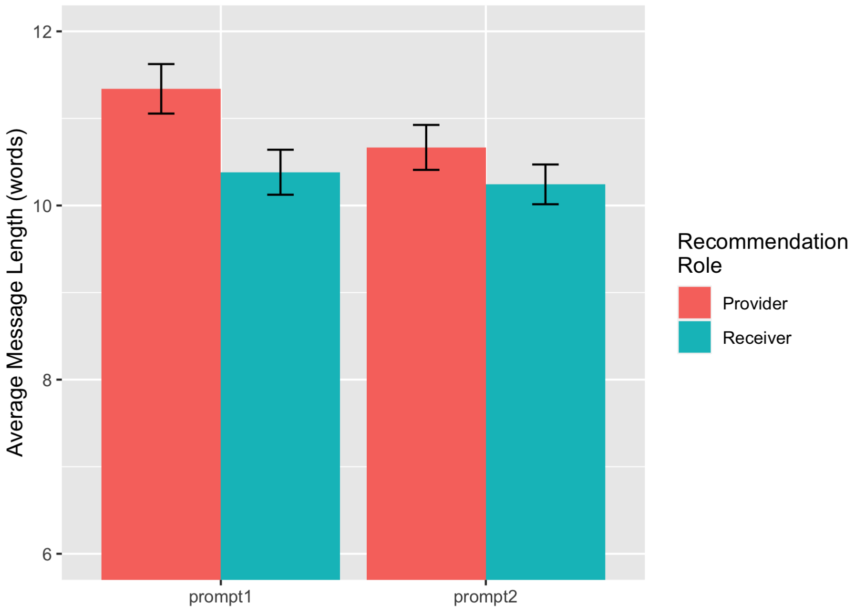

Figure 5.3
Word counts differ significantly depending on the recommendation role of the participant. Participants played different roles in each half of the conversation, referred to as Prompt 1 and Prompt 2.