Entrainment in Human-to-Human Dialogue and its Application in End-to-End Dialogue Systems

by

Mingzhi Yu

B.S. in Computer Science, University of Delaware, 2015

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Mingzhi Yu

It will be defended on

November 5th 2021

and approved by

Diane Litman, The Department of Computer Science, University of Pittsburgh

Erin Walker, The Department of Computer Science, University of Pittsburgh

Adriana Kovashka, The Department of Computer Science, University of Pittsburgh

Shuang Ma, Microsoft AI&Research, Microsoft Corp., Redmond

Entrainment in Human-to-Human Dialogue and its Application in End-to-End Dialogue Systems

Mingzhi Yu, PhD

University of Pittsburgh, 2021

Entrainment is a linguistic phenomenon in which people mimic each other in their conversations. It occurs in a wide range of linguistic dimensions. Entrainment has been exploited in various natural language processing tasks related to dialogue, such as dialogue outcome prediction and dialogue response generation. However, only a few studies have attempted to incorporate entrainment into neural network-based dialogue systems systematically. The present thesis aims to build a neural network-based end-to-end response generation model capable of generating diverse responses by leveraging lexical entrainment, a type of entrainment based on text features. We first demonstrate an automatic entrainment measure relying on conventional similarity metrics based on a bag-of-words approach. Then we show an alternative neural network-based approach to perform the same core similarity measure for entrainment quantification. Lastly, we proposed an end-to-end dialogue response generation model that controls entrainment degree to aid response diversity. We will focus on investigating the effect of incorporating lexical entrainment in the end-to-end dialogue response generation model.

Table of Contents

List of Tables

List of Figures

Preface

This thesis concludes my Ph.D. study at the University of Pittsburgh from 2016 to 2021.

I want to thank my advisor, Dr. Diane Litman, for being my mentor and teacher over the past 6 years. Dr. Litman guides me through the journey of my study with her genuine advice and thoughtful suggestion. During these years of working with her, she showed me how to be a hard-working researcher who always looks for a precise answer with rigorous academic thinking. I sincerely appreciate her mentorship and responsive help over these years. I would also like to thank my dissertation committee: Dr. Erin Walker, Dr. Adriana Kovashka, and Dr. Ma, from whom I received so many insightful comments and, of course, challenging but inspiring questions for this dissertation. I also want to thank all the peers in the PETAL lab led by Dr. Litman and the Facet lab led by Dr.Walker. I feel thankful for the research feedbacks given by my labmates and the encouragement we gave to each other.

I want to express my deepest gratitude to my parents. They gave me tremendous supports and love. Without my parents, I would have never had the opportunity to study computer science. I would also like to thank my husband Dengchao, who has been very understanding and supportive since day 1 I started my Ph.D. Without my family, this dissertation would have never existed.

I would also thank my friends for we share so many good memories: Xiaoyu Liang, Kaiyu Shen, Wei Guo, Fangzhou Cheng, Haoran Zhang, Xue Ning, Zhipeng Luo, Fan Zhang, Wenchen Wang, Nannan Wen, Mingda Zhang, Keren Ye, Zinan Zhang, Xiaozhong Zhang, Longhao Li, Tazin Afrin, Luca Lugini, and Ahmed Magooda.

1.0 Introduction

A long-term goal of automatic dialogue systems is to build a human-like conversational agent. A critical type of conversational agent is designed explicitly for non-goal-oriented dialogue. Since the advent of data-driven neural dialogue models, many researchers have developed end-to-end model solutions for non-task-oriented dialogue. Perhaps the most popular type of recent end-to-end model is the sequence to sequence (S2S) model. However, the S2S model can suffer from a *safe response* issue, that is, the dialogue system tends to generate trivial responses such as "I don't know" or "yes". There are many studies that attempt to address this issue. In this study, we proposed a novel approach that leverages a linguistic phenomenon–entrainment–to build a dialogue response generation system capable of generating natural conversation.

Entrainment is the phenomenon in which individuals unconsciously mimic each other in their conversations. It can be found in various aspects of language such as using the same lexical items to describe the same concepts [10], converging in the speech pitch and intensity [52], or having a similar speech or writing style [33]. There has been much study of this phenomenon across a wide range of linguistic dimensions such as acoustic and prosodic [52, 58], lexical [9] and syntactical [8]. As entrainment has been found as an indicator in team collaboration, and thus, it has also been frequently exploited in interdisciplinary studies combining psychology and linguistics [33, 84]. Furthermore, previous works in dialogue systems showed that entrainment leads to some promising improvements in the performance of systems, for example, the word error rate (WER) for spoken dialogue systems [61], and BLEU scores–a scale for evaluating language generation by text alignment [80]–for the response generation system [44]. In this study, we will leverage entrainment in a generative dialogue model that produces the next utterance appropriate for a given dialogue context. The model targets to generate diverse responses with entrainment on the fly. Figure 1 shows a chit-chat example between a human user and a conversational agent. The first response "I don't know" is an example of a dialogue system generating a generic response lacking diversity due to the safe response issue. If we incorporate entrainment into the system, the agent can generate a response such as the second response. The second response, "What is a lottery", exhibits lexical entrainment in echoing the word "lottery" from the context.

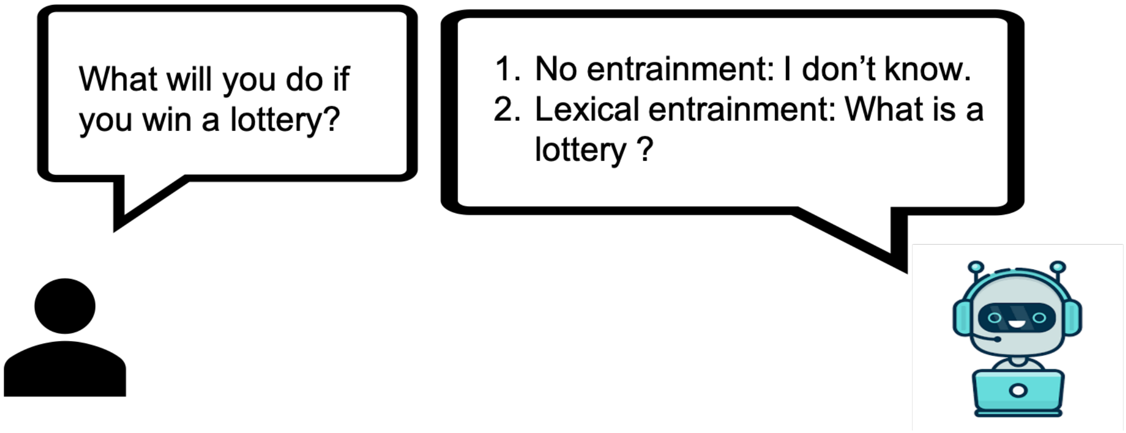

What will you do if you win a lottery?
1. No entrainment: I don't know.
2. Lexical entrainment: What is a lottery ?

Figure 1: An example of a chit-chat between a human user and a conversational agent.

One challenge of building an entrainable dialogue response generation model is to develop an accurate matching algorithm that can measure the degree of entrainment. Although many works have focused on this goal, most of these systems are based on bag-of-words paradigm that relies on matching the word count of specific lexical items [83, 17]. The common method of this measure is to define a set of linguistic markers and compare their word usages. In Chapter 4 of this study, we follow a similar method to measure lexical entrainment based on the bag-of-words approach. We modify an existing bag-of-words approach to calculate lexical entrainment and applied that altered method to measure entrainment in a multiparty dialogue corpus. We validate this measure by predicting dialogue success and failure using entrainment. Furthermore, we also investigate how team characteristics impact entrainment.

The core instrument of the above bag-of-words approach is a similarity measure between conversational partners on linguistic markers. This can lead to concerns of inadequate linguistic representation [74]. To address this concern, recent studies in lexical entrainment start to use the linguistic representation in high dimensional spaces [74, 44]. In Chapter 5, we proposed a neural network-based matching model to perform the core similarity measure for entrainment. Our model aims at automatically learning and matching global linguistic features between dialogue context and responses by utilizing a novel attention-based component in our neural network model. The novel component projects the input representation to a universal latent space and then generalizes global linguistic features such as style, structure, and shared semantics across all input sequences. We first validate this model against a popular dialogue response matching benchmark, and we further investigate the effect of leveraging the global linguistic features. By using our model to perform the similarity measure for entrainment, we observed a stronger entrainment signal in a corpus-based entrainment study compared to the conventional bag-of-words approach used in Chapter 4.

In Chapter 6, we will further incorporate entrainment into an end-to-end dialogue response generation model to generate more natural and diverse responses. We will utilize the neural network-based similarity measure in Chapter 5 and train our model following a two-stages training strategy. We will examine whether entrainment can be a useful tool to address the *safe response* issue of sequence-to-sequence dialogue response generation models. The reported model is an entrainable end-to-end dialogue response generation model that can generate diverse responses on the fly.

1.1 Contribution

This work has a broader impact on the dialogue community, for it proposes a dialogue response generation model by leveraging entrainment. The approach introduces new insights to incorporate linguistics signals into data-driven end-to-end dialogue models. The proposed study also adds some novel efforts to improve the end-to-end models for dialogue response matching and generation.

For the linguistics community, this work introduces a neural method to measure linguistic entrainment computationally. We investigate three topics: the existence of multiparty lexical entrainment, prediction of dialogue success by entrainment, and the relationship between team characteristics and lexical entrainment. We then propose a data-driven method for the entrainment similarity measure to aid representation learning and obviate the need for hand-crafted features. Our approach is an unsupervised methodology that can be applied to other unstructured dialogue datasets for entrainment analysis.

For the machine learning community, we propose introducing an attention-based architecture used in other research areas, such as speech synthesis, into dialogue response matching tasks. The architecture aids input representation by generalizing universal features across individual inputs under minimal supervision. Furthermore, we incorporate entrainment into a Transformer model to perform the dialogue response generation task.

1.2 Research Questions and Hypotheses

The final goal of the present thesis is to construct and evaluate an entrainable dialogue generation model. Each chapter is devoted to one or more specific hypotheses as described below.

Chapter 4: We proposed a lexical entrainment measure based on an existing bag-of-words approach based linguistic style matching. We predicted the dialogue success and failure of teams corpus using this measure. We also investigate the relationship between lexical entrainment and team characteristics by testing the following hypotheses:

- **Hypothesis 1**: The entrainment measure can strengthen the prediction of dialogue success and failure beyond team characteristics.

- **Hypothesis 2**: The entrainment measure is significantly related to team characteristics, i.e., team size, ethnicity, age, and gender diversity.

Chapter 5: We proposed a new similarity measure for lexical entrainment using a neural network model. Our model is a dialogue response matching model that matches a dialogue context and response. Compared to the bag-of-words approach, our approach is context-aware, and it further leverages global linguistic features generalized from inherent input representation. The global linguistic features represent a high-level abstraction of input, such as sentence structures or shared semantics. The following are our hypotheses:

- **Hypothesis 3**: Learning the global linguistic features, the model will achieve better dialogue response matching than the baseline models that do not leverage the global features. We evaluated our model by an extrinsic evaluation related to a dialogue response matching task.
- **Hypothesis 4**: Using the neural network-based similarity measure for entrainment calculation, we will observe stronger group entrainment signals in a corpus-based entrainment analysis. Those signals are more predictive of dialogue success and failure.

Chapter 6: This chapter aims to build a dialogue system that can generate diverse responses by leveraging entrainment. We showed that incorporating entrainment into an end-to-end dialogue generation system can improve the variability of generated responses. Our hypotheses are:

- **Hypothesis 5** Our response generation model will generate responses with entrainment.
- **Hypothesis 6** Our model will generate more diverse responses. The overall response quality should be satisfactory considering both fluency and diversity.

1.3 Outline

In Chapter 2, we review the literature on linguistic entrainment. We discuss different entrainment dimensions, entrainment frameworks, and entrainment features in the field of computational linguistics.

In Chapter 3, we discuss the datasets used in this thesis. In Chapter 4, we first experiment with an existing bag-of-words approach of the similarity measure. We conduct a corpus-based study on lexical entrainment in multiparty dialogue. We demonstrate lexical entrainment in multiparty dialogue, and investigate how it is associated with different team characteristics, including team size, gender composition, and ethnic composition. We then show how lexical entrainment is predictive of the group relationship, which can be viewed as an essential indicator of dialogue success in the non-task-oriented dialogue.

In Chapter 5, we introduce a neural network model to perform as the core similarity measure for lexical entrainment. The model is a type of dialogue response matching model that matches a dialogue context and response. We first validate the model design in a popular dialogue response matching task. Then we perform a corpus-based entrainment analysis by using our model to perform the similarity measure for entrainment. We show that our approach results in a stronger entrainment signal compared to a baseline approach following the bag-of-word paradigm.

In Chapter 6, we propose an approach to incorporate lexical entrainment into a neural end-to-end dialogue response generation model. We first validate our model on generating responses with entrainment. Then we evaluate the diversity of generated responses by controlling entrainment degree. The evaluation includes both automatic metrics and human judgment. Automatic evaluation suggests that our model results in a good improvement in diversity with good overall quality, but the human evaluation only shows marginally improved overall quality rather than diversity, which implies that there is no outstanding advantage to use entrainment for the safe response issue.

The last Chapter 7 is a summary of this dissertation. We list a set of hypotheses and the corresponding conclusions in this dissertation. Additionally, we also discuss the limitation and future applications.

2.0 Linguistic Entrainment in Conversations

In this chapter, we review the literature in linguistic entrainment with a specific focus on the field of computational linguistics. In addition to the overview in this chapter, each chapter contains a section with a review of works specifically related to the chapter.

2.1 Entrainment in Linguistic Dimensions

Researchers have found substantial evidence for entrainment in many linguistic dimensions. Table 1 contains five trendy linguistic dimensions that previous studies have focused on, along with some related works. Acoustic-prosodic entrainment entails matching specific speech features, such as speech pitch, intensity, and accent. Lexical entrainment occurs when conversational partners develop or choose the same terms to describe the same objects. Linguistic style entrainment is the matching of language style unrelated to the actual conversational content. Syntactic entrainment is the coordination in the sentence syntactic structure. The research presented a mixture of different types of entrainment, including lexical, semantic, syntactic, and style.

Table 1: Linguistic entrainment dimensions and related studies.

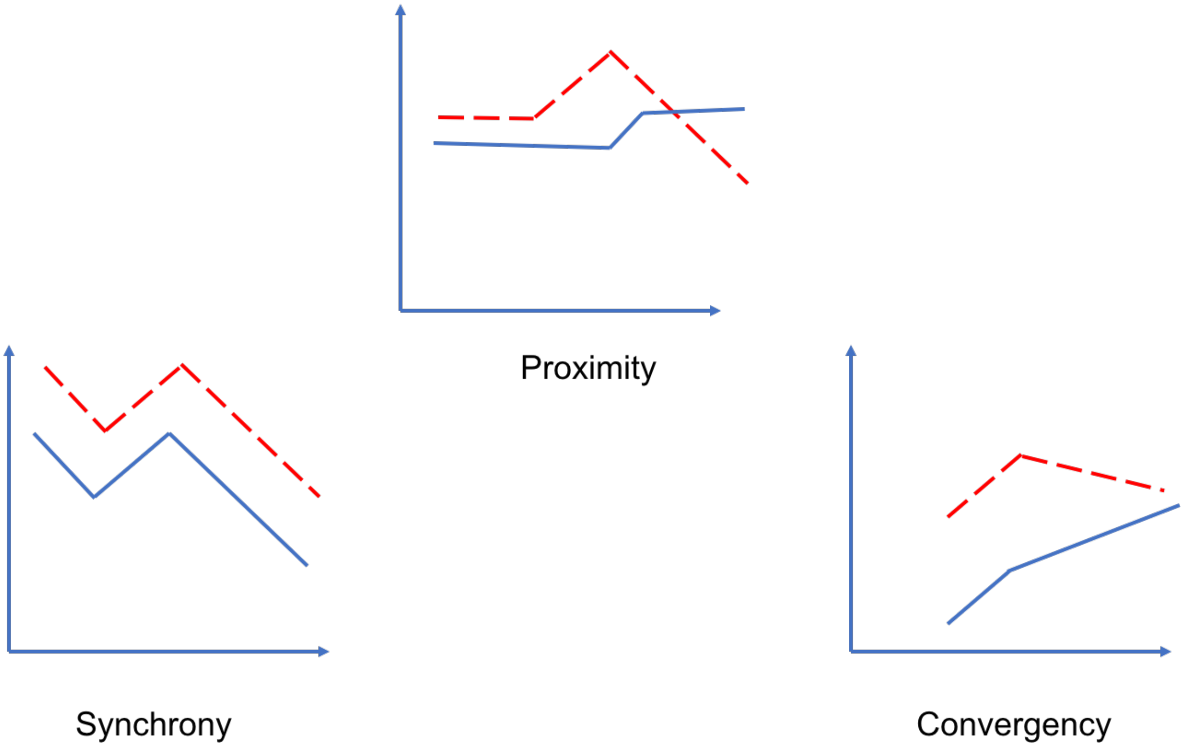
Dimensions	Related Studies
Acoustic-Prosodic	[50, 52, 116, 65, 7, 47, 48, 34]
Lexical	[10, 9, 116, 76, 90, 108, 103]
Syntactic	[8, 103, 16, 92, 13]
Linguistic Style	[83, 33, 19, 18, 77, 106]
Speech dynamics	[53, 51, 41, 24, 69]

2.2 Framework to Model Entrainment

Levitan and Hirschberg [50] introduce a framework to model entrainment. They categorize entrainment into three types: **Proximity**, **Convergence**, and **Synchrony**. In their work, they measure entrainment at the global and local levels. The local-level measure concerns changes in feature values over short periods. The counterpart is the global measure. Overall, **Proximity** reflects the similarity across conversational partners. **Convergence** entails the increase in partner similarity. **Synchrony** is the accordance across partners. Figure 2 illustrates these three types. Many recent studies follow this framework to define a set of rule-based measures to quantify entrainment [117, 90, 28, 52, 127, 58]. Recently, Wynn and Borrie [123] expand this framework and propose eight entrainment types depending on three factors: **Class**, **Level**, and **Dynamicity**. Table 2 shows theses factors and their corresponding values. The definitions of **Proximity** and **Synchrony** are consistent with [50]. **Local** level measures entrainment at turn exchanges. **Global** level measures entrainment over some time session. **Static** dynamicity measures entrainment as a static variable. In contrast, **Dynamic** dynamicity is the change of entrainment over time. **Convergence** belongs to this category. In this thesis, entrainment measures are in the class of **Local Static Proximity** and **Global Dynamic Proximity**. In Chapter 4, we first measure a lexical similarity across conversational partners per turn. The similarity is modeled as a static variable per turn rather than a time sequence. Additionally, we also measure Convergence over intervals. Therefore, we view Chapter 4 as an investigation of **Local Static Proximity** and **Global Dynamic Proximity**. For the similar reasons, Chapter 5 is also related to **Local Static Proximity** and **Global Dynamic Proximity** because in Chapter 5, we just replace a bag-of-words similarity measure with a neural approach. The Chapter 6 incorporates entrainment in a response generation task for each turn. Thus it belongs to the category of **Local Static Proximity**.

2.3 Entrainment Measures

Studies have developed various entrainment measures for each dimension. In the extant corpus-based studies, many measures are based on hand-crafted features. Table 3 lists distinct features



Proximity

Synchrony

Convergency

Figure 2: Three classes of entrainment are defined by Levitan and Hirschberg [50]. The x-axis represents time, and the y-axis represents the feature value. The dashed line and solid line show feature values of two individual interlocutors over time.

Table 2: Classification factors of entrainment defined by Wynn and Borrie [123].

Factors	
Class	Proximity or Synchrony
Level	Local or Global
Dynamicity	Static or Dynamic

Table 3: A list of features and some related studies

Dimension	Features
acoustic-prosodic	intensity, pitch, voice quality, speaking rate [52, 58, 116, 65]
lexical	count of high-frequency words , topic words [90, 76]
syntactic	count of key verbs or phrases indicating syntactic structure [8, 103, 16, 92]
linguistic style	count of function words [33, 127, 85]
speech dynamics	pause duration, count of turn-taking type, utterance length [24, 53, 51, 41]

along with their related studies. Intensity, pitch, voice quality, and speaking rate are common features used frequently in acoustic-prosodic entrainment. For lexical and linguistic style entrainment, perhaps the most popular feature is the occurrence count of a set of predetermined linguistic markers, such as high-frequency words, topic words, and function words. Similarly, in syntactic dimensions, measures often focus on the usage proportion of a key phrase set that epitomizes the syntactic structure of sentences. Some example features are pause duration, count of the turn-taking type, and utterance length in speech dynamics. At the time of this thesis, no latent structures have been found among those lower-level entrainment features [117].

Features can be extracted at different levels. Some popular choices of extraction levels are by turns [19, 17], multiple turns [74], interpausal units (IPUs) and time intervals [90, 127, 50] that are proportional to the conversation. In Chapter 4 and 5, features are extracted by IPUs because data used in these 2 chapters are transcribed based on IPUs. Entrainment is measured as convergence based on time intervals in Chapter 5. In Chapter 6, features are extracted by conversational turns because data used in this chapter is based on conversational turns.

2.4 Evaluate Entrainment Measures

Early experiment-based studies examined entrainment measures by setting experimental conditions and control groups [10, 8, 30]. A representative study was performed by Brennan and Clark [10]. In their experiment, these researchers developed three experiments where partners can gradually develop lexical entrainment. In more recent corpus-based studies, evaluation of entrainment measures is mostly extrinsic. For example, existing works have attempted to compare entrainment between conversational partners and non-partners [50, 90]. Another popular stream of approach includes associating entrainment to other interpersonal behaviors in dialogue, such as group relationships [127], positive or negative effects [74], being liked by partners [52], and dialogue success [44]. Additionally, using entrainment to distinguish artificial fake conversation and real conversation is also a common practice [75, 116, 88]. Our study evaluates our entrainment measure by predicting dialogue success, correlating with existing measures, and distinguishing between fake and real conversation.

3.0 Dataset Overview

There are many human dialogue datasets available to train data-driven dialogue systems. Many of them have been used in open tasks to build dialogue systems. In this study, we specifically focus on constrained corpora in which the conversations are limited to specific topics. Compared to a chit-chat corpus, conversations in a constrained corpus are limited to particular topics. Compared to the fully task-oriented conversation, conversations in the constrained corpus are less structured and have ambiguous dialogue states because the conversations are somewhat spontaneous. These features make the constrained corpus especially useful for building a non-task-oriented dialogue model. We further restrict the scope of this study to a particular set of datasets depending on specific interests and data availability. We treat written and spoken dialogue as equivalently, without preference for spoken or written dialogue. In this section, we first introduce the datasets for the study. The details of each dataset, such as description, dataset characteristics, and other statistics, are discussed. We then discuss the natures and drawbacks of each dataset when using them for building data-driven models. We also justify the reasons for selecting each dataset. Other task-dependent data uses, such as processing, are elaborated in each chapter. Overall, we select four public constrained corpora. Two **essential** corpora where lexical entrainment has been demonstrated will be used as the major data source in this thesis. Two **non-essential** corpora will be used in the ablation studies of Chapter 5 to support our further investigation about a proposed model component. The size of these corpora differs depending on our usage. Table 4 shows the statistics of each dataset. Dialogue examples are included in the Appendix A.

Table 4: Dataset statistics

Roles in this thesis	Names	Multiparty or Dyadic	Numbers of Utterances	Numbers of Conversations	Numbers of Speakers	Avg. # turns/IPUs per dialogue	Used in Chapters	Dialogue Type
Essential	Teams Corpus	Multiparty	∼66, 000	124	213	532 IPUs	4, 5	Spoken
Essential	Wikipedia Talk Page	Multiparty	∼391,000	125,292	38,462	3 turns	6	Written
Non-Essential	Ubuntu Dialogue Corpus	Dyadic	∼7,000,000	930,000	Unknown	8 turns	5	Written
Non-Essential	Douban Corpus	Dyadic	∼7,000, 000	1,000, 000	Unknown	7 turns	5	Written

3.1 Teams Corpus (Essential)

Teams Corpus [58] is a small-scale **multiparty** spoken dialogue dataset. It consists of 128 multiparty conversations elicited from 213 native speakers of American English. Each group of speakers participated in a collaborative game called Forbidden Island. The conversation during the game is recorded and transcribed. This corpus can be used for training a small neural network models. It is also used in this thesis for examining new approaches to measure entrainment, because entrainment has already been demonstrated in this corpus by other previous work. So we can easily find baselines to compare to in our study. The dataset is also handy for performing case studies for specific tasks. Furthermore, this dataset also provides information about speaker persona and a survey about group relationships (See the Appendix B). A drawback of using this dataset is that it contains a significantly smaller volume of data than other more frequently used corpora to build dialogue systems. Therefore, it is challenging to fit this small dataset into a large model. We select this dataset because there are many entrainment-related studies based on it, which allows us to establish comparisons with other works. Figure 3 is a conversation excerpt.

3.2 Wikipedia Talk Page Corpus ((Essential)

Wikipedia Talk Page Corpus is a collection of written conversations among Wikipedia editors. The dataset is also publicly available at ConvoKit [1]. Conversations in this dataset include the interactions concerning edits on Wikipedia articles and the discussion on the open nomination for admins election. Multiple editors can post comments on the same topic, which leads to a **multiparty** conversation. Using this dataset, Danescu-Niculescu-Mizil et al. [19] finds users entrain more to admins, and admins entrain more to non-admins. Overall, lexical entrainment has been demonstrated in this corpus. Although this is a medium-size dataset compared to the Ubuntu Dialogue Corpus, it is large enough to train our proposed neural model. We used this model to train dialogue generation models (see Chapter 6). Figure 4 shows a dialogue excerpt.

[1] https://convokit.cornell.edu/

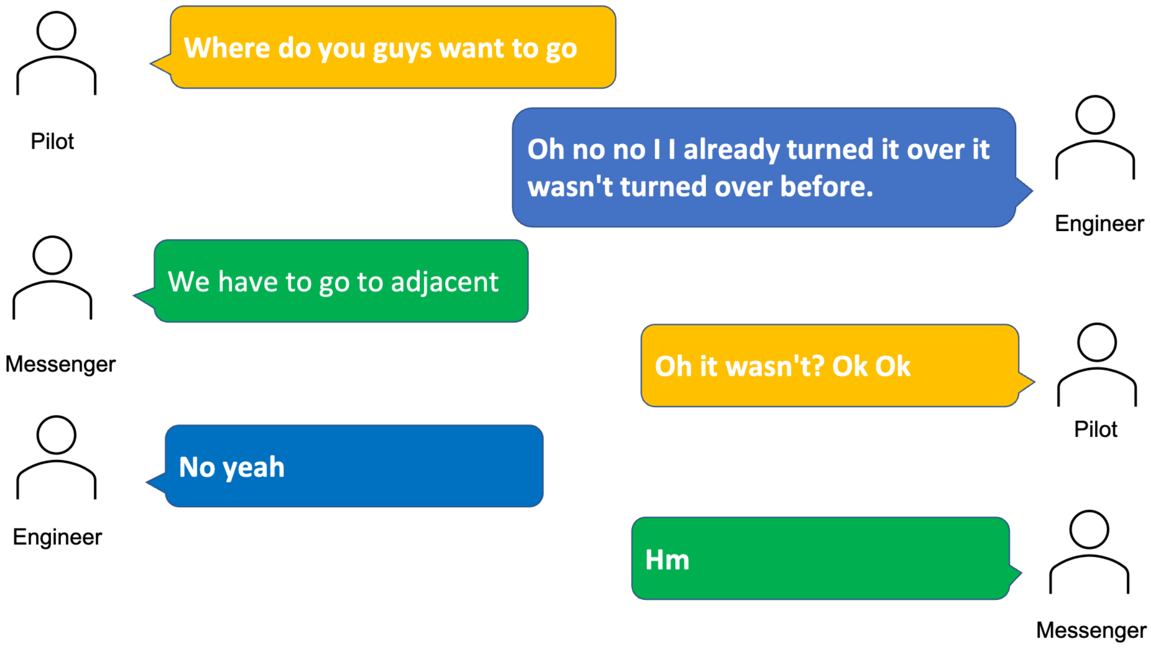
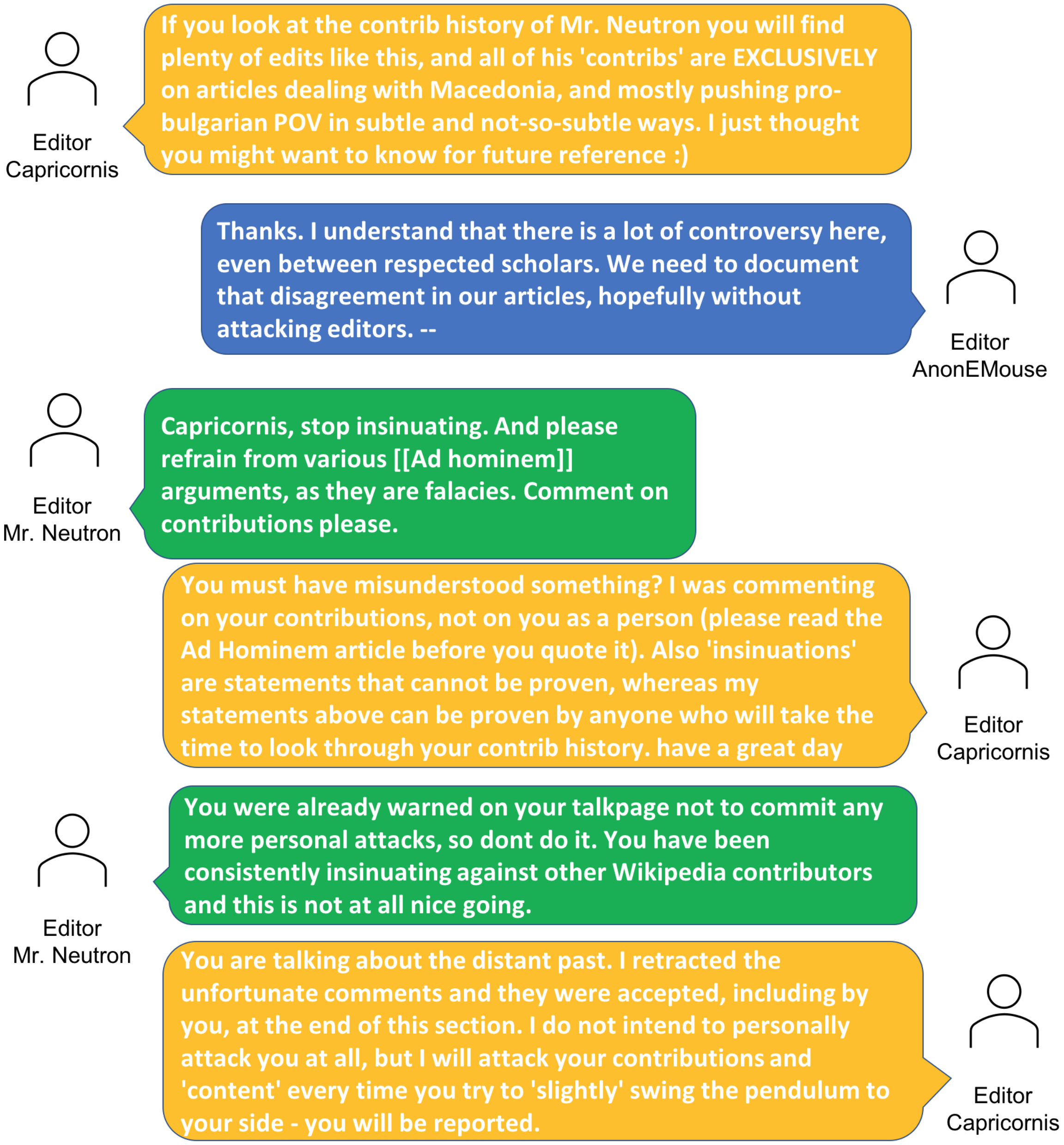

Pilot
Where do you guys want to go
Oh no no I I already turned it over it wasn't turned over before.
Engineer
Messenger
We have to go to adjacent
Oh it wasn't? Ok Ok
Pilot
Engineer
No yeah
Hm
Messenger

Figure 3: A dialogue excerpt from the Teams Corpus

3.3 Ubuntu Dialogue Corpus V1 (Non-essential)

The Ubuntu Dialogue Corpus (V1) [62] is a large-scale written dialogue dataset. The dataset is popular for studies in which neural dialogue systems are constructed [140, 97, 63]. These data are extracted from the discussion threads in the Ubuntu IRC channel. In this channel, users can ask technical questions about the Ubuntu system, and other users can join the discussion to answer questions. The Ubuntu Dialogue Corpus contains many disentangled **dyadic** dialogues that can be used to train a data-driven neural dialogue model. Because these data mainly contain technical topics related to Ubuntu systems, they are particularly well-suited to train a dialogue model for technical support. However, several drawbacks remain for their usage. First, the dataset contains many out-of-vocabulary (oov) words such as file paths, commands, typos, and acronyms. These oov can lead to a large token size for the neural network model training, causing potential problems to construct a meaningful embedding space for input. We chose Ubuntu datasets for two primary reasons. The first is that it is easy to find comparable baseline models because many researchers



Editor
Capricornis

If you look at the contrib history of Mr. Neutron you will find plenty of edits like this, and all of his 'contribs' are EXCLUSIVELY on articles dealing with Macedonia, and mostly pushing pro-bulgarian POV in subtle and not-so-subtle ways. I just thought you might want to know for future reference :)

Editor
AnonEMouse

Thanks. I understand that there is a lot of controversy here, even between respected scholars. We need to document that disagreement in our articles, hopefully without attacking editors. --

Editor
Mr. Neutron

Capricornis, stop insinuating. And please refrain from various [[Ad hominem]] arguments, as they are falacies. Comment on contributions please.

Editor
Capricornis

You must have misunderstood something? I was commenting on your contributions, not on you as a person (please read the Ad Hominem article before you quote it). Also 'insinuations' are statements that cannot be proven, whereas my statements above can be proven by anyone who will take the time to look through your contrib history. have a great day

Editor
Mr. Neutron

You were already warned on your talkpage not to commit any more personal attacks, so dont do it. You have been consistently insinuating against other Wikipedia contributors and this is not at all nice going.

Editor
Capricornis

You are talking about the distant past. I retracted the unfortunate comments and they were accepted, including by you, at the end of this section. I do not intend to personally attack you at all, but I will attack your contributions and 'content' every time you try to 'slightly' swing the pendulum to your side - you will be reported.

Figure 4: A dialogue excerpt from the Wikipedia Talk Page Corpus. Text from each editor is shown in the same color.

have used them. The second is that it can provide a relatively large amount of dialogue data to train neural models. Figure 6 shows a dialogue excerpt between two users.

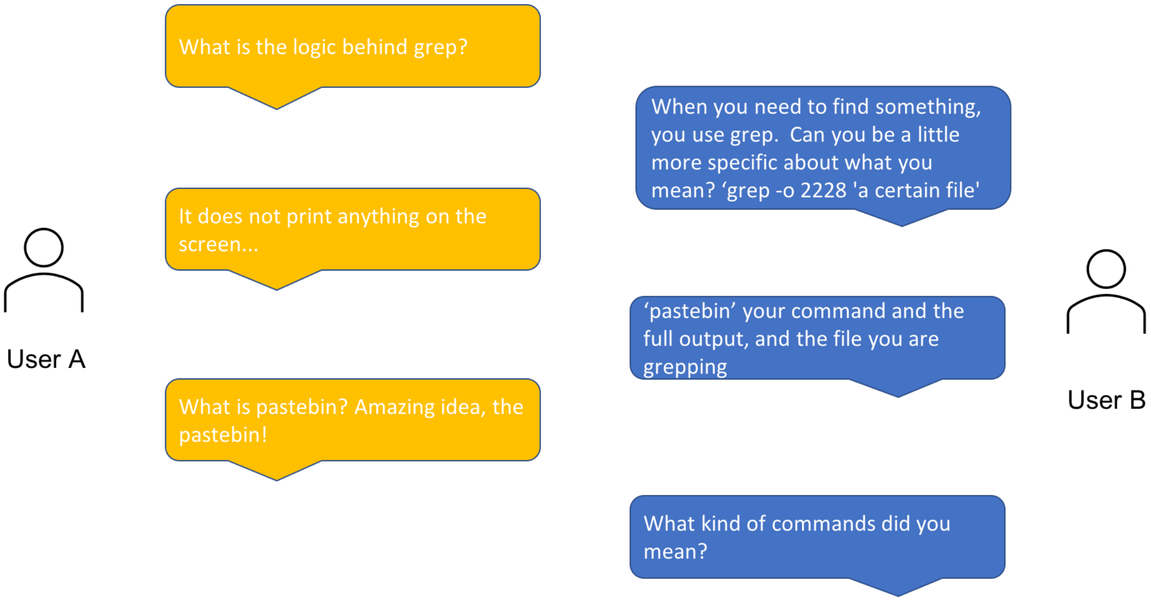
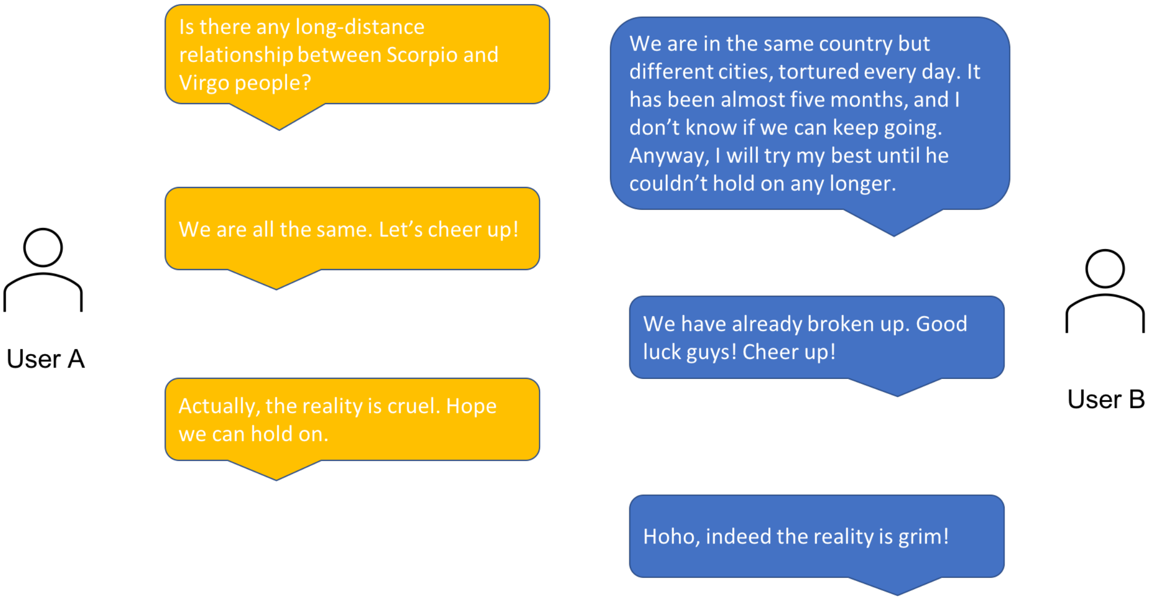

What is the logic behind grep?
When you need to find something, you use grep. Can you be a little more specific about what you mean? ‘grep -o 2228 'a certain file'
It does not print anything on the screen...
User A
‘pastebin’ your command and the full output, and the file you are grepping
User B
What is pastebin? Amazing idea, the pastebin!
What kind of commands did you mean?

Figure 5: An excerpt from Ubuntu Dialogue Corpus

3.4 Douban Corpus (Non-essential)

The Douban Corpus is a Chinese conversational written dialogue corpus provided in Zhou et al. [140]. Douban Corpus is a popular dataset frequently used as a benchmark corpus for the response selection task. The corpus contains 1 million **dyadic** Chinese dialogues crawled from the Douban group, a popular social network in China. It covers a wide range of chit-chat topics such as entertainment and daily life. These characteristics make the Douban Corpus ideal to train or test non-goal-oriented dialogue systems. In this work, for the same reason of using Ubuntu, we included Douban Corpus mainly to investigate a proposed model component and to establish a comparison with other baseline models in Chapter 5.



Is there any long-distance relationship between Scorpio and Virgo people?

We are in the same country but different cities, tortured every day. It has been almost five months, and I don’t know if we can keep going. Anyway, I will try my best until he couldn’t hold on any longer.

We are all the same. Let’s cheer up!

We have already broken up. Good luck guys! Cheer up!

Actually, the reality is cruel. Hope we can hold on.

Hoho, indeed the reality is grim!

Figure 6: An excerpt from Douban Corpus. The example is copied from a prior study of Lin et al. [57]. The original text is in Chinese. To show an example, the authors translate it into English.

4.0 Measuring Multiparty Lexical Entrainment and Investigate its Relationship to Team Characteristics and Dialogue Success

4.1 Introduction

In this chapter, we first follow the popular word count based approach of similar measure that can be used to quantify entrainment. Specifically, we design this measure by adapting linguistic style features from Pennebaker and King [83]. We apply our measure on Teams Corpus, a publicly available spoken dialogue corpus. Since Teams Corpus is a multiparty dialogue corpus, we further design our measure by extending the measure for multiparty entrainment from Litman et al. [58]. Our goal is to use lexical entrainment to predict the dialogue success indicated by perceived team social outcomes in Teams Corpus.

However, prior studies have found some team characteristics can impact entrainment [27, 33]. Therefore, by using our measure, we then further investigate whether group or team characteristics relate to multiparty entrainment since multiple individuals simultaneously engage in the same conversation. While dyad research has analyzed entrainment and gender composition [52, 73], relationships between team characteristics and multiparty entrainment could be more complex, given the increasing number of person-to-person and person-to-team communications. Three types of team characteristics are investigated: gender composition as in prior work, as well as team size and diversity. Meanwhile, other studies have shown that team characteristics can also impact on team processes [79, 99, 26]. For instance, team size is negatively correlated with team conflict [3]. Therefore, we hypothesize that both entrainment and team characteristics, specifically team size, gender, age and ethnic diversity of a team, are associated with the perception of team social outcomes, which is an indicator of dialogue success and failure. Then we use hierarchical regression models to examine the contribution of multiparty entrainment in explaining perceived team social outcomes above and beyond team characteristics.

In general, we propose a lexical entrainment measure based on a popular approach. We predict the dialogue success and failure of Teams Corpus using our entrainment measure. We also investigate the relationship between lexical entrainment and team characteristics.

The following are hypotheses we are going to examine in this Chapter (The numbering of the hypotheses are based on Chapter 1.2):

1. **Hypothesis 1 (H1)**: Our entrainment measure can strengthen the prediction of dialogue success and failure beyond team characteristics.

2. **Hypothesis 2 (H2)**: Entrainment is significantly related to team characteristics, i.e. team size, ethnic, age and gender diversity.

This work was published at the 32nd International Florida Artificial Intelligence Research Society Conference.

4.2 Dataset

In this study, we use the Teams Corpus mentioned in chapter 3. The freely available Teams Corpus [58] consists of 47 hours of audio and transcriptions from 62 teams (35 three-person, 27 four-person). The audio files are manually segmented and transcribed at the level of inter-pausal units (IPUs), based on a pause length of 200 milliseconds. Each team consists of American native speakers from 18 to 67 years old who played two rounds of the cooperative board game Forbidden Island. 213 individuals (79 males, 134 females) were assigned to the teams and given one of four game roles: Engineer, Messenger, Pilot, Explorer.

The corpus also includes survey data. A pre-game survey collected personal information such as age, gender, and eight options for ethnicity. While each participant could choose multiple options, in this work we categorize each speaker into nine exclusive categories: Caucasian (150), East Asian (12), South Asian (11), Pacific Islander (0), Black (15), Native American (0), Hispanic (3), Middle Eastern (2), and Multiple Ethnicity (20) for participants who chose more than one of the other categories. The gender data yields seven types of team gender composition: 0% female (2), 25% female (4), 33% female (7), 50% female (9), 66% female (18), 75% female (10), 100% female (12). Participants also took post-game surveys to evaluate team processes. These surveys contained a series of self-report questions on team cohesion, satisfaction, and other team social outcome constructs.

The work presented here are based only on the data related to the first of the two games, as only these transcriptions were available when the work is published. We are not planning to replicate the experiments on the second games since the main focus of this thesis is the neural network approach for entrainment measure. The goal of this Chapter is to served as a baseline for the next Chapter 5. In the next Chapter, we will utilize both game one and two. Before computing entrainment, we further processed these transcripts by removing punctuation, converting all words to lower case, and removing a list of interjections, e.g., 'hmm', that are not discussed in linguistic style [83]. We then concatenated all the processed IPU transcriptions for each speaker.

4.3 Linguistic Style Features

Before computing entrainment, we first extracted linguistic style features for each speaker in each transcript using LIWC2007 [85], a computational application for text analysis that includes a dictionary mapping a list of words to 64 psychological and linguistic categories. We used this dictionary to label each word in each speaker's concatenated IPU transcripts with potentially multiple LIWC categories. The final number of occurrences of each category was then converted into percentage.

In our study we only focused on a limited subset of LIWC categories, namely function words. The first reason is that function words reflect the speaker's psychological state and convey information about the interactive process [15]. Function words represent a high-level linguistic difference in style. Second, in contrast to content words, function words do not rely on any specific task domain [33] and have a very high frequency in daily speech [93]. Using function words as features can alleviate feature sparsity. Since a considerable number of studies about linguistic style have used function words [33, 18, 72], we directly adopted the 9 LIWC categories as function words in [33].

Figure 7 shows how we used LIWC to create function word features from a transcript excerpt. After the transcript preprocessing and speaker IPU concatenation discussed above, LIWC scored each speaker's input text and generated the category percentages for each of the 64 categories. For instance, the Engineer uttered 24 words in this excerpt but only one word belongs to the category negate. Thus, the category percentage for negate is 1/24 = 4.20%. Since one word may belong to multiple categories, the sum of category percentages for the 64 categories may exceed 100.

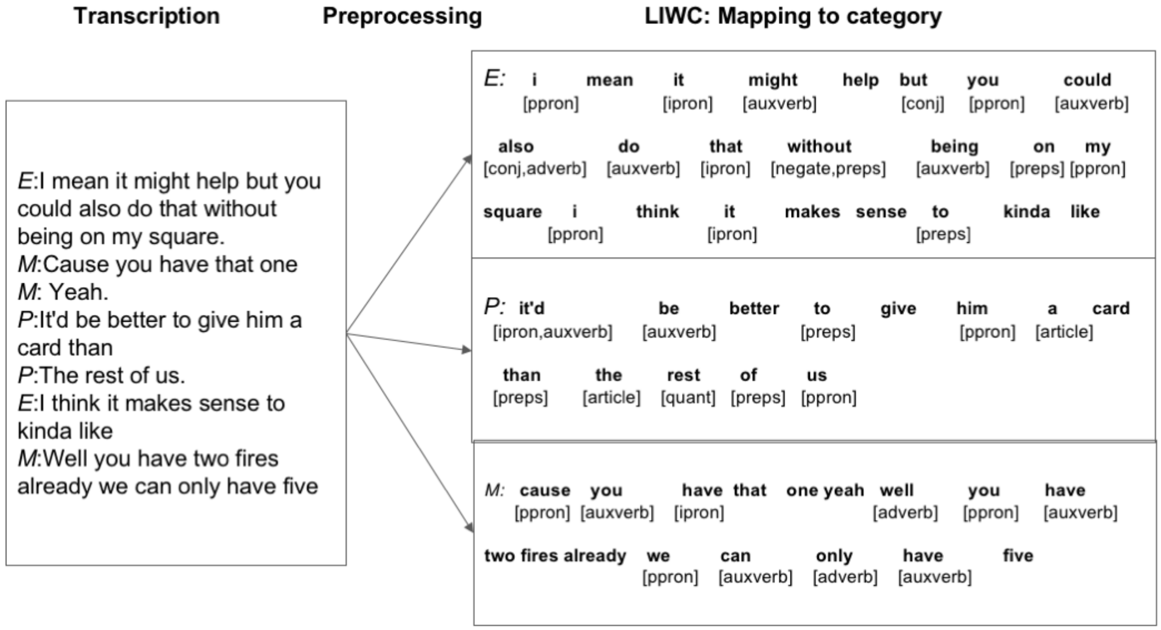

Transcription
Preprocessing
LIWC: Mapping to category
E:I mean it might help but you could also do that without being on my square.
M:Cause you have that one
M: Yeah.
P:It'd be better to give him a card than
P:The rest of us.
E:I think it makes sense to kinda like
M:Well you have two fires already we can only have five
E: i mean it might help but you could
[ppron] [ipron] [auxverb] [conj] [ppron] [auxverb]
also do that without being on my
[conj,adverb] [auxverb] [ipron] [negate,preps] [auxverb] [preps] [ppron]
square i think it makes sense to kinda like
[ppron] [ipron] [preps]
P: it'd be better to give him a card
[ipron,auxverb] [auxverb] [preps] [ppron] [article]
than the rest of us
[preps] [article] [quant] [preps] [ppron]
M: cause you have that one yeah well you have
[ppron] [auxverb] [ipron] [adverb] [ppron] [auxverb]
two fires already we can only have five
[ppron] [auxverb] [adverb] [auxverb]

Figure 7: Using LIWC to create function word features. Each tag corresponds to a LIWC function word category. negate: negation, conj: conjunctions, preps: prepositions, ppron: personal pronouns, ipron: impersonal pronous, article: article, adverb: adverbs, quant: quantifiers, auxverb: auxiliary verbs.

4.4 Measures of Team Linguistic Style Entrainment

There are various methods to directly calculate multiparty entrainment using linguistic style. Some text-based studies have proposed probabilistic frameworks in linguistic style matching based on pairwise comparisons between speakers [72, 18]. However, compared to their data, our data has a lower density of reciprocated interactions per pair. The number of conversations between speakers is insufficient for constructing such probability models. Addressee identification to create appropriate pairs is also not straightforward. [33] developed a method to perform linguistic style matching based on multiparty speech, but they only focused on a global measure rather than on the degree of change. Recently, [58] proposed a method to compute multiparty entrainment on acoustic-prosodic features based on the same Teams Corpus as used here. Their method highlighted feature change over time, which is more relevant to linguistic style entrainment.

For each feature, they calculated the difference between a pair of speakers as the absolute difference of feature values, and the team difference as the average difference over all pairs. In our study, linguistic style is a single feature with multiple categories, so we converted their calculation of pair differences by summing up all the category differences. Moreover, we weighted category differences by the frequency of categories. More specifically, $TDiff_{unw}$ (unweighted team difference) converts the team difference in [58] to deal with multiple feature categories. $TDiff_w$ (weighted team difference) extends $TDiff_{unw}$ by weighting the category differences similarly to that in Gonzales et al. [33]. We calculated both $TDiff_{unw}$ and $TDiff_w$ for each pair of speakers and then averaged over all pairs. The formulas are shown in Equations 1, 2, and 3, where F, K, and $|team\ size|$ respectively refer to the function word category set, an arbitrary function word category, and the team size. $KDiff_{ij}$ refers to the weighted category difference of category K between speakers i and j.

$$TDiff_{unw} = \frac{\sum_{\forall i \neq j \in team}(\sum_{K \in F}(|K_i - K_j|)}{|team\ size| * (|team\ size| - 1)} \tag{1}$$

$$TDiff_{w} = \frac{\sum_{\forall i \neq j \in team}(\sum_{K \in F}(|KDiff_{ij}|)}{|team\ size| * (|team\ size| - 1)} \tag{2}$$

$$KDiff_{ij} = \frac{|K_i - K_j|}{K_i + K_j}, KDiff_{ij} = 0 \text{ if } K_i, K_j = 0 \tag{3}$$

Litman et al. [58] then define convergence, a type of entrainment measuring increase in feature similarity, by comparing the $TDiff$ of two non-overlapping temporal intervals of a game as in Equation 4. C_{ij} and $TDiff$ refer to the team's convergence and the weighted (or unweighted) team differences, respectively. Assuming the game is divided into n disjoint temporal intervals, i and j refer to two predetermined temporal intervals in chronological order.

$$C_{ij} = TDiff_i - TDiff_j, i < j \in n \tag{4}$$

However, this definition leaves two unanswered questions. First, the measure of convergence allows negative values that represent divergence, which is the tendency that team members speak differently. Second, it requires the researcher to hand pick temporal intervals that are not guaranteed to result in an optimal measurement of entrainment. Hence, we derived four new variables of convergence (see Equations 5 and 6): Max and Min calculating the maximum and minimum positive C_{ij}, and absMax and absMin calculating the absolute maximum and minimum $|C_{ij}|$.

$$Max \text{ or } Min = Max\{C_{ij} > 0\} \text{ or } Min\{C_{ij} > 0\} \tag{5}$$

$$absMax \text{ or } absMin = Max\{|C_{ij}|\} \text{ or } Min\{|C_{ij}|\} \tag{6}$$

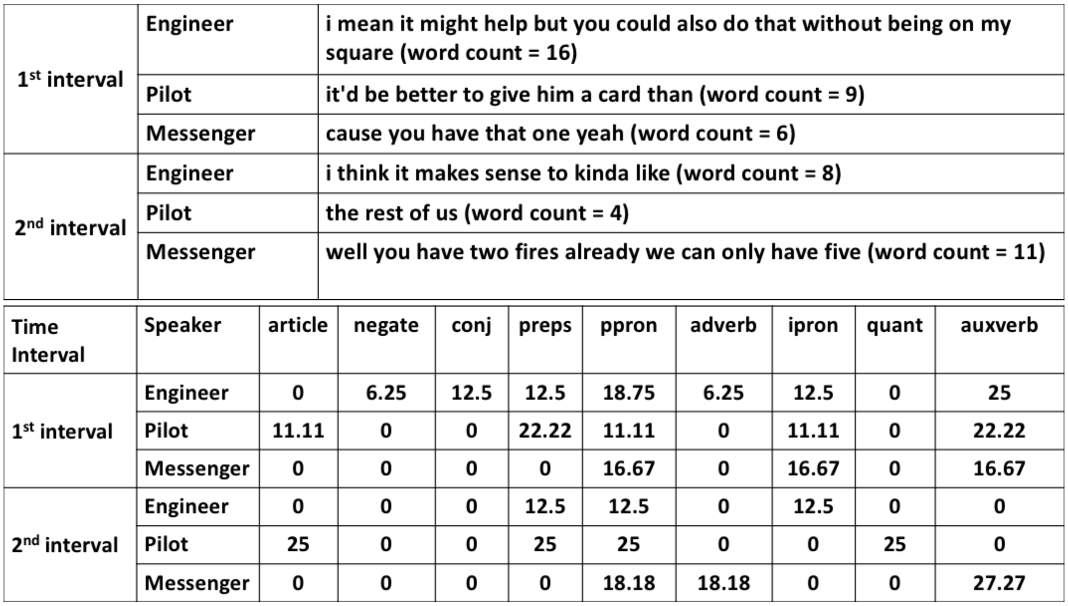
Rather than two fixed intervals, we iterated over all two arbitrary temporal intervals in chronological order and conducted the comparison. Consequently, the Max and Min only measure maximum and minimum convergence so that they directly reflect the decrement of $TDiff$ between two intervals. The absMax and absMin measure the maximum and minimum magnitude of the change of $TDiff$ in the entire conversation. Unlike the Min and Max, the absMax and absMin are determined by the values of convergence or divergence. We added the absMax and absMin beyond Min and Max so that they reflect the overall fluctuation ranges of $TDiff$, which might also be an important aspect of entrainment. Therefore in total, we defined eight measures of team entrainment: **unweighted** and **weighted Max, Min, absMin**, and **absMax convergence**.

The parameter n in Equation 4 determines the length of temporal intervals being compared. Many studies defined n as two so that the conversation is evenly divided into two halves [50, 90]. Since Litman et al. [58] previously found that in the Teams corpus the highest acoustic-prosodic convergence occurred within the first and last three minutes, we used this finding to define our n. We evenly divided each game, which was limited to 30 minutes, into ten intervals, so each interval is less than three minutes. Since our focus is on measure development in this work, methods for optimally tuning this temporal parameter are left for future work.

We will use Figure 7's excerpt to illustrate our calculations. Assuming n is set to two, we first divide the excerpt into two time intervals. Assuming that the temporal midpoint of the excerpt occurs after the fourth IPU, the first interval includes the first through fourth IPUs. The second interval includes the fifth through seventh IPUs. For each speaker, all IPUs in each interval are concatenated and input to LIWC. The interval division and LIWC category percentage output are shown in Figure 8. Based on Equation 1, the unweighted pair difference between the Engineer and Pilot in the first interval is calculated as the sum of the absolute differences of all categories, which is equivalent to $|0 - 11.11| + |6.25 - 0| + |12.5 - 0| + |12.5 - 22.22| + |18.75 - 11.11| + |6.25 - 0| + |12.5 - 11.1| + |0 - 0| + |25 - 22.22| = 57.64$. Similarly, the pair differences between the other two pairs (Engineer and Messenger, Pilot and Messenger) are 52.08 and 50. The unweighted team difference is the average of these pair differences, which is 53.24. The weighted team difference is calculated using Equation 2, with the pair difference now being normalized by the frequency of each category. For instance, the absolute difference between Engineer and Pilot of negate is $|6.25 - 0| = 6.25$. This number is less than the absolute difference of $|18.75 - 11.11| = 7.64$ for the category ppron. However, the occurrence of negate is less common than ppron in the speech of Engineer and Pilot. The weighted difference of negate is $|6.25 - 0|/(6.25 + 0) = 1$, which is now greater than the weighted difference of ppron which is $|18.75 - 11.11|/(18.75 + 11.11) = 0.26$.

4.5 Measures of Team Characteristics

This work focuses on the following team characteristics: **team size**, **gender diversity** (Blau's index and female percentage), **ethnic diversity**, and **age diversity**. Note that the female percentage measures the numerical female dominance in a team, while gender diversity indicates the variability of gender composition. Diversity of age, which has continuous values, is measured by the population standard deviation. Diversity of ethnicity and gender with categorical values is measured by Blau's index of heterogeneity [5] as in Equation 7, where P_k is the proportion of a specific category k.



1st interval	Engineer	i mean it might help but you could also do that without being on my square (word count = 16)
	Pilot	it'd be better to give him a card than (word count = 9)
	Messenger	cause you have that one yeah (word count = 6)
2nd interval	Engineer	i think it makes sense to kinda like (word count = 8)
	Pilot	the rest of us (word count = 4)
	Messenger	well you have two fires already we can only have five (word count = 11)

Time Interval	Speaker	article	negate	conj	preps	ppron	adverb	ipron	quant	auxverb
1st interval	Engineer	0	6.25	12.5	12.5	18.75	6.25	12.5	0	25
	Pilot	11.11	0	0	22.22	11.11	0	11.11	0	22.22
	Messenger	0	0	0	0	16.67	0	16.67	0	16.67
2nd interval	Engineer	0	0	0	12.5	12.5	0	12.5	0	0
	Pilot	25	0	0	25	25	0	0	25	0
	Messenger	0	0	0	0	18.18	18.18	0	0	27.27

Figure 8: The interval division and LIWC category percentage. Top: The input text of each speaker per interval. Bottom: The corresponding LIWC category percentage.

$$Ethnic/Gender\ Diversity_{\ Blau's} = 1 - \sum {P_k}^2 \qquad (7)$$

4.6 Measures of Perceived Team Social Outcomes

We assessed the perception of team social outcomes using the existing self-reported post-game survey responses. The survey contains scales related to team processes and team conflict. Team processes consist of the perceptions of team cohesion, general team satisfaction, potency/efficacy, and perceptions of shared cognition [118, 110, 35, 31]. These four measures were strongly correlated with each other. Thus, we aggregated them into a single scale by averaging their z-scored scale composites, Cronbach's α = 0.78. Team conflict consists of task, process and relationship conflict. These three types of conflict reflect the topic of the conflict, be it about the task at hand, work processes, or interpersonal values and personal relationships. Process conflict is consistently negatively related to performance, but task conflict is positively related to performance under some conditions [21]. Therefore, we kept these three types of conflict as individual variables. Overall, we thus have four measures of perceived team social outcomes: **team processes, task conflict, process conflict** and **relationship conflict**.

4.7 Results

4.7.1 Relating Team Characteristics and Entrainment

We first tested the relationship between linguistic style entrainment and team characteristics with continuous values (gender, ethnic and age diversity) using Spearman rho correlations. There was a significant positive correlation between unweighted convergence Min and gender diversity, $(r(62) = .22, p < .05)$. This correlation indicated that teams with greater gender diversity had higher minimum convergence than teams with less gender diversity.

We then performed one-way ANOVA tests between linguistic style entrainment and the categorical team characteristics, i.e., percentage of females and team size. The unweighted absMax was found to significantly vary with female percentage for the 7 conditions (see corpus section), F(6,55) = 2.79, p = .019. Tukey HSD post hoc tests indicated that the 25% condition (N = 4, M = 40.15, SD = 13.263) was significantly different with the 50% condition (N = 9, M = 19.56, SD = 9.435), 66% condition (N = 18, M = 19.39, SD = 9.407) and 75% condition (N = 10, M = 18.92, SD = 8.117). The mean of the 25% condition was larger than all other three conditions. This finding suggests that the maximum magnitude of the change of unweighted team differences in the 4-person team with one female was greater than other mixed-gender teams with more than one female.

4.7.2 Predicting Perceived Team Social Outcomes

We predicted four measures of perceived team social outcomes: team processes (MIN = -2.57, MAX = 1.51, M = 0.00, SD = 0.80); task conflict (MIN = 1.00, MAX = 3.33, M = 1.75, SD = 0.46); process conflict (MIN = 1.00, MAX = 3.00, M = 1.58, SD = 0.41) and relationship conflict

(MIN = 1.00, MAX = 1.75, M = 1.15, SD = 0.20). A hierarchical linear regression (HLR) model allows us to consider the impact of team characteristics on the perception of team social outcomes, and then examine the significance of multiparty entrainment as a predictor beyond or controlling for team characteristics. Two models, Model 1 (M1) and Model 2 (M2), were constructed. Team size and team diversity (gender, ethnic and age) were entered simultaneously as independent team characteristic variables (IVs). multiparty entrainment was entered into M2 as an IV beyond the team characteristics. Only variables of multiparty entrainment that significantly contributed to the model were selected in M2. The dependent variable (DV) of each HLR was the variable describing the perceived team social outcomes.

Table 5: Predicting team outcomes using hierarchical regression. Age: Age Diversity, Ethnic: Ethnicity Diversity, Gender: Gender Diversity, % Female: Percentage of female, w absMax: weighted convergence absMax, unw absMax: unweighted convergence absMax. β: standardized Beta. * if p <0.05, ** if p <0.01.

DV		**IV**	**Model 1 (β)**	**Model 2 (β)**
Task		Age	0.04	-0.03
		Ethnic	0.19	0.19
		Gender	-0.04	-0.11
		%Female	-0.21	-0.30
		Team size	0.24	0.25*
		unw absMax		-0.31*
	R^2		0.13	0.22
	F		1.60	2.54*
Process		Age	0.06	0.04
		Ethnic	0.25	0.29*
		Gender	-0.08	-0.13
		%Female	-0.08	-0.14
		Team size	0.34**	0.31*
		w absMax		-0.36**
	R^2		0.16	0.28
	F		2.10	3.57**

Significant HLR models are shown in Table 5. The M2 predicting the task and process conflict were both significant, but no M1 was significant. In the HLR predicting task conflict, no team characteristics contributed significantly to M1. Introducing variables of multiparty entrainment to M2 explained an additional 9.2% variation in task conflict and the ΔR^2 was significant, $\Delta F(1, 55) = 6.46, p < 0.05$. Unweighted absMax and team size were both significant contributors to task conflict in M2. The negative association between unweighted absMax and task conflict suggested that the higher maximum magnitude of the change of team difference signaled less team conflict. Meanwhile, the positive association between team size and task conflict in M2 added evidence to previous findings that team size is positively associated with team conflict [3, 99].

In the HLR predicting process conflict, team size contributed significantly to M1. Adding multiparty entrainment (the weighted absMax) to M2 explained an additional 12.2% of the variability in process conflict, and the ΔR^2 was significant, $\Delta F(1, 55) = 9.36, p < 0.01$. multiparty entrainment along with team size and ethnic diversity were important predictors to M2. Team size and ethnic diversity were both positively associated with process conflict. We observed a negative association between the weighted absMax and process conflict. This finding implied that the higher maximum magnitude of the change of team difference signaled less process conflict.

Overall, we found a negative association between maximum magnitude of the change of team difference and team conflict, specifically process and task conflict. Team size and ethnic diversity both had effects on team conflict. Maximum magnitude of the change of team difference was a significant predictor in team conflict.

To determine whether the team characteristics had a significant impact on the conflict variables above and beyond the effect for entrainment, we switched the IVs in M1 and M2. Variables of entrainment were entered into M1 stepwise and then the team characteristics that had shown significance in the previous HLR were entered into M2 (see Table 6). We observed similar findings in that both M1 and M2 significantly predicted task and process conflict. The maximum magnitude of the change of the team difference was significantly negatively associated with task and process conflict. Team size and, for process conflict, ethnic diversity were significantly related to conflict above and beyond entrainment.

4.8 Conclusions

We first proposed a method to measure multiparty linguistic style entrainment by converting and extending methods developed in prior studies of linguistic style matching and team acoustic-prosodic entrainment. We then examined the relationship between multiparty entrainment and team characteristics. Our analysis implies that teams with greater gender diversity had greater

Table 6: Flipped HLR : Entrainment was stepwise entered in M1. Team characteristics showing significance in prior HLRs were entered in M2.

DV		**IV**	**Model 1 (β)**	**Model 2 (β)**
Task		unw absMax	-0.27*	-0.27*
		Team size		0.25*
	R^2		0.07	0.13
	F		4.77*	4.55*
Process		w absMax	-0.34**	-0.34**
		Team size		0.32**
		Ethnic		0.26*
	R^2		0.12	0.26
	F		7.87**	6.69**

minimum convergence than teams with less gender diversity, similarly to the findings in [52, 73] that mixed-gender pairs generally entrain more in dyadic conversations. Moreover, the 4-person teams with more than one female had a higher maximum magnitude of change in team difference. Perhaps the existence of a female subgroup reconciled the team difference in these teams. In conclusion, different gender compositions affect the entraining behaviors of teams. These findings show that gender plays an important role for linguistic entrainment in human interactions. They also reveal a need to study the underlying process of multiparty entrainment with different granularity levels.

Next, we predicted the perception of team social outcomes by team characteristics and variables of entrainment with hierarchical regression models. The experimental results indicated that the maximum magnitude of the change of the team difference was negatively associated with team conflict. Adding this variable of entrainment beyond team characteristics resulted in statistically significant improvements in model prediction. Finally, by entering entrainment variables in the first rather than second model, we showed that entrainment was significantly negatively associated with the task and process conflict, both when controlling for team characteristics and when not. Although the overall models did not account for a large variance, the base model of only team characteristics was improved significantly by adding entrainment. In sum, we found that entrainment is a promising feature to predict team social outcomes. In terms of broader impact, we can now possibly evaluate the success of team conversations using linguistic style entrainment. Additional interdisciplinary research building on our findings could test whether entrainment mediates the effects of team characteristics on social and task outcomes in different settings.

Generally, this chapter examines the existence of lexical entrainment in the Teams Corpus based on a popular bag-of-words similarity measure. In the section 4.4, the bag-of-word similarity measure consists of a series of word counting and mathematical operations on these word counts. More specifically, the similarity measure here refers to the weighted or unweighted team differences calculated in Equation 1 and 2. Regardless of being weighted or unweighted, the team differences are based on the word count difference between function word categories, making this algorithm fall into a bag-of-words paradigm. This approach shows the possibility to quantify lexical entrainment computationally utilizing a set of predefined features. Furthermore, the results imply that lexical entrainment can signal the dialogue success indicated by the group relationship in a multiparty corpus.

In this work, we choose to use a statistical model as the starting point of our entrainment study. The statistical model allows us to examine impact factors of entrainment and its effectiveness in predicting dialogue outcomes. Another reason is that many earlier prior studies about entrainment in Teams Corpus are based on statistical models. Thus we attempt to follow their practice to perform our initial investigation. In the following chapters, thanks to the advance of deep learning with neural network models in recent years, we propose two other models that combine more machine learning techniques.

5.0 A Neural Network-Based Linguistic Similarity Measure for Entrainment in Conversations

5.1 Introduction

In the last chapter, we introduce a bag-of-words approach to quantify entrainment. During our experiment, we notice that our algorithms can suffer from a feature sparsity issue, to address which we must carefully choose the number of time intervals. Thus we are seeking a better solution to quantify entrainment. One possible leverage is in the similarity measure because the core instrument of entrainment measures is a linguistic similarity measure between conversational partners [10, 9, 116, 76, 90, 108, 103]. Most current measures are built upon the bag-of-words model that relies on linguistic markers such as function words or high-frequency words [90, 76, 33, 127, 85]. In Chapter 4 we also show a similarity measure following the bag-of-words paradigm. However, linguistic markers are insufficient to catch context, irony, sarcasm, or other word semantics [83]. Sparsity caused by low-level word usage raises reliability concern for this type of measure [129]. For example, in the bag-of-words similarity measure described in the last chapter, we observe frequent zero word counts for many function word categories. Figure 8 illustrates the example intermediate steps of the similarity measure, which results in many zero counts for multiple function word categories. While more advanced measures in recent entrainment studies are starting to utilize word representation enriched with semantics such as word embeddings [74], the basic comparison granularity is still single words isolated from the conversation flow. Although the usage of conventional linguistic marks such as function words have been validated in many studies, [83, 33, 19], the extraction of these hand-crafted features can be expensive when being deployed in large-scale systems.

Neural network models are data-driven and are highly self-governing. Therefore we propose an alternate approach using neural networks to perform the similarity measures of entrainment calculation. Using neural network-based models allows us to decouple the entrainment similarity measure from the bag-of-words paradigm. Specifically, input sequences can be represented by high-dimensional vectors embedded with semantic meaning. Beyond word-level information, using sequential architectures such as Long-Short Term Memory Network (LSTM), the model can learn structural dependencies of input at all levels. Feature extraction is also fully automated in neural-based models.

Our model is based on dialogue response matching models, which particularly fit the conversation scenarios because they target matching the dialogue context and a response (See Sec 5.2). Besides that, naturally, we can adopt the dialogue response matching task benchmark as an extrinsic evaluation of the similarity measure. Current state-of-the-art dialogue response matching models mainly focus on learning from the inherent meaning of word representation per dialogue. However, the conventional similarity measure for entrainment often leverages corpus-level linguistic features that can be shared across dialogues, such as corpus topics [90], high-frequency words in the corpus [76] and language style reflected by a pre-defined set of function words [33]. Therefore, to simulate this mechanism, we introduce an attention-based architecture to our neural dialogue response matching model to generalize global features shared across all input dialogues. The architecture is based on Global Style Token (GST) that was originally proposed in the speech synthesis task to generate speech with appropriate styles given its context [115]. GST claims the features learned by GST are global because GST is shared across all inputs. Therefore, here we follow the terminology to call the features "global" in our study.

GST is designed to generalize shared representations across all inputs. In a prior study in multilingual models, the shared representation is interpreted as common semantics concepts [114]. The representation generated by GST is agnostic to the actual content and input forms, leading to a better generalization in representation for unstructured data. We collectively call GST in our model *shared stylebook* as their parameters are shared across all dialogues. The "style" in our shared stylebook describes linguistic features beyond lexical with a broader definition. We aim at leveraging neural networks and the stylebook to learn richer language information such as sentence structure and semantics[1]. Thus we can leverage these features in the similarity metrics for entrainment to promote a more comprehensive text-based entrainment beyond the lexical. We will refer to this type of comprehensive text-based entrainment also as "lexical entrainment".

To validate our proposed approach, we examine two specific hypotheses (The numbering of

[1] Note here a previous study based on bag-of-words approaches [117] has demonstrated there is no latent relation between entrainment features.

the hypotheses is based on Chapter 1.2):

1. **Hypothesis 3**: Learning the global linguistic features, the model will achieve better dialogue response matching than the baseline models that do not leverage the global features. We evaluated our model by an extrinsic evaluation related to a dialogue response matching task.

2. **Hypothesis 4**: Using the neural network-based similarity measure for entrainment calculation, we will observe stronger group entrainment signals in a corpus-based entrainment analysis. Those signals are more predictive of dialogue success and failure.

In conclusion, leveraging global linguistic features in neural dialogue response matching models improves model performance. Furthermore, we observe that in a corpus-based entrainment study, our neural network-based similarity measure leads to a stronger entrainment signal.

This work has been submitted to Arvix and will be submitted to an incoming NLP conference.

5.2 Chapter-specific Related Work

5.2.1 Matching Dialogue Response Selection

Matching between the dialogue context and responses is a trendy task in building retrieval-based dialogue systems. The neural network-based models received the most attention in recent years. Early studies focus on single-turn interactions that only considers the dialogue context as a single query by concatenating all previous turns [126, 62, 113]. Later studies are more interested in learning multi-turn interactions so that the multiple turns in the context are all used as separate queries [140, 63, 105]. Recent studies show increasing interests in using pre-trained language models such as BERT [121, 20]. Our work focuses on building a single-turn dialogue response matching model. *Compared to the existing single-turn model, our model provokes learning global linguistic features.*

5.2.2 Style Response Generation/Selection

Another closely related research topic is dialogue style generation or selection. One typical strategy to generate stylized dialogue responses is to employ two separate training stages for response generation and style controlling. Works in style controlling use different approaches such as pre-training stylized language models [78], fine-tuning model with styled corpus [2], using adversarial training [135], and learning a shared latent space between a response and stylized sentences [29]. Generating personalized [56] or emotional responses [137] are also in the same category since they all require require controlling some type of style. Our study specifically focuses on dialogue style matching, which has been viewed as a subtask in some style generation models [66, 78]. *Compared to previous studies, rather than a defined style, our dialogue style matching model focuses on matching the individual dialogue context and response.*

5.2.3 Linguistic Entrainment

There has been substantial evidence for entrainment in many linguistic dimensions, such as acoustic-prosodic entrainment [50, 52, 116, 65], lexical [10, 9, 116], and syntactic entrainment [8, 103, 16, 92]. To evaluate entrainment measures, early studies often set experimental conditions or control groups [10, 8, 30]. In the later corpus-based studies, evaluations are mostly extrinsic such as comparing entrainment between conversational partners and non-partners [50, 90], associating entrainment to other interpersonal behaviors in dialogue such as group relationships [127], positive or negative effects [74], being liked by partners [52], and dialogue success [44]. Here we evaluate our entrainment measure by predicting dialogue success reflected by social outcomes, and by correlating it with prior measures of entrainment from the corpus.

5.3 Data

In this study, we will focus on Teams Corpus so that we can directly compare our new approach with the bag-of-words approach described in Chapter 4.

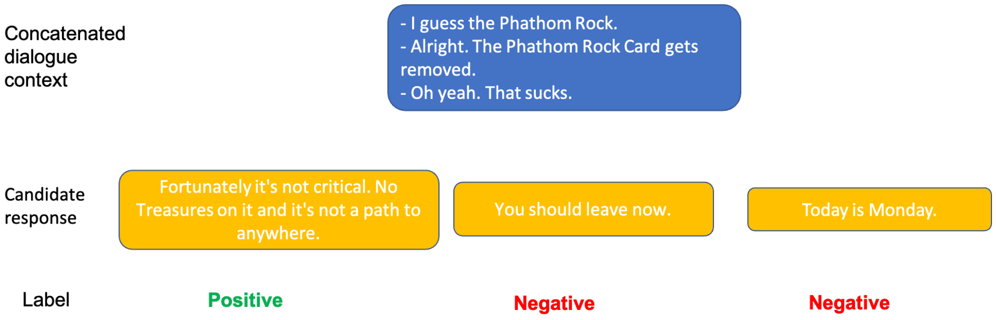

Concatenated dialogue context
- I guess the Phathom Rock.
- Alright. The Phathom Rock Card gets removed.
- Oh yeah. That sucks.
Candidate response
Fortunately it's not critical. No Treasures on it and it's not a path to anywhere.
You should leave now.
Today is Monday.
Label
Positive
Negative
Negative

Figure 9: A problem example for the dialogue response matching task. 1 positive and 2 negative input examples for the dialogue response matching task. Multiple turns are concatenated into 1 single turn and used as the input of our model.

5.4 Model

Our model is a neural dialogue response matching model. It measures the matching between a dialogue context and a response, and it can be used as a similarity measure for entrainment. We train and evaluate the model following the standard framework of dialogue response matching task defined in the following .

5.4.1 Problem Formalization and Data

Given a dialogue context, response matching models determine whether an utterance is proper as a response. Formally, each train and test example is a triplet (**C**, R, y) where **C** is the dialogue context, R is a response, and y is a label indicating whether R is proper for **C**. Given a dialogue **D** $= u_1, u_2, ..., u_n$ where u_i is the utterance for i-th turn, we can extract a dialogue context **C** = $u_1, u_2, ..., u_{n-1}$, a ground truth response **R** = u_n, and we can randomly sample false responses R′ from the same corpus. Therefore, we can formulate our task as a binary classification task to determine $y \in (0, 1)$ for each (**C**, R, y) as $y = 1$ indicating the ground truth. A candidate response is positive when $y = 1$ and negative when $y = 0$. Figure 9 shows 3 input examples.

5.4.2 Model Design

Following the practice in prior works [140, 126, 121, 20, 63, 105], we train our model with a binary classification objective. We adopted a representation-matching-aggregation framework used in previous works [140, 122]. Figure 10 is the model illustration. Note that state-of-the-art dialogue response matching models are mostly multi-turn models. Our model is single-turn because multi-turn models substantially benefit from learning turn interactions by complicated models. To avoid that and focus on our goal in this study, we choose to follow a simpler single-turn model design.

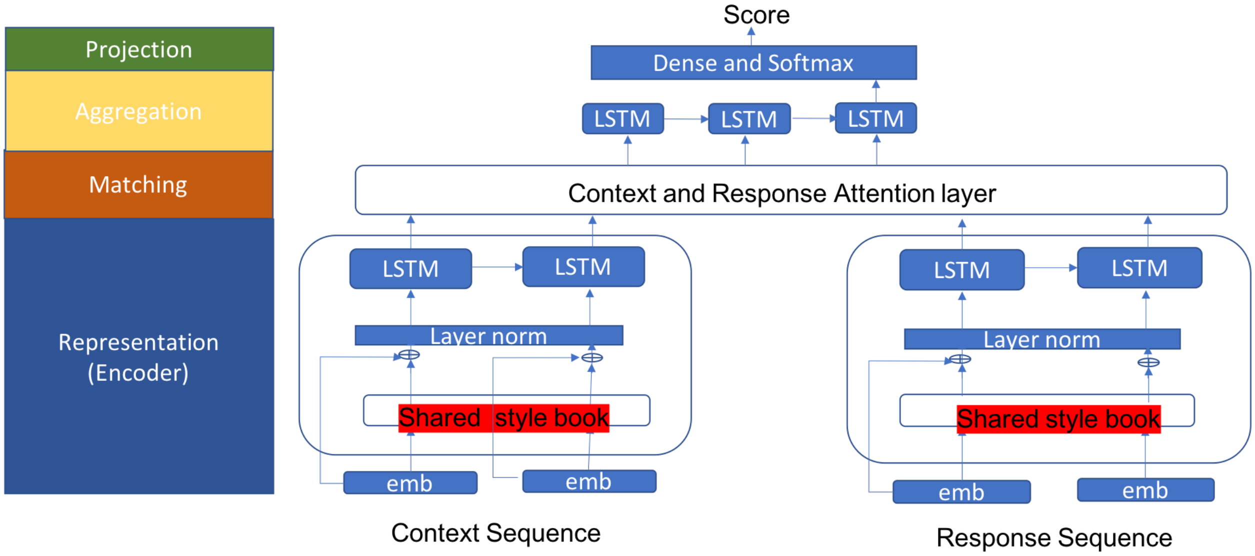


Figure 10: Model illustration. The part highlighted with red is the stylebook.

5.4.2.1 Representation (Encoder) Encoder consisted with several parts:

Embedding Layer

The embedding layer transforms our input of subword tokens to high-dimensional continuous representations. Given a dialogue context **C** and a response candidate **R**, the representations are **C** = $[e_{c,w_1},...e_{c,w_n}]$ and **R** = $[e_{r,w_1},...,e_{r_{wn}}]$, where e_{c,w_i} and e_{r,w_i} represents the embeddings of the i-th token of **C** and **R** respectively. Here $\mathbf{C} \in \mathbb{R}^{n_c \times d}$ and $\mathbf{R} \in \mathbb{R}^{n_r \times d}$ where n_c, n_r and d denotes the number of tokens in the context, the number of tokens in the response, and the embedding size, respectively.

Shared stylebook

The stylebook consists of a set of randomly initialized global key-value pairs. **Unlike the self-attention [109] that the key (K) and value (V) are the linear transformation of input query (Q) itself, our K and V are global for all Q. This is the reason we could call the features learned by the stylebook as global features.** The stylebook is followed by a multi-head scaled dot-product attention [109] that performs as a similarity metric between the key-value set and the input embeddings. Equations 8 and 9 define the attention function where the query (Q) is the input embeddings of the encoders. Specifically, Q is equivalent to **C** in the context encoder or **R** in the response encoder. V denotes value consisted of randomly initialized weights so that $V \in \mathbb{R}^{T \times d_v}$ where T and d denote the size of the stylebook and its dimension. Here we let d be the same as the embedding dimension because we will apply a residual connection later. Key (K) is a linear transformation of V.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

$$\begin{gathered} \text{MultiHead}(Q, K, V) = \text{Concat}(head_1, .., head_n) \\ \text{where } head_i = \text{Attention}(Q_i, K_i, V_i) \end{gathered} \tag{9}$$

For each head *i* in *n* heads, we have Q_i, K_i, V_i that $Q_i \in \mathbb{R}^{n_q \times d_i}$, $K_i \in \mathbb{R}^{T \times d_i}$,$V_i \in \mathbb{R}^{T \times d_i}$, where n_q, T and d_i are the query length, the size of the stylebook and the size of each head. The output of the attention layer is a similarity matrix $M_{\text{style}} \in \mathbb{R}^{n_q \times d}$ where $n_q = n_c$ for context and $n_q = n_r$ for the response. We can view this similarity matrix as style embeddings for they represent the contribution of input embeddings on each type of “style” in the stylebook. We employ a residual connection and layer normalization (Add&Norm) after the attention. Thus the final output is a hybrid embedding vector that combines the inherent and style embeddings, which is denoted as

$C_{hybrid} \in \mathbb{R}^{n_c \times d}$ for the context and $R_{hybrid} \in \mathbb{R}^{n_r \times d}$ for the response.

LSTM Layer

We choose to use an LSTM to learn the dependencies and temporal relationships between input features. LSTMs are a popular type of RNN to model sequential inputs for its prominent ability to control the short-term or long-term information. In our case, the inputs for this layer are hybrid embeddings from the stylebook, and the outputs are hidden states for each time step denoted by $H_c \in \mathbb{R}^{n_c \times d_h}$ for **C** , and $H_r \in \mathbb{R}^{n_r \times d_h}$ for **R**, where d_h is the number of hidden units. We will use H_c and H_r as the final context and response encodings generated from the encoders.

5.4.2.2 Matching This layer performs the matching between context and response encodings. We use the scaled dot-product attention [109] to measure the similarity between context encodings H_c and response encoding H_r. Specifically the query Q is the response encoding H_r, and the key-value pairs are from context encoding H_c. This allows a response to query the most related context information stored in value. Thus, each element in the resulting matrix reflects the similarity between the response and context until the i-th text segment. The layer output is a similarity matrix $M_{r,c}$, which $M_{r,c} \in \mathbb{R}^{n_r \times d}$.

5.4.2.3 Aggregation Similar to prior works in neural matching networks [122, 140, 63], we use an aggregation layer to aggregate matching across segments. Our model aggregates all the segmental matching given by $M_{r,c}$ using an LSTM layer. We use the last hidden state h_{n_r} from the aggregation layer as the sequence-level matching.

5.4.2.4 Projection The output vector h_{n_r} will be fed into a dense layer followed by a softmax layer. The output probability is used as the matching score g between the context **C** and a response candidate **R**. Formally, the g is calculated as in Equation 10.

$$g(C, R) = \text{softmax}(\mathbf{W}h_{n_r} + \mathbf{b}) \tag{10}$$

where **W** and **b** are learned parameters.

5.5 Measuring Entrainment

5.5.1 Train the Matching Model

We firstly train our matching model on the dataset. We create a Teams Corpus dataset for dialogue response matching task (see Section 5.4.1). We sampled examples from each dialogue. To make an example, we extract the previous 5 turns as the dialogue context and the following turn as the ground truth responses. This process results in 107,420 positive instances. We split positive instances in train, validation, and test based on a ratio of 6:2:2. Then for each positive instance, we randomly sampled 9 false responses for validation and test sets, and 1 false response for the train set. This operation results in a dataset of 129K, 215K, 215K examples in train, validation, and test set.

5.5.2 Measuring Entrainment as Convergence

Our approach to measure entrainment in Teams Corpus is based on Yu et al. [127] (Chapter 4). For each conversation in the corpus, we first split it into 10 equivalent time intervals. For an utterance *i* in the interval *j*, we use above model to score the similarity between *i* and the dialogue context **C** consisted of the previous 5 turns. Equation 11 shows the calculation. *g(C, i)* denotes the model generated matching score between context **C** and *i* (see Section 5.4.2.4). Then we average the similarity score over the total *n* utterances spoken by a speaker during interval *j*. Yu et al. [127] use a bag-of-words based similarity score to quantify group difference, and then calculate convergence. Note that the baseline has 2 types of bag-of-words similarity scores depending on different algorithms, but we do not worry about them here because we will replace the bag-of-words score with our neural one. Defined in Equation 12, team difference (*TDiff*) is the averaged similarity difference for pair-wise speakers supposing there are *m* speakers speak in the interval *j*. Shown in Equation 13, entrainment is measured as the convergence, which indicates the increase in partner similarity, between 2 arbitrary intervals *q* and *p* with *q* being earlier than *p*. To obviate the need to select time intervals, 4 types of convergence variables are derived from C_{pq}: *Max*, *Min*, *absMax*, *absMin*. The calculation formulas of *Max* and *absMax* are shown in Equation 14. *Min* and *absMin* are calculated similarly. To summarize, *compared to Yu et al. [127] in Chapter 4, we* *use the group difference (Equation 1 and 2) and convergence formula (Equation 4, 5 and 6), but with a more advanced model-generated similarity score.*

$$Score_{speaker} = \frac{\sum_i^n g(C, i)}{n} \tag{11}$$

$$TDiff_j = \frac{\sum_{\forall a,b \in m}(|Score_a - Score_b|)}{|m| * (|m| - 1)} \tag{12}$$

$$C_{qp} = TDiff_q - TDiff_p, q < p <= 10 \tag{13}$$

$$Max = Max\{C_{ij} > 0\}, absMax = Max\{|C_{ij}|\} \tag{14}$$

5.6 Experiments

5.6.1 Hypothesis 3 (H3)

We hypothesize that leveraging global features will aid input representation, leading to a more robust model in matching dialogue responses. We train 2 models on Teams Corpus to examine this hypothesis: one is our proposed model, and another one is a baseline model with the stylebook removed. Next, we determine whether our model outperforms the baseline model.

5.6.1.1 Evaluation Metrics We follow the standard metrics of dialogue response matching task to evaluate Recall@1 (R@1), Recall@2 (R@2), and Recall@5 (R@5). The k in the Recall@k means that the true positive response is among the first k ranked candidates.

5.6.1.2 Implementation Details Our model is implemented in Pytorch and trained using 3 GPUs. We use pre-trained English byte pair embeddings (bpemb) from BPEmb [36]. Model configuration is tuned on the validation set. The embedding dimension is 300. The maximum token length is 40 for the context and 20 for the response. The size of the stylebook is set to 500. Encoders are shared between the context and the responses. The LSTM layer in encoders has 1024 hidden units. The aggregation LSTM layer has 128 hidden units. All multi-head attention used in this model have 4 heads. Models are trained in mini-batches with a size of 128. The learning rate is 0.0001. We use Adam optimizer. The loss function is cross-entropy. We train a maximum of 10 epochs and optimize training at the R@1 on the validation set.

5.6.1.3 Model Performance Table 7 shows the evaluation results. Our proposed model with the stylebook overall outperforms the baseline. Without the stylebook, the model performance decays a margin. The R@1, R@2, R@5 decrease 3.7%, 5.1%, and 3.5%, respectively. We also compare the number of model parameters. We observe only a minimal growth of parameter size. Beyond Teams data, we further test our stylebook model on another 2 dialogue response matching datasets and similarly observe an improvement in the model performance. Further experiment details are given in the Appendix.

Table 7: Model performance on Teams Corpus for the dialogue response matching task. Size: model size

	R@1	R@2	R@5	Size
Our model	**24.9%**	**41.8%**	**74.7%**	12.4M
- stylebook	21.2%	36.7%	71.2%	12.2M

5.6.1.4 Understanding the Stylebook In this chapter, we propose to leverage a new neural component. In this section, we focus on this proposed architecture and attempt to justify its usage. Although it is generally a challenging task to perform model interpretation on neural structures, in this section, we will conduct three ablations studies, including a visualization, a robustness test, and a SOTA comparison, to understand the proposed stylebook to our full extent.

Visualization

In this section, to understand what the stylebook learns and its impact on the model, we dive deep into the stylebook by adopting different strategies. We conduct a series of studies to visualize the stylebook and style embeddings (see Section 5.4.2.1) generated from the stylebook. Intuitively, style embeddings are generic input embeddings conditioned on style tokens. Since the stylebook and its embeddings are both high dimensional vectors, we utilize a feature reduction algorithm, t-distributed stochastic neighbor embeddings (t-sne)[107] to project vectors to a 2D or 3D space, depending on which space gives us a more intuitive view.

We first attempt to visualize the stylebook itself by extracting the **V** from the key-value pair of the stylebook. **V** is the trained value metrics defined in the scale dot attention, which has been explained in Section 5.4.2.1. According to our model configuration, V contains 500 style tokens, and each token is represented with a 300-dimensions vector. Ideally, each token will represents an individual "style". Thus each token should be separable when being projected to the space. We expect to see a minimum overlap between tokens. Figure 11 shows the projection. We can see from the figure that each token is well separated.

Then, we attempt to see how the style embeddings represent inputs by generalizing the global linguistic features from the inherent input embeddings. Unlike the visualization of stylebook's **V**, which is a static weight metric once trained, we need to project input samples to obtain embeddings. Thus, we hand-label 130 utterances from the Teams Corpus with 13 categories of styles based on our intuition. For example, utterances claiming acknowledgment such as "yeah" and "yes" are categorized as *Acknowledgment*. We collect 10 utterances for each category. Please see the Appendix C for the description of the 13 categories. We extract the averaged style embeddings of the 130 utterances. [2] And we then project them into a 3D space using t-sne with a perplexity of 5 and learning rate of 1 [3]. Figure 12 shows the results. Each data point in the figure represents an utterance, and it is displayed in the figure by its category in a distinct background color. Figure 12 is a clustering overview that shows 2 major clusters. Figure 13 and 14 are the focus views of the first (Cluster 1) and the second cluster (Cluster 2) respectively. Cluster 1 mainly contains short utterances such as questions, acknowledgments, and questions starting with "What". As a counterpart, Cluster 2 mainly contains long utterances such as long questions consisting of more than 2 short questions. In each cluster, utterances belonging to the same category are more likely to be located

[2] The style embeddings here are extracted from a smaller model with 300 LSTM hidden units.

[3] We use an embedding projector provided in Smilkov et al. [98].

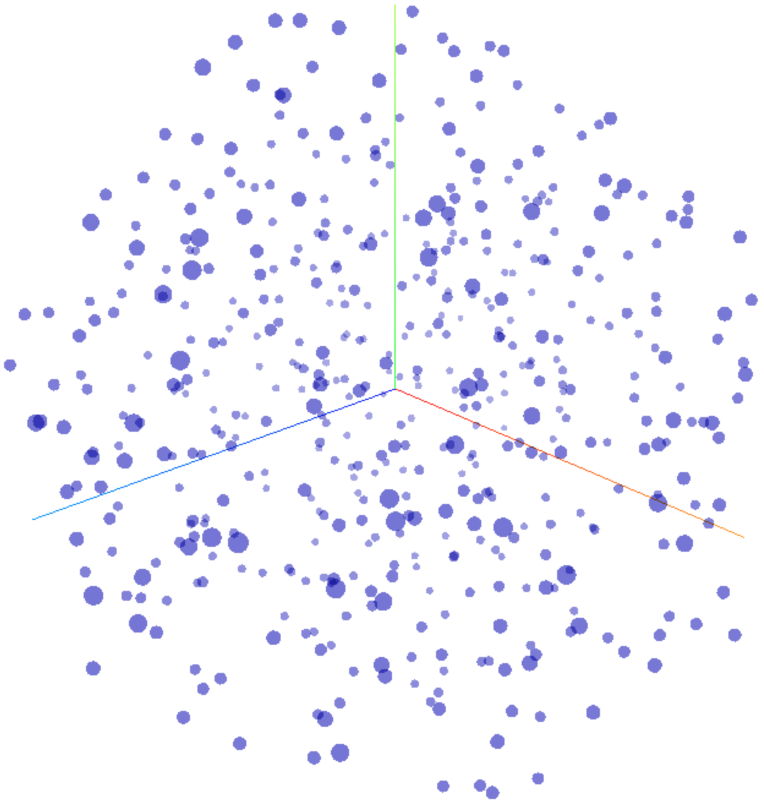


Figure 11: The 3D projection of stylebook V. Each data point in the figure indicates a style token. In total, there are 500 tokens.

closely, indicating that the style embeddings can identify utterances with similar characteristics.

Beyond Teams Corpus, to further support our analysis, we also conduct a similar visualization experiment in an external corpus. We sample 10000 labeled utterances from a publicly available dialogue dataset – the Switchboard Dialog Act Corpus (SwDA)[102]. Each utterance is labeled by its dialogue act. Dialogue acts are sentence-level labels indicating the types of utterances, similar to the categories of styles we used in Teams Corpus. There are over 40 dialogue acts in SwDA. Some examples of dialogue acts are Acknowledgment, Statement-non-opinion, and Yes-No-Question. For more details about the corpus and labels, please see SwDA descriptions [4]. We then conduct the same experiment to visualize the style embeddings of those utterances. Note that we directly apply the stylebook trained on Teams Corpus to Switchboard without any finetuning. Thus, the style embeddings extracted from Switchboard utterances may be less robust than Teams Corpus. We only use Switchboard here for an exploratory purpose. Figure 15 shows the visualization. We observe clear clusters of Acknowledgment and Statement-non-opinion. Utterances labeled with the same dialogue acts such as Declarative Yes-No-Question, Yes-No-Question, and Conventional-closing also tend to be located closely. Overall, the findings are similar to what we observe in Teams Corpus. The style embeddings seem to generate embeddings that can differentiate utterances with various styles.

[4] https://compprag.christopherpotts.net/swda.html

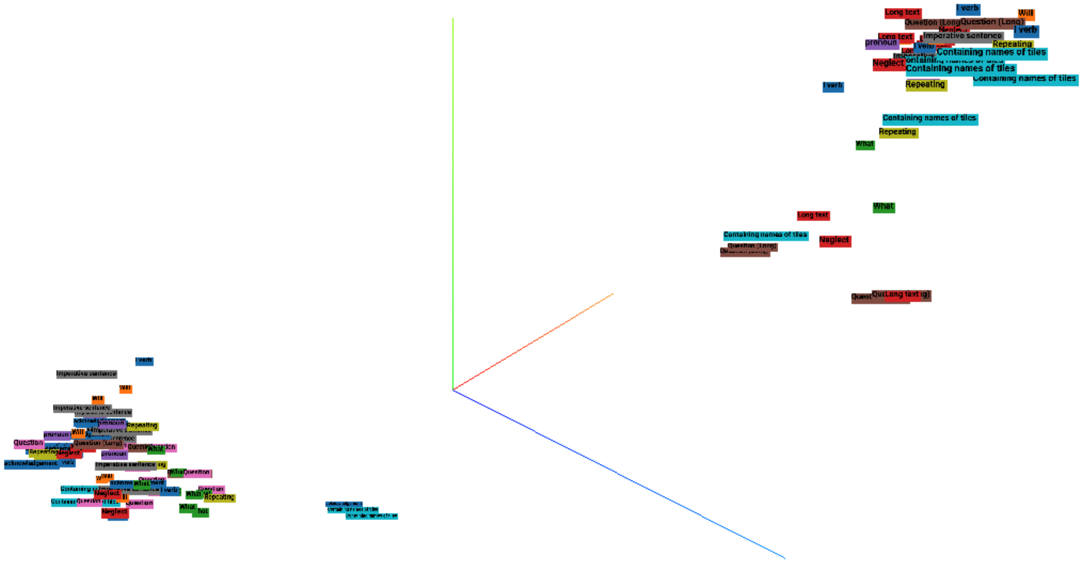


Figure 12: The clustering overview shows 2 major clusters. Each data point in the figure represents an utterance, and it is displayed in the figure by its category in a distinct background color.

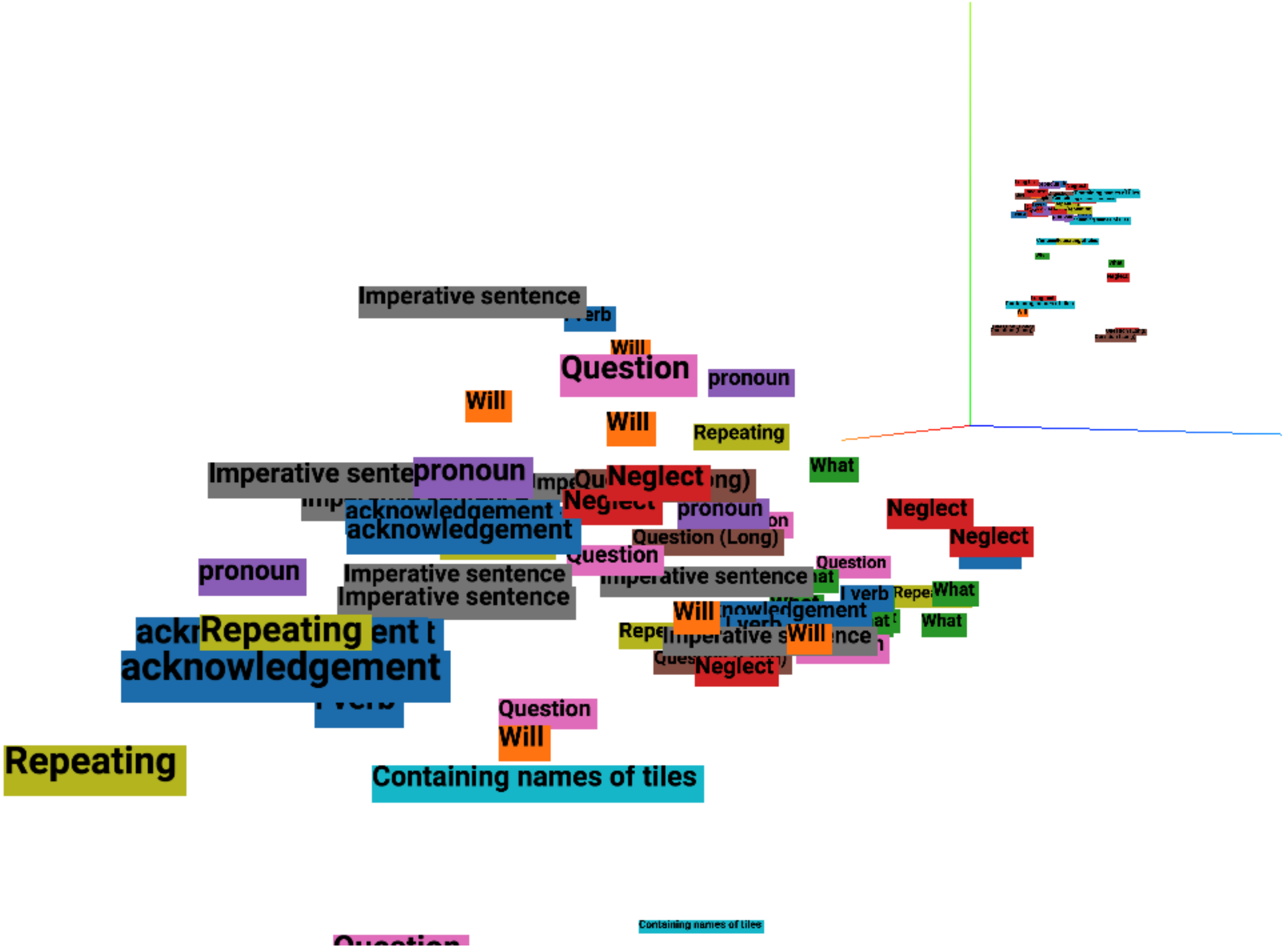


Figure 13: A focus view of Cluster 1, which contains many short sequences.

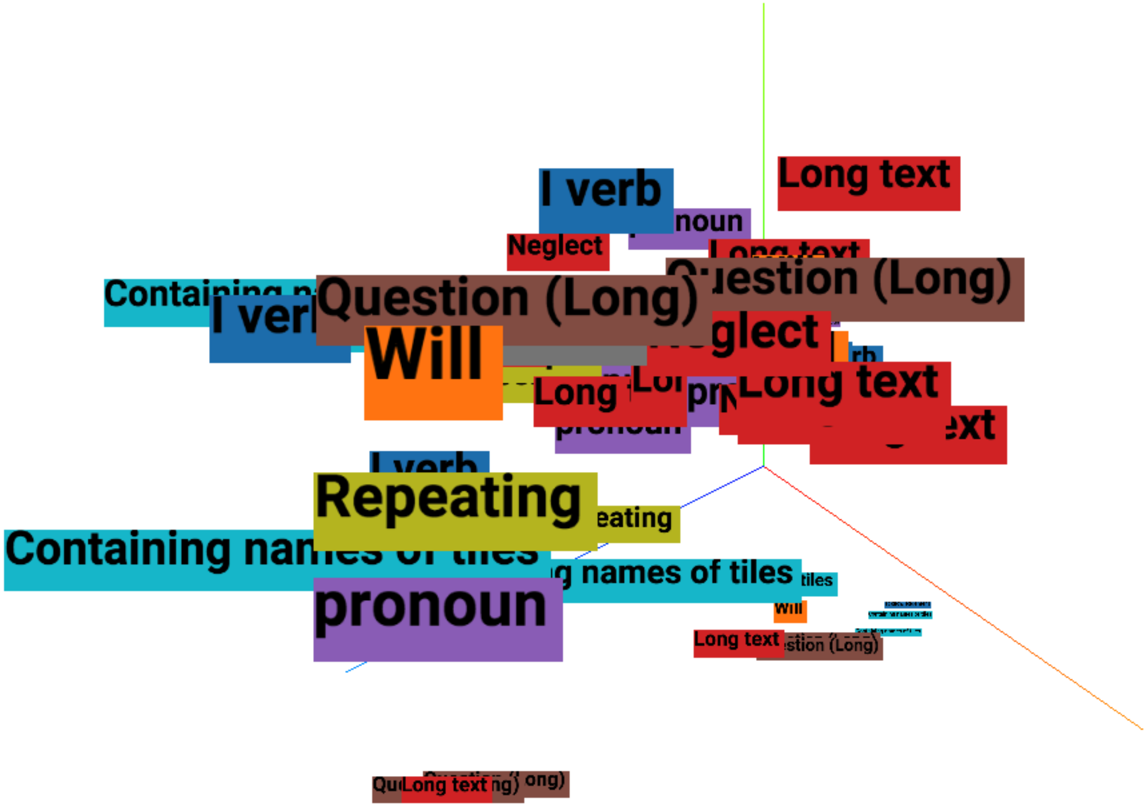


Figure 14: A focus view of Cluster 2, which contains many long sequences.

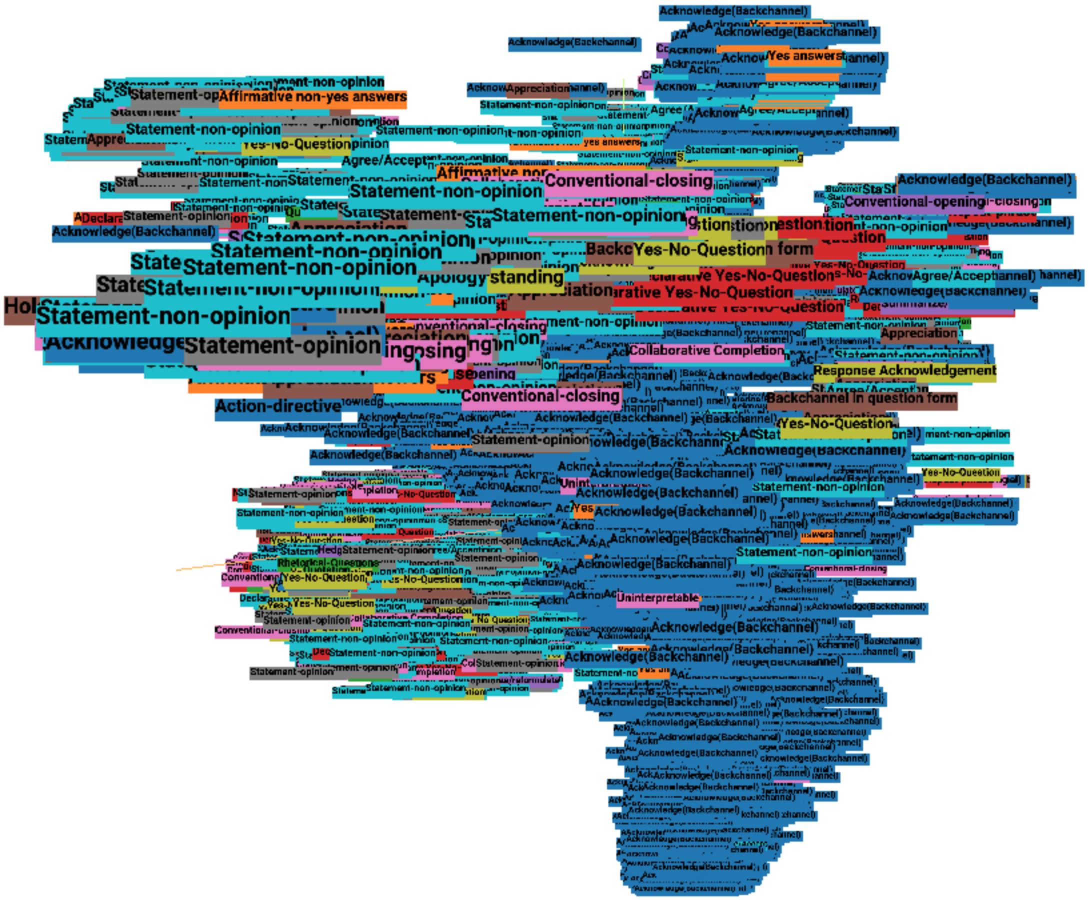


Figure 15: Clustering overview of SwAD. Each data point in the figure represents an utterance, and it is displayed in the figure by its dialogue act in a distinct background color. Please note that the colored labels in this figure for SwDA are different from those in Figure 12 for the Teams Corpus even though the color schemes look the same for these 2 figures.

Sentence-level visualization highlights the relationship of sequences. We then further visualize the style embeddings for each token, which emphasizes the semantics of tokens. Furthermore, since the stylebook embeddings are derived from the inherent bpemb embeddings, we can check token semantics by comparing the visualization of bpemb embeddings and the stylebook embeddings, thus understanding what the stylebook has generalized.

In total, there are 1408 sub-word pieces and 278 unique tokens. Figure 16 shows an overview of its t-sne 2D projection. Overall we also observe some clusters. It is not easy to label each cluster based on the tokens since they are just fragments of words, but we notice a small gathering of very short prefix or suffix tokens. We then compare the original bpemb embeddings of tokens to our style embeddings. Figure 18 shows the overview of 2D projection. Compared to the style embeddings in Figure 16, bpemb embeddings in Figure 17 contains more clusters with clearer edges. We also obverse smaller clusters of very short prefix or suffix tokens in bpemb (See the zoom-in view of this smaller cluster in Figure 17). But compared to stylebook embeddings, there are 2 well-separated clusters in bpemb, both containing short prefix or suffix, indicating bpemb gives us a finer-grained clustering.

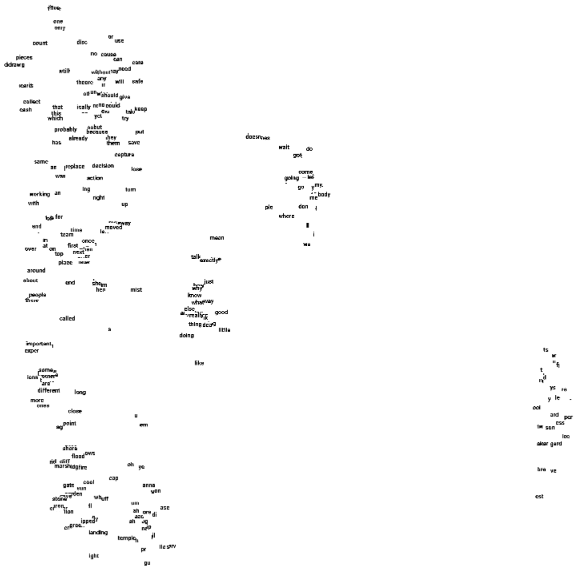


Figure 16: A 2D projection of style embeddings per input token. Each data point is a token labeled by its content.

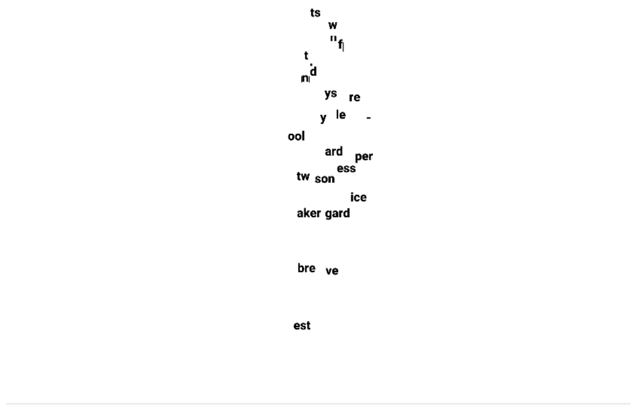


Figure 17: A zoom-in view of the smaller cluster of very short prefix or suffix tokens.

probably
theoretically
action
could
already
around
capture
cash
lose
turn
left
once
working
same
people
important
like
standing
light
march
green
then
walk
do
going
good
talk
i
where
we
capacity
mean
called
sh
ch
pr
result

Figure 18: A 2D projection of bpemb embeddings for each input token piece. Each data point is a token piece labeled by its content.

Then we examine a token's 10 nearest neighbors in bpemb embeddings and our style embeddings because any change of neighbors can imply a change of semantic meaning in different embedding spaces. Our observations fall into 2 categories: "Mostly unchanged" and "Mostly changed". We categorize a token as "Mostly Unchanged" when the number of changed neighbors is no more than 5. Here we show 8 representative tokens and list their 10 nearest neighbors in Figure 20. Note that here all tokens are sub-words. Thus in this study, we can't directly examine these tokens by their part of speech tags. In the examples, we cherry-pick tokens that are not split into sub-words to facilitate understanding. Generally, using style embeddings leads to a rerank of nearest neighbors or changes in neighbors. Our observation found that "Mostly Changed" tokens are related to dialogue content rather than function words. For example, the words "decision" and "mist" have actual meaning compared to the words "he" and "could" that act as function words to form a grammatic structure. This implies that the stylebook adds more contextualized information to the generic embeddings. Meanwhile, we also notice a change of neighbors for those very short tokens such as prefix and suffix, implying stylebook is also sensitive to the long or short form of inputs. This further explains why the stylebook distinguishes between long and short sentences in our analysis of sentence clustering. It also explains why the clustering of short tokens using style embeddings slightly derivatives from using the bpemb embeddings.

Additionally, to dive deep into the change of nearest neighbors per token after using style embeddings, we implement a K nearest neighbors (KNN) algorithm [5] and apply that to both bpemb and style embeddings. Before we apply KNN, we use a principal component analysis (PCA) to reduce the embedding dimensions to 3 so that embeddings have the same dimensions as the dimensions of the projection space in the previous visualization experiment. In total, 191 tokens belong to the "Mostly Changed" category, and 87 tokens belong to the "Mostly Unchanged" category. This implies that the stylebook curates the semantics of most of the tokens embeddings. Figure 19 depicts the percentage of tokens by the nearest neighbors changed. Most of the tokens have more than 5 changed neighbors.

Robustness Testing

So far, we have shown that using the stylebook in our model improves the matching accuracy

[5] We use a KNN implementation in Sklearn (https://scikit-learn.org/stable/)

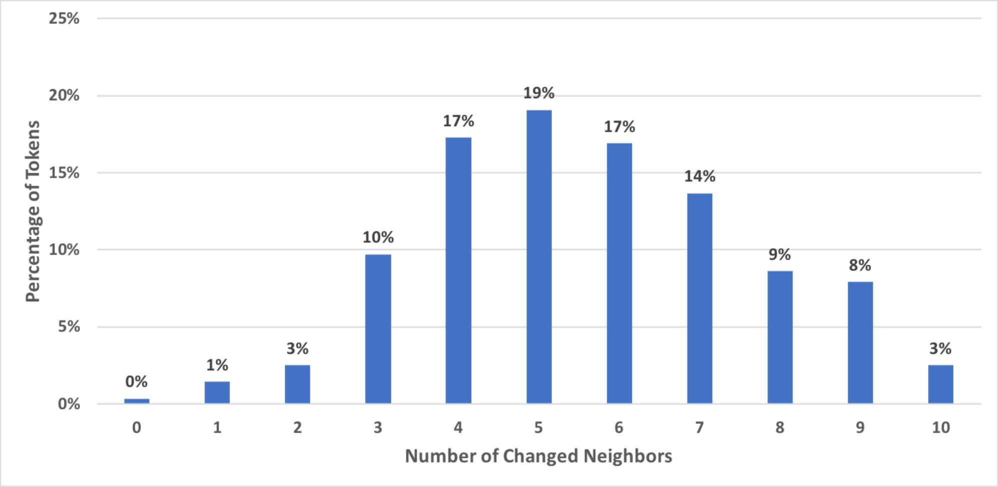
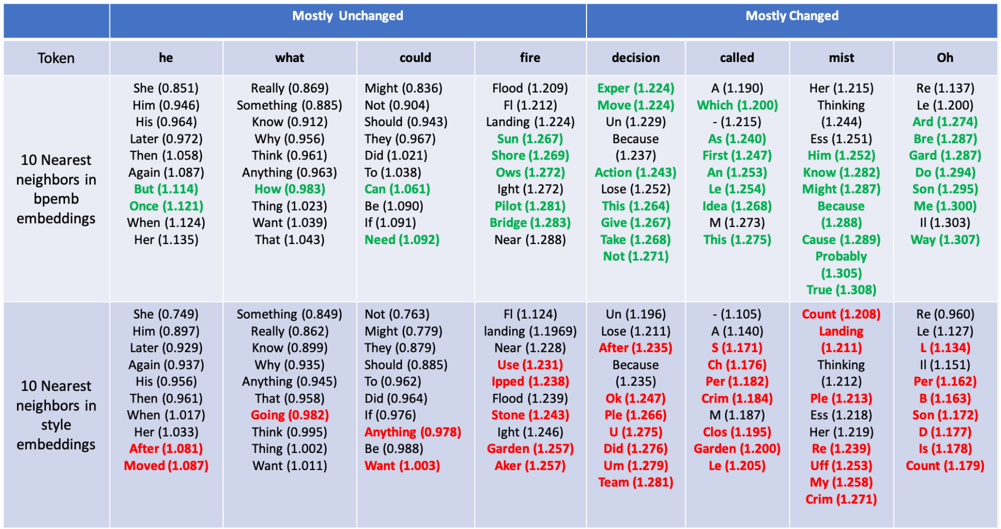


Figure 19: The percentage of tokens by their nearest neighbors changed. The percentage is not cumulative.

between context and responses on Teams Corpus. Visualization also helps us understand what the stylebook learns and how it aids input representation with the global "style". But before we jump to a conclusion on H3, we want to ensure the conclusion can be applied to more general cases on other datasets beyond Teams Corpus. Therefore, we conduct a robust test to test the performance on other datasets.

We applied our model to 2 more dialogue response matching datasets, Ubuntu and Douban. Following the same strategy of creating dialogue response matching datasets for our model (see Section 5.5.1), we construct datasets for Ubuntu [62] and Douban [139] (See Chapter 3). Ubuntu focuses on a large number of dyadic conversations, whereas Douban focuses on dyadic conversations in a foreign language (Chinese). We choose to use a preprocessed version of data provided in [140]. The dataset provides a train, a validation, and a test set. Table 8 shows the statistics of the resulted datasets for Ubuntu and Douban. Note that the validation set of Douban contains dialogues with only 2 candidates. The test set contains dialogues that have multiple or no ground-truth answers. We follow the practice in prior works to remove all positive and negative dialogues from the test set [140, 122, 105].

Model configuration is similar to Teams Corpus with some minor adjustments. The LSTM



	Mostly Unchanged				Mostly Changed			
Token	**he**	**what**	**could**	**fire**	**decision**	**called**	**mist**	**Oh**
10 Nearest neighbors in bpemb embeddings	She (0.851) Him (0.946) His (0.964) Later (0.972) Then (1.058) Again (1.087) **But (1.114)** **Once (1.121)** When (1.124) Her (1.135)	Really (0.869) Something (0.885) Know (0.912) Why (0.956) Think (0.961) Anything (0.963) **How (0.983)** Thing (1.023) Want (1.039) That (1.043)	Might (0.836) Not (0.904) Should (0.943) They (0.967) Did (1.021) To (1.038) **Can (1.061)** Be (1.090) If (1.091) **Need (1.092)**	Flood (1.209) Fl (1.212) Landing (1.224) **Sun (1.267)** **Shore (1.269)** **Ows (1.272)** Ight (1.272) **Pilot (1.281)** **Bridge (1.283)** Near (1.288)	**Exper (1.224)** **Move (1.224)** Un (1.229) Because (1.237) **Action (1.243)** Lose (1.252) **This (1.264)** **Give (1.267)** **Take (1.268)** **Not (1.271)**	A (1.190) **Which (1.200)** - (1.215) **As (1.240)** **First (1.247)** **An (1.253)** **Le (1.254)** **Idea (1.268)** M (1.273) **This (1.275)**	Her (1.215) Thinking (1.244) Ess (1.251) **Him (1.252)** **Know (1.282)** **Might (1.287)** **Because (1.288)** **Cause (1.289)** **Probably (1.305)** **True (1.308)**	Re (1.137) Le (1.200) **Ard (1.274)** **Bre (1.287)** **Gard (1.287)** **Do (1.294)** **Son (1.295)** **Me (1.300)** Il (1.303) **Way (1.307)**
10 Nearest neighbors in style embeddings	She (0.749) Him (0.897) Later (0.929) Again (0.937) His (0.956) Then (0.961) When (1.017) Her (1.033) **After (1.081)** **Moved (1.087)**	Something (0.849) Really (0.862) Know (0.899) Why (0.935) Anything (0.945) That (0.958) **Going (0.982)** Think (0.995) Thing (1.002) Want (1.011)	Not (0.763) Might (0.779) They (0.879) Should (0.885) To (0.962) Did (0.964) If (0.976) **Anything (0.978)** Be (0.988) **Want (1.003)**	Fl (1.124) landing (1.1969) Near (1.228) **Use (1.231)** **Ipped (1.238)** Flood (1.239) **Stone (1.243)** Ight (1.246) **Garden (1.257)** **Aker (1.257)**	Un (1.196) Lose (1.211) **After (1.235)** Because (1.235) **Ok (1.247)** **Ple (1.266)** **U (1.275)** **Did (1.276)** **Um (1.279)** **Team (1.281)**	- (1.105) A (1.140) **S (1.171)** **Ch (1.176)** **Per (1.182)** **Crim (1.184)** M (1.187) **Clos (1.195)** **Garden (1.200)** **Le (1.205)**	**Count (1.208)** **Landing (1.211)** Thinking (1.212) **Ple (1.213)** Ess (1.218) Her (1.219) **Re (1.239)** **Uff (1.253)** **My (1.258)** **Crim (1.271)**	Re (0.960) Le (1.127) **L (1.134)** Il (1.151) **Per (1.162)** **B (1.163)** **Son (1.172)** **D (1.177)** **Is (1.178)** **Count (1.179)**

Figure 20: 10 Nearest neighbors of the 8 tokens we selected. The parenthesized numbers show the euclidean distance between neighbors and the token piece. We categorize a token as “Mostly Unchanged” when the number of changed neighbors is no more than 5. Otherwise, we categorize a token as “Mostly Changed”. Green neighbors in bpemb embeddings are replaced by red neighbors in style embeddings.

Table 8: Data statistics of Ubuntu and Douban corpus.

Corpus	Split	Number of instances	positive : negative instances
Ubuntu	Train	1.00M	1:1
	Valid	0.5M	1:9
	Test	0.5M	1:9
Douban	Train	1.00M	1:1
	Valid	50K	1:1
	Test	10K	-

layer in encoders has 1 layer with 300 hidden units. Model is trained in mini-batches with a size of 256. The learning rate is 0.001. The maximum context length is 160 and 80 for the response. For Douban Corpus, we use Chinese bpemb with a vocabulary size of 100K. We train a maximum of 5 epochs. The samples provided in the Douban validation set only have 2 candidate responses. This is different from the Teams Corpus and Ubuntu, whose test set contains dialogue samples with 10 candidate responses. Thus we don't directly optimize the validation set R@1 for Douban, and we optimize the validation loss instead.

We again conduct the same experiment to exam the effectiveness of the stylebook. Table 9 shows the evaluation results. Without the stylebook, the model performance degrades. The R@1 decreases 1.9% and 2.6% percents for Ubuntu and Douban, respectively. The R@2 also decreases 1.6% for Ubuntu and 1.7% percent for Douban. This indicates that leveraging the stylebook can effectively improve model performance. Meanwhile, we also notice that compared to the R@1 and R@2, the performance degradation in R@5 caused by removing the stylebook is slightly smaller. For Douban, the R@5 is even higher without the stylebook. One possible explanation is that the baseline already achieves a high R5, so the performance improvement can only be marginal after using the stylebook.

After all, we conclude that the usage of the stylebook is robust regardless of the data type and size. We test our model on a small dataset (Teams Corpus), a large dataset (Ubuntu), and even a

Table 9: The effectiveness of our stylebook. Evaluation results of the ablation study on the effectiveness of our stylebook on both the Ubuntu and Douban Corpus. # Params: the number of parameters.

	Ubuntu				Douban			
	R1	R2	R5	# Params	R1	R2	R5	# Params
Our model	**0.688**	**0.810**	**0.943**	**4.5M**	**0.227**	**0.383**	0.719	**31.5M**
- stylebook	0.669	0.794	0.937	4.2M	0.201	0.366	**0.722**	31.2M

foreign language dataset (Douban). The results suggest that using the stylebook can improve the matching accuracy between context and response.

The SOTA Comparison

After the above studies, we have found some convincing evidence about the effectiveness of learning global features using the stylebook. Another remaining question is whether learning global linguistic features is worthy. It is unclear how the stylebook can improve matching compared to other state-of-art baselines leveraging other strategies. Therefore, we establish a comparison between our model and a set of state-of-art baselines on Ubuntu and Douban frequently used to benchmark dialogue response matching tasks. We select the baselines based on the method leverages, including basic single-turn models, advanced single-turn matching models, and multi-turn matching models. Our model belongs to advanced single-turn models.

- **Basic single-turn models** only considers the context as a single query [126, 62, 113], including TF-IDF, RNN, CNN, LSTM and BiLSTM in early works [62].
- **Advanced single-turn matching models** More advanced single-turn models with LSTM variants including MV-LSTM [111], Match-LSTM [112], Attentive-LSTM [104].
- **Multi-turn matching models** Models that takes multiple turns and consider the dependencies across multi-turn context, including Multi-view [139], DL2R[126], SMN[122], DAM[140], IOI[105] , MSN[128]

Table 10 shows the evaluation results of our models as well as other baselines. We copied the results of baseline models from existing literature. For Douban, our model significantly outperforms all single-turn baselines in terms of R@1 and R@2. Although multi-turn models generally work better than single-turn models, our model still outperforms the multi-turn models Multi-view in R@1 and R@2.

Therefore, we conclude that the stylebook, which learns global features, is competitive to other single-turn models variants, although leveraging the stylebook does not beat multi-turn models. However, our model framework is based on single-turn. To obtain a fairer comparison, we need to embed the stylebook in the multi-turn models and then examine its performance, which can be future work.

5.6.2 Hypothesis 4 (H4)

We hypothesize that our neural network-based measures will capture a stronger entrainment signal compared to the bag-of-words measures. A recent study [89] on Teams Corpus suggests that more robust entrainment measures carrying stronger signals will lead to a more robust prediction of dialogue outcomes. Thus, we examine this hypothesis with an **extrinsic** evaluation to predict dialogue success on Teams Corpus.

5.6.2.1 Baseline Models The baseline is a bag-of-word approach from Yu et al. [127] on the Teams Corpus, which is proposed in Chapter 4. Following Chapter 4, we only use on Game 1.

5.6.2.2 Validate our Measure Before the prediction, we first validate our similarity measures to ensure they convey some linguistic signals associated with entrainment. Thus we calculate the Pearson correlations between our baseline convergence variables and their baseline counterparts. Note that the baseline approach provides 2 types of convergence variables of being weighted and unweighted based on different bag-of-words algorithms. Furthermore, to investigate the impact of the stylebook in our model, we remove the stylebook and examine the correlations again. The results are shown in Table 11. We first find that our *Max* and *absMax* are strongly correlated to the baseline convergence variables. This finding suggests that the neural model can be used as

Table 10: Evaluation results of the dialogue response matching task. We group the models horizontally based on single-turn, advanced single-turn and multi-turn models. Underlined numbers show the best single-turn performance in literature. Our results are shown in bold underlined text. Italic numbers show the state-of-the-art multi-turn model performance.

	Ubuntu			Douban		
	R@1	R@2	R@5	R@1	R@2	R@5
TF-IDF[62]	0.410	0.545	0.708	0.096	0.172	0.405
RNN[62]	0.403	0.547	0.819	0.118	0.223	0.589
CNN[62]	0.549	0.684	0.896	0.121	0.252	0.647
LSTM[62]	0.638	0.784	0.949	0.187	0.343	0.720
BiLSTM[43]	0.630	0.780	0.944	0.184	0.330	0.716
MV-LSTM[111]	0.653	0.804	0.946	0.202	0.351	0.710
Match-LSTM[113]	0.653	0.799	0.944	0.202	0.348	0.720
Attentive-LSTM[104]	0.633	0.789	0.943	0.192	0.328	0.718
Multi-view [139]	0.662	0.801	0.951	0.202	0.350	0.729
SMN[122]	0.726	0.847	0.961	0.233	0.396	0.724
DAM[140]	0.767	0.874	0.969	0.254	0.410	0.757
IOI [105]	0.796	0.894	0.974	0.269	0.451	0.786
MSN [128]	*0.800*	*0.899*	*0.978*	*0.295*	*0.452*	*0.788*
Ours model	**0.688**	**0.810**	**0.943**	**0.227**	**0.383**	**0.719**

Table 11: The Pearson correlations between our baseline convergence variables and their baseline counterparts. Rows only show our variables that have at least one significant correlation. Not sig: not significant. * if p<0.05, ** if p<0.01

		Baseline					
		Weighted			Unweighted		
		absMax	Max	Min	absMax	Max	Min
Ours	absMax	0.426**	-	-	0.316*	-	-
	Max	-	0.419**	-	-	0.351**	-
	Min	-	-	Not sig	-	-	.290*
-stylebook	absMax	0.257*	-	-	Not sig	-	-

a similarity measure for entrainment due to its correlation. On the other hand, the correlation becomes much weaker if we eliminate the stylebook from our model. Intuitively, this finding implies that the stylebook may contribute to capture the linguistic signal related to entrainment.

5.6.2.3 Evaluation Method We follow the baseline approach to evaluate entrainment measures by predicting dialogue success using a regression model. Entrainment measures are used as the independent variables (IVs) to predict dialogue success measures as the dependent variables (DVs). The DVs are entered into the model stepwise. We construct both IVs and DVs strictly following the baseline. DVs are 4 social outcome scales extracted from Teams Corpus surveys: **Team Processes**, **Task Conflict**, **Process Conflict** and **Relationship Conflict**. **Team Processes** is an aggregated scale of team cohesion, general team satisfaction, potency/efficacy, and perceptions of shared cognition [118, 110, 35, 31]. Conflict scales reflect the conflicts in completing tasks, work processes, and interpersonal relationships. IVs are convergence variables in the Equation 14.

5.6.2.4 Predicting Dialogue Success Table 12 shows standardized coefficients (β), R^2 and F value of a regression model with statistical significance. Here we construct 3 models: **Baseline** is the baseline model that predicts DVs by bag-of-words entrainment measures. The result of **Base-**

line is copied directly from the previous work. **Ours** predicts DVs by our neural network-based entrainment measures. Additionally, **No Stylebook** predicts DVs by our neural network-based entrainment measures with no stylebook removed from the model structure. Results show that overall both our neural model **Our** and **No Stylebook** are stronger in predicting all DVs reflecting all **Conflicts** variables. For explaining variation in **Task Conflict**, **Baseline** only acheives significant R^2 of 7%, but using entrainment measures from **Ours** and **No Stylebook**, the resulted R^2 is highly significant. **No Stylebook** achieves the highest R^2 improvement of 7% compared to the **Baseline**. We have the same finding for **Process Conflicts**. Although the improvement in R^2 between **Baseline** and our models are smaller, R^2 of **Ours** and **No Stylebook** are highly significant. More notably, using our neural entrainment measures, we can predict **Relationship Conflict**, which was not predictable by the **baseline**. Both **Our** and **No Stylebook** achieve significant 8.0% and 11% R^2 for **Relationship Conflict**. **No Stylebook** is a highly significant regression model. Therefore, we conclude that our neural-based entrainment measures are stronger in predicting all DVs reflecting Conflicts compared to the baseline model. Beyond the performance improvement, we found that **No Stylebook** performed better than **Our** having the stylebook in its model structure. This implying that the improvement in regression was not caused by using the stylebook. We also noticed that the most predictive IV across all 3 models is *absMax*, which represents the maximum magnitude of convergence. Also, negative entrainment coefficients reveal that a higher convergence signals less conflict in the conversation. This finding is aligned with existing findings.

Here we also include a prediction with team characteristics as team characteristics are also reported in the baseline in the last Chapter as important factors to dialogue success and failures. We use the identical team characteristics used in the baseline. We used team size, female percentage, age diversity, gender diversity, and ethnic diversity. Table 13 shows results. The findings are very similar to the findings when not including team characteristics in prediction. Compared to the baseline, our model improves the predicting performance with team characteristics. Additionally, we find team size is likely to be a significant coefficient for all predictions.

5.7 Case Study

We perform a case study on Teams Corpus by comparing the similarity score g (see the Equation 10) generated by our model to investigate how it reflects entrainment and whether the shared stylebook has any impact on that. Based on our interpretation, we cherry-pick several Teams Corpus dialogue examples that exhibit different types of entrainment according to our understanding. Table 25 shows the dialogue context, the ground truth response, and the scores generated from our neural model when using and not using the shared stylebook.

The model with the stylebook assigns higher matching scores to cases 1 to 4. In case 1, similarly to the context, the response contains an exclamation immediately following a phrase starting with “it’s”. This example can be interpreted as a case of structure entrainment because the same sentence structure is repeated in the response. In case 2, following “we’re dead” in context, a speaker immediately said “we lost” with a similar meaning in game playing. Such entrainment is beyond the lexical. We consider this as entrainment in semantics. Case 3 is another case of structure entrainment. In its context, speakers used very similar phrase structures such as “you haven’t done”, ”I haven’t done”, and “you are done”. In the response, a speaker followed such a structure and said “I’m done”, which was not the same phrase used in context but had a similar structure. This is also a case beyond simple lexical matching. Case 4 is interesting. There are 4 treasures to be collected in the Forbidden Island game, and each treasure has its own shape. Speakers develop their terms to refer to those treasures during the conversation. We have seen in transcriptions some examples of terms: chalice, cup, stone, flame, and lion. We view this as a type of concept entrainment because speakers mutually agree on a concept during a conversation. In this case 4, the word “treasure” is used in the context, and following that, a speaker refers that to “lion”. Using the stylebook seems to help the model catch such connection better than not using the stylebook.

The model without the stylebook assigns higher matching scores to cases 5 to 7. In case 5, the phrases “this one” and “that one” are frequently used in the context. The response also contains “that one”, and more notably, the speaker chooses to say “the lion one” when there is a simpler alternation “the lion”. We interpret this example as a phrase entrainment. Case 6 and 7 are both cases of phrase entrainment because the same terms are repeated in the response following the context. The model would assign higher scores to those cases without the curation of the stylebook.

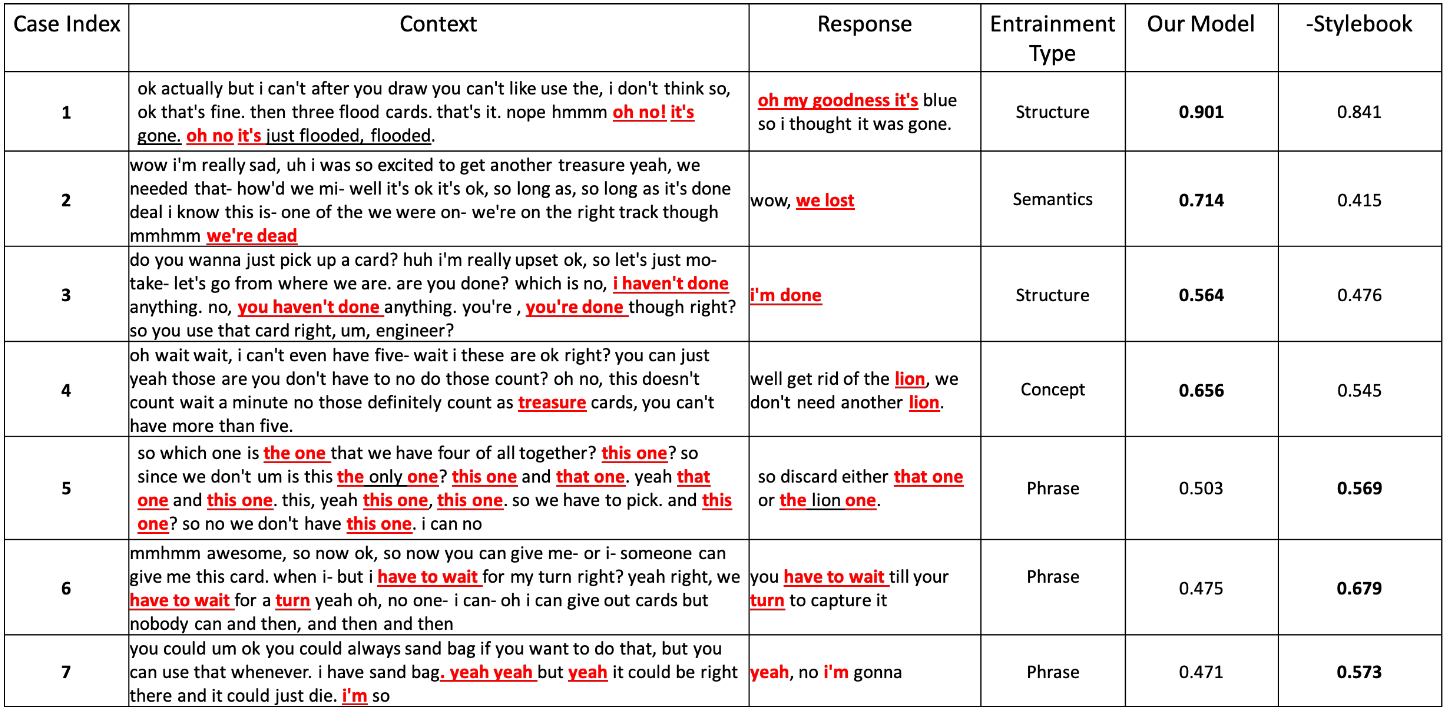
In this case study, we found the model without stylebook inclines to catch lexical entrainment based on replication in terms. On the other hand, the model with the stylebook prefers language similarity beyond merely lexical, such as similarity in sentence structures, semantics, and concept in terms. Those features are what we intend to address in this study by proposing a neural model to perform the similarity calculation for entrainment.

Table 12: 3 Regression models on Teams Corpus. Baseline ent, Our ent, No Stylebook ent denote the entrainment convergence variables derived by the baseline approach, our neural model, our neural model after removing the stylebook, respectively. w. absMax:weighted absMax, unw. absMax:unweighted absMax, * if p <0.05, ** if p <0.01.

DV		**IV**	**Baseline (β)**	**Our(β)**	**No Stylebook(β)**
Task Conflict		Baseline ent (*unw. absMax*)	-0.27*	-	-
		Our ent (*absMax*)	-	-0.35**	-
		No Stylebook ent (*absMax*)	-	-	-0.37**
	R^2		0.07	0.12	**0.14**
	F		4.77*	8.10**	9.73**
Process Conflict		Baseline ent (*w. absMax*)	-0.34**	-	-
		Our ent (*Max*)	-	-0.34**	-
		No Stylebook ent (*absMax*)	-	-	-0.37**
	R^2		0.12	0.12	**0.14**
	F		7.87**	7.85**	9.44**
Relationship Conflict		Baseline ent	-	-	-
		Our ent (*absMax*)	-	-0.28*	-
		No Stylebook ent (*absMax*)	-	-	-0.33**
	R^2		-	0.08	**0.11**
	F		-	5.20*	7.21**

Table 13: 3 hierarchical regression models with team characteristics. Baseline ent, Our ent, No Stylebook ent denote the entrainment convergence variables derived by the baseline approach, our neural model, our neural model after removing the stylebook, respectively. w absMax:weighted absMax, unw absMax:unweighted absMax, * if p <0.05, ** if p <0.01.

DV		**IV**	**Baseline** (β)	**Our**(β)	**No Stylebook**(β)
Task Conflict		Age diversity	-0.03	0.07	0.07
		Ethnic diversity	0.19	0.25	0.19
		Gender diversity	-0.11	-0.04	-0.01
		%Female	-0.30	-0.19	-0.21
		Team size	0.25*	0.18	0.14
		Baseline ent (unw absMax)	-0.31*	-	-0.35
		Our ent (absMax)	-	-0.34**	-
		No Stylebook ent (absMax)	-	-	-0.35**
	R^2		0.22	**0.23**	**0.23**
	F		2.54*	2.78*	2.78*
Process Conflict		Age diversity	0.04	0.07	0.01
		Ethnic diversity	0.29*	0.31*	0.30*
		Gender diversity	-0.13	-0.10	-0.14
		%Female	-0.14	-0.07	-0.23
		Team size	0.31*	0.29*	0.37**
		Baseline ent (w absMax)	-0.36**	-	-
		Our ent (Max)	-	-0.35**	-
		No Stylebook ent (absMin)	-	-	-0.38**
	R^2		0.28	0.27	**0.29**
	F		3.57**	3.41**	3.70**
Relationship Conflict		Age diversity	-	0.19	0.19
		Ethnic diversity	-	0.18	0.13
		Gender diversity	-	0.08	0.09
		%Female	-	0.09	0.08
		Team size	-	0.30*	0.27*
		Our ent (absMax)	-	-0.26*	-
		No Stylebook ent (absMax)	-	-	-0.27*
	R^2		-	**0.21**	**0.21**
	F		-	2.45*	2.46*



Case Index	Context	Response	Entrainment Type	Our Model	-Stylebook
1	ok actually but i can't after you draw you can't like use the, i don't think so, ok that's fine. then three flood cards. that's it. nope hmmm **oh no! it's** gone. **oh no it's** just flooded, flooded.	**oh my goodness it's** blue so i thought it was gone.	Structure	**0.901**	0.841
2	wow i'm really sad, uh i was so excited to get another treasure yeah, we needed that- how'd we mi- well it's ok it's ok, so long as, so long as it's done deal i know this is- one of the we were on- we're on the right track though mmhmm **we're dead**	wow, **we lost**	Semantics	**0.714**	0.415
3	do you wanna just pick up a card? huh i'm really upset ok, so let's just mo- take- let's go from where we are. are you done? which is no, **i haven't done** anything. no, **you haven't done** anything. you're , **you're done** though right? so you use that card right, um, engineer?	**i'm done**	Structure	**0.564**	0.476
4	oh wait wait, i can't even have five- wait i these are ok right? you can just yeah those are you don't have to no do those count? oh no, this doesn't count wait a minute no those definitely count as **treasure** cards, you can't have more than five.	well get rid of the **lion**, we don't need another **lion**.	Concept	**0.656**	0.545
5	so which one is **the one** that we have four of all together? **this one**? so since we don't um is this **the** only **one**? **this one** and **that one**. yeah **that one** and **this one**. this, yeah **this one**, **this one**. so we have to pick. and **this one**? so no we don't have **this one**. i can no	so discard either **that one** or **the** lion **one**.	Phrase	0.503	**0.569**
6	mmhmm awesome, so now ok, so now you can give me- or i- someone can give me this card. when i- but i **have to wait** for my turn right? yeah right, we **have to wait** for a **turn** yeah oh, no one- i can- oh i can give out cards but nobody can and then, and then and then	you **have to wait** till your **turn** to capture it	Phrase	0.475	**0.679**
7	you could um ok you could always sand bag if you want to do that, but you can use that whenever. i have sand bag. **yeah yeah** but **yeah** it could be right there and it could just die. **i'm** so	**yeah**, no **i'm** gonna	Phrase	0.471	**0.573**

Figure 21: Context-response matching examples. The previous 5 turns (IPUs) are concatenated as the context. Our model and -Stylebook show the similarity scores from our model before and after removing the stylebook.

5.8 Conclusion and Future Work

We present a neural dialogue response matching model designed explicitly as a similarity measure for lexical entrainment. We propose a novel architecture, the shared stylebook, to generalize the global shared linguistic features across dialogues. We perform several ablation studies to understand the impact of the stylebook and the underlying meaning of its embeddings. The results suggest that the shared stylebook improves the model performance in a dialogue response matching task. We visualize the style embeddings and observe some sentence clustering by their characteristics or styles. We perform ablation studies to understand our model. We find our similarity measure is strongly correlated with an existing bag-of-words entrainment measure. Meanwhile, removing the stylebook will weaken the correlation, implying that the stylebook is vital for generating meaningful entrainment measures.We then conduct an extrinsic evaluation to compare our measure and the bag-of-words measure in dialogue outcome prediction. Our measure leads to a more robust prediction model with a stronger entrainment signal. On the other hand, the improvement in prediction is not caused by using the stylebook. One possible explanation is that the dialogue success or failure is more correlated with specific types of lexical entrainment than others. Our ablation studies on the stylebook demonstrate how the stylebook assists input embeddings in the semantic space. Our case study further reveals that the model with stylebook prefers entrainment in sentence structure, semantic, and concepts. Without the stylebook, the model seems more biased to low-level lexical entrainment such as phrase repetitions. Thus we suppose that the dialogue success and failure in Teams Corpus are possibly associated more with low-level lexical entrainment. In the future, we aim at evaluating our model utility for computing entrainment in other types of corpora.

6.0 Dialogue Response Generation with Entrainment

6.1 Introduction

In Chapter 5, we propose a neural network-based similarity measure for entrainment by leveraging the global linguistics features. In this chapter, as the ultimate goal of this dissertation, we further propose incorporating entrainment into an end-to-end dialogue response generation model capable of generating more natural and diverse responses. We will utilize the neural network-based similarity measure in Chapter 5 and train a Transformer-based [109] dialogue response generation model. Our model can be trained in an end-to-end manner with minimum supervision.

Generating natural and engaging dialogue responses is the ultimate goal of dialogue response generation models. Most of the recent data-driven approaches of response generation are built upon neural sequence-to-sequence (S2S) models [95, 67, 56]. However, sequence-to-sequence(S2S) models often suffer from oversimplified training objectives, leading to generate trivial and generic responses [56, 101]. This issue is known as the *safe response* issue of neural dialogue response generation models. Many approaches have been proposed to solve this problem, such as topics modeling [124], learning dialogue history [95], and user profiling [56]. *In this study, we propose to address the safe response by leveraging linguistic entrainment in the dialogue, a phenomenon that conversational partners tend to mimic each other in their languages.* Prominent evidence has demonstrated the prevalence of linguistic entrainment in human dialogue [9, 19]. According to its definition, a high degree of entrainment in a response implies the response is in accordance with the dialogue context. Intuitively, incorporating entrainment in responses generation can enrich a response with information more related to a dialogue context.

Entrainment has been widely exploited in many previous studies for different purposes. It has been used in automatic spoken dialogue systems to encourage users to adopt the system-preferred lexical terms [81, 61]. It is also used as leverage in dialogue response generation systems to improve the coherence between generated responses and dialogue context [23, 56]. Recently researchers have started to gain more interest in building an entrainable dialogue system [44].

This study is also motivated by creating an entrainable dialogue system, but *compared to other* *studies in this realm, our study focuses on exploring the effectiveness of entrainment in improving response diversity*. We incorporate entrainment into a Transformer-based S2S model and train the model in two stages. In the first stage, we train a dialogue response matching model that scores the degree of entrainment between the context and response. In the second stage, we train a Transformer-based dialogue response generation model that generates responses given the context and a degree of entrainment. Our approach allows us to manipulate the degree of entrainment during model inference and further control response diversity. Our approach is a flexible strategy to generate entrained responses on the fly. Also, compared to other related studies in the area of dialogue response generation, our evaluation concentrates on understanding the feasibility of using entrainment as a promoter of response diversity. For this purpose, we conduct both automatic and human evaluations.

Precisely, we have 2 hypotheses(The numbering of the hypotheses are based on Chapter 1.2):

1. **Hypothesis 5**: we predict that our response generation model will generate responses with entrainment.

2. **Hypothesis 6**: we assume that our model will generate more diverse responses. Meanwhile, it is crucial that responses are fluent enough to interpret. Therefore, we also need to ensure a satisfactory response quality by considering both response diversity and fluency.

Our results show that for H5, our model is more sensitive to the fake conversation containing less entrainment. The responses generated by our model are more similar to context in terms of lexical usage and semantics. For H6, our model with a high level of entrainment generates responses that achieve the best diversity and fluency in automatic metrics. The human evaluation shows that our model with a low level of entrainment generates better responses in terms of both fluency and diversity by a small margin compared to other baselines. A statistical test on degrees of entrainment suggests that there is no significant difference in response diversity between degrees.

6.2 Chapter-specific Related Work

6.2.1 Response Generation Models

A response generation model is the core component for constructing a generative-based dialogue system. Compared to its counterpart, the response matching model that matches a dialogue context with a response, the response generation model generates responses given the dialogue context. So far, the most common approach for this task is built upon S2S models [55, 124, 95]. Besides that, other more advanced approaches such as generative adversarial training (GAN) [46, 25], reinforcement learning (RL) [25], pre-trained models [133] and knowledge graphs (KG) [125] have also recently started to attract attention. We use the S2S architecture as our model backbone and focus on exploring the utility of linguistic entrainment in this context. Specifically, we use vanilla Transformer [109] as it is a generic state-of-the-art S2S model that has been widely used in many neural models. *We are interested in building our model from scratch on an entrainment-verified corpus. Thus, in our study, we attempt to avoid pre-trained language models such as Bert [22] because they are normally trained on enormous text data that can be unverified for entrainment. We also do not fine-tune models here because we will adjust the internal architecture in the Transformer to couple with entrainment.*

6.2.2 Generating Diverse Responses

Many response generation models have attempted to improve response diversity. For S2S models, Li et al. [55] propose a new objective function using Maximum Mutual Information (MMI). Zhao et al. [134] leverage a discourse-level latent variable and propose a new type of bag-of-words-based loss. Zhou et al. [136] build a mechanism-aware model with a joint attention mechanism constructed with a biased probability distribution. Other than S2S, approaches in KG have started to emerge recently. Moon et al. [71] introduce a knowledge graph that enables their model to learn the entities related to dialogue context. Xu et al. [125] further propose an approach combined with KG and RL. *Our model belongs to the category of S2S. Compared to other existing work, we attempt to capture a latent linguistic signal as a supplementary condition to the model, which further allows our model to generate a controlled response.*

6.2.3 Lexical Entrainment in Dialogue Response Generation

Several attempts have been made to incorporate entrainment into a generative dialogue system. Early studies mainly focused on modeling entrainment in rule-based systems [11, 13]. A few similar task-oriented systems encouraged language priming by aligning language used by the system and the user [39, 81, 61]. Later studies are more interested in automatically modeling entrainment by machine-learning techniques [56, 23]. Recent studies have shown a trend to leverage entrainment in reinforcement learning [44]. *Our approach is in the vein of automatically modeling entrainment. Compared to other studies, our work focuses on controlling entrainment as a supplementary condition to boost response diversity.*

6.2.4 Stylized/Conditional Response Generation

Stylized or Conditional dialogue generation aims at generating responses based on specific styles or conditions. This area of research is related to our study because linguistic entrainment can be interpreted as a unique "style." Unlike stylized dialogue generation that requires a well-defined style, dialogue generation with entrainment is more flexible because the target styles are case-by-case depending on each dialogue. There have been many works in stylized dialogue generation [78, 29]. Controllable dialogue response generation is also related to stylized dialogue generation [56, 132, 131]. *To summarize, compared to those existing works, the style of our model depends on individual conversations.*

6.3 Modeling Approach

Our approach is a two-staged training performed by a pipeline that consists of two parts: entrainment scoring and response generation. The pipeline is shown in Figure 22. A processing example is shown in Figure 23. A context and response pair is entered into the first stage, and the matching model generates a score (0.87 in Figure 23) as the degree of entrainment. Then, given the same context and the degree of entrainment, the second stage model generates a corresponding response. Responses used in training are ground-truth responses.

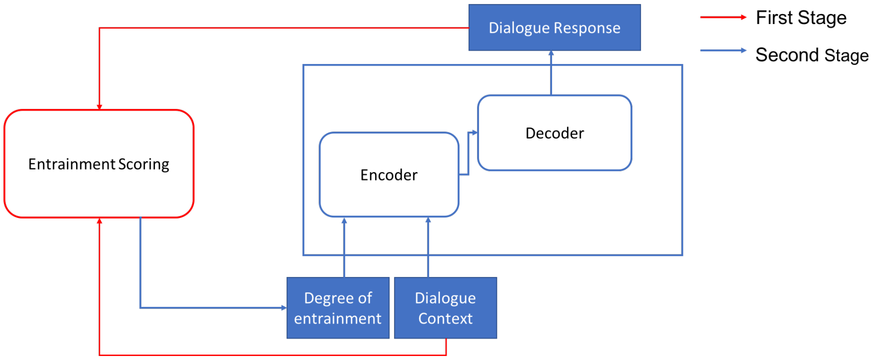


Figure 22: The two-stage training. Red and blue arrows show the procedures at the 1st and 2nd stages, respectively.

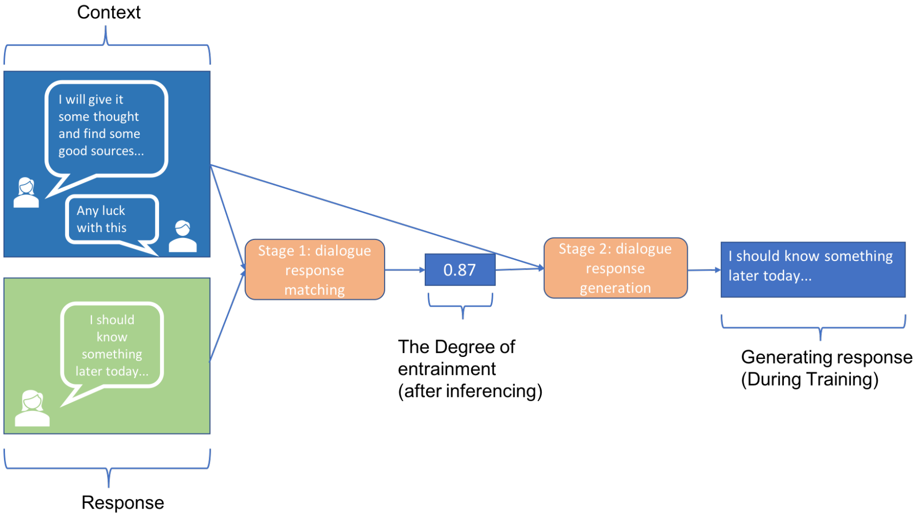

Context
I will give it some thought and find some good sources...
Any luck with this
I should know something later today...
Response
Stage 1: dialogue response matching
0.87
The Degree of entrainment (after inferencing)
Stage 2: dialogue response generation
I should know something later today...
Generating response (During Training)

Figure 23: An example of the training procedure.

6.3.1 Entrainment Scoring

Entrainment scoring is at the first stage, where a dialogue response matching model scores the degree of entrainment between a dialogue context and a response. We formally define the task as the following: given a dialogue context **C**, the goal of the model is to learn the degree of entrainment **E** between a response **R** and **C**. Many dialogue response matching models rely on a similarity score to determine whether a response is proper for the given dialogue context. Rather than the final output, the similarity score is often handled as an intermediate output. We will extract and utilize the similarity score as our entrainment score **E** between a dialogue context and response. We utilize the response matching model introduced in Chapter 5. This model focuses on learning and leveraging entrainment between context and response, and we have demonstrated the connection between their model similarity score and entrainment. We train the model following its task setting where the input is a triplet (**C**, **R**, y) where **C** is the dialogue context, **R** is a response, and y is a label of either 0 or 1 for a randomly sampled response and the ground truth, respectively. The model is trained with a binary classification objective. Later we can apply the model to context-responses pairs and then extract the similarity score as **E** for the following second stage training. We call this **E** as the **First Stage Degree**, which will be used to train the second-stage model.

6.3.2 Response Generation

Response generation is at the second stage, where a dialogue response generation model generates a response given a dialogue context and the degree of entrainment from the first stage. Formally, we define the task as the following: given a dialogue context **C** and the degree of entrainment **E**, the objective of the model is to generate a response **R**. The generation model is a vanilla Transformer model [109] with an encoder-decoder S2S architecture. The Transformer is a powerful attention-based neural architecture in sequence to sequence modeling. Many state-of-the-art language models are derived from the Transformer, such as Bert [22], and GPT-2 [87] that both use the Transformer as their basic building blocks. As an initial attempt to incorporate entrainment into the neural generation model, we intentionally start from the basic block to develop a more relevant intuition about the benefits of entrainment. Focusing on the basic building block also give us flexibility in the future to integrate entrainment in other Transformer-based model.

To incorporate entrainment into the Transformer model, we scale the encoder-decoder attention by the degree of entrainment **E** where **E** is a scale between 0 and 1. Originally, Scaled Dot-Product Attention contains three essential matrices: the queries **Q**, the keys **K**, and the values **V**. **Q** is the context encodings generated from encoders, and **K** and **V** are linear transformations from the response encodings generated from the decoders. In our modification, we add entrainment **E** and scale both **K** and **V** by it. The intuition behind the scaled attention by entrainment is that less entrainment in a response implies fewer similarities between the context and the response. Equation 15 shows our modified attention calculation after adding **E**.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{\mathbf{Q} \cdot (E\mathbf{K})^T}{\sqrt{d_k}}) \cdot E\mathbf{V} \tag{15}$$

6.4 Data

We use a constrained non-goal-oriented dialogue corpus for this study. Constrained corpora with limited topics are popular to train S2S models. To ensure that entrainment exists in the data, we select a dialogue corpus that has been used in other entrainment-related studies. Wikipedia Talk Page[1] is a collection of conversations from Wikipedia editor's talk pages, where Wikipedia editors discuss the changes made to Wikipedia articles or projects. For more details please see Section 3.2. Previously, Danescu-Niculescu-Mizil et al. [19] have constructed a public dialogue corpus from the Wikipedia Talk Page, and they have demonstrated the existence of entrainment between speakers in this corpus. We use the same corpus here. In the following section, we introduce two datasets constructed from Wikipedia Talk Page, specifically designed for the two tasks in our two-stage training. Note that there are many typos, urls, and links in this corpus. To reduce the complexity of datasets, we clean the corpus by replacing all links, urls, numbers, email addresses with special tokens.

[1] https://convokit.cornell.edu/documentation/wiki.html

6.5 Model Training and Evaluation Metrics

6.5.1 Entrainment Scoring

6.5.1.1 Dataset We first construct a dialogue response matching dataset for entrainment scoring. To make an example, we use 2 conversational turns as a dialogue context and the following turn as the ground truth response. Compared to other datasets used in previous chapters, we use less turns as context because utterances in Wikipedia Talk Page are long. This results in 129,928 positive instances. We split positive instances to train, validation, and test sets based on a ratio of 6:2:2. Then for each positive instance, we randomly sampled 9 false responses for validation and test sets and 1 false response for the train set. The final dataset contains approximately 156K, 260K, 260K examples in the train, validation, and test sets.

6.5.1.2 Automatic Evaluation Metrics We apply an extrinsic evaluation of entrainment scores generated by our first-stage model. Previously, in the response generation task, candidate responses are often ranked by how well a response candidate matches the given dialogue context. Evaluation metrics often focus on Recall@k, which indicates that the true positive response is among the first k ranked candidates. In this work, we use entrainment scores to rank candidates because our entrainment score is also the similarity score extracted from the matching model. We evaluate Recall@1 (R@1), Recall@2 (R@2), and Recall@5 (R@5) as previous studies. A high Recall value indicates good entrainment scores.

6.5.1.3 Implementation Details Our model is implemented in Pytorch. We follow the default model configuration. The maximum token length is 200 for the context and 100 for the response [2]. We train a maximum of 20 epochs and optimize the R@1 on the validation set. The best model obtained after training achieves a 0.385, 0.577, and 0.857 of R@1, R@2, and R@5 on the test set, respectively.

[2] We crop the context from the beginning and response from the end.

6.5.2 Response Generation

6.5.2.1 Dataset Using the best model trained from the first stage for entrainment scoring, we obtain the entrainment score of each utterance in Wikipedia Talk Page corpus given the previous 2 turns. The entrainment score generated is the **First Stage Degree** to train the generation model. This results in a total of 129,928 context-response-degree triplets [3]. We split positive instances into train, validation, and test set. This results in a dataset that contains approximately 129K, 500, 500 examples in train, validation, and test set, respectively.

6.5.2.2 Automatic Evaluation Metrics We use the following automatic evaluation metrics for our generation task. Depending on the hypothesis, we may use one or multiple metrics.

- **Perplexity (PPL)** computes how likely the probability distribution of a model will predict a sequence. It is a popular metric to evaluate the dialogue generation models [44, 95, 131, 124]. A lower value of PPL indicates a better generation quality. As an important metric of natural language generation, PPL can reflect the matching between a probability distribution of the training set and the test set. Note that generic responses lacking diversity could also lead to good PPL.

- **BLEU** [80] is a prevalent automatic evaluation metrics for language generation tasks [134, 56, 32]. It computes the co-occurrences of n-grams in the reference and predicted sequences. We specifically focus on BLEU-2 that only considers uni-grams and bi-grams. A higher value of BLEU indicates a better quality.

- **Distinct-1, Distinct-2, Distinct-3** denote the proportion of the unique unigram, bigram, and trigram in the generated responses, respectively. They are used as evaluation metrics in many prior works to reflect lexical diversity of generated responses [55, 124, 64, 29]. In the following sections, we refer to Distinct as the group of Distinct-1, 2, 3. A high value of Distinct indicates more informative, engaging, and diverse responses.

[3] Note that parts of context-response pairs are used as training data for the entrainment scoring model, but entrainment scores are not ground truth labels.

6.5.2.3 Implementation Details Our model is implemented in Pytorch and trained using 3 GPUs. Our inputs are subword pieces tokenized by the pre-trained English byte pair embeddings [37]. Model configuration is tuned on the validation set. The embedding dimension is 512. The maximum token length is 200 for the context and 100 for the response. The batch size is 32. The numbers of encoder and decode layers are both 4. The size of hidden units in the encoder and decoder block are both 2048. The dropout rate is 0.1. All multi-head attention used in this model has 4 heads. The learning rate is 0.0001. We use Adam optimizer. The loss function is cross-entropy. We train a maximum of 150 epochs and optimize training at the BLEU-2 on the validation set. During decoding, we use a beam search with a beam size of 5. The generated sequence length is set to 100.

6.6 Experiments

6.6.1 Hypothesis 5 (H5)

In H5, we hypothesize that our model can generate responses with entrainment. We will examine this hypothesis using two evaluation methods. One method will follow a conventional methodology for entrainment evaluation, and another will be a more direct approach utilizing strategies in computational linguistics.

6.6.1.1 Evaluation Method 1 In previous studies related to entrainment, researchers often validate their entrainment measures by examining whether there exists a statistical difference of entrainment between real and fake conversations [74, 42, 116]. Intuitively, natural conversations are supposed to have a higher degree of entrainment than fake artificial conversations. Thus, inspired by this methodology, we adopt a similar strategy to test whether our model tends to generate responses with entrainment. Specifically, we examine the model bias indicated by PPL in real *Real* and fake *Fake* conversations. If our model inclines to generate dialogue responses with entrainment, PPL should differ when testing on the data that show entrainment, i.e., *Real*, versus the data that show less entrainment, i.e., *Fake*.

The *Fake* contains fake conversations. We make a fake conversation from a real conversation in the test set by keeping the context but replacing function words in the response. Function words have been used as a linguistic marker to validate entrainment in the Wikipedia Talk Page [19]. We define a set of function words based on Linguistic Inquiry Word Count (LIWC) [85], which assigns each function word with a word category. We use 9 categories that have been utilized as linguistic markers for entrainment [127, 33]. Each function word in the real conversation has 15$\%$ of a chance to be replaced by another function word from the same word category. Hence, we preserve the coherence between a context and response by keeping most dialogue content identical to a real conversation. In the meantime, *we reduce the level of entrainment by removing common linguistic markers between the context and response*. H1 is further decomposed into 2 sub hypotheses:

1. Hypothesis 5.1 **(H5.1)** hypothesizes that our model will achieve better PPL in *Real* compared to the baseline model that has no bias for entrainment. H5.1 is formally formulated as $PPL_{Ours,Real} < PPL_{Baseline,Real}$.

2. Hypothesis 5.2 **(H5.2)** hypothesizes that compared to the baseline, our model will suffer more when we change the test set from *Real* to *Fake*, H5.2 is formulated as Diff($PPL_{Ours,Real}$, $PPL_{Ours,Fake}$) $>$ Diff($PPL_{Baseline,Real}$, $PPL_{Baseline,Fake}$) where Diff(x,y) denotes the difference between x and y. Note that here we don't directly compare a model's PPL on *Real* and *Fake*, because we can always expect a PPL increase in *Fake* resulted from a degradation in sentence quality after manipulating the function words.

Baseline Model

Our baseline model is a vanilla Transformer model [109] that does not utilize entrainment.

Results

Table 15 shows the results. For H5.1, our model achieves worse PPL in *Real* compared to the baseline model. This indicates that our model is weaker when being tested on the ground truth responses compared to the baseline. H5.1 is not supported. One possible explanation is that by introducing entrainment into the model, we also bring in training noise. In the meantime, propagated er-

Table 14: Evaluation results for H5 method 1. The numbers shown in the table are PPL. Bold numbers indicate the results that support our hypothesis.

	Baseline	Ours
Real	99.62	112.55
Fake	260.95	304.88
Difference	161.33	**192.33**

rors from the first stage can flow through the pipeline due to stage interdependence. For H5.2, compared to the baseline, our model suffers more in the PPL when we change the test set from *Real* to *Fake*, where Diff($PPL_{Ours,\mathbf{Real}}$, $PPL_{Ours,\mathbf{Fake}}$) $>$ Diff($PPL_{Baseline,\mathbf{Real}}$, $PPL_{Baseline,\mathbf{Fake}}$). This indicates that our model is more sensitive to fake conversations containing misused function words and less entrainment, while the baseline is more robust. Also, baseline performing better in *Fake* provides evidence that compared to the baseline model, our model relies more on function words that convey entrainment. H5.2 is supported.

Evaluation Method 1 mainly depends on PPL, which can raise some concern about the reliability of H5.1 conclusion because generic responses lacking diversity could also lead to good PPL. Therefore, we introduce Evaluation Method 2 using a more computational strategy.

6.6.1.2 Evaluation Method 2 One possible method directly computes the language overlap as entrainment between generated responses and their context. Due to the fact entrainment can occur in many dimensions such as language structure, styles, and lexical as described in related works, here we limit our scope in lexical entrainment for its simplicity. Specifically, we will evaluate the lexical similarity between generated responses and their context. We adopt text similarity measures, including the cosine similarity of TF-IDF and Word Mover Distance (WMD) [40] of word embeddings. TF-IDF similarity focuses on the usage of lexical items. Word Mover Distance measures the minimum distance between two vectors in a common semantic space and thus can indicate semantic similarity. Formally, H5 is further stated as:

1. Hypothesis 5.3 **(H5.3)** hypothesizes that our model will achieve better TF-IDF similarity compared to the baseline model that does not incorporate entrainment. A successful proof of this hypothesis implies that our model tends to use the same lexical items used in context.

2. Hypothesis 5.4 **(H5.4)** hypothesizes that our model will achieve better WMD compared to the baseline model that does not incorporate entrainment. A successful proof of this hypothesis implies that our model tends to use lexical items having similar semantic meanings.

Baseline Model

Our baseline model is a vanilla Transformer model [109] that does not utilize entrainment. Our model performs decoding with the First Stage Degree.

Results

Table 15 shows the results. For embeddings used with WMD, we experiment different pre-trained embedding models, including word2vec Googlenews [70], fastText [6], glove Giga, and glove twitter [86]. FastText contains fewer out-of-vocabulary words because it is a sub-word-based model. Glove Giga and word2vec Googlenews are large models trained on a large amount of general data from Wikipedia and Google News. Glove twitter contains embeddings for internet vocabulary, similar to the language used in Wiki Talk Page. TF-IDF is trained on the context of each context-response pair [4].

The results first show ground-truth responses always achieve the best similarity. Other than that, our model outperforms the vanilla Transformer in all methods. Overall, our model generates more similar responses to the context in terms of word usage and semantics by incorporating entrainment. H5.3 and H5.4 are both supported.

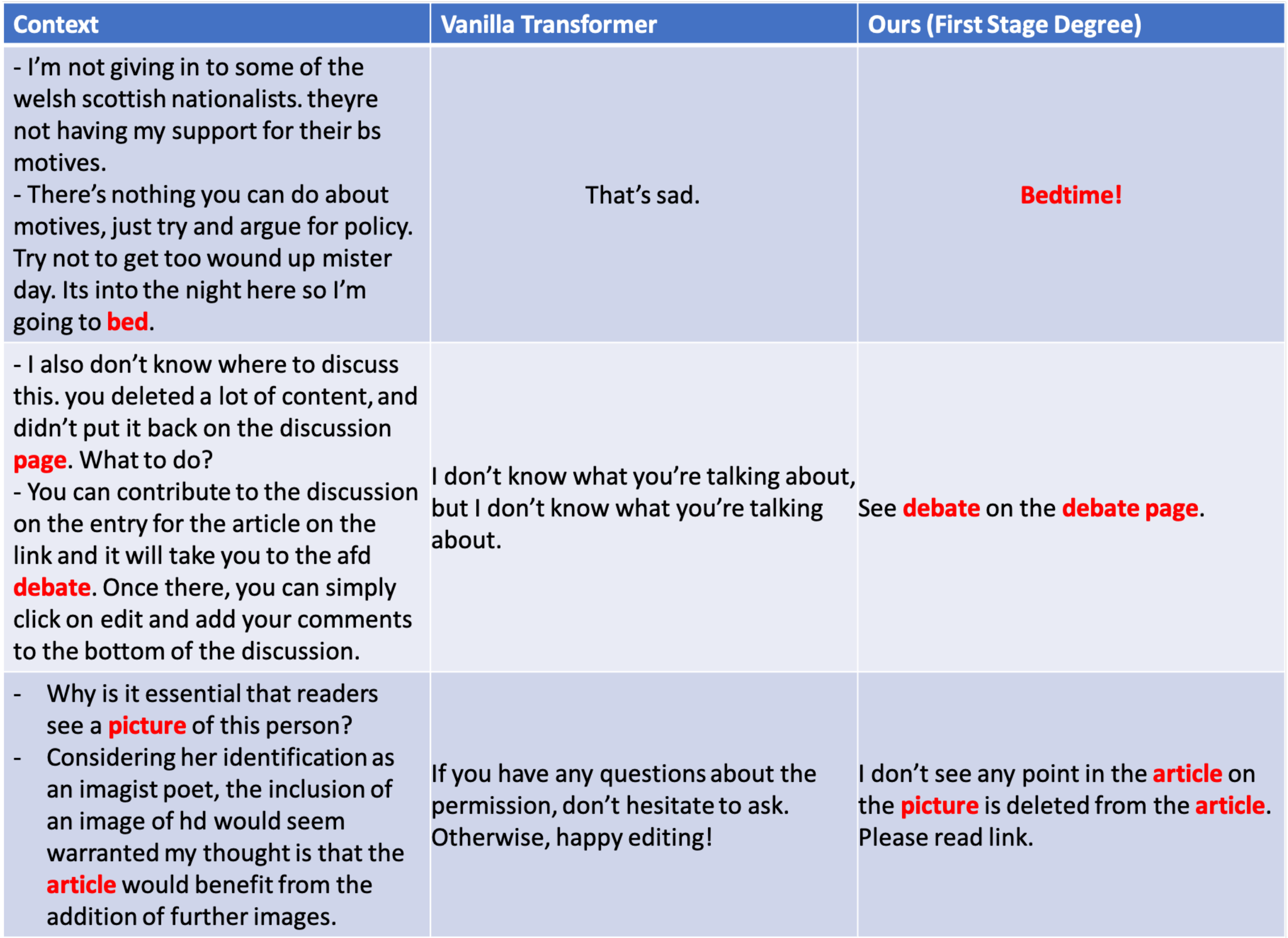
Figure 24 shows 3 examples. The baseline and our model both generate responses based on the same context. The red text in the table shows the part that is considered as lexical entrainment in our point of view. In the first example, the word “bed” is mentioned in the context, and our model generates “bedtime” as “bed” and “bedtime” can be closed in semantic space. In the second and third examples, the same words, e.g., the “debate” in the second example, are used in responses

[4] We use WMD, and IF-IDF provided in Gensim[91]

Table 15: Evaluation results for H5 method 2. Columns of method show models and similarity algorithms. Parenthesized numbers are the embeddings dimensions. Bold numbers indicate the best results. Bold and underlined numbers indicate the best results except for the ground-truth. A greater TF-IDF cosine indicates a greater similarity, while a smaller WMD indicates greater similarity.

Method	Ground-truth	Vanilla Transformer	Ours (First Stage Degree)
TF-IDF cosine	**0.181**	0.123	<u>**0.126**</u>
FastText WMD (300)	**0.827**	0.951	<u>**0.935**</u>
Glove Giga WMD (300)	**0.843**	0.975	<u>**0.953**</u>
Glove Twitter WMD (200)	**0.777**	0.827	<u>**0.809**</u>
Googlenews WMD (300)	**0.981**	1.100	<u>**1.086**</u>

generated by our model. The vanilla Transformer does not show any advantage in generating similar lexical items in these 3 examples.



Context	Vanilla Transformer	Ours (First Stage Degree)
- I'm not giving in to some of the welsh scottish nationalists. theyre not having my support for their bs motives. - There's nothing you can do about motives, just try and argue for policy. Try not to get too wound up mister day. Its into the night here so I'm going to **bed**.	That's sad.	**Bedtime!**
- I also don't know where to discuss this. you deleted a lot of content, and didn't put it back on the discussion **page**. What to do? - You can contribute to the discussion on the entry for the article on the link and it will take you to the afd **debate**. Once there, you can simply click on edit and add your comments to the bottom of the discussion.	I don't know what you're talking about, but I don't know what you're talking about.	See **debate** on the **debate page**.
- Why is it essential that readers see a **picture** of this person? - Considering her identification as an imagist poet, the inclusion of an image of hd would seem warranted my thought is that the **article** would benefit from the addition of further images.	If you have any questions about the permission, don't hesitate to ask. Otherwise, happy editing!	I don't see any point in the **article** on the **picture** is deleted from the **article**. Please read link.

Figure 24: Generated responses from the baseline vanilla Transformer and our model decoded with First Stage Degree entrainment. Bold and red text shows the parts that could consider as lexical entrainment.

6.6.2 Hypothesis 6 (H6)

In hypothesis 6 (H6), we assume that our model will generate more diverse responses. The overall response quality should be satisfactory considering both fluency and diversity.

6.6.2.1 Evaluation Method To examine response diversity, we split our evaluation procedure into two parts focusing on two sub hypotheses:

1. Hypothesis 6.1 (**H6.1**) hypothesizes incorporating entrainment in the model will lead to a more diverse response generation. For evaluation, baselines will generate dialogue responses given a dialogue context. Our models will generate a corresponding response given the same dialogue context and a supplementary degree of entrainment. We experiment with the First Stage Degree of the ground truth response and three other levels of degrees of entrainment: LOW, MID, and HIGH, indicating the degree of entrainment of 0.33, 0.66, and 0.99, respectively. So in the First Stage Degree, every context is bound with a distinct entrainment score, while in the same level, every context is bound with an equivalent entrainment score. We compare the automatic evaluation metrics among the baselines and our models. The evaluation metrics focus on response diversity reflected by Distinct. To exam the overall response quality, we also evaluate response fluency reflected by BLEU-2. We follow Li et al. [55] to not use PPL as a fluency measure because generic responses lacking diversity could also lead to good PPL, which is aligned with the interest of this work. Additionally, we conduct a human evaluation on response diversity and fluency (see Human Evaluation).

2. Hypothesis 6.2 (**H6.2**) hypothesizes that response diversity will vary depending on different degrees of entrainment, and there exists a linear relationship between the degree of entrainment and response diversity. The goal is to understand how the degree of entrainment can impact response diversity. Given a dialogue context, we let our model generate responses by assigning the three different levels of degrees of entrainment mentioned in H6.1. Similarly to H6.1, the evaluation metrics focus on response diversity reflected by Distinct and response fluency reflected by BLEU-2. Again, we conduct a human evaluation on response diversity and fluency (see Human Evaluation). We predict that there is a statistically significant difference

in diversity among the three degrees of entrainment.

6.6.2.2 Baseline Models We compare our model with the following neural response generation models. All of the models do not explicitly incorporate entrainment:

1. **Transformer**: the vanilla Transformer model that our model is based on [109]. We have used this model as the baseline in **H5**. **Transformer** is an attention-based neural model.

2. **HRED**: a hierarchical encoder-decoder model based on RNN [97]. It is one of the state-of-the-art S2S dialogue response generation models for non-task-oriented dialogue. **HRED** is widely used as a neural baseline model in dialogue response generation task [130, 94, 82, 44].

3. **VHRED**: a variational encoder-decoder model extended from **HRED** based on RNN [95]. Similar to **HRED**, **VHRED** is also used as a popular neural baseline model for response generation task [130, 82, 44].

6.6.2.3 Automatic Evaluation Results Table 16 shows the results. We first investigate H6.1, where we hypothesize that responses generated by our entrainable model are more diverse than those generated by the baselines. We organize the table into 2 categories: Models and Levels. Models focus on models with different architectures. Levels focus on our models with different degrees of entrainment.

In Models, we compare different model architectures. We found transformer-based models, i.e., ours and the vanilla Transformer, in general, outperform the RNN-based models, i.e., HRED and VHRED. We compare our model coupled with the First Stage Degree to the vanilla Transformer. The result reveals that our model marginally improves the Distinct-2, Distinct-3, and BLEU-2 beyond the vanilla transformer model. This also validates the effectiveness of using entrainment in vanilla Transformer. In Levels, we further investigate the utility of different degrees of entrainment in improving the best results in Models. HIGH achieves the approximately 1.27, 0.3, 3.81, 7.52 percentage improvement in terms of BLEU-2 and Distinct-1,2,3, respectively.

In the union of Models and Levels, our model HIGH achieves the highest Distinct and BLEU-2. Therefore, in summary, the results show that incorporating entrainment in the model can lead

Table 16: Evaluation results for H6. BLEU-2 and Distinct are calculated on decoding outputs. **Bold text** shows the best results in each section of the table. *Italic text* shows the results that are better than the vanilla Transformer model.

		BLEU-2	Distinct-1	Distinct-2	Distinct-3
Models	HRED	2.67%	10.08%	35.76%	56.08%
	VHRED	1.95%	11.02%	38.37%	58.06%
	Vanilla Transformer	2.69%	**16.24**%	42.59%	58.09%
	Ours (First Stage Degree)	***2.99%***	15.70%	***43.01%***	***58.81%***
Levels	Ours (LOW)	0.63%	12.30%	31.75%	42.01%
	Ours (MID)	2.08%	*16.50%*	42.48%	56.89%
	Ours (HIGH)	***4.26%***	***16.54%***	***46.82%***	***66.33%***

to a more diverse response generation, and in the meantime, the overall quality considering both diversity and fluency is satisfactory. H6.1 is supported. Note that the BLEU-2 of HIGH indicates that our model-generated responses are more similar to the ground truth.

For H6.2, where we hypothesize that response diversity will vary depending on different degrees of entrainment, we see that all performance metrics tend to increase as the degree of entrainment increases.HIGH has the highest diversity and fluency. H6.2 is supported.

6.6.2.4 Human Evaluation Results Researchers have argued that automatic evaluation metrics only weakly correlate with human judgments in dialogue response generation tasks [59]. Thus we further conduct a human evaluation for H2 on both response quality and diversity. We recruit 5 human judges, 4 from Amazon Mechanical Turk (AMT) and 1 from the CS department of a university. We mix the source of recruitment to include judges from different backgrounds, since the corpus contains some technical content as well as other topics such as politics, celebrities, and text editing. All judges demonstrated adequate English proficiency by passing a pre-screening English proficiency test designed by us.

Additionally, we include the English test and an illustration of the evaluation interface in the

Appendix D.

We select 100 cases of dialogue context from the test set. Given a context, we ask human judges to rate candidate responses generated from the 7 models in Table 16. Additionally, we add the ground truth as a candidate response to verify the quality of human judgments by assuming that the ground truth will always achieve the highest score. Hence, in total, 8 candidate responses are shown to each human judge per dialogue context. We randomly shuffle the responses in the display so that the human judges can not easily distinguish the origin of responses. Following a popular rating practice [94, 78, 59], we employ a 5-point scale with 0 being the lowest and 4 being the highest to rate response fluency and diversity. Separating fluency and diversity enables us to perform a fine-grained quality analysis. The scale is explained in Table 17. We exclude 1 AMT judge from our study because the judge has a Cohen's Kappa $\kappa < 0.2$ with other judges. The same practice has been employed in a previous work [59] to increase the reliability of the human evaluation. Hence, 4 judges are used for our evaluation, with 6400 ratings in total [5]. The average fluency and diversity κ across each pair of 4 judges are 47.6 (moderate agreement) and 33.33 (fair agreement), respectively [6].

Table 17: The 5-point scale used in our human evaluation. Rating criteria are based on a prior study of Serban et al. [94], but modified to address more on response diversity. Specifically, we added description about whether responses are informative and generic.

Rating	**Fluency**	**Diversity**
0	Incomprehensible	Not Relevant
1	Non-Native English	Little Relevance, Little information, Generic
2	Disfluent English	Much Relevance, Some information, Generic
3	Good English	Most Relevant, Informative, Not Generic
4	Flawless English	All Relevant, Informative, and Interesting.

Table 18 shows the rating results. The ground truth indeed achieves the highest fluency and diversity. We first investigate H6.1 on diversity. Evaluating response diversity seems to be a

[5] 6400 ratings = 100 contexts * 8 responses * 2 categories * 4 raters.

[6] According to the κ guidelines in Landis and Koch [49].

difficult task. In general, the Cohen's κ for models' **diversity** are lower than **fluency**. In both Models and Levels, the vanilla Transformer achieves the highest average **diversity**. Meanwhile, LOW ranks second, and HIGH ranks third. Note that HIGH achieves the highest diversity in automatic evaluation. In the diversity of HIGH, the proportion of 3-points and 4-points responses sums up to 11.3%, which is larger than that of the vanilla Transformer (10.0%) and LOW (7.8%). Beyond the numbers in the table, we observe that in many cases, HIGH tends to generate long sentences containing more diverse n-grams and repeated phrases. This implies that HIGH has a bias towards diversity rather than fluency, which shows a typical trade-off between fluency and diversity of models in the task of generating diverse responses.

To examine overall response quality, we check response fluency as additional evidence. In both Models and Levels, VHRED and our model LOW both achieve the highest average **fluency**. LOW has a higher κ than VHRED, indicating judges have a stronger agreement on the **fluency** of LOW. In Models that compare different models architectures, Transformer-based models, i.e., Vanilla Transformer and our model using the First Stage Degree of entrainment, perform worse than the RNN-based models, i.e., HRED and VHRED. But later in Levels, by adjusting the degree of entrainment, our model LOW and MID can improve some fluency beyond the vanilla Transformer, leading to a comparable fluency between LOW and RNN-based models.

Because the best fluency and diversity are not achieved by the same model, to determine the best response quality, we sum the average fluency and diversity for all models and levels. LOW obtains the highest total score, indicating the best response quality. Conclusively, H6.1 is only partially supported in human evaluation because our model is not the most diverse model, but it has the best overall quality.

We then investigate H6.2. We perform a one-way ANOVA test between the average **diversity** and **fluency** of Levels. No statistically significant difference is found in **diversity**. No linear relationship between degree of entrainment and response **diversity**. **fluency** is statistically significantly different among groups (F(2,297)=12.13, $p<0.01$). Post Hoc LSD Tests reveal that HIGH **fluency** is smaller than other Levels (both $p<0.01$). Therefore, We consider H6.2 is not supported or only partially supported on response **fluency**.

Table 18: A summary of our human evaluation. 400 ratings are given to each model or level over all 4 judges for each of fluency and diversity. **Bold numbers** show the best performance. Underlined numbers show the best performance except for the ground truth. Avg.: Average. Kappa: Averaged pairwise weighted quadratic Cohen's κ among judges per model or level.

		Fluency							Diversity							Avg. Fluency + Avg. Diversity
		Score Distribution (%)					Avg. Score	Kappa	Score Distribution (%)					Avg. Score	Kappa	
		0	1	2	3	4			0	1	2	3	4			
Models	HRED	2.0	6.0	10.3	61.0	20.8	2.93	35.0	48.3	25.0	15.5	8.8	2.5	0.92	19.4	3.85
	VHRED	2.0	5.5	7.3	65.3	20.0	**2.96**	21.6	50.3	22.0	17.5	5.3	5.0	0.93	24.6	3.89
	Vanilla Transformer	7.8	7.8	8.0	56.8	19.8	2.73	50.9	41.8	24.8	23.5	8.5	1.5	**1.03**	28.6	3.76
	Ours (First Stage Degree)	3.0	11.8	12.3	59.8	13.3	2.69	58.8	43.8	26.0	22.8	5.5	2.0	0.96	31.5	3.65
Levels	Ours (LOW)	1.8	4.8	7.8	67.5	18.3	**2.96**	44.1	42.5	23.0	26.8	6.8	1.0	1.01	25.2	**3.97**
	Ours (MID)	4.5	7.8	9.3	63.5	15.0	2.77	54.5	44.5	26.3	23.3	4.8	1.3	0.92	29.0	3.69
	Ours (HIGH)	8.8	17.0	11.3	49.3	13.8	2.42	52.2	46.3	24.3	18.3	8.3	3.0	0.98	25.1	3.40
Ground Truth		1.0	5.3	8.8	43.8	41.3	**3.19**	27.9	25.0	16.0	13.3	13.3	32.5	**2.12**	35.6	**5.31**

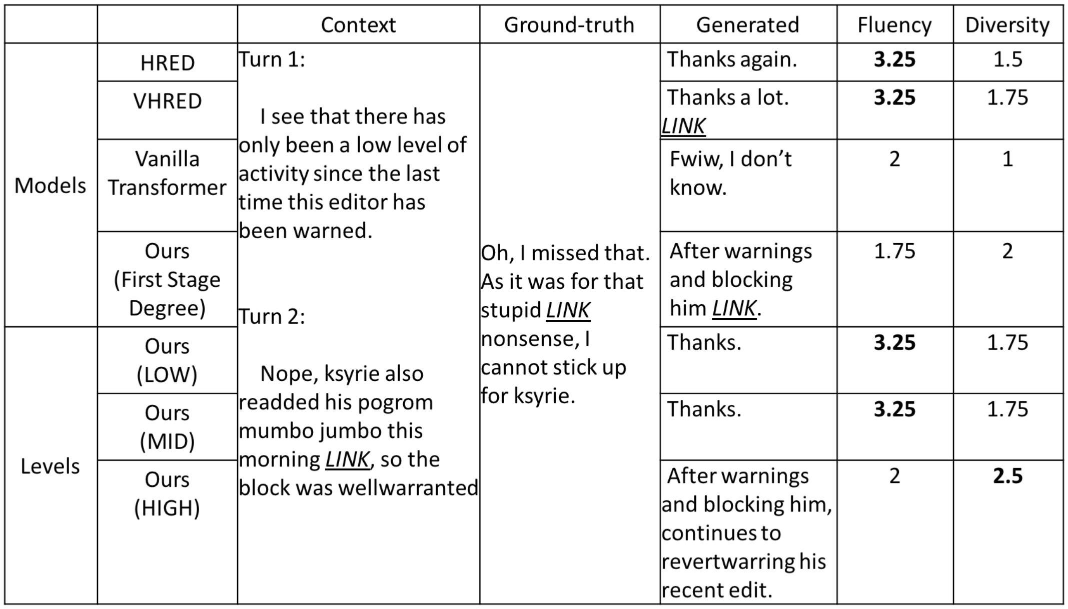
6.7 Case Study

We perform a case study on some generated responses by different models and levels. Figure 25 shows 2 cases. Case 1 shows that our models incorporating entrainment can improve both response fluency and diversity. We observe that RNN-based models, i.e., HRED and VHRED, achieve a better response fluency than Transformer-based models in Models. By controlling entrainment in Levels, LOW and MID can improve fluency beyond the vanilla Transformer, and they achieve a comparable fluency as RNN-based models. Note that responses that achieve the highest fluency are all generic. This further shows that the judges have correctly construed our fine-grained metrics to avoid confusion between fluency and diversity during evaluation. HIGH achieves the highest diversity. The different bias of LOW and HIGH between fluency and diversity exhibits a trade-off between fluency and diversity by using different levels of entrainment. Low favors fluency, while High favors diversity. Case 2 is an example where our model fails to vary response diversity by controlling the degree of entrainment. In Case 2, VHRED and our models, including First Stage Degree, LOW, MID, and HIGH, achieve the highest fluency. Although incorporating entrainment aids diversity beyond RNN-based models and the vanilla Transformer, the diversity at different levels is equivalent. We also observe that our model with First Stage Degree generates the same response as HIGH. This implies that the First Stage Degree of entrainment of ground truth is close to HIGH. No generated responses are close to the ground truth.

6.8 Conclusion and Future Work

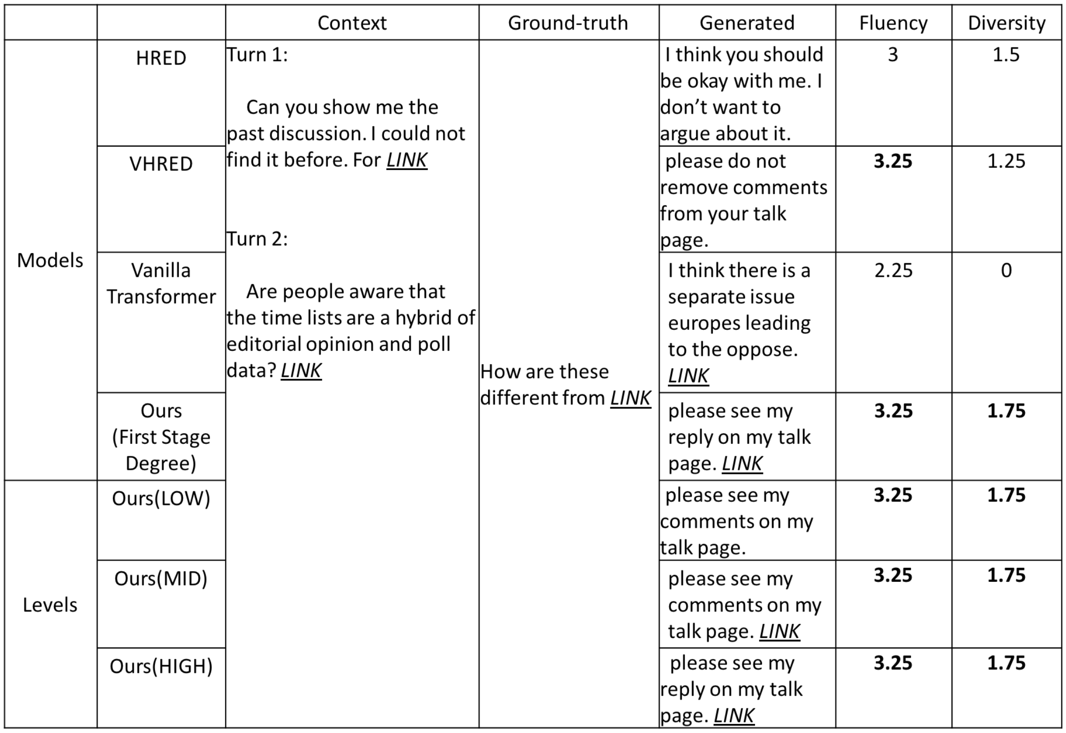
Motivated by the *safe response* issue of current S2S dialogue response generation models, we propose an approach to improve response diversity by incorporating linguistic entrainment into the vanilla Transformer model. Our approach is a two-staged pipeline consisting of two neural models for dialogue response matching and generation. Our model focuses on generating responses by jointly learning the dialogue context and the degree of entrainment. The automatic evaluation suggests that our model results in a good improvement in **diversity** with a good overall quality, but the human evaluation only shows marginally improved overall quality rather than **diversity**, which implies that there is no outstanding advantage to use entrainment for the *safe response* issue. Our statistical analysis in human evaluation shows there is no linear relationship between the degree of entrainment and diversity, but the test reveals that the highest degree of entrainment has the worst fluency. In the future, we intend to experiment with our approach on other corpora.

In this work, we only use one strategy by scaling the attention between encoder and decoder to incorporate entrainment into the Transformer model. Beyond that, there are other potential strategies. For instance, if we view entrainment as a condition code of the input, we can perform code-switching by appending the entrainment score to the input. We can also utilize the entrainment score in an auxiliary training task for multi-task learning, a more advanced strategy requiring a more dedicated model design. Many new strategies can be explored in future work. For these new strategies, our model can serve as a baseline to benchmark them.



		Context	Ground-truth	Generated	Fluency	Diversity
Models	HRED	Turn 1: I see that there has only been a low level of activity since the last time this editor has been warned. Turn 2: Nope, ksyrie also readded his pogrom mumbo jumbo this morning *LINK*, so the block was wellwarranted	Oh, I missed that. As it was for that stupid *LINK* nonsense, I cannot stick up for ksyrie.	Thanks again.	**3.25**	1.5
	VHRED			Thanks a lot. *LINK*	**3.25**	1.75
	Vanilla Transformer			Fwiw, I don't know.	2	1
	Ours (First Stage Degree)			After warnings and blocking him *LINK*.	1.75	2
Levels	Ours (LOW)			Thanks.	**3.25**	1.75
	Ours (MID)			Thanks.	**3.25**	1.75
	Ours (HIGH)			After warnings and blocking him, continues to revertwarring his recent edit.	2	**2.5**

(a) Case 1



		Context	Ground-truth	Generated	Fluency	Diversity
Models	HRED	Turn 1: Can you show me the past discussion. I could not find it before. For *LINK* Turn 2: Are people aware that the time lists are a hybrid of editorial opinion and poll data? *LINK*	How are these different from *LINK*	I think you should be okay with me. I don't want to argue about it.	3	1.5
	VHRED			please do not remove comments from your talk page.	**3.25**	1.25
	Vanilla Transformer			I think there is a separate issue europes leading to the oppose. *LINK*	2.25	0
	Ours (First Stage Degree)			please see my reply on my talk page. *LINK*	**3.25**	**1.75**
Levels	Ours(LOW)			please see my comments on my talk page.	**3.25**	**1.75**
	Ours(MID)			please see my comments on my talk page. *LINK*	**3.25**	**1.75**
	Ours(HIGH)			please see my reply on my talk page. *LINK*	**3.25**	**1.75**

(b) Case 2

Figure 25: Case study. **Bold text** show the highest scores in the corresponding column. Fluency and Diversity are averaged fluency and diversity in human evaluation.

7.0 Summary

7.1 Conclusions

In this thesis, we present 3 studies on quantifying, learning, and utilizing linguistic entrainment in dialogue with different strategies in natural language processing. We first start our study in understanding entrainment in human-to-human conversations. We propose an approach to quantifying entrainment in conversations by matching function words between speakers. Additionally, to further understand the nature of entrainment, we investigate several impact factors related to demographics, including team size, gender, age, and ethnic composition of the conversational group. There are only a few studies that have looked at the connection between multiparty entrainment and team characteristics. Although research has shown that speaker profiles such as gender can impact entrainment, team characteristics are sometimes overlooked in entrainment studies. So beyond proposing the approach to quantify entrainment, our work also contributes to reveal the relationship between team characteristics and multiparty entrainment, which is a research question that is not well studied. We found that gender is a significant impact factor. Most importantly, we experimented with entrainment as a predictor of dialogue success and observed a statistically significant predicting model. This promising result shows that entrainment is a potential indicator of dialogue success.

During the above primitive research, we realize that current entrainment quantification approaches are tightly tied to the bag-of-words paradigm. Bag-of-words approaches can cause feature sparsity and ignore semantics, syntax, and styles. Thus, we propose a new approach to automatically learn and score entrainment with a neural network model, which allows us to decouple entrainment measure from the bag-of-words paradigm and further consider language semantics and forms. Our proposed model is a dialogue context-response matching model with a new attention module named 'stylebook". It attempts to generalize global features from inputs, mimicking the mechanism of using predefined features in bag-of-words approaches. We found that the stylebook improves model performance in matching. Representation visualization shows that the stylebook learns information related to context and forms, which aid input representations beyond inherent embeddings. The entrainment measures obtained from the model are highly correlated with existing bag-of-words measures. Despite the fact that using the stylebook does not aid the prediction, our neural-based measures lead to a more robust prediction model for dialogue success. To our knowledge, this piece of work is the first attempt to utilize the neural dialogue response matching model as the similarity measure for entrainment. The stylebook that we proposed to use in the model is another important novelty. Our work contributes to data-driven entrainment quantification by machine learning, which can overcome the disadvantages of bag-of-words entrainment measures.

Generating entrainment scores in a neural model paves the way to build an end-to-end automatic dialogue system with entrainment. Then in our last study, we built such an entrainable dialogue system that would address the "safe response" issues of current sequence-to-sequence response generation models. Our system incorporates entrainment degrees scored by the model introduced in the previous study into a vanilla Transformer during training, leading to a controllable response generation with entrainment. Responses generated by our model show a greater lexical similarity compared to the base model with no entrainment incorporated. Our model also outperforms vanilla Transformer and the other 2 RNN-based models in automatic evaluation metrics. However, the advantage is marginal in the human evaluation, for which our model is not optimized. Although our attempt to build such an entrainble response generation model is not completely successful, our work still contributes to providing a new potential solution to the safe response issue of S2S response generation model. Leveraging linguistic entrainment is a novel strategy that only a few previous studies have considered. Our proposed generation model uses entrainment in a post-hoc manner, so it is flexible and can be easily integrated into other Transformer-based model.

Table 19 is a summarizing of the hypotheses in this study and our conclusions:

7.2 Limitations

1. **Addressee and Addressee in dialogue**

   Datasets used in this thesis have no labels for addressee and addresser. Two essential datasets: Teams Corpus (Chapter 4 and 5) and Wikipedia Page Talk (6) are both multiparty speaker

Table 19: A summarizing of the hypotheses in this study and our conclusions

		Hypothesis	Conclusion
H1		Our bag-of-words entrainment measure can strengthen the prediction of dialogue success and failure beyond team characteristics.	Supported
H2		Our bag-of-words entrainment measure is significantly related to team characteristics, i.e., team size, ethnicity, age, and gender diversity.	Supported
H3		Leveraging global features by the stylebook will aid input representation, leading to a more robust model in matching dialogue responses.	Supported
H4		Our neural network-based measures will capture a stronger entrainment signal compared to the bag-of-words measures.	Supported
H5		Our model can generate responses with entrainment	
	H5.1	our model will achieve better PPL in *Real* compared to the baseline model that has no bias for entrainment.	Not supported
	H5.2	Compared to the baseline, our model will suffer more when we change the test set from *Real* to *Fake*	Supported
	H5.3	Our model will achieve better TF-IDF similarity compared to the baseline model that does not incorporate entrainment.	Supported
	H5.4	Our model will achieve better WMD compared to the baseline model that does not incorporate entrainment.	Supported
H6		Our model will generate more diverse responses. The overall response quality should be satisfactory considering both fluency and diversity.	
	H6.1	Incorporating entrainment in the model will lead a more diverse response generation.	Supported
	H6.2	Response diversity will vary depending on different degrees of entrainment, and there exists a linear relationship between the degree of entrainment and response diversity.	Partially Supported.

datasets. Dialogue units are interpausal units or conversational turns. Thus multiple conversation threads can be entangled in consecutive dialogue units, leading to bad coherence. Such datasets will introduce noise to our model training and thus harm model performance. This issue can be addressed by data labeling in the future.

2. **Human Evaluation**

   Wikipedia Talk Page data used in Chapter 6 is extracted from online written forums. It contains context related to various topics such as pop culture, history, and technical discussions. We found that the human evaluation is challenging to general audiences who might not have related backgrounds. In this study, we only conduct a small human evaluation as an initial attempt. Future studies should consider conducting more human evaluations.

3. **Evaluation of entrainment**

   Evaluations of entrainment measures are extrinsic in this thesis. Most entrainment studies also adopt extrinsic evaluations due to the complexity and ambiguity of directly annotating entrainment (See 5.2). Therefore, our approaches are not directly optimized for entrainment. Our methods are constrained in unsupervised approaches without guidance from annotations or labels. Extrinsic evaluations of entrainment depend on corpora. In Chapter 4 and 5, entrainment evaluations are designed specifically for Teams Corpus. We didn't include external datasets because that requires additional information such as dialogue success measured by the same strategies used in Teams Corpus.

4. **Evaluation of generated responses**

   During our study, we found the implications learned from automatic and human evaluations are not completely aligned. For example, in 6, our model (HIGH) achieves a good improvement in **Bleu** and **Distinct** compared to other models, but it performs poorly in human evaluations. We are aware of the gap between human evaluation and automatic metrics according to a previous study [59]. Due to the current limitation of model training objectives, we have to optimize our model on standard automatic metrics, which might contribute to the performance discrepancies between human and automatic evaluation.

7.3 Future Applications of Entrainable Dialogue Systems

The focus of this study is to utilize entrainment to improve the response quality of automatic dialogue systems. For this purpose, we present a series of approaches that specifically concentrate on response diversity, and some of these approaches show promising results. Dialogue systems combined with entrainment can be applied in various scenarios. Here we provide a few examples of their potential applications in the real world.

1. **Open-Domain Chatbots**

   Open-domain chatbots are not limited to a specific task. It has become a trendy research topic in recent years. Some successful implementations are XiaoIce [138] and MILABOT [96]. Researchers often train open domain chatbots with a large number of human conversations, hoping chatbots can learn and mimic how humans interact. Developing an engaging and fun open-domain conversation requires many skills beyond just building the language model. Entrainment exhibited naturally in human conversation can be used as an additional component to blend into building recipes. The present thesis is also an example of using entrainment in an open-domain chatbot.

2. **Spoken Language Understanding**

   Spoken language understanding (SLU) is crucial in spoken dialogue systems for task-specific visual assistants, such as Google Home, Apple Siri, and Amazon Alexa [38]. Automatic speech recognition (ASR) and natural language understandings (NLU) are two essential components. Recently, as the technical advance of SLU, interests in ASR and NLU are turning to multi-turn interpretations [1, 45, 119, 14]. Because entrainment is a context-dependent signal, we can consider it an additional condition in building ASR and NLU, thus improving the overall SLU performance. Entrainment in spoken features can also be utilized in ASR. There have been some successful attempts to use speech entrainment in spoken dialogue systems [54, 60].

3. **Learning Companions**

   Visual collaborative learning facilitates the knowledge understanding of students through intelligent systems [68, 100]. Research has shown that learning with visual companions can benefit the learning outcomes. Mirroring is a strategy that has been widely discussed in visual companions studies [120, 4, 12]. Essentially, linguistic entrainment is a type of linguistic mirroring. Thus, entrainable dialogue systems can be potentially employed to design more efficient visual learning companions.

8.0 Bibliography

[1] Waheed Ahmed Abro, Guilin Qi, Huan Gao, Muhammad Asif Khan, and Zafar Ali. Multi-turn intent determination for goal-oriented dialogue systems. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[2] Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. Generating stylistically consistent dialog responses with transfer learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–412, 2017.

[3] Allen C Amason and Harry J Sapienza. The effects of top management team size and interaction norms on cognitive and affective conflict. *Journal of management*, 23(4):495–516, 1997.

[4] Ivon Arroyo, Beverly Park Woolf, James M Royer, and Minghui Tai. Affective gendered learning companions. In *Artificial Intelligence in Education*, pages 41–48. IOS Press, 2009.

[5] Peter Michael Blau. *Inequality and heterogeneity: A primitive theory of social structure*, volume 7. Free Press New York, 1977.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[7] Stephanie A Borrie, Nichola Lubold, and Heather Pon-Barry. Disordered speech disrupts conversational entrainment: a study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in psychology*, 6:1187, 2015.

[8] Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25, 2000.

[9] Susan E Brennan. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96: 41–44, 1996.

[10] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482, 1996.

[11] Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. Modelling alignment for affective dialogue. In *Workshop on adapting the interaction style to affective factors at the 10th international conference on user modeling (UM-05)*, 2005.

[12] Winslow Burleson. *Affective learning companions: Strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance*. PhD thesis, Massachusetts Institute of Technology, 2006.

[13] Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. In *Empirical methods in natural language generation*, pages 85–104. Springer, 2009.

[14] Yun-Nung Chen, Ming Sun, Alexander I Rudnicky, and Anatole Gershman. Leveraging behavioral patterns of mobile applications for personalized spoken language understanding. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 83–86, 2015.

[15] Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, 1:343–359, 2007.

[16] Alexandra A Cleland and Martin J Pickering. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230, 2003.

[17] Cristian Danescu-Niculescu-Mizil. A computational approach to linguistic coordination. 2012a.

[18] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM, 2011.

[19] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM, 2012.

[20] Kenichi Yokote Makoto Iwayama Kenji Nagamatsu Dario Bertero, Takeshi Homma. Model ensembling of esim and bert for dialogue response selection. In *Proceedings of workshop of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[21] Frank RC De Wit, Lindred L Greer, and Karen A Jehn. The paradox of intragroup conflict: a meta-analysis. *Journal of Applied Psychology*, 97(2):360, 2012.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[23] Ondřej Dušek and Filip Jurčíček. A context-aware natural language generator for dialogue systems. *arXiv preprint arXiv:1608.07076*, 2016.

[24] Stanley Feldstein et al. Personality and simultaneous speech. 1974.

[25] Shaoxiong Feng, Hongshen Chen, Kan Li, and Dawei Yin. Posterior-gan: Towards informative and coherent response generation with posterior generative adversarial network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7708–7715, 2020.

[26] David M Fisher, Suzanne T Bell, Erich C Dierdorff, and James A Belohlav. Facet personality and surface-level diversity as team mental model antecedents: implications for implicit coordination. *Journal of Applied Psychology*, 97(4):825, 2012.

[27] Heather Friedberg, Diane Litman, and Susannah BF Paletz. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 404–409. IEEE, 2012.

[28] Ramiro H Gálvez, Lara Gauder, Jordi Luque, and Agustin Gravano. A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 215–224, 2020.

[29] Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. Structuring latent spaces for stylized response generation. *arXiv preprint arXiv:1909.05361*, 2019.

[30] Simon Garrod and Anthony Anderson. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218, 1987.

[31] Josette MP Gevers, Christel G Rutte, and Wendelien Van Eerde. Meeting deadlines in work groups: Implicit and explicit mechanisms. *Applied psychology*, 55(1):52–72, 2006.

[32] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wentau Yih, and Michel Galley. A knowledge-grounded neural conversation model. *arXiv preprint arXiv:1702.01932*, 2017.

[33] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19, 2010.

[34] Agustín Gravano, Štefan Beňuš, Rivka Levitan, and Julia Hirschberg. Three tobi-based measures of prosodic entrainment and their correlations with speaker engagement. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 578–583. IEEE, 2014.

[35] Richard A Guzzo, Paul R Yost, Richard J Campbell, and Gregory P Shea. Potency in groups: Articulating a construct. *British journal of social psychology*, 32(1):87–106, 1993.

[36] Benjamin Heinzerling and Michael Strube. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. *arXiv preprint arXiv:1710.02187*, 2017.

[37] Benjamin Heinzerling and Michael Strube. Bpemb: Tokenization-free pre-trained subword embeddings in 275 languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

[38] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[39] Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 151–164. Springer, 2016.

[40] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. Supervised word mover's distance. *Advances in neural information processing systems*, 29, 2016.

[41] Joseph Jaffe and Stanley Feldstein. *Rhythms of dialogue*, volume 8. Academic Press, 1970.

[42] Mahaveer Jain, John McDonough, Gahgene Gweon, Bhiksha Raj, and Carolyn Rose. An unsupervised dynamic bayesian network approach to measuring speech style accommodation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 787–797, 2012.

[43] Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*, 2015.

[44] Seiya Kawano, Masahiro Mizukami, Koichiro Yoshino, and Satoshi Nakamura. Entrainable neural conversation model based on reinforcement learning. *IEEE Access*, 2020.

[45] Young-Bum Kim, Sungjin Lee, and Ruhi Sarikaya. Speaker-sensitive dual memory networks for multi-turn slot tagging. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 541–546. IEEE, 2017.

[46] Xiang Kong, Bohan Li, Graham Neubig, Eduard Hovy, and Yiming Yang. An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. *DEEP-DIAL*, 2019.

[47] Spyros Kousidis, David Dorran, Yi Wang, Brian Vaughan, Charlie Cullen, Dermot Campbell, Ciaran McDonnell, and Eugene Coyle. Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[48] Spyros Kousidis, David Dorran, Ciaran Mcdonnell, and Eugene Coyle. Convergence in human dialoguestime series analysis of acoustic feature. In *Proceedings of SPECOM*, page 2, 2009.

[49] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[50] Rivka Levitan and Julia Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[51] Rivka Levitan, Agustín Gravano, and Julia Bell Hirschberg. Entrainment in speech preceding backchannels. 2011.

[52] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 11–19. Association for Computational Linguistics, 2012.

[53] Rivka Levitan, Stefan Benus, Agustin Gravano, and Julia Hirschberg. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposia*, 2015.

[54] Rivka Levitan, Stefan Benus, Ramiro H Gálvez, Agustín Gravano, Florencia Savoretti, Marian Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. In *Interspeech*, volume 16, pages 1166–1170, 2016.

[55] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[56] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[57] Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9220–9229, 2020.

[58] Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, 2016.

[59] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[60] José Lopes, Maxine Eskenazi, and Isabel Trancoso. Automated two-way entrainment to improve spoken dialog system performance. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8372–8376. IEEE, 2013.

[61] José Lopes, Maxine Eskenazi, and Isabel Trancoso. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112, 2015.

[62] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015.

[63] Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. Constructing interpretive spatio-temporal features for multi-turn responses selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 44–50, 2019.

[64] Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. Multi-task learning for speaker-role adaptation in neural conversation models. *arXiv preprint arXiv:1710.07388*, 2017.

[65] Nichola Lubold and Heather Pon-Barry. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12, 2014.

[66] Liangchen Luo, Jingjing Xu, Junyang Lin, Qi Zeng, and Xu Sun. An auto-encoder matching model for learning utterance-level semantic dependency in dialogue generation. *arXiv preprint arXiv:1808.08795*, 2018.

[67] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[68] Ioannis Magnisalis, Stavros Demetriadis, and Anastasios Karakostas. Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE transactions on Learning Technologies*, 4(1):5–20, 2011.

[69] Joseph D Matarazzo and Arthur N Wiens. *The interview: Research on its anatomy and structure*. Transaction Publishers, 2017.

[70] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[71] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, 2019.

[72] Arjun Mukherjee and Bing Liu. Analysis of linguistic style accommodation in online debates. *Proceedings of COLING 2012*, pages 1831–1846, 2012.

[73] Laura L Namy, Lynne C Nygaard, and Denise Sauerteig. Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21 (4):422–432, 2002.

[74] Md Nasir, Sandeep Nallan Chakravarthula, Brian Baucom, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. Modeling interpersonal linguistic coordination in conversations using word mover's distance. *arXiv preprint arXiv:1904.06002*, 2019.

[75] Md Nasir, Brian Baucom, Craig Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. Modeling vocal entrainment in conversational speech using deep unsupervised learning. *IEEE Transactions on Affective Computing*, 2020.

[76] Ani Nenkova, Agustin Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*, pages 169–172. Association for Computational Linguistics, 2008.

[77] Kate G Niederhoffer and James W Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

[78] Tong Niu and Mohit Bansal. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389, 2018.

[79] Charles A O'Reilly, David F Caldwell, and William P Barnett. Work group demography, social integration, and turnover. *Administrative science quarterly*, pages 21–37, 1989.

[80] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[81] Gabriel Parent and Maxine Eskenazi. Lexical entrainment of real users in the let's go spoken dialog system. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[82] Yookoon Park, Jaemin Cho, and Gunhee Kim. A hierarchical latent structure for variational conversation modeling. *arXiv preprint arXiv:1804.03424*, 2018.

[83] James W Pennebaker and Laura A King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[84] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[85] James W Pennebaker, Roger J Booth, and Martha E Francis. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*, 2007.

[86] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[87] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[88] Zahra Rahimi and Diane Litman. Weighting model based on group dynamics to measure convergence in multi-party dialogue. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 385–390, 2018.

[89] Zahra Rahimi and Diane Litman. Entrainment2vec: Embedding entrainment for multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8681–8688, 2020.

[90] Zahra Rahimi, Anish Kumar, Diane Litman, Susannah Paletz, and Mingzhi Yu. Entrainment in multi-party spoken dialogues at multiple linguistic levels. *Proc. Interspeech 2017*, pages 1696–1700, 2017.

[91] Radim Řehřek, Petr Sojka, et al. Gensim—statistical semantics in python. *Retrieved from genism. org*, 2011.

[92] David Reitter, Frank Keller, and Johanna D Moore. Computational modelling of structural priming in dialogue. In *Proceedings of the human language technology conference of the naacl, companion volume: Short papers*, pages 121–124, 2006.

[93] Elizabeth Rochon, Eleanor M Saffran, Rita Sloan Berndt, and Myrna F Schwartz. Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and language*, 72(3):193–218, 2000.

[94] Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[95] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[96] Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*, 2017.

[97] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron

Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*, 2016.

[98] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016.

[99] Ken G Smith, Ken A Smith, Judy D Olian, Henry P Sims Jr, Douglas P O'Bannon, and Judith A Scully. Top management team demography and process: The role of social integration and communication. *Administrative science quarterly*, pages 412–438, 1994.

[100] Amy Soller, Alejandra Martinez, Patrick Jermann, and Martin Muehlenbrock. From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. *International Journal of Artificial Intelligence in Education*, 15(4):261–290, 2005.

[101] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[102] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[103] Svetlana Stoyanchev and Amanda Stent. Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 189–192, 2009.

[104] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*, 2015.

[105] Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, 2019.

[106] Paul J Taylor and Sally Thomas. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, 1(3):263–281, 2008.

[107] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[108] Mija M Van der Wege. Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60(4):448–463, 2009.

[109] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[110] Ruth Wageman, J Richard Hackman, and Erin Lehman. Team diagnostic survey: Development of an instrument. *The Journal of Applied Behavioral Science*, 41(4):373–398, 2005.

[111] Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. Match-srnn: Modeling the recursive matching structure with spatial rnn. *arXiv preprint arXiv:1604.04378*, 2016.

[112] Shuohang Wang and Jing Jiang. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*, 2015.

[113] Shuohang Wang and Jing Jiang. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*, 2016.

[114] Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. Multilingual neural machine translation with soft decoupled encoding. *arXiv preprint arXiv:1902.03499*, 2019.

[115] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.

[116] Arthur Ward and Diane Litman. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Workshop on Speech and Language Technology in Education*, 2007.

[117] Andreas Weise and Rivka Levitan. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 297–302, 2018.

[118] Hein Wendt, Martin C Euwema, and IJ Hetty van Emmerik. Leadership and team cohesiveness across cultures. *The Leadership Quarterly*, 20(3):358–370, 2009.

[119] Yue Weng, Sai Sumanth Miryala, Chandra Khatri, Runze Wang, Huaixiu Zheng, Piero Molino, Mahdi Namazifar, Alexandros Papangelis, Hugh Williams, Franziska Bell, et al. Joint contextual modeling for asr correction and language understanding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6349–6353. IEEE, 2020.

[120] Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009.

[121] Shuangzhi Wu, Yufan Jiang, Xu Wang, Wei Miao, Zhenyu Zhao, Xie Jun, and Mu Li. Enhancing response selection with advanced context modeling and post-training. In *Proceedings of workshop of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[122] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network:

A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*, 2016.

[123] Camille J Wynn and Stephanie A Borrie. Classifying conversational entrainment of speech behavior: An updated framework and review. *PsyArXiv*, 2020.

[124] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. *arXiv preprint arXiv:1606.08340*, 2016.

[125] Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9338–9345, 2020.

[126] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64, 2016.

[127] Mingzhi Yu, Diane Litman, and Susannah Paletz. Investigating the relationship between multi-party linguistic entrainment, team characteristics and the perception of team social outcomes. In *The Thirty-Second International Flairs Conference*, 2019.

[128] Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 111–120, 2019.

[129] Peter B Zeldow and Dan P McAdams. On the comparison of tat and free speech techniques in personality assessment. *Journal of personality assessment*, 60(1):181–185, 1993.

[130] Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. Modeling topical relevance for multi-turn dialogue generation. *arXiv preprint arXiv:2009.12735*, 2020.

[131] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[132] Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*, 2019.

[133] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *ACL, system demonstration*, 2020.

[134] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*, 2017.

[135] Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. Stylized dialogue response generation using stylized unpaired texts. *arXiv preprint arXiv:2009.12719*, 2020.

[136] Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. Mechanism-aware neural machine for dialogue response generation. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[137] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*, 2017.

[138] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

[139] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view response selection for human-computer conversation. In *Pro-*

ceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 372–381, 2016.

[140] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, 2018.

Appendix A Dialogue Examples

A.1 An Example Excerpt For Teams Corpus Game 1

Engineer: Ok I’m going to

Engineer: shore up these two.

Messenger: Good move.

Engineer:Then we got one and then I guess I can also

Engineer:Can I use my powers twice in one play

Pilot: Mm

Messenger:yes

Engineer:OK well I guess yeah cause we (–)

Messenger: Well the Pilot’s limited to once per turn.

Pilot: yeah

Engineer: Ok and then I have (–) two treasure cards.

Pilot: two treasure cards

Messenger: Two Treasure cards.

Pilot: Mmhmm

Engineer: yeooh

Engineer: Ok so, let me see sorry

Messenger: uh oh

Engineer: so i move this up one tick

Messenger: and then these are gonna get shuffled

Engineer: Mmhmm

Messenger: If we had any sandbags we’d wanna use them while these were getting shuffled.

Engineer: (–) shuffle and then discard this into the treasure

Engineer: Um and I guess I have to, I’m still picking two cards I guess then.

Messenger: yeah

Pilot: Mmhmm

Messenger: That's the worst part. It starts out easier but then it starts sinking faster and faster.

Engineer: Engineer: So Iron Gate and Phanthom Rock.

Pilot: Iron Gate

Pilot: umm

Messenger: That's the Iron Gate

Pilot: Ok

Messenger: Phathom Rock

Pilot: Phathon Rock, oh um

Engineer: Aw, oh my gosh.

Messenger: Oh sinking already

Engineer: Ok

Pilot: I guess the Phathom Rock

Messenger: Alright. The Phathom Rock Card gets removed

Engineer: Oh yeah. That sucks.

Messenger: Foruntately it's not critical. No Treasures on it and it's not a path to anywhere.

Pilot: Mmhmm

Messenger: Ok that was your turn. So

Engineer: Ok

A.2 An Example For Wikipedia Talk Page

Turn 1: And I'll be along to copy-edit the blurb, thanks.

Turn 2: Consulting with the principal editor on substantive matters, and making parallel changes to the lede, with no substantive change if consensus cannot be accomplshed.–

Turn 3: You are "not" the owner of the article. Please read the text the edit box, and note [[WP:OWNERSHIP]]. You are perfectly welcome to discuss changes that are made, at the article talk page or, if concerning the TFA blurb, on user talk pages. Your aggressive behaviour last time was unacceptable.

Turn 4: I am simply repeating what Raul told you.–

Turn 5: Raul can say what he likes, but he needs consensus, and there is none for your aggressive ownership. Looks like there will be another fight coming up.

Turn 6: Raul is the featured article director, confirmed by the community. If you wish to override him, I'm not quite sure how you do it, but certainly community consensus is involved.–

A.3 An Example For Ubuntu Corpus

Turn 1: Cool also I feel like I m helping the project by using it more traffic to the site etc

Turn 2: Did you send output for when you run metasploit? Yeah I gues you are :D

Turn 3: Hhaa ok man I'll use ubuntus paste site now.

Turn 4: Yeah ill get the output error and re paste bin now gimme a sec.

Turn 5: Oh see you do n't need rvm number number you just need ruby number number – rvm is ruby s version manager. I have an idea hold on

Turn 6: sudo apt-get install ruby number number

Turn 7: Did you have it before you installed metasploit ?

Turn 8: I already got it mate.

Turn 9: Nope.

Turn 10: Don't think so can rlly remeber.

Turn 11: I'd remove metasploit make sure ruby is number number then install metasploit again – Do it in verbose mode in case it fails

Turn 12: That'll at least give us more insight into why it s failing to notice ruby if it does n't fix it altogether.

Turn 13: Ok thanks how would I remvoe it in verbose?

Turn 14: Ok thanks man what command do I use :/

Turn 15: The latest run file.

Turn 16: Try just doing it again with the run file.

Turn 17: Wait a sec metasploit hasn't worked once yet right ok will do :) nope.

A.4 An Example For Douban Corpus in Chinese with a English Translation

A.4.1 Chinese

Turn 1: 清明节去见家长了结果没有见面礼求jms分析.第一次见未来的儿媳妇就是要给钱的楼上.

Turn 2: 关键是人家父母第一次见儿子女朋友也不能就确定这是未来的儿媳妇吧,而不能确定是不是未来儿媳妇也不等于就是不喜欢这个姑娘,所以第一次见面给不给钱只能说明当地的风俗习惯而代表不了这家人真正的态度.这个对的你们的关系还未完全定下来.

Turn 3: 夷? 我们这边是第一次见面要给姑娘钱, 我不懂结果人家也没给, 我觉得反正没订婚拿不合适, 但我家人有些介意bf, 说他家是等订婚给, so我都能理解. 不过的确人家一问没有心里不舒服那是一定的.

Turn 4: 他们家算的够精的啊.

A.4.2 English

The English Translation is from Google, and further amended by me:

Turn 1: I went to see his parents on Ching Ming Festival, but I didn't receive gift money. I am asking you sisters for some opinions. Replied to the last post: they are supposed to give gifts for the first time to meet their future daughter-in-law.

Turn 2: The key is that the parents cannot be sure that this is the future daughter-in-law when they first meet their son's girlfriend, and this does not mean that they don't like the girl. Local customs can explain it, but it does not represent the true attitude of this family. This relationship between you and your boyfriend has not yet been fully settled.

Turn 3: Hm? Our local tradition for the first meeting is to give the girl money. I didn't know about this, and his parents didn't give that to me. I don't think it would be appropriate to take that money before engagement. Still, my family is a little bit worried about my boyfriend, saying that his family is waiting for the formal engagement, which is understandable. But, indeed, people do feel uncomfortable when they ask.

Turn 4: His family is so canny.

Appendix B Participants Survey (Post-Game) for Teams Corpus

Q: Please use the 1 to 5 scale to answer the questions below. Please choose the number that fits best.

	1: None (1)	2 (2)	3 (3)	4 (4)	5: A lot (5)
How frequently did you have disagreements within your group about the task you were working on? (1)	❍	❍	❍	❍	❍
How often were there disagreements about who should do what in your group? (2)	❍	❍	❍	❍	❍
How much relationship tension was there in your group? (3)	❍	❍	❍	❍	❍
How much conflict of ideas was there in your group? (4)	❍	❍	❍	❍	❍
How much conflict was there in your group about task responsibilities? (5)	❍	❍	❍	❍	❍
How often did people get angry while working in your group? (6)	❍	❍	❍	❍	❍
How often did people in your group have conflicting opinions about the task you were working on? (7)	❍	❍	❍	❍	❍
How often did you disagree about resource allocation in your group? (8)	❍	❍	❍	❍	❍
How much emotional conflict was there in your group? (9)	❍	❍	❍	❍	❍

Q: Please use the following scale to rate your agreement on each item.

	1: Highly Inaccurate (1)	2 (2)	3 (3)	4 (4)	5: Highly Accurate (5)
Working together energizes and uplifts members of our team. (1)	❍	❍	❍	❍	❍
There is a lot of unpleasantness among members of this team. (2)	❍	❍	❍	❍	❍

The longer we work together as a team, the less we do. (3)	○	○	○	○	○
Every time sometime attempts to correct a team member whose behavior is not acceptable, things seem to get worse rather than better. (4)	○	○	○	○	○
My relations with other team members are strained. (5)	○	○	○	○	○
I very much enjoy talking and working with my teammates. (6)	○	○	○	○	○
The chance to get to know my teammates is one of the best parts of working on this team. (7)	○	○	○	○	○

Q: Please use the following scale to rate your agreement on each item.

	1: Strongly disagree (1)	2 (2)	3 (3)	4 (4)	5: Strongly agree (5)
I enjoy the kind of work we do in this team. (1)	○	○	○	○	○
Working on this team is an exercise in frustration. (2)	○	○	○	○	○
Generally speaking, I am very satisfied with this team. (3)	○	○	○	○	○

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: There is not a friendly atmosphere among people. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)

○ F: There is a friendly atmosphere among people. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People in my group do not trust each other. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People in my group trust each other (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People are not warm and friendly. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People are warm and friendly. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People do not treat each other with respect. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People treat each other with respect. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People do not work well together as a team. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People work well together as a team. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People do not cooperate with each other. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People cooperate with each other. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People are not willing to share resources. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People are willing to share resources. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: People almost never speak well of the group. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: People almost always speak well of the group. (6)

Q Thinking of your team, please choose the letter A through F that best matches for each item.

○ A: The people are not proud to belong to the group. (1)
○ B (2)
○ C (3)
○ D (4)
○ E (5)
○ F: The people are proud to belong to the group. (6)

Q: Please choose the number from 1 to 5 that fits best, from 1 (to no extent) to 5 (to a great extent).

	1: To no extent (1)	2: To a limited extent (2)	3: To some extent (3)	4: To a considerable extent (4)	5: To a great extent (5)
This team has confidence in itself. (1)	❍	❍	❍	❍	❍
This team believes it can become unusually good at its tasks. (2)	❍	❍	❍	❍	❍
This team expects to be a high-performing team. (3)	❍	❍	❍	❍	❍
This team feels it can solve any problem it encounters. (4)	❍	❍	❍	❍	❍
This team believes it can be very productive. (5)	❍	❍	❍	❍	❍
This team can get a lot done when it works hard. (6)	❍	❍	❍	❍	❍
No task is too tough for this team. (7)	❍	❍	❍	❍	❍

Q: Please use the following scale to rate your agreement on each item.

	Strongly Disagree (1)	Disagree (2)	Neither Agree nor Disagree (3)	Agree (4)	Strongly Agree (5)
In my group, we have similar thoughts about the best way to proceed. (1)	❍	❍	❍	❍	❍
In my group, we eventually agree on what to do. (2)	❍	❍	❍	❍	❍
In my group, we have similar ideas about how to go about winning the game. (3)	❍	❍	❍	❍	❍

Appendix C Categories of test examples

we hand-label 130 utterances from the Teams Corpus with 13 categories of styles based on our intuition. The following is a list of these categories and their corresponding description.

- **Acknowledgment** Utterances demonstrating acknowledgment of dialogue content, such as “yeah”, “yes” and “ok”.
- **Long Text** Utterances containing at least 10 words. For example, “And then that’s all the action I’m gonna take this turn. There’s really not much else to do.”
- **Question** Utterances that are questions such as “Can I have two of those please?”
- **Containing names of tiles** Utterances that contain at least one name of tiles, a game tool in the board game of Forbidden Island. For example “Cave of Shadows which is what I’m on.”.
- **What** Utterances that are questions starting with the word “What”. For example, “So what cards do we have?”
- **Question(Long)** Utterances that are long questions consisted of more than two short questions. An example is “Both of them? These two?”.
- **Repeating** Utterances that contain at least two repeated words or phrases such as “You’re right. You’re right.”.
- **Will** Utterances that demonstrate some anticipation and plan using the word “Will”. For example, “Now I guess I’ll be collecting these!”
- **Pronoun** Utterances that contain pronouns such as “it”, “he” or “she”. A valid example is “Pilot, it’s your turn right now.”
- **Imperative sentence** Utterances that makes requests or commands. An example is “Discard the lion.”.
- **I verb** Utterances that have a structure of “I” + a verb. An example is “I think so.”
- **Negatives** Utterances that states something is false or incorrect, such as “there’s no point.”.
- **..., right?** Affirmation Utterances that ask for affirmation, and have a structure of something + “, right?”. A valid example will be “You moved and then flipped, right?”

Appendix D Wikipedia Talk Page Human Evaluation

D.1 English Proficiency Test

English Test

Please choose a grammatically correct sentence.

○ I am not edit this page with you.

○ No, you didn't conversation with the common sense of humorous reflection.

○ I've added a little more to the information on the April Fool's page.

○ Yes, and I think that lots of other characters with the Firefox.

Please choose a grammatically correct sentence.

○ I do not understand your motives, and I am frustrated.

○ That‘s one of the listing portuguese is under the creative commons.

○ I don't know, but I'll keep an eye out for anyway. Cheers!

○ Unfortunately, the only way I can find the time to do it, is there?

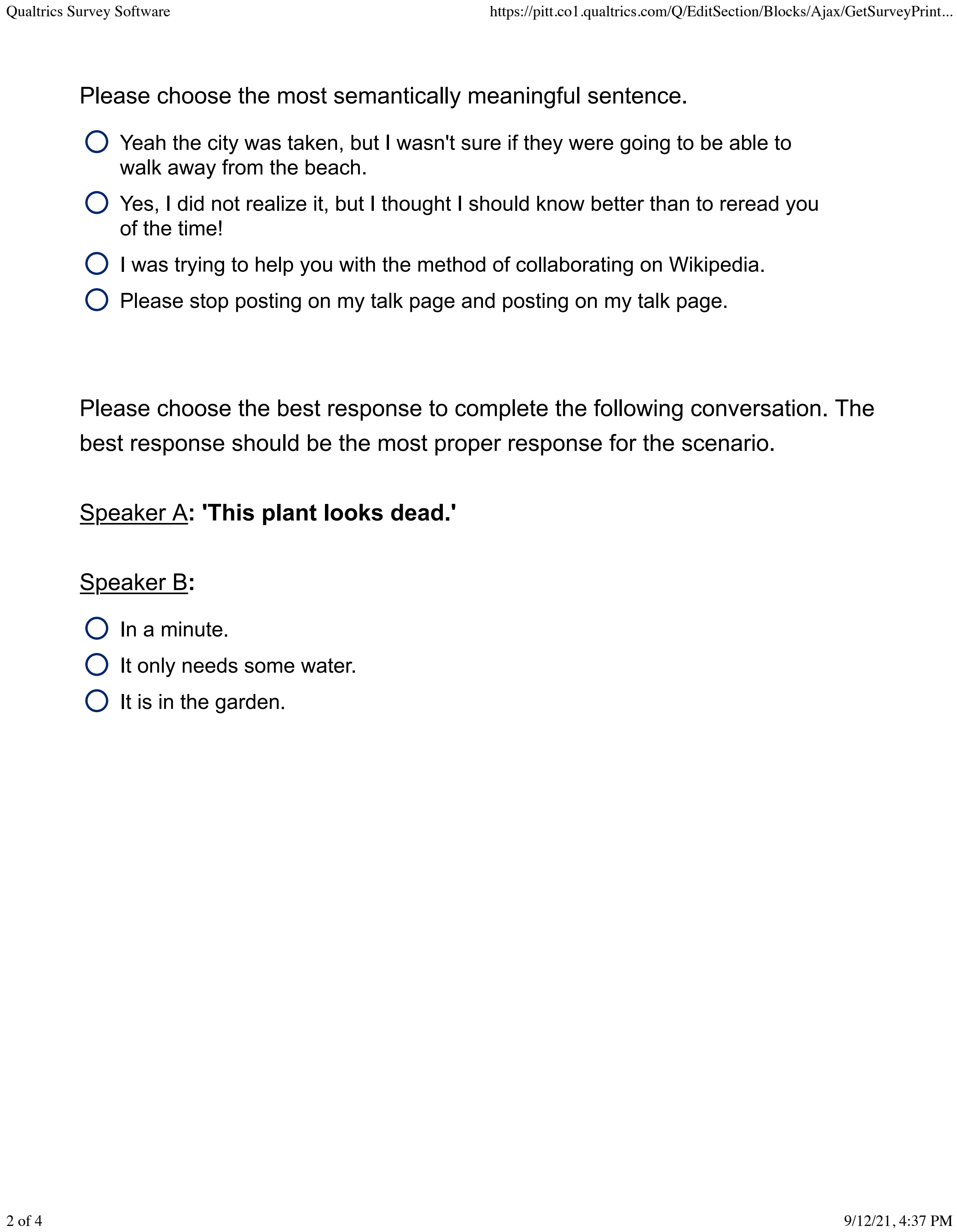
Please choose the most semantically meaningful sentence.

○ I do not speak french painting.

○ Sorry, I don't know how to do that, but I'm not sure how to do that.

○ The edit summary gave me a few days ago, so I'm going to take it.

○ I have no interest in supporting your claims. Furthermore, I have no interest in attacking Wikipedia policies and guidelines.



Please choose the most semantically meaningful sentence.

○ Yeah the city was taken, but I wasn't sure if they were going to be able to walk away from the beach.

○ Yes, I did not realize it, but I thought I should know better than to reread you of the time!

○ I was trying to help you with the method of collaborating on Wikipedia.

○ Please stop posting on my talk page and posting on my talk page.

Please choose the best response to complete the following conversation. The best response should be the most proper response for the scenario.

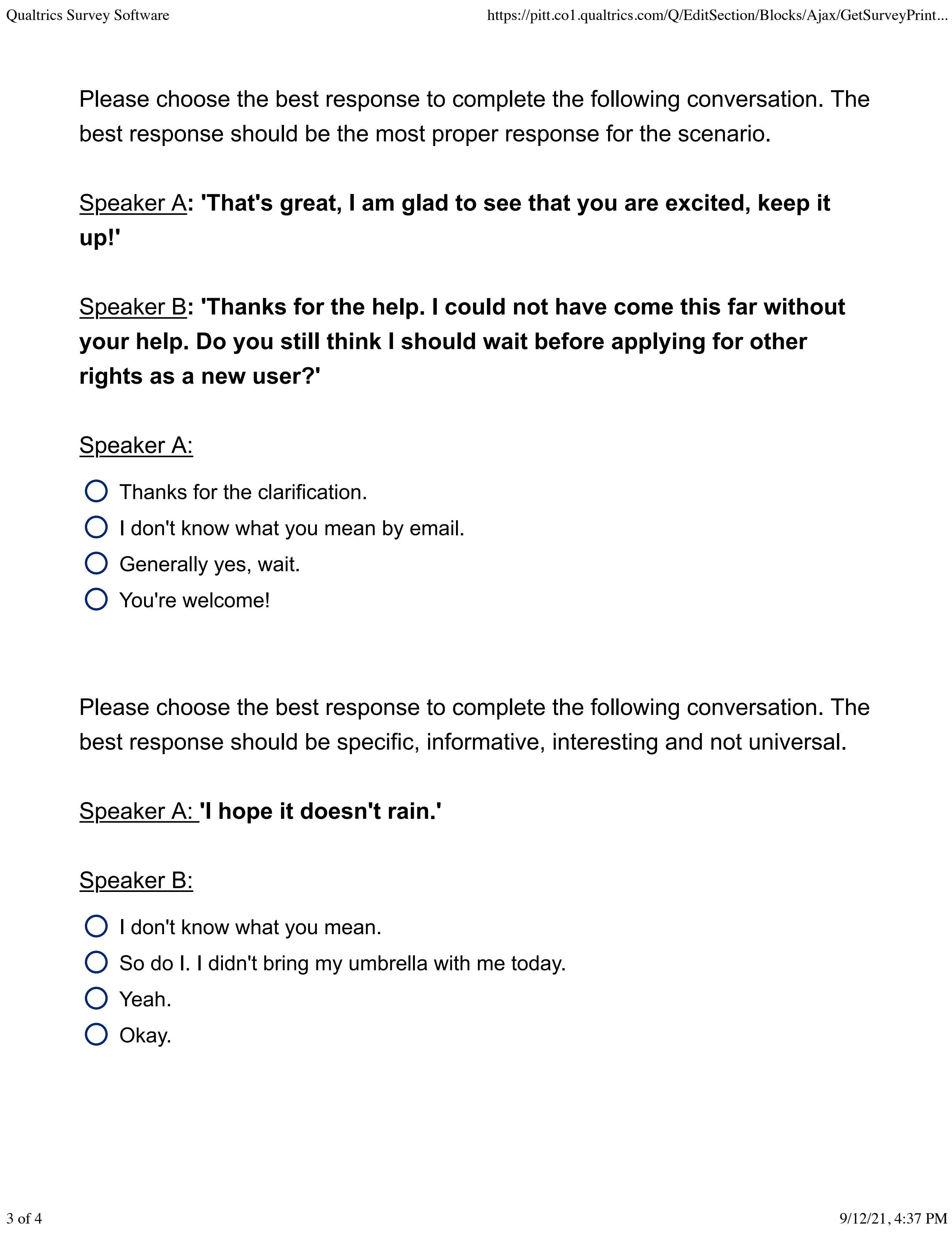
Speaker A: **'This plant looks dead.'**

Speaker B**:**

○ In a minute.

○ It only needs some water.

○ It is in the garden.



Please choose the best response to complete the following conversation. The best response should be the most proper response for the scenario.

Speaker A**: 'That's great, I am glad to see that you are excited, keep it up!'**

Speaker B**: 'Thanks for the help. I could not have come this far without your help. Do you still think I should wait before applying for other rights as a new user?'**

Speaker A:

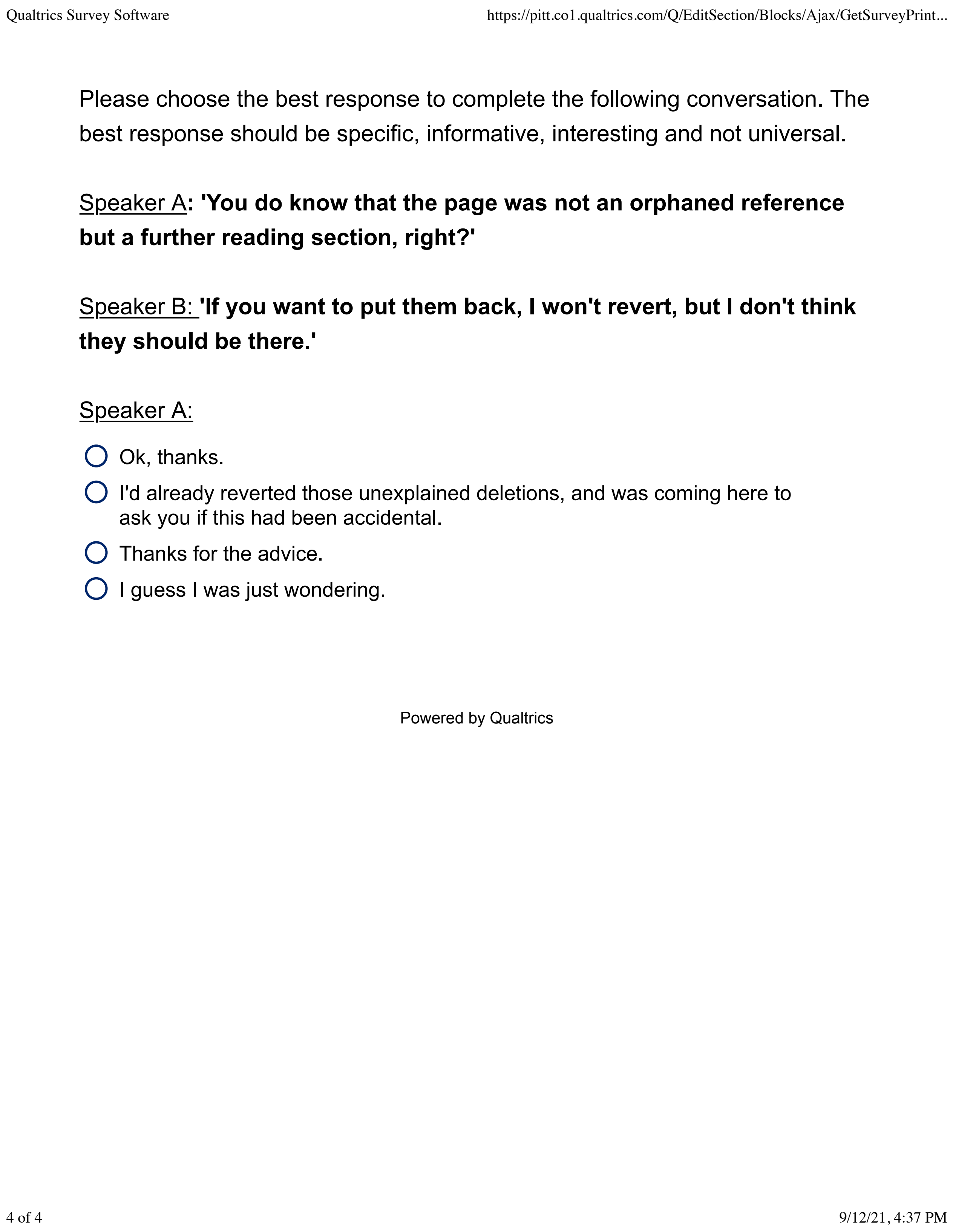
○ Thanks for the clarification.
○ I don't know what you mean by email.
○ Generally yes, wait.
○ You're welcome!

Please choose the best response to complete the following conversation. The best response should be specific, informative, interesting and not universal.

Speaker A: **'I hope it doesn't rain.'**

Speaker B:

○ I don't know what you mean.
○ So do I. I didn't bring my umbrella with me today.
○ Yeah.
○ Okay.



Please choose the best response to complete the following conversation. The best response should be specific, informative, interesting and not universal.

Speaker A**: 'You do know that the page was not an orphaned reference but a further reading section, right?'**

Speaker B: **'If you want to put them back, I won't revert, but I don't think they should be there.'**

Speaker A:

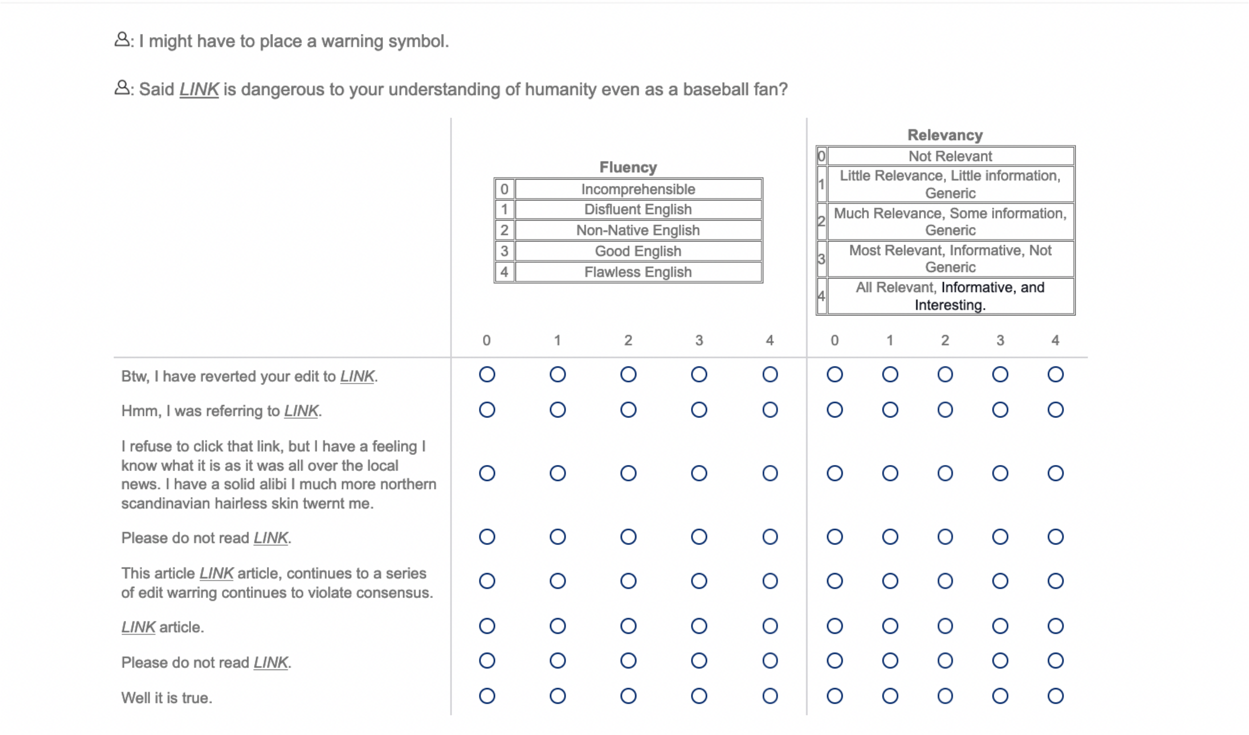
○ Ok, thanks.

○ I'd already reverted those unexplained deletions, and was coming here to ask you if this had been accidental.

○ Thanks for the advice.

○ I guess I was just wondering.

Powered by Qualtrics



👤: I might have to place a warning symbol.

👤: Said *LINK* is dangerous to your understanding of humanity even as a baseball fan?

Fluency

0	Incomprehensible
1	Disfluent English
2	Non-Native English
3	Good English
4	Flawless English

Relevancy

0	Not Relevant
1	Little Relevance, Little information, Generic
2	Much Relevance, Some information, Generic
3	Most Relevant, Informative, Not Generic
4	All Relevant, Informative, and Interesting.

	Fluency					Relevancy				
	0	1	2	3	4	0	1	2	3	4
Btw, I have reverted your edit to *LINK*.	○	○	○	○	○	○	○	○	○	○
Hmm, I was referring to *LINK*.	○	○	○	○	○	○	○	○	○	○
I refuse to click that link, but I have a feeling I know what it is as it was all over the local news. I have a solid alibi I much more northern scandinavian hairless skin twernt me.	○	○	○	○	○	○	○	○	○	○
Please do not read *LINK*.	○	○	○	○	○	○	○	○	○	○
This article *LINK* article, continues to a series of edit warring continues to violate consensus.	○	○	○	○	○	○	○	○	○	○
LINK article.	○	○	○	○	○	○	○	○	○	○
Please do not read *LINK*.	○	○	○	○	○	○	○	○	○	○
Well it is true.	○	○	○	○	○	○	○	○	○	○

D.2 User Interface for Evaluation