Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3140–3149
Marseille, 20-25 June 2022

UgChDial: A Uyghur Chat-based Dialogue Corpus for Response Space Classification

Zulipiye Yusupujiang, Jonathan Ginzburg
Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle
zulipiye.yusupujiang@linguist.univ-paris-diderot.fr, yonatan.ginzburg@u-paris.fr

Abstract

In this paper, we introduce a carefully designed and collected language resource: UgChDial – a Uyghur dialogue corpus based on a chatroom environment. The Uyghur Chat-based Dialogue Corpus (UgChDial) is divided into two parts: (1). Two-party dialogues and (2). Multi-party dialogues. We ran a series of 25, 120-minutes each, two-party chat sessions, totaling 7323 turns and 1581 question-response pairs. We created 16 different scenarios and topics to gather these two-party conversations. The multi-party conversations were compiled from chitchats in general channels as well as free chats in topic-oriented public channels, yielding 5588 unique turns and 838 question-response pairs. The initial purpose of this corpus is to study query-response pairs in Uyghur, building on an existing fine-grained response space taxonomy for English. We provide here initial annotation results on the Uyghur response space classification task using UgChDial.

Keywords: Dialogue Corpus, Response Space Classification, Uyghur, Less-Resourced Languages

1. Introduction

Characterizing the response space of questions is of great importance, as it is a critical element in developing a dialogue system that interacts in a natural way. It also provides a fundamental benchmark for dialogue/question theories. Building on detailed corpus analysis in English (the British National Corpus (Burnard, 2007) and three other more genre-specific corpora (BEE (Rosé et al., 1999), MapTask (Anderson et al., 1991), Cornell Movie (Danescu-Niculescu-Mizil and Lee, 2011))) and in Polish (Pezik, 2014), as well as on formal dialogue semantic analysis in the framework of KoS (Ginzburg, 2012; Łupkowski and Ginzburg, 2016), (Ginzburg et al., 2019; Ginzburg et al., 2022) propose that the class of responses to a question q_1 can be partitioned into three main categories:

(1) a. q(uestion)–specific: responses directly or indirectly about or subquestions of q_1;

b. MetaCommunicative: responses directly about or subquestions of a question defined in part from the *utterance* of q_1;

c. Evasion: responses directly about or subquestions of a question that is distinct from q_1 and arises from some other component of the context:

1. Ignore (address the situation, but not the question; e.g., *Anon: on the Sunday before you killed the animals, you didn't in fact feed them. Why was that? Harry: Only water.* (BNC));
2. Change the topic (e.g., *Nicola: Come on, let's get dressed. Which pants are you wearing? Oliver: What's he got on his mouth?* (BNC));
3. Motive ('Why do you ask?');
4. Difficult to provide a response ('I don't know').

The taxonomy is introduced in detail in Section 5.

In light of these studies, we aim to address the challenge of characterizing the response space to questions in a low-resource language – Uyghur. The Uyghurs are Turkic ethnic groups native to the Xinjiang Uyghur Autonomous Region in Northwest China. There are approximately 20 million Uyghurs around the world. The majority (12-15 million) live in Northwest China, and there are also large diasporic communities living in other Turkic countries such as Kazakhstan (223,100), Kyrgyzstan (60,210), Uzbekistan (55,220), and Turkey (60,000). Besides these, there are increasing number of Uyghurs in many western countries, such as the United States, Canada, Australia, France, Sweden, Netherlands, and Germany etc. [1] Uyghur is an agglutinative language with a rich morphological structure that belongs to the Turkic language group. Several studies have been conducted on Uyghur from a linguistic perspective, such as orthography, phonetics, lexical studies, morphology, syntax, and semantics. However, to our knowledge, there is no research, either theoretical or empirical, on Uyghur dialogue, let alone on the research and development of dialogue systems. Constructing a Uyghur dialogue corpus thus plays a pivotal role in establishing theoretical and empirical research on Uyghur dialogue, along with the development of Uyghur dialogue systems. Therefore, this study seeks to build a Uyghur dialogue corpus that will help address these research gaps.

[1] There has been some disputes concerning how to estimate the population of Uyghurs, see also https://en.wikipedia.org/wiki/UyghursPopulation

The current work is motivated by two main aims: (1). To investigate the response space of questions in Uyghur dialogue based on the fine-grained response space taxonomy introduced in (Ginzburg et al., 2019; Ginzburg et al., 2022), thereby enabling a comparative study with that of English and other languages. (2). To provide a high-quality Uyghur dialogue resource for developing a dialogue system for Uyghur.

To collect naturally generated conversations in Uyghur, we implemented an open-source chatroom platform on a secure server and invited native speakers to participate in the project. Obtaining a high-quality dialogue corpus to serve our purpose was challenging. One of the most crucial concerns was how to ensure that the contexts would be sufficiently varied to ensure the potential occurrence of a wide range of categories for the response space taxonomy. This requires a careful design of the collection methods, topics for discussion, and legal issues. To resolve these concerns, we collected the data in three main ways: first, we recruited native speakers (mainly Uyghurs living in Turkey, France, Germany, and Netherlands) to discuss certain topics or scenarios we created for this study in the chatroom, and collected two-party dialogues as a result. Second, we invited volunteers to a chatroom and encouraged them to have spontaneous and topically unrestricted discussions in a default public channel. This resulted in multi-party chitchat dialogues. Third, we created other public channels relating to topics such as education, language, games, politics, arts, and literature, etc. We announced a time in advance for a discussion in such topic-oriented public channels, so that native speakers who were interested could join and discuss freely with each other. This provided another source for multi-party dialogues.

The data collection process is still ongoing, and the data collected so far were mainly generated from the first two steps mentioned above. Up to now, we conducted 25 chat sessions for the two-party dialogues, and one session of the two-party dialogues is available at `https://osf.io/n24ur/` for reference. As for the multi-party chitchat dialogues, there are approximately 20 dialogues of different lengths. The third step for collecting multi-party dialogues in various topic-oriented public channels is currently being optimized.

In the following section, we give a brief overview of literature regarding the creation of dialogue corpora in English. Details of the method and design for collection process are presented in Section 3, including the recruitment process, legal concerns, and the chatroom setup. We also provide detailed statistical results on the collected corpus in Section 4. Following this, we provide the response space annotation results on the collected Uyghur dialogue corpus in Section 5. We then conclude by outlining the findings and thoughts for future studies in Section 6.

2. Related Work

A considerable amount of literature has been published on collecting dialogue datasets for building data-driven dialogue systems. Most of these studies focused on collecting English dialogue data, and applied different strategies according to the purpose of their research. In what follows, we adopt the analytic scheme proposed by Serban et al. (2015). They start by identifying the nature of the interaction: whether it is between a human and a machine, or between two humans. Since the former are not relevant for the current paper, we do not discuss them here.

Serban et al. (2015) divide human-human interaction into **spoken dialogue** and **written dialogue**. The **spoken dialogue** corpora are further broken down into three subcategories:

- *spontaneous spoken corpora*, which record spontaneous and unplanned natural spoken dialogue between participants, such as the **Switchboard dataset** (Godfrey et al., 1992). In the Switchboard corpus, participants had unrestricted conversations about a given casual topic, and it has been widely used, especially for tasks such as dialogue act modelling (Stolcke et al., 2000).

- *constrained spoken corpora*, in which participants are assigned conversational topics in advance and asked to stay on topic. For example, the **HCRC Map Task Corpus** (Anderson et al., 1991), **the Green Persuasive Dataset** (Douglas-Cowie et al., 2007), and the **Corpus of Professional Spoken American English** (CPSAE) (Barlow, 2000).

- *scripted corpora*, which are usually dialogues from movies and TV shows. The **Cornell Movie-Dialogue Corpus** (Danescu-Niculescu-Mizil and Lee, 2011), the **Movie DiC Corpus** (Banchs, 2012), and the **Corpus of American Soap Operas** (Davies, 2012) represent corpora belonging to this category.

Apart from the *scripted corpora*, most of the other spoken dialogue corpora listed above were recorded in special settings. These require a significant amount of work to transcribe and post-process the recorded data. In contrast, human-human written dialogue corpora help reduce the work on transcription, as they are often collected from micro-blogging platforms and online chatroom conversations. As above, Serban et al. (2015) also divided the **written dialogue** corpora into *spontaneous* and *constrained* written corpora. The first computer-mediated corpus was the **NPS Internet Chatroom Conversation Corpus** (Forsythand and Martell, 2007), which was built from English spontaneous conversations generated in age-specific chat rooms. Besides, the **Twitter Corpus** (Ritter et al., 2010) was built with 1.3 million post-reply pairs extracted from Twitter micro-blogging conversations, and used for unsupervised approach for modeling dialogue acts. Another

example of spontaneous written corpora is the **NUS SMS Corpus** (Chen and Kan, 2013), which involves collecting SMS messages between two users.

As for the *constrained written* corpora, the **Settlers of Catan Corpus** (Afantenos et al., 2012) is a collection of conversations resulted from 40 game sessions. This corpus has been used for modeling negotiations and strategic dialogue. Another game corpus is the **Cards corpus** (Djalali et al., 2012), a collection of conversations between two players playing the "Cards world" web-based game. Potts (2012) used this dataset to study locative question-answers pairs for identifying task dependence phenomena in real task-oriented dialogue.

Another source, similar to the method used in this study, are dialogues collected in chatroom environments. Shaikh et al. (2010) built a multi-party English chat corpus for modeling social phenomena in discourse, such as agenda control, influence, and leadership in online chat conversations. They also provided annotation of the collected dialogues with communication links, dialogue acts, local topics, and meso-topics [2]. An additional example is the **Ubuntu Dialogue Corpus** (Lowe et al., 2015), a dialogue corpus derived from the Ubuntu IRC channels logs[3], where users ask a question about a problem and other users reply in a multi-party setting. From these, the Ubuntu Dialogue Corpus extracted task-specific two-person conversations. This large dataset is applicable in developing a technical support system. The **Ubuntu Dialogue Corpus** is different from the earlier **Ubuntu Chat Corpus** (Uthus and Aha, 2013) which was collected for research on multi-participant chat analysis.

Similar to the survey conducted by Serban et al. (2015), Mahajan and Shaikh (2021) provided a collection of available English multi-party dialogue corpora, and built a taxonomy for multi-party dialogue corpora according to their source type. In addition, they propose desiderata for future data collection for developing multi-party dialogue systems. The taxonomy was divided into three big subcategories: *spoken unscripted*, *spoken scripted*, and *written*. Here we only focus on the *written* category, since we also collect *written* Uyghur dialogue in this study. The written corpora were further classified into four different subcategories: *synchronous chat*, *synchronous game*, *asynchronous forum*, and the *asynchronous microblog*. The corpus we have built for Uyghur dialogue lies within the *synchronous chat* category. Mahajan and Shaikh (2021) list 5 corpora as synchronous chat: the **NPS Internet Chatroom Conversation Corpus** (Forsythand and Martell, 2007), the **Ubuntu Dialogue Corpus** (Lowe et al., 2015), the original **Ubuntu Chat Corpus** (Uthus and Aha, 2013), the **Molweni Corpus** (Li et al., 2020), and the **MPC: Multi-party English Chat Corpus** (Shaikh et al., 2010).

All the studies reviewed here focus on collecting either two-party or multi-party dialogues, depending on their research purposes. Since we are interested in studying the response space of questions in both cases, we decided on setting up an open-source chatroom system — **Rocket.Chat**[4]— where we can collect both two-party and multi-party dialogues.

3. Data Collection Design

As mentioned earlier, the Uyghur dialogue corpus we are compiling consists of both two-party and multi-party dialogues. The two-party dialogues are conversations between two participants on a given topic, whereas the multi-party dialogues obtained so far are collections of free chitchats among multiple participants in public channels of the chatroom. At this stage, we have collected 50 hours of chat dialogues from 25 two-party chat sessions, and each session lasted for 120 minutes. We have also collected approximately 20 multi-party dialogues of free, unrestricted chitchats. In this section, we introduce the chatroom setups, recruitment process, legal concerns, and designed topics.

3.1. Chatroom Setup

We implemented **Rocket.Chat** – an open-source, fully customizable communication platform, on a secure server based in France in order to follow the European data protection regulation – GDPR[5]. This platform is remotely accessible through all devices, such as web browsers, computers, and smartphones. Participants can also access our server via the Rocket.Chat mobile application by entering the domain name specific to this project. Uyghur has three different writing systems: Uyghur Arabic-based script (UEY), Uyghur Latin-based script (ULY), and Uyghur Cyrillic script (UKY). We decided to restrict use to (UEY), and hence we implemented the Yulghun Uyghur online keyboard[6] on the message box so that the participants only type in the standard Uyghur Arabic-based script by default.

3.2. Legal Concerns

Given the sensitivity of the target group and other ethical considerations, we implemented several methods to ensure the security and anonymity of the participants. First, the server is hosted on servers physically located at the Laboratoire de Linguistique Formelle (LLF) of the University of Paris Cité, France, and the whole data collection procedures and data are protected by CNRS Data Protection Delegate[7]. The server architecture uses

[2] *meso-topics* are main topics which will persist through a number of turns and become the focus of a part of the conversation (Shaikh et al., 2010)

[3] https://irclogs.ubuntu.com/

[4] https://github.com/RocketChat/Rocket.Chat

[5] https://gdpr.eu/

[6] https://www.yulghun.com/news/vkb.html; "Yulghun" is a Uyghur noun which means "Populus euphratica". It is a special desert poplar in the Tarim Basin.

[7] https://intranet.cnrs.fr/protection_donnees/reseauxdecontact/Pages/il.aspx

an anonymizing reverse proxy to anonymize the IPs, and only the LLF's IT and Multimedia Service (SIM) team can access them for security and troubleshooting purposes. Secondly, following GDPR, a consent form is available for the participants upon registration. In this way the participants are well-informed about the aim of the project, terms, conditions, and their rights to withdraw their consent at any time. Meanwhile, we also collected demographic information such as gender, education level, and age range of participants for statistical purposes. Thirdly, participants were cautioned not to use a login that reveals their identity and not to send identifiable personal information during the chat. Finally, we manually anonymized all the demographic and personal information provided by participants. In this way, we ensure that the users do not reveal their identity, and that the collected dialogues do not contain any identifying element before publishing the final corpus for scientific use only.

3.3. Subjects

There are two kinds of participants: volunteers and recruited subjects. Volunteers participated in the free chats in public channels for collecting multi-party dialogues. In contrast, the recruited subjects were assigned to various topics for discussion or asked to perform some tasks in pairs to generate two-party dialogues. All participants are native Uyghur speakers who live in the diaspora, mostly living in Turkey, France, Germany, and Netherlands, etc. We first posted on social media platforms announcements about this project and invited volunteers to join in. Approximately 120 native speakers have registered on our platform so far. Next, we sent out a recruitment message on these platforms to recruit eligible native speakers. 55 people responded to our recruiting message. For our research, we selected 16 participants based on their language competence, communication skills, as well as typing speed in Uyghur Arabic-based script (UEY). There are 8 males and 8 females recruited and compensated for their time (10 euros per hour).

3.4. Chat Sessions

We conducted experiments for two-party dialogues and multi-party dialogues in two separate phases. For our research aim, characterizing the response space of questions, we need to collect conversations that are as natural as possible and that potentially cover the fine-grained response space taxonomy proposed by Ginzburg et al. (2019; Ginzburg et al. (2022). Therefore, we need dialogues from a broad range and on various topics. The following subsections present the details of chat sessions we have organized to date.

3.4.1. Two-party Dialogues

As an extension to the initial design of scenarios presented in our earlier work (Yusupujiang and Ginzburg, 2020), we created 16 different scenarios and topics which leads to different conversational situations. There are three main types of such topics:

- **Role-playing scenarios:** the underlying idea here is to let the participants get involved in the conversations as smoothly as possible, and most importantly, to collect various dialogues on different settings and topics. As a result, the possibility of collecting question-response pairs that further our aim will increase. There are some scenarios requiring participants to act in a controversial situation, such as *police vs. criminal*, *debtor vs. debtee*, *sales person vs. a customer with complaint*, etc. In such circumstances, people tend to have more argumentative conversations and result in more evasive responses as `change of the topic`, `ignore`, `difficult to provide answer`, etc., as well as `indirect answers`. On the other hand, we designed some cooperative or advice-giving role-playing scenarios, from which we can study the response space in friendly and cooperative circumstances. Topics include planning a vacation together, discussing children's education, advising a newly pregnant friend, conversations about exam preparation, etc. We expect to get more response types such as `direct answers`, `clarification response`, `acknowledgments`, `dependent question`, etc. from this type of situations.

- **Open discussions:** in addition to role playing, we wanted the participants to be themselves and express their opinions on the topics provided. The participants talked about their ideal society, life during the pandemic, the current Turkish economy, and some conversations about food and different cuisines. During the experiment, participants were entirely autonomous regarding their conversational style and language choices, so we encouraged them to present their true thoughts. We expect the collected conversations to be very similar to the spontaneous ones in real life. Therefore, we assigned topics according to participants' real-life situations. For instance, the topic about the Turkish economy was given to the participants from Turkey, and the topic on advising a pregnant friend was assigned to two females who have children, etc.

- **Direction giving:** we had two sessions on this highly cooperative direction giving task. In this task, participants were asked to sketch out a detailed travel plan to the current location of their partner. This task was done in two rounds so that each participant could take both roles. We expected to collect dialogues similar to that of from the **HCRC Map Task Corpus** (Anderson et al., 1991).

3.4.2. Multi-party Dialogues

The multi-party dialogues collected hitherto were obtained by a somewhat different method. First, we created several channels in our chatroom, and native speakers participated in volunteering. We have a general default channel in which users are allowed to chat on any topic at any time. That has resulted in several spontaneous conversations among participants. However, it is not avoidable to have noisy dialogues in such an environment, as multiple conversations can simultaneously occur. Secondly, we have created some topic-related public channels that participants can join in those channels of their choice. These topics include education, daily life, games, politics, literature, psychology, languages, music, arts, etc. We invited native speakers in advance through social media and asked them to chat during a specific time on channel-related topics. However, these topics were not strictly specified in advance, so participants could freely discuss them with each other. On average, there were 3-4 people online at the same time participating in such discussions.

4. Data Statistics

Several statistical analyses were used to study the characteristics of the collected Uyghur Chat-based Dialogue Corpus (UgChDial). As mentioned earlier, this is an ongoing project, and thus the statistics presented in this paper represents only a part of the final larger dataset. Table 1 shows the overall size of the corpus in terms of turns, words, and QR-pairs (Question-Response pairs). We did not count punctuations, URL links, mentions, tags, and emoticons as words, but they were included in turn counts. Besides, we calculated the number of emoticons separately. As shown in Table 1, there are two main parts of the corpus: two-party dialogues and multi-party dialogues. We have collected 7323 turns, 48796 words, 593 emoticons, and annotated 1581 QR-pairs from the two-party dialogues. The average number of words per turn is 6,66 in the two-party-dialogues.

	Two-Party	MP-Chat	MP-Topic
Total words	48796	28934	8774
Total turns	7323	4142	1446
Avg.Words/Turn	6.66	6,99	6,07
Emoticons	593	1345	212
QR-pairs	1581	620	218

Table 1: Overall size of the collected Uyghur dialogue corpus. `MP-Chat`: Multi-party Chitchat; `MP-Topic`: Multi-party Topic-oriented Dialogues

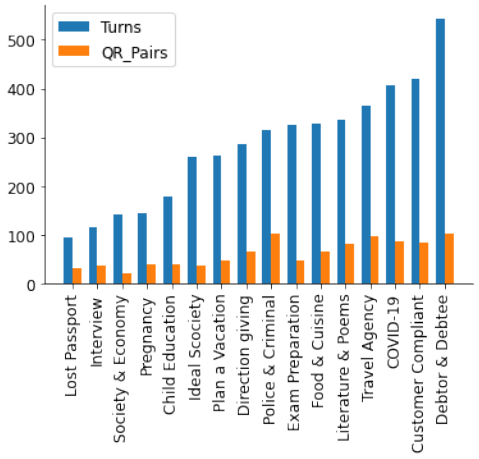
Table 1 also presents the statistical results on the collected multi-party dialogues in terms of words, turns, emoticons, question-answer pairs, and average words per turn. There are two main paths for collecting multi-party dialogues: *chitchats* and *topic-oriented dialogues*. We have collected 4142 turns from chitchats in general channels, whereas 1446 turns from dialogues in topic-oriented public channels. In addition, we collected 1345 emoticons from chitchats and 212 from topic-oriented dialogues. We have annotated 620 and 218 QR-pairs from chitchats and topic-oriented dialogues, respectively. The average number of words per turn is similar in both cases— 6,99 and 6,07.

4.1. Various Topics of Two-party Dialogues

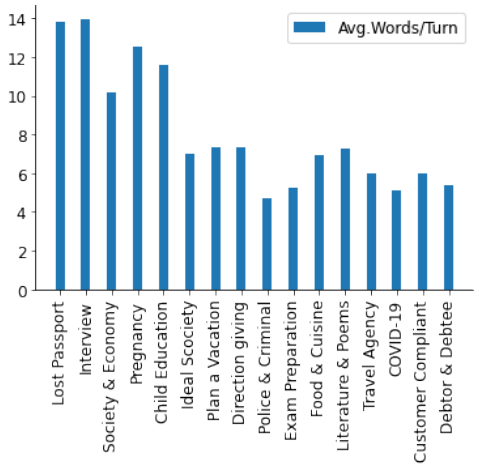
Figure 1 illustrates the statistical results for each conversational topic we have designed for collecting two-party dialogues. There are 16 unique topics used across 25 sessions, and thus some topics were used twice during our experiments. Therefore, we averaged the number of turns collected from sessions which used the same topics. As mentioned in Section 3, we divided the topics into three main sessions: *role-playing scenarios*, *open discussion*, and *direction giving*. The role-playing scenarios include `lost passport`, `interview`, `pregnancy`, `child education`, `plan a vacation`, `police & criminal`, `debtor & debtee`, `exam preparation`, `literature & poem`, `travel agency`, and `customer complaint`. The open discussion category contains topics on `society & economy`, `ideal society`, `Food & Cuisine`, and `COVID-19`. In addition, we had two sessions for the `direction giving` tasks, in which participants sketched a plan for traveling to the partner's location.
All two-party dialogue sessions were conducted for a continuous duration of 120 minutes. However, we can observe from Figure 1 that the average turns per topic, average words per turn, and the number of QR-pairs collected differ across the various scenarios. There could be many reasons for this disparity, and one of the main reasons is that some topics require participants to reflect and think well before responding in the chat. For example, in the scenario about looking for a lost passport with the help of police, the police need to ask detailed questions to help the passport owner recall how and where s/he lost it. Likewise, the passport owner also needs to describe as many situations as possible to the police. This naturally led to longer sentences in each message (about 14 words per turn) but fewer turns (less than 100 turns) in two hours. By contrast, in scenarios which require participants to act in a controversial position, such as `police & criminal` and `debtor & debtee`, participants tend to use shorter sentences or more non-sentential utterances, and often do not need much thinking time. As a result, more messages (about 400-550 turns) were produced with fewer words (around 5-6 words per turn) in each message, and we also obtained more QR-pairs from such scenarios.

4.2. Discussion on Topic Types of Two-party Dialogues

Following the above discussion, we further divided the *role-playing scenarios* and *open discus-*

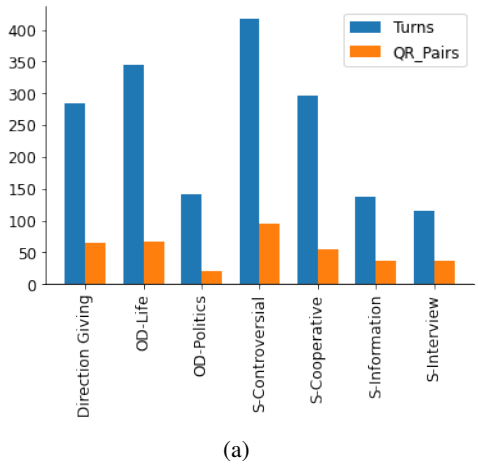


(a)

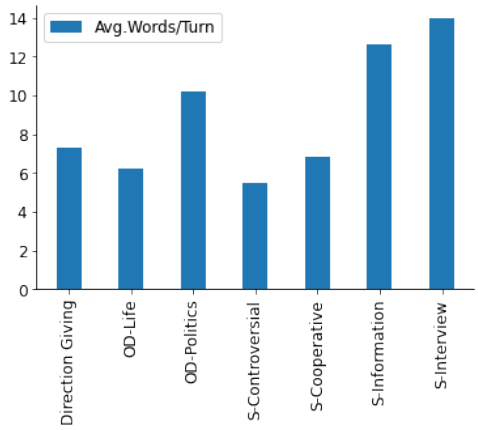


(b)

Figure 1: (a) `Average turns and QR-pairs` resulted by **each conversational topic** (b) `Average words per turn` from **each conversational topic** in two-party dialogues.



(a)



(b)

Figure 2: (a) Average turns and QR-pairs resulted **by different topic types** (b) Average words per turn **in different topic types** in two-party dialogues. The prefixes **S-** and **OD-** refer to **Scenario** and **Open Discussion**, respectively.

sions into smaller categories. As Table 3 in the Appendix shows, the *role-playing scenarios* were divided into `S-Controversial`, `S-Cooperative`, `S-Information`, and `S-Interview`. Besides, the *open discussions* were further subcategorized into `OD-Politics` and `OD-Life`. It is apparent from Figure 2 that role-playing scenarios with controversial settings generated the most turns and QR-pairs, around 420 turns and 100 QR-pairs in average. However, scenarios which aim at providing information or interviewing resulted in fewer turns, only about 115-140 turns and 30-35 QR-pairs in average. We can also observe from the figure that open discussions on more formal topics such as politics, economics, etc., resulted in the fewest QR-pairs, only around 20 QR-pairs in average. Furthermore, the direction giving tasks, open discussions on general daily life-related topics, and also cooperative scenario topics generated similar number of turns and QR-pairs, approximately 290-350 turns and 55-65 QR-pairs in average. By comparing the results in Figure 2 (a) and (b), we see that the number of turns is inversely correlated with the average number of words per turn.

5. Annotations

One of the purposes of building the Uyghur chat-based dialogue corpus is to study the response space of questions in dialogue. As explained in the introduction, we can respond to a question in different ways. Ginzburg et al. (2022) created a full taxonomy for response space with 9 unique classes and one OTHER class. Table 4 presents this full response space taxonomy, among which, the first three classes: Direct answer (DA), Indirect answer (IND), and Dependent question (DP) are responses specific to the initial question. Furthermore, Clarification response (CR) and Acknowledgement (ACK) are meta-communicative responses. However, Motivation, Ignore, Change the topic (CHT), and

the Difficult to provide an answer (DPR) are evasion responses.

Table 2 shows that we have collected 2419 question-answer pairs from the current UgChDial corpus, among which 1581 QR-pairs were collected in two-party dialogues, and 838 QR-pairs were from multi-party conversations. We have subsequently annotated the responses in each QR-pairs based on the fine-grained response space taxonomy developed in (Ginzburg et al., 2022) and presented in Table 4. We provide the overall distribution of response classes in the UgChDial corpus in Table 2. All response classes were annotated by the first author, a native Uyghur speaker who has good experience in response class annotation.

- **Two-party dialogues:** we can learn from Table 2 that the fine-grained response space taxonomy covered 99.7% of all two-party dialogues in UgChDial corpus, as only 0.3% of responses were classified as OTHER. The most frequent response type is the direct answer, which takes up 52% of the responses. The next biggest response class is the indirect answers (IND=19%), followed by the third most common class, change the topics (CHT=10%). Besides, the less frequent response classes are MOTIV, DP, CR, and ACK.

- **Multi-party dialogues:** the overall distribution of response classes in multi-party dialogues is similar to the two-party dialogues. The first two most frequent classes are direct answers (DA=55%) and indirect answers (IND=27%). However, there are less CHT class in multi-party dialogues than in the two-party dialogues (5% vs. 10%).

For comparison, we have also included the response space distribution in the British National Corpus (BNC) reported by Ginzburg et al. (2022). Table 2 shows that the overall trend of the distributions in the BNC and UgChDial response space seem broadly similar: the direct answers account for more than 50% of the overall response types in both corpora. On the other hand, some marked differences exist, regarding the frequency of indirect answers (IND), Clarification Response (CR), and Change Topic utterances (CHT). For now, we reserve judgement as to the source of these differences, given that the corpora differ in medium (spoken v. chat) and in collection methods (BNC—largely from real-life situations, UgChDial—collected in a chatroom environment by using carefully designed scenarios and topics.). We hope to collect comparable English data, which will allow for a more systematic comparison.

Inter-annotator agreement: To examine the reliability of the annotation, we invited another native Uyghur speaker to annotate one of the two-party dialogue sessions. This annotator underwent several training sessions on response-type annotation with the first author. Synchronous chat-based conversations, such as UgChDial, often lead to interlaced conversations, as participants tend to be typing at the same time. This makes it difficult for annotators to identify question-response pairs. Therefore, the main annotator manually added the turn ID of the corresponding question of each response in the conversation. This facilitated the annotation process for new annotators. There are 131 QR-pairs from this double annotated two-party dialogue session. The inter-annotator reliability Cohen's κ score and Krippendorff's $alpha$ score between two annotators is 0.7464 and 0.7461 respectively.

	Two-party	**Multi-party**	**BNC**
DA	52% (816)	55% (457)	64.1% (393)
IND	19% (303)	27% (223)	9.8% (60)
DP	1% (19)	0.6% (5)	1.3% (8)
CR	3% (47)	1.1% (9)	7% (43)
ACK	1.5% (23)	2% (19)	3.1% (19)
CHT	10% (165)	5% (46)	2.3% (14)
MOTIV	1% (17)	0.5% (4)	0.3% (2)
IGNORE	7% (112)	5% (46)	4.2% (26)
DPR	5% (74)	2.6% (23)	7.3% (45)
OTHER	0.3% (5)	0.7% (6)	0.5% (3)
Total	1581	838	613

Table 2: Distribution of response classes in UgChDial corpus comparing to the data for BNC corpus reported in (Ginzburg et al., 2022)

5.1. Coarser Response Space Taxonomy and Topic Types

We noted in Section 3 that we created 16 different scenarios and topics to collect various response classes. In addition, we further divided these topics into 7 subcategories based on their types, as shown in Table 3. Thus, we are interested in the distribution of the different response classes across the 7 different topic types. What's more, we are interested in studying the classification with a coarser taxonomy with only 4 distinct response classes, namely, Direct Answer, Indirect Answer, Clarification Response, and Evasion. All response classes which belong to neither Direct Answer, Indirect Answer, nor Clarification Response, were merged and classified as Evasion. We think that this is a particularly practical taxonomy for dialogue system design while retaining a modicum of semantic richness.

Figure 3 illustrates the distribution of each response category in the coarser response space taxonomy across seven different topic types. As presented in the figure, the direct answers accounted for 60%-70% of the total responses in conversational topics such as direction giving, open discussion about daily life, cooperative scenarios, and information providing scenarios. The common characteristic of such topics is that participants need to provide as precise information as possible, and conversations often occurred in a friendly and cooperative mood. That leads to more direct answers

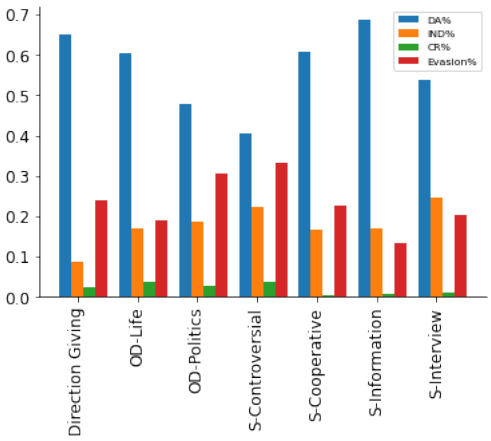


Figure 3: Distribution of the four main classes (DA, IND, CR, Evasion) in the coarser taxonomy across different topic types. The prefixes **S-** and **OD-** refer to **Scenario** and **Open Discussion**, respectively.

and fewer evasive responses (approximately 13%-25%) to questions. The proportion of the indirect answers is also relatively low in such topics, only around 8%-17%. By contrast, the direct answers constituted only 40% of the total responses in the `S-Controversial` category. In such cases, participants tend to chat in a more argumentative and unfriendly mood due to the controversial characteristics of such topics. As a result, we collected more evasive responses (around 35%) from such scenarios. Another interesting result came from the interview scenario, in which one speaker played the role of a minister, and the other played the role of a journalist. In which, indirect responses constitute about 25% of the total responses, which is highest among all topics. As for open discussions about politics or economics, the composition of direct answers is below 50%, but indirect answers (19%) and evasive responses (30%) are higher compared to the topics of collaboration, direction giving, and providing information. In addition, the graph also shows that clarification response accounted for a significantly lower percentage of responses to all topic types.

6. Conclusions and Future Work

The main goal of the current study was to build a Uyghur dialogue corpus with the aim of studying the response space of questions in dialogue. We deployed and customized an open source chatroom system - Rocket.Chat to collect Uyghur chat-based dialogue data. As it is an ongoing project, we presented information on the data gathered so far in this paper. There are two main parts of the Uyghur Chat-based Dialogue Corpus (UgChDial): (1). Two-party dialogues, and (2). Multi-party dialogues. We conducted a series of 25 two-party chat sessions of 120 minutes each, which amounted to 7323 turns and 1581 question-response pairs. We created 16 unique scenarios and topics for collecting these two-party dialogues. The multi-party dialogues were collected from chitchats in general channels as well as from free chats in topic-oriented public channels, yielding 5588 unique turns and 838 question-response pairs for multi-party dialogues. A sample of the corpus is available at `https://osf.io/n24ur/` for reference. In addition, we annotated the responses to questions based on the fine-grained response space taxonomy, and discussed the annotation results in Section 5.

The major contributions of this study are three-fold: (1). To our knowledge, this study presents the first dialogue corpus for Uyghur. Thus, it lays the groundwork for future research into Uyghur dialogue studies, and provides a language resource for developing dialogue systems for Uyghur; (2). The data collection methods and the conversational topics and scenarios created for collecting two-party dialogues are replicable for building dialogue corpora for other languages. Therefore, this study provides useful insights for constructing dialogue corpora for other low-resource languages; (3). This paper presents detailed statistics and analysis on the response space classification of Uyghur dialogues, which lays the foundation for future comparative studies on the characterization of response space across languages. In addition, we demonstrated our findings about the relations between different types of topics and the distribution of response classes.

A natural progression of this study is to optimize the data collection of multi-party conversations and create more diverse topics and scenarios for collecting more two-party Uyghur dialogues. Once the collection and annotation process is complete, the whole corpus will be available to the research community. In addition, we intend to apply the same methodology and scenario topics to collect two-party conversations in English. That will allow us to conduct a comparative study of the response spaces in Uyghur and English chat dialogues.

7. Acknowledgements

We acknowledge the support by a public grant overseen by the French National Research Agency (ANR) as part of the program *Investissements d'Avenir* (reference: ANR-10-LABX-0083). It contributes to the IdEx Université Paris Cité- ANR-18-IDEX-0001. In addition, we would like to thank Achille Falaise, Paruke Litifu, Doriane Gras, Loïc Liégeois, and Alexandre Roulois for their technical support and thoughtful suggestions.

Appendix

8. Bibliographical References

Afantenos, S., Asher, N., Benamara, F., Cadilhac, A., Dégremont, C., Denis, P., Guhe, M., Keizer, S., Lascarides, A., Lemon, O., et al. (2012). Developing a corpus of strategic conversation in the settlers of

Topic Type	Topic (number of sessions)
S-Controversial	Police-Criminal(2);
	Customer Complaint(2);
	Debtor&Bebtee(2);
	Travel Agency(1)
S-Cooperative	Plan a vacation(2);
	Literature-Poems(1);
	Exam Preparation(1)
S-Information	Child Education(2);
	Lost Passport(1);
	Pregnancy(1)
S-Interview	Interview(2)
OD-Life	COVID-19(2);
	Food-Cuisine(2);
	Ideal Society(1)
OD-Politics	Society-Economy
Direction Giving	Direction giving(2)

Table 3: Grouping of two-party dialogue topics by topic type

catan. In *SeineDial 2012-The 16th Workshop On The Semantics and Pragmatics Of Dialogue*.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Banchs, R. E. (2012). Movie-dic: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–207.

Barlow, M. (2000). *Corpus of Spoken, Professional American-English*. Rice University.

Lou Burnard, editor. (2007). *Reference guide for the British National Corpus (XML Edition)*. Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk/XMLedition/URG/`.

Chen, T. and Kan, M.-Y. (2013). Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*, 47(2):299–335.

Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

Davies, M. (2012). The corpus of american soap operas: 100 million words, 2001–2012.

Djalali, A., Lauer, S., and Potts, C. (2012). Corpus evidence for preference-driven interpretation. In *Logic, Language and Meaning*, pages 150–159. Springer.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., et al. (2007). The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer.

Forsythand, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

Ginzburg, J., Yusupujiang, Z., Li, C., Ren, K., and Łupkowski, P. (2019). Characterizing the response space of questions: a corpus study for english and polish. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 320–330.

Ginzburg, J., Yusupujiang, Z., Li, C., Ren, K., Kucharska, A., and Łupkowski, P. (2022). Characterizing the response space of questions: data and theory. *Dialogue and Discourse (under review)*. `https://drive.google.com/file/d/1AieL7JERQhJnTPlbgn1P_YPDaLP8gGJl/view`.

Ginzburg, J. (2012). *The interactive stance*. Oxford University Press.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Li, J., Liu, M., Kan, M.-Y., Zheng, Z., Wang, Z., Lei, W., Liu, T., and Qin, B. (2020). Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Łupkowski, P. and Ginzburg, J. (2016). Query responses. *Journal of Language Modelling Vol*, 4(2):245–292.

Mahajan, K. and Shaikh, S. (2021). On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352.

Potts, C. (2012). Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th west coast conference on formal linguistics*, pages 1–20. Cascadilla Proceedings Project.

Pezik, P. (2014). Spokes search engine for Polish conversational data. CLARIN-PL digital repository.

Ritter, A., Cherry, C., and Dolan, W. B. (2010). Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Rosé, C. P., Eugenio, B. D., and Moore, J. D. (1999). A dialogue-based tutoring system for basic electricity and electronics. In Susanne P. Lajoie et al., edi-

Full-Taxonomy	Description
Direct Answer **(DA)**	the response directly offers an answer to the question.
Indirect Answer **(IND)**	the answer to the question can be indirectly inferred from this utterance.
Dependent questions **(DP)**	the answer to the original question depends on the answer to this query response.
Clarification Response **(CR)**	the speaker asks for extra information to confirm (s)he understood the question correctly, requires additional information to understand it better, or provides some information to clarify/correct misinformation from the previous utterance.
Acknowledgement **(ACK)**	the speaker acknowledges that (s)he heard the question, such as mhm, aha,... etc.
Motivation **(MOTIV)**	a query response about the motivation of asking the initial question.
Ignore **(IGNORE)**	Ignore: the utterance does not relate to the question, but to the situation.
Change the topic **(CHT)**	the utterance signals that the speaker does not want to answer the question, instead (s)he changes the topic, and gives an evasive response.
Difficult to provide an answer **(DPR)**	the speaker indicates that (s)he does not know the answer, or it is difficult for her/him to provide an answer, so points at a different information source,
OTHER	Utterance that does not fit in any of the categories above.

Table 4: Full Response Space Taxonomy proposed in (Ginzburg et al., 2022) and used in this paper

tors, *Artificial intelligence in education*, pages 759–761. IOS, Amsterdam.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Shaikh, S., Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Taylor, S. M., and Webb, N. (2010). Mpc: A multi-party chat corpus for modeling social phenomena in discourse. In *LREC*.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Uthus, D. C. and Aha, D. W. (2013). The ubuntu chat corpus for multiparticipant chat analysis. In *2013 AAAI Spring Symposium Series*.

Yusupujiang, Z. and Ginzburg, J. (2020). Designing a gwap for collecting naturally produced dialogues for low resourced languages. In *Workshop on Games and Natural Language Processing*, pages 44–48.