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

Zulipiye Yusupujiang Jonathan Ginzburg

Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle

Abstract

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, and one session of the two-party dialogues is available at `https://osf.io/n24ur/` for reference. 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.

Why this study?

1. **Motivation**: This work is motivated by two main aims:
   - To investigate the response space of questions in Uyghur dialogue based on the fine-grained response space taxonomy introduced in [2, 1], thereby enabling a comparative study with that of English and other languages.
   - To provide a high-quality Uyghur dialogue resource for developing a dialogue system for Uyghur.
2. **Importance**: Constructing a Uyghur dialogue corpus plays a pivotal role in establishing theoretical and empirical research on Uyghur dialogue, along with the development of Uyghur dialogue systems.

Introduction to 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. Here are the main distribution of Uyghur population:

- Xinjiang Uyghur Autonomous Region in Northwest China (12-15 million)
- Kazakhstan (223,100)
- Kyrgyzstan (60,210)
- Uzbekistan (55,220)
- Turkey (60,000)
- Western countries such as the United States, United Kingdom, Canada, Australia, France, Sweden, Germany, and Netherlands, etc.

Uyghur is an agglutinative language with a rich morphological structure that belongs to the Turkic language group. The typical word order in a Uyghur sentence is Subject-Object-Verb (SOV), e.g., *Men Uyghurche Oquymen.* "I Uyghur Study." Uyghur has three different writing systems: (1). Uyghur Arabic-based script **(UEY)**, (2). Uyghur Latin-based script **(ULY)**, and (3). Uyghur Cyrillic script **(UKY)**.

Data Collection Design

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.
2. **Legal Concerns**:
   - the server is hosted on servers physically located at the Laboratoire de Linguistique Formelle (LLF) of the University Paris Cité, France, and the whole data collection procedures and data are protected by CNRS Data Protection Delegate.
   - following GDPR, a consent form is available for the participants upon registration.
   - participants were cautioned not to use a login that reveals their identity and not to send identifiable personal information during the chat.
   - all the demographic and personal information provided by participants were manually anonymized if there is any.
3. **Subjects**: All participants (approximately, 120 registered users) are native Uyghur speakers who live in the diaspora, mostly living in Turkey, France, Germany, and Netherlands, etc. There are two kinds of participants: volunteers and recruited subjects (8 males, 8 females). The recruited subjects were compensated for their time.

Chat Sessions: Two-party Dialogues

As an extension to the initial design of scenarios presented in our earlier work [3], We created 16 different scenarios and topics for two-party dialogues. There are three main types of such topics:

- **Role-playing scenarios:** the aim 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, including both controversial and cooperative scenarios.
- **Open discussions:** in addition to role playing, we wanted the participants to be themselves and express their opinions on the topics provided. 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.
- **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.

Chat Sessions: Multi-party Dialogues

There are also two different ways of collecting multi-party dialogues, 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.
- 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, etc. We invited native speakers in advance through social media, and asked them to chat during a specific time on channel-related topics (on average, 3-4 people online at the same time).

Data Statistics

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.

	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

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.

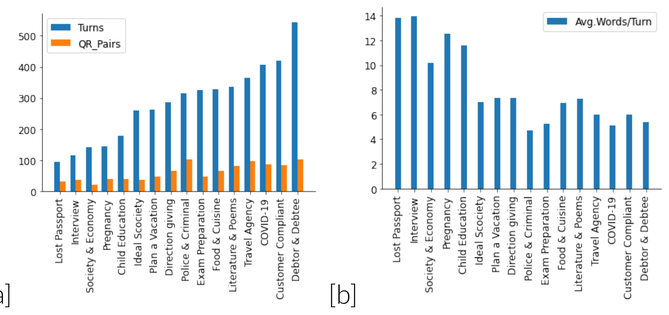

Figure 1. (a) `Average turns and QR-pairs` resulted by **each conversational topic** (b) `Average words per turn` from **each conversational topic** in two-party dialogues.

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. By contrast, some scenarios 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.

Discussion on Topic Types of Two-party Dialogues

Following the above discussion, we further divided the *role-playing scenarios* and *open discussions* into smaller categories. As Table 2 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`.

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 2. Grouping of two-party dialogue topics by topic type

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.

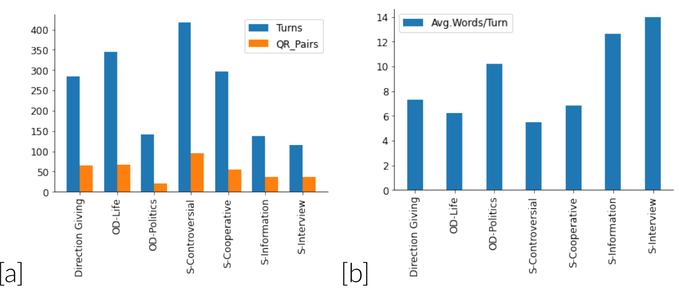

Figure 2. (a) Average turns and QR-pairs resulted **by different topic types** (b) Average words per turn **in different topic types** in two-party dialogues. The prefixes **S-** and **OD-** refer to **Scenario** and **Open Discussion**, respectively.

Response Space Annotation

Response Space Annotation Results of UgChDial Corpus

Table 3 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.

	Two-party	Multi-party	BNC
Direct Answer	52% (816)	55% (457)	64.1% (393)
Indirect Answer	19% (303)	27% (223)	9.8% (60)
Dependent Question	1% (19)	0.6% (5)	1.3% (8)
Clarification Response	3% (47)	1.1% (9)	7% (43)
Acknowledgement	1.5% (23)	2% (19)	3.1% (19)
Change the Topic	10% (165)	5% (46)	2.3% (14)
Motivation	1% (17)	0.5% (4)	0.3% (2)
Ignore	7% (112)	5% (46)	4.2% (26)
Difficult to Provide An Answer	5% (74)	2.6% (23)	7.3% (45)
OTHER	0.3% (5)	0.7% (6)	0.5% (3)
Total	1581	838	613

Table 3. Distribution of response classes in UgChDial corpus comparing to the data for BNC corpus reported in [1]

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. 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.

Coarser Response Space Taxonomy and Topic Types

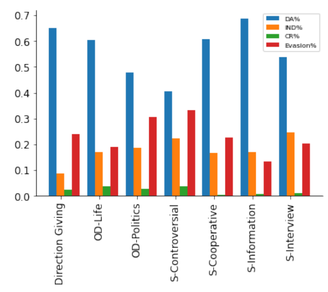

Figure 3. Distribution of the four main classes (DA, IND, CR, Evasion) in the coarser taxonomy across different topic types. The prefixes **S-** and **OD-** refer to **Scenario** and **Open Discussion**, respectively.

Contribution

The major contributions of this study are three-fold:

- We presented the first freely available 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;
- 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.
- The detailed statistics and analysis on the response space classification of Uyghur dialogues lays the foundation for future comparative studies on the characterization of response space across languages.

References

[1] Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, Aleksandra Kucharska, and Paweł Łupkowski. Characterizing the response space of questions: data and theory. *Dialogue and Discourse (under review)*, 2022. `https://drive.google.com/file/d/1AieL7JERQhJnTPlbgn1P_YPDaLP8gGJl/view`.

[2] Jonathan Ginzburg, Zulipiye Yusupujiang, Chuyuan Li, Kexin Ren, and Paweł Łupkowski. 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, 2019.

[3] Zulipiye Yusupujiang and Jonathan Ginzburg. Designing a gwap for collecting naturally produced dialogues for low resourced languages. In *Workshop on Games and Natural Language Processing*, pages 44–48, 2020.

zulipiye.yusupujiang@linguist.univ-paris-diderot.fr yonatan.ginzburg@u-paris.fr