On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods

Khyati Mahajan and Samira Shaikh
Department of Computer Science
University of North Carolina at Charlotte
kmahaja2,samirashaikh@uncc.edu

Abstract

We present a comprehensive survey of available corpora for multi-party dialogue. We survey over 300 publications related to multi-party dialogue and catalogue all available corpora in a novel taxonomy. We analyze methods of data collection for multi-party dialogue corpora and identify several lacunae in existing data collection approaches used to collect such dialogue. We present this survey, the first survey to focus exclusively on multi-party dialogue corpora, to motivate research in this area. Through our discussion of existing data collection methods, we identify desiderata and guiding principles for multi-party data collection to contribute further towards advancing this area of dialogue research.

1 Introduction

To say research in conversational agents and natural language generation has seen an explosive growth in recent years would be an understatement, as evidenced by the increasing number of papers published on this topic. However, most current research in this area has focused on two-party or dyadic conversations. This focus is important, since many open questions remain with dialogue systems in dyadic settings, such as modeling long-term dialogue context modeling and infusion of knowledge, persona and empathy (Li et al., 2016; Hedayatnia et al., 2020; Liu et al., 2020)

Nevertheless, there is still a pressing need to focus on more naturally occurring conversations which consist of more than two speakers (Kirchhoff and Ostendorf, 2003), also known as *multi-party dialogue*. Humans naturally tend to work in groups and teams. Conversational agents capable of working in multi-party dialogue situations stand to advance the future of work, since they can be integrated into teams, e.g., in surgery, search and rescue, or manufacturing and design. The settings for such agents could be informal (e.g. chatroom assistants) or formal (e.g. meeting assistants) settings. Particularly, with conversational assistants such as Amazon Alexa, there is a push to develop AI to understand multiple users and act as teammates (Winkler et al., 2019; Seeber et al., 2020).

At the same time, methods and models built for two-party cannot simply be generalized for multi-party conversations. Some challenges that are unique to multi-party dialogue include speaker identification (figuring out who is speaking), turn-taking (understanding whether to respond or not) and tailoring the content of the response to each agent or person (Sibun, 1997).

Several of these challenges can be approached through data-driven methods (Hawes et al., 2009; de Bayser et al., 2019). Given that corpora are the currency for data-driven methods, and facilitate further research on building data-driven multi-party dialogue systems, we present this systematic survey of existing corpora for multi-party dialogue. We describe how these corpora (Section 3) were collected (Section 4) along with the tasks that are undertaken on these corpora. Our key goal is to identify desiderata that could help guide data collection efforts towards making research in multi-party dialogue more mature (Section 5).

Our survey follows prior efforts in systematic reviews of dialogue corpora (Serban et al., 2018), evaluation of chatbots (Venkatesh et al., 2018; Deriu et al., 2020), and NLG evaluation (Howcroft et al., 2020). Gatt and Krahmer (2018) provide a meticulous survey of the state-of-the-art in Natural Language Generation, however they do not include a separate discussion on corpora. The systematic review of dialogue corpora conducted by Serban et al. (2018) does not primarily focus on multi-party corpora. Deriu et al. (2021) provide a systematic survey on the evaluation of dialogue systems, which includes a section of datasets and

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 338–352
July 29–31, 2021. ©2021 Association for Computational Linguistics

benchmarks, but again the focus is not primarily towards multi-party dialogue systems. Consequently, the goal of this article is to make the following contributions:

- presenting a comprehensive listing of a large number of available multi-party dialogue corpora, and organize these into a taxonomy. To accomplish this goal, we start from a collection of over 300 published papers.
- presenting a detailed overview of data collection methods for multi-party dialogue, especially the need for specialized equipment and environments.
- providing recommendations for collecting new useful datasets, to advance research in this area.

Our intent is that with an up-to-date synthesis of available resources, and by drawing attention to the challenges particular to multi-party dialogue, we can provide insights of exploiting recent data-driven techniques to address these challenges.

2 Method

Selection Criteria: Similar to recent work in systematic review of relevant literature, we followed the PRISMA method to identify, screen and include articles for this survey (Howcroft et al., 2020; Reiter, 2018). We searched Google Scholar and Semantic Scholar for the keywords *multi-party dialogue* and variations thereof (e.g., *multi-party, multiparty conversation*). We began by considering all papers that appeared in conferences and journals which focus on NLP and NLG, including all × CL venues as well as AI conferences and venues (e.g., AAAI, IJCNLP, Interspeech). We then iterated through the references and citations of these papers, and included any relevant articles that were missed through keyword search. This identification step resulted in 362 papers overall.

As part of our screening process, we limit the discussion to corpora that (a) have already been used in existing research in conversational systems; (b) which have a text component, and focus on the English language; and (c) which include *multiple speakers in the majority of conversations*, finally resulting in 343 papers. We release our annotated references to the 343 papers on Github[1]. Unsurprisingly, we found that majority of corpora papers were published in LREC and SIGDIAL venues, in addition to *ACL venues.

Organizing corpora by genre: Next, we organized all included corpora into a new taxonomy (Figure 1). Corpora are first categorized by whether they include *Spoken* or *Written* dialogue. Spoken corpora are further divided as *unscripted* vs. *scripted*. Within these type-based divisions, the corpora are then arranged by their main sources. The *unscripted spoken corpora* are thus arranged into 4 main categories - *informal discourse* mainly consisting of informal interactions such as radio talk shows, *formal discourse* mainly consisting of formal interactions such as debates, *spontaneous speech* mainly consisting of spontaneous interactions such as teenage talk, and *meetings and interviews* mainly focused on data from sources such as TV interviews. Similarly, the *scripted spoken corpora* are arranged into scripts and dialogues from *plays*, *movies* and *TV series*. Lastly, the *written corpora* are arranged into four categories- *synchronous* mainly consisting of *chatroom talk*, and *online game-playing forums* with users mainly conversing about game progression; and *asynchronous* mainly consisting of posts made on online *forums* and short text messages on *microblog* websites with character limits for posts.

Tables 1 and 2 present additional details about each corpus, including the name and source citation, topics presented, quantitative details such as number of dialogues, words, total length, and speakers, as well as whether they are multi-modal. All the available corpora have been used for data-driven research on multi-party dialogue. We thus include the *Task Descriptions* each corpus has been used for in the past. These tasks range from machine reading comprehension and turn-taking to speaker-identification.

3 Existing Corpora for Multi-Party Dialogue

In the subsections below, we outline the descriptions of each corpus.

3.1 Spoken Corpora

Spoken corpora is the most prevalent type of corpora available for multi-party dialogue. Spoken corpora presented in this paper are further divided into two main categories (Table 1) - *unscripted* which refers to spontaneous, unplanned dialogues; and *scripted* which refers to planned dialogue such as TV and movie scripts. The distinction between scripted and unscripted is made to allow for dif-

[1] https://tiny.one/mpd-references

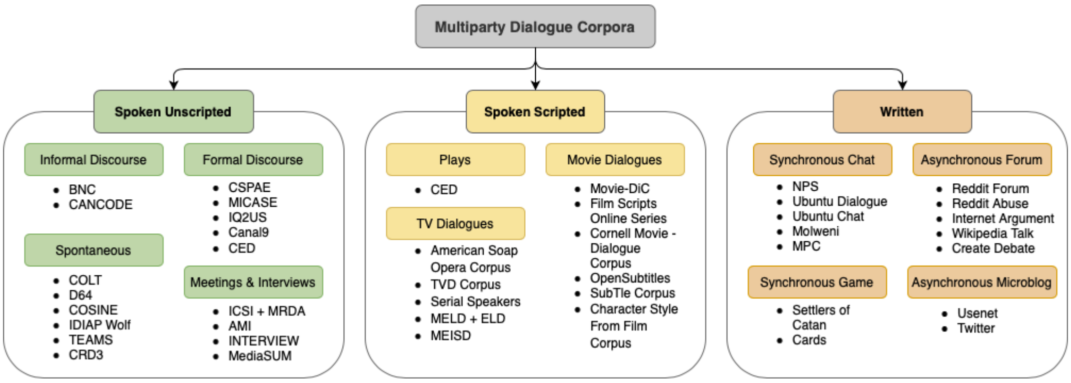

Figure 1: Taxonomy of available Multi-party Corpora, organized by source type.

ferent modelling tasks, since scripted dialogue displays an absence of hesitations, repetitions and other normal non-fluency features.

3.1.1 Unscripted Spoken Corpora

One of the earliest multi-party spoken corpora is the British National Corpus (**BNC**) (Leech, 1992), originally created by the Oxford University press in 1980s-1990s. Covering a wide range of genres, including some written conversations, as well as POS-tagged data (Leech et al., 1994), it is important as a generalized multi-party conversation corpus. It has been used to study social differentiation in the use of English vocabulary (Rayson et al., 1997), word frequency differences in spoken vs written text (Leech et al., 2001), and amplifiers such as "very" and "so" in the English language (Xiao and Tao, 2007).

The Cambridge and Nottingham Corpus of Discourse in English (**CANCODE**) (McCarthy, 1998) focuses on interpersonal communication conversations in various settings such as hair salons and restaurants. It has been used to study language use for teaching in classrooms (O'keeffe et al., 2007), and is a resource for linguistic features of discourse. This corpus is not openly available anymore.

A more informal, casual English corpus is the Bergen Corpus of London Teenage Language (**COLT**) (Stenström and Breivik, 1993), which was recorded in secret to document spontaneous conversations and teenage language. It has been used to study trends in teenage language evolution (Stenström et al., 2002), and is an excellent resource for spontaneous informal multi-party interaction.

The **D64** Multimodal corpus (Oertel et al., 2012) is another addition to spontaneous multi-party dialogue, focusing on recording multi-modal dynamic interactions without specifying a topic.

The COnversational Speech In Noisy Environments (**COSINE**) (Stupakov et al., 2012) corpus introduces data collected in noisy environments, extending the challenges faced in multi-party dialogue such as turn-taking, and has been used to evaluate such systems (Raffensperger et al., 2012).

The **IDIAP Wolf** corpus (Hung and Chittaranjan, 2010) focuses on group behavior in a competitive role-playing game setting, with a pre-condition of bad faith interactions similar to the "werewolf" or "mafia" game that makes it a unique corpus. It has been used in the AIWolfDial task to help train game-playing AI (Kano et al., 2019). While specific instances of lying are not annotated, the "werewolf" of each game is annotated in the corpus.

On the flip side, the **TEAMS** corpus (Litman et al., 2016) where teams of three or four speakers play two rounds of a cooperative board game, provides a novel resource for studying team entrainment and participation dominance. Rahimi and Litman (2020) use it to build a novel graph-based vector representation of multi-party entrainment, gaining insights into the dynamics of the entrainment relations.

Recently, the Critical Role Dungeons and Dragons Dataset (**CRD3**) (Rameshkumar and Bailey, 2020) was released, which is a game-based corpus set in an open-ended scenario. The paper also provides an abstractive summarization benchmark and evaluation, based on each dialogue's summary.

Within formal settings, one of the oldest corpus is the Corpus of Spoken, Professional American-English (**CSPAE**) (Barlow, 2000), consisting of two main components. The first is White House press conferences, and the second is transcripts of meetings on national tests involving statements, discussions, and questions. In the past, it has proved a valuable resource for studying idioms and their usage (Liu, 2003). It is available as a paid resource.

The Michigan Corpus of Academic Spoken English (**MICASE**) (Simpson-Vlach and Leicher, 2006) includes academic speech from university settings. It also comes with abstracts for each transcript, and has been used in online speech summarization (Murray and Renals, 2007).

Debate-based settings are also ideal candidates for multi-party corpora building, and thus the Intelligence Squared Debates (**IQ2US**) (Yang et al., 2010) are an important source. They follow an Oxford-style debating structure, and contain structured data making for a great resource for debate and argumentation analysis (Zhang et al., 2016).

Canal9 (Vinciarelli et al., 2009) is another debate corpus, consisting of political debates. It includes a rich set of socially relevant annotations, and has been used in tasks such as conflict detection (Kim et al., 2012). A historic debate corpus is the Trial Proceedings component of the Corpus of English Dialogues (**CED**) (Kytö and Walker, 2006), which has been used to study signalling function in discourse (Lenker, 2018).

Supplementing formal discourse in debate corpora are formal meeting corpora, with 2 corpora that have become really important for studying multi-party decision-making and discussions of actions to take are the **ICSI** meeting corpus (Janin et al., 2003), which also has Meeting Recorder Dialogue Act (MRDA) annotations (Shriberg et al., 2004); and the multi-modal **AMI** meeting corpus (Renals et al., 2007). ICSI has been used to further study multi-party language modeling (Ji and Bilmes, 2004), and AMI has been used to build summarization for meetings (Zhu et al., 2020).

Recent additions include data from interviews, such as the **INTERVIEW** (Majumder et al., 2020) and **MediaSum** (Zhu et al., 2021) corpora. They include transcripts from interviews on channels such as National Public Radio NPR and CNN.

3.1.2 Scripted Spoken Corpora

Scripted spoken corpora consist of pre-defined scripts such as those for plays, movies, and TV series. These are inherently different as they are not spontaneous, and have pre-defined roles for speakers as well as information on when the dialogues turns are taken. Some corpora are actually labelled with this information, while others are simply transcript-like (Table 1).

One of the earliest available scripted spoken corpora is a second component of the Corpus of English Dialogue **CED** (Kytö and Walker, 2006) focusing on Prose Fiction. It has been used to study language styles in Shakespeare's plays in the context of contemporaneous plays (Demmen, 2012).

The **Movie-DiC Corpus** (Banchs, 2012) consists of a wide range of American movie scripts, along with context descriptions. It has even been used to generate parallel corpora for dialogue translation (Wang et al., 2016). The Film Scripts Online Series corpus includes British movie scripts, but is not available online.

The **Cornell Movie-Dialogue Corpus** (Danescu-Niculescu-Mizil and Lee, 2011) contains metadata associated with each movie script, and has been used to generate emotionally aligned responses to dialogue (Asghar et al., 2020).

The **Character Style From Film Corpus** (Walker et al., 2012a) is another resource contributing towards guided text generation by providing character styles, created from the archive IMSDB. It has been used to generate stylistic dialogue for narratives (Xu et al., 2018).

Both the **OpenSubtitles** (Tiedemann, 2012) and **SubTle corpus** (Ameixa and Coheur, 2013) are based on the OpenSubtitles site. They are corpora of plain scripts, but the website continues to contribute as a resource for more data (Lison and Tiedemann, 2016; Lison et al., 2018).

Bridging the sources of movie and TV scripts is the **Corpus of American Soap Operas** (Davies, 2013) which focuses on informal language, and has been used to study cultural representation differences in American soap operas (Khaghaninejad et al., 2019).

A TV series corpus including data from shows like *The Big Bang Theory* and *Game of Thrones*, supplemented by crowd-sourced contributions for tasks such as summarization is the **TVD** Corpus (Roy et al., 2014). It has been used to build models for speaker identification (Knyazeva et al., 2015). The **Serial Speakers** (Bost et al., 2020) dataset supplements data from both the aforementioned TV serials by also including the *House of Cards* and additional annotations.

Recently, the Multimodal EmotionLines Dataset (**MELD**) (Poria et al., 2019) corpus has been presented by extending the (**ELD**) (Hsu et al., 2018), with audio-visual modality along with text. It has been used as a resource for Dialogue Act Classification (Saha et al., 2020). The **MEISD** (Firdaus et al., 2020) dataset is build further with TV scripts from 10 series, adding *Friends*, *How I Met Your Mother*, *The Office*, *House M.D.*, *Grey's Anatomy*, *Castle*, *Breaking Bad* to the aforementioned series.

3.2 Written Corpora

Written corpora for multi-party have often resulted from online chatroom discussions, like the **NPS** Chat Corpus (Forsythand and Martell, 2007), which is shared as a part of the NLTK (Loper and Bird, 2002), and is one of the first Computer-Mediated corpora.

The Ubuntu IRC chatroom has also contributed to corpora such as the **Ubuntu Dialogue** Corpus (Lowe et al., 2015) and **Ubuntu Chat** Corpus (Uthus and Aha, 2013), which were collected as users asked questions relating to Ubuntu on the forum, and other users answered the questions. They have been used to train end-to-end dialogue systems (Lowe et al., 2017). The **Molweni** corpus (Li et al., 2020) builds on the Ubuntu Chat Dialogue corpus, and adds annotations for machine reading comprehension and dscourse parsing.

Another corpus based on chatroom data is the Multi-Party Chat (**MPC**) Corpus (Shaikh et al., 2010) which presents an annotated corpus based on four levels with communication links, dialogue acts, local topics and meso-topics, and has been used to understand user roles and modeling leadership and influence (Strzalkowski et al., 2012).

Game-playing corpora such as the **Settlers of Catan** Corpus (Afantenos et al., 2012) and **Cards** Corpus (Djalali et al., 2011) are great informal additions to chatroom corpora, with a competitive environment albeit in an informal setting. They have been used for tasks such as training models for negotiation dialogues (Cadilhac et al., 2013).

Online forums such as Reddit, and Wikipedia have also contributed to such corpora. These notably include the **Reddit** (Chang et al., 2020) corpus which has also been extended into larger corpora (Baumgartner et al., 2020).

There have also been argumentative corpora obtained from online interactions, like the **Reddit Domestic Abuse Corpus** (Schrading et al., 2015) taken from subreddits specific on domestic abuse, allowing for discourse analysis on this subject.

Debate and agreement corpora such as the **Internet Argument** Corpus (Walker et al., 2012b), **Agreement in Wikipedia Talk Pages** (Andreas et al., 2012) and Agreement by Create Debaters (Rosenthal and McKeown, 2015), from debate and discussion forums online such as CreateDebate also contribute towards argumentation in dialogue research (Rakshit et al., 2018).

Additionally, there have been corpora obtained from social media such as UseNet and Twitter. These include the **UseNet** Corpus (Shaoul and Westbury, 2007, 2011), a platform which is considered a precursor to more recent forums; and the **Twitter** Corpus (Ritter et al., 2010), which was intended to help model dialogue acts.

3.3 Special Mentions

This section includes special mentions of corpora as well as frameworks and toolkits that do not fall under our previous categories.

There are very few corpora which have focused on **human-machine** dialogue for multi-party interactions. The only such corpora existing to the best of our knowledge is the Mission Rehearsal Exercise (**MRE**) Corpus (Robinson et al., 2004), which presents a dataset built as audio face-to-face sessions between human trainees and virtual agents. The main theme of the multimodal dataset is decision-making for a platoon-leader in a peace-keeping mission, with the trainee acting as a lieutenant. The corpora has about 30K words, 2K utterances, and a total of 55 speakers. Traum et al. (2008) also introduce another 3-party negotiation dialogue corpus, called the Stabilization and Support Operations (**SASO-EN**) corpus, which grew out of experiments on the MRE corpus (Lee et al., 2007), focusing on eye-gaze behavior in 3-party negotiation. In an example scenario, the data consists of a human user who plays the role of a captain whose mission is to move a local clinic to a safer location by negotiating with the doctor and mayor of the city.

FriendsPersona (Jiang et al., 2020) is a another scripted spoken multi-party corpus, which focuses on annotated personalities of scripted characters based on the Big Five personality traits, consisting of 711 conversations from the TV show Friends. It was recently introduced, and has already been used towards personality detection tasks (Christian et al., 2021; Yang et al., 2021).

In the formal meeting and lecture space, the **IDIAP meeting** corpus (Jovanovic et al., 2006) is another extension under the AMI project (AMI and ICSI were discussed in Section 3.1.1), which focuses on addressing behavior in multi-modal, multi-party, face-to-face conversations. The corpus additionally contains hand-annotated dialogue acts, adjacency pairs, addressees and gaze directions of meeting participants. The Computers in Human Interaction Loop (**CHIL**) is another corpus (Mostefa et al., 2007) which provides numerous synchronized audio and video streams of real lectures and meetings, captured in multiple recording sites over a period of 4 years, focusing on human interaction in smart rooms. However, this corpus is a paid resource, available via ELRA[2]. Connected to formal spoken corpora, but focusing on the question-answering task in multi-party dialogue is the recently introduced **QAConv** corpus (Wu et al., 2021), with 34k questions taken from about 28k dialogues, with around 26k words and 32 speakers consisting of conversations taken from email, panels and other formal communication channels.

There are also several corpora, especially multimodal, which have been transcribed, but we could not find the statistics. These include the **VACE multimodal** meeting corpus (Chen et al., 2005), which investigates the interaction among speech, gesture, posture, and gaze in meetings. Another corpus is the **MULTISIMO** corpus (Koutsombogera and Vogel, 2018), towards modeling of collaborative aspects of multimodal behavior in groups that perform simple tasks between 2 people, supported by a facilitator. Mana et al. (2007) also present the **Mission Survival Corpora** (MSC) 1 and 2, a multi-modal corpus of multi-party meetings, automatically annotated using audio-visual cues (speech rate, pitch and energy, head orientation, hand and body fidgeting). Due to the limited information available, we do not add these corpora to the tables or the taxonomy.

A variation of the **Machines Talking to Machines** framework (Shah et al., 2018) allows a simulated user bot and a domain-agnostic system bot to converse to exhaustively generate dialogue "outlines", i.e. sequences of template utterances and their semantic parses, which can then be contextually rewritten by crowdworkers to maintain saliency and coherence while preserving meaning. We include the framework in this survey as it could contribute to collecting data for multi-party dialogue by extending it to include more simulated users and bots.

We also make special mention of the Convokit tool (Chang et al., 2020), which is a toolkit for downloading corpora for dialogues. It allows the downloads to follow standard format for all available corpora. It also provides the functionality to load custom datasets in a similar format, making it easier to work with multiple corpora at once.

4 Data Collection Methods

Several methods of data collection have been used to collect the aforementioned corpora. We organize these into three main categories and discuss in detail below.

Aggregated from various sources: BNC, CANCODE, and MICASE employ the aggregation method to build the corpora. They pull information from various sources, including text from sources such as newspapers, journals, publicly available government meetings, radio phone-ins, academic writings, seminars, advising sessions etc. These corpora incorporate multiple types of speech, and often include speech surrounding multiple topics (especially BNC and CANCODE, MICASE mainly focuses on academic settings to collect data). They are thus great candidates for studying language semantics and have been employed to study large-scale vocabularies (McCarthy et al., 2010) and word sense disambiguation (Roberts and Erklärung, 2012) in the past.

Transcribed from pre-recorded media: Single (or double) source origins, such as COLT, CRD3, and IQ2, maintain focus on certain themes, such as formal meeting data. These are not collected within specialized environments, but consist of either transcribed speech recorded in the wild, transcribed interviews & meetings, and online forum or social media data. This category also includes scripted corpora, which are usually collections of various scripts & dialogues from plays, movies and TV series, such as TVD and SubTle. Having a set theme allows these corpora to be used for generating themed text such as MELD being used for character identification as a part of the 2018 SemEval challenge (Choi and Chen, 2018).

Collected in specialized environments: Most multi-modal corpora employ specialized environments or equipment to collect data that can be synchronized across multiple modalities. Most focus on data collection using *audio*, which can then be transcribed. Specialized room environments with studio-quality recording (ICSI, AMI), close-talking mics (ICSI, IDIAP Wolf, TEAMS), and a combination of far- and close-field mics (COSINE,

[2] https://tiny.one/chil-data

	Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
UNSCRIPTED INFORMAL	**Aggregated from various sources**							
	British National Corpus (BNC)	Informal	854	10M	100 hrs*	23466	✓	word sense disambiguation, morphological & syntactic analysis
	CANCODE	Informal	-	5M	550 hrs*	-	×	language learning, POS tagging
	Collected in specialized environments							
	D64 Corpus	Natural	2	70K*	8 hrs	5	✓	involvement detection, studying silence and overlap in conversation
	COSINE	Natural	10	160K	42 hrs	3.69 per session	✓	recognition of speech and speakers in noisy environments
	IDIAP Wolf Corpus	Game	15	60K*	7 hrs	8-12 groups	✓	group performance in task-based interaction, implicit communication
	TEAMS corpus	Game	116K	3M	47 hrs	3-4/ game	✓	entrainment, speaker transitions, personality identification & team dynamics
	Transcribed from pre-recorded media							
	COLT corpus	Natural	100	500K	55 hrs	31	×	teenage talk trends
	CRD3	Game	159	5M	-	72	✓	character-action interactions in role playing games
UNSCRIPTED FORMAL	**Aggregated from various sources**							
	MICASE	Academic	152	1.7M	200 hrs	1571	✓	male/female adjective use, academic discourse and vocabularies, English language learning
	Collected in specialized environments							
	AMI Meeting Corpus	Formal	175	900K*	100 hrs	4-5 per meeting	✓	recognizing socio-economic roles, decision and action detection, summarization, dialogue act tagging
	ICSI MRDA	Meetings	75	795K	72 hrs	3-10 per meeting	✓	speaker overlap, summarization, speaker identification
	Transcribed from pre-recorded media							
	Intelligence Squared Debates	Debates, predecided	108	1.8M	200 hrs*	3-5 per debate	✓	predictive models of debates, discourse modeling
	CSPAE	Politics, education	200	2M	220 hrs*	400+	×	speech style and gender distinctions, speech variation between written and spoken corpora
	CED (1560-1760)	Movies, formal	-	1.2 M	-	-	×	early English language variations and changes over time
	MediaSum	Interview	463K	720M	-	6.5 per dialogue	✓	dialogue summarization
	INTERVIEW corpus	Interview	105K	126.7M	10K	184K	✓	follow-up question generation
	Canal9	Political Debates	70 debates	-	43 hrs	5 per debate	✓	speaker identification, turn-taking, conflict detection
SCRIPTED SPOKEN	**Transcribed from pre-recorded media**							
	Movie-DiC	Movie dialogues	132K	6M	-	1-7 per dialogue	×	turn taking, speaker identification, emotional dialogue generation
	Cornell Movie Dialogue Corpus	Movie dialogues	220K	9M	-	9035	×	
	Film scripts online series	Movie scripts	263K	16M	1500 scripts	2-6 per script*	×	(information unavailable)
	OpenSubtitles	Movie subtitles	337M	2.5G	-	2-6 per script*	×	turn taking, speaker identification, emotional dialogue generation
	SubTle corpus	Movie subtitles	3.35M	20M	6184 movies	2-6 per script*	×	
	Character Style from Film Corpus	Movie subtitles	151K	9.6M	862 movies	2-6 per script*	×	
	American Soap Opera Corpus	TV dialogues	1.2M	100M	-	10-12 per script	×	
	TVD corpus	TV dialogues	10K	600K	-	2-6 per script	✓	
	MELD	TV dialogues	1400	109K	13.6 hrs*	400	✓	
	Serial Speakers	TV dialogues	106K	682K	130 hrs	6 per script*	✓	
	MEISD	TV dialogues	1000	50K unique	22 hrs	4072	✓	

Table 1: Further details for all spoken corpora. Starred (*) numbers are approximated from available information.

Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
NPS Chat Corpus	Informal chat	15	100M			×	part-of-speech tagging, dialogue act recognition
Ubuntu Dialogue Corpus	Ubuntu OS Chatroom	930K	100M	-	-	×	speaker identification, discourse parsing, machine comprehension, response selection
Ubuntu Chat Corpus	Ubuntu OS Chatroom	10655	2B	-	-	×	language learning, POS tagging
Molweni	Ubuntu OS Chatroom	10K	24K	200 hrs	3.5 per dialogue	×	machine reading comprehension, discourse parsing
MPC Corpus	Informal chatroom	14	58K	-	5 per session	×	turn-taking, speaker identification, detecting influence & leadership, group behavior
Settlers of Catan	Informal, game-playing	21	-	-	2-6 players	×	modeling bargaining, negotiation, trading dialogue, risk-management in dialogue, action identification
Cards Corpus	Informal, game-playing	1266	282K	-	-	×	goal-driven dialogue, event knowledge based questioning
Reddit Corpus	Informal forum	84979	76M-414M*	-	521K	Maybe	discourse, cyberbully detection, exploring incel language
Reddit Domestic Abuse Corpus	Abusive forum	21333	19M-303M	-		×	language biases, detecting harassment
Internet Argument Corpus	Political forum	11000	73M	-	-	×	summarization, rhetoric and sarcasm, stance detection
Agreement in Wikipedia Talk Pages	Informal	822	110K	-	-	×	linguistic tracing of manipulations, dialog act recognition, social act recognition, conflict detection, speaker identification
Agreement by Create Debaters	Informal	10000	1.4M	-	-	×	constructive disagreement, sarcasm, rumor classification, stance identification
Twitter Corpus	Informal microblog	1.3M	125M	-	-	×	dialogue act recognition, author and topic identification, event discovery
UseNet Corpus	Informal microblog	47860	7B	-	-	×	modeling and analyzing text written on mobile devices

Table 2: Further details for all written corpora. Starred (*) numbers are approximated from available information.

AMI) have provided better data collection for corpora, allowing for annotations of speech activity and pauses as well. Another popular data collection method focuses on *video*, such as motion sensing (D64), and video cams (IDIAP Wolf, TEAMS, AMI), which supplement speech data well by also allowing for annotation of head movement, gesture, and eye-gaze tracking.

There are also multiple projects that emulate online social media platforms for controlled data collection, such as the Truman platform and Community Connect (Mahajan et al., 2021).

5 Desiderata for Data Collection

Given the multitude of corpora available and the modeling tasks that need to be undertaken to develop conversational agents for multi-party dialogue, we outline here **three** key criteria for future efforts in data collection:

1. Participant balance and tracking: We find from the tasks identified in Tables 1 and 2 that speaker and addressee identification are important open tasks in multi-party dialogue modeling. Consequently, corpora should contain sufficient information, in the data or in the metadata, to track participants within dialogues and across dialogues, if possible. Where possible, participants should be balanced in terms of age, gender and ethnicity and other demographic factors, so as to not preferentially model any specific type of language use.

2. Signal to Noise ratio: The corpora should contain a sufficiently high number of texts as possible, however, these should be of sufficiently high quality. Particularly, for data that are scraped from the web (e.g. Twitter or Reddit), it is possible for the noise to drown out important signals in the data. It is important to document all considerations and assumptions made when collecting the data. In most cases, specific details are outlined for data that are collected under specialized settings, and extreme care is taken to synchronize collection across modalities. We encourage a similar level of attention to detail when data are aggregated from existing sources. When possible, data collection studies should be preregistered so that researchers can describe their hypotheses, methods, and analyses beforehand (Nosek et al., 2019).

3. Ethical Considerations: Creating corpora focusing on multiple speakers requires multiple considerations to protect personally identifiable information (PII), while making sure that the corpus is annotated well to allow for usability. Especially in the case of multi-modal corpora, where eye-gaze and head movements have been used as features for tasks such as turn-taking, there are important guidelines to consider since it is not possible to remove PII easily (Benedict et al., 2019).

6 Discussion

The three desiderata listed above provide us with a set of guidelines for thinking about the challenges for thoughtful data collection. This (potentially non-exhaustive) list of questions is inspired by the current movement in several research fields to pre-register studies in advance (Nosek et al., 2019; Vilhuber, 2020) and the needs for datasheets for datasets (Gebru et al., 2018).

Research Questions and Hypotheses:

- What is/are the research question(s) that the data can help answer? How are the research questions operationalized for multi-party settings?
- What phenomena are being studied? How will the phenomena be measured? Does the phenomena apply to each participant, multiple participants in multi-party conversation or to the conversation overall?

Data Collection:

- Will the corpus contain enough examples of the phenomena under study? How will you know if the corpus contains examples of the phenomena?
- Are number of speakers in the corpus adequate to study the phenomena?
- Are the data sources representative? Do they prefer certain demographics or certain forms over others, especially marginalized groups?
- For multi-modal corpora, which non-verbal cues are available? Are text annotations available, such as start/end times for turns, who a speaker is looking at, when pauses occur, etc?
- If data are sampled from existing sources, how are selection criteria determined? Are they justified?

Ethical Considerations and PII:

- Has PII been eliminated as much as possible, especially where inclusion of such data is not necessary and does not affect the quality of the data?
- Has informed consent to release data been obtained from all parties, especially where PII could not be removed, and the full extent of release and its possible consequences conveyed to participants?
- If speaker metadata is removed for preserving PII, are all the data where a speaker is being referred to also converted with a similar scheme?

7 Conclusion and Future Work

We present a systematic review and a taxonomy of available corpora for multi-party dialogue. We also identify key tasks that are typically conducted through the use of these corpora and we review how existing corpora are collected. To ensure that data-driven models that are developed using these and any future corpora, are high quality, we advance three critical desiderata, that lead us to several guiding principles. While we attempt to be as comprehensive as possible, there are certain **limitations** of this present article. We recognize that our review focuses entirely on English language data and models. Certainly, corpora exist in other languages, e.g. in Chinese and French (Riou et al., 2015; Liu et al., 2012). We also do not provide any detail about the modeling tasks, e.g. turn taking. Extending our review to include additional languages and detailed description of modeling tasks is indeed part of a future, larger publication.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, et al. 2012. Developing a corpus of strategic conversation in the settlers of catan. In *SeineDial 2012-The 16th Workshop On The Semantics and Pragmatics Of Dialogue*.

David Ameixa and Luísa Coheur. 2013. From subtitles to human interactions : introducing the subtle corpus. Technical report.

Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. 2012. Annotating agreement and disagreement in threaded discussion. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 818–822, Istanbul, Turkey. European Language Resources Association (ELRA).

Nabiha Asghar, Ivan Kobyzev, Jesse Hoey, Pascal Poupart, and Muhammad Bilal Sheikh. 2020. Generating emotionally aligned responses in dialogues using affect control theory. *arXiv preprint arXiv:2003.03645*.

Rafael E. Banchs. 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, Jeju Island, Korea. Association for Computational Linguistics.

Michael Barlow. 2000. *Corpus of Spoken, Professional American-English*. Rice University.

J. Baumgartner, Savvas Zannettou, B. Keegan, Megan Squire, and J. Blackburn. 2020. The pushshift reddit dataset. In *ICWSM*.

Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. 2019. Learning multiparty turn-taking models from dialogue logs. *arXiv preprint arXiv:1907.02090*.

Catherine Benedict, Alexandria L Hahn, Michael A Diefenbach, and Jennifer S Ford. 2019. Recruitment via social media: advantages and potential biases. *Digital health*, 5:2055207619867223.

Xavier Bost, Vincent Labatut, and Georges Linares. 2020. Serial speakers: a dataset of TV series. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4256–4264, Marseille, France. European Language Resources Association.

Anaïs Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 357–368, Seattle, Washington, USA. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

L. Chen, R. Rose, Ying Qiao, I. Kimbara, Fey Parrill, Haleema Welji, Tony X. Han, J. Tu, Zhongqiang Huang, M. Harper, Francis K. H. Quek, Yingen Xiong, D. McNeill, Ronald Tuttle, and T. Huang. 2005. Vace multimodal meeting corpus. In *MLMI*.

Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana. Association for Computational Linguistics.

Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z Zamli. 2021. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1):1–20.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Mark Davies. 2013. *Corpus of american soap operas*. Brigham Young University.

Jane Demmen. 2012. *A corpus stylistic investigation of the language style of Shakespeare's plays in the context of other contemporaneous plays*. Lancaster University.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, pages 1–56.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

A. Djalali, S. Lauer, and Christopher Potts. 2011. Corpus evidence for preference-driven interpretation. In *Amsterdam Colloquium on Logic, Language and Meaning*.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

T. Hawes, J. Lin, and P. Resnik. 2009. Elements of a computational model for multi-party discourse: The turn-taking behavior of supreme court justices. *J. Assoc. Inf. Sci. Technol.*, 60:1607–1615.

Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and D. Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. In *INLG*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hayley Hung and Gokul Chittaranjan. 2010. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. *Proceedings of the 18th ACM international conference on Multimedia*.

A. Janin, D. Baron, Jane Edwards, D. Ellis, D. Gelbart, N. Morgan, Barbara Peskin, T. Pfau, E. Shriberg, A. Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 1:I–I.

Gang Ji and Jeffrey Bilmes. 2004. Multi-speaker language modeling. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 133–136, Boston, Massachusetts, USA. Association for Computational Linguistics.

Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13821–13822. AAAI Press.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23.

Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, Issei Tsunoda, Shoji Nagayama, Dolça Tellols, Yu Sugawara, and Yohei Nakata. 2019. Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations. In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.

Mohammad Saber Khaghaninejad, Mehrnoosh Dehbozorgi, and Mohammad Amin Mokhtari. 2019. Cultural representations of americans, europeans, africans and arabs in american soap operas: A corpus-based analysis. *Language & Translation*, 7(3):133–141.

S. Kim, F. Valente, and A. Vinciarelli. 2012. Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5092.

Katrin Kirchhoff and Mari Ostendorf. 2003. Directions for multi-party human-computer interaction research. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*, pages 7–9.

E. Knyazeva, Guillaume Wisniewski, H. Bredin, and François Yvon. 2015. Structured prediction for speaker identification in tv series. In *INTERSPEECH*.

Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Merja Kytö and Terry Walker. 2006. *Guide to A corpus of English dialogues 1560-1760*. Acta Universitatis Upsaliensis.

Jina Lee, Stacy Marsella, David Traum, Jonathan Gratch, and Brent Lance. 2007. The rickel gaze model: A window on the mind of a virtual human. In *International workshop on intelligent virtual agents*, pages 296–303. Springer.

G. Leech. 1992. 100 million words of english:the british national corpus (bnc). *Second Language Research*, 28:1–13.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Geoffrey Leech, P Rayson, and A Wilson. 2001. Word frequencies in written and spoken english: based on the british national corpus.

U. Lenker. 2018. ‘there’s an issue there . . . ’: Signalling functions of discourse-deictic there in the history of english. *Language Sciences*, 68:94–105.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. 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, Austin, Texas. Association for Computational Linguistics.

D. Liu. 2003. The most frequently used spoken american english idioms: A corpus analysis and its implications. *TESOL Quarterly*, 37:671–700.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Ting Liu, Samira Shaikh, Tomek Strzalkowski, Aaron Broadwell, Jennifer Stromer-Galley, Sarah Taylor, Umit Boz, Xiaoai Ren, and Jingsi Wu. 2012. Extending the MPC corpus to Chinese and Urdu - a multiparty multi-lingual chat corpus for modeling social phenomena in language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2868–2873, Istanbul, Turkey. European Language Resources Association (ELRA).

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

R. Lowe, Nissan Pow, I. Serban, Laurent Charlin, C. Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue Discourse*, 8:31–65.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Khyati Mahajan, Sourav Roy Choudhury, Sara M. Levens, Tiffany Gallicano, and S. Shaikh. 2021. Community connect: A mock social media platform to study online behavior. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: A large-scale open-source corpus of media dialog. *arXiv preprint arXiv:2004.03090*.

N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro. 2007. Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. *The Medical Roundtable*, pages 9–14.

Diana McCarthy, B. Keller, and R. Navigli. 2010. Getting synonym candidates from raw data in the english lexical substitution task. In *Proceedings of the 14th euralex international congress*.

Michael McCarthy. 1998. *Spoken language and applied linguistics*. Cambridge University Press.

Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M Chu, Ambrish Tyagi, Josep R Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, et al. 2007. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language resources and evaluation*, 41(3):389–407.

Gabriel Murray and S. Renals. 2007. Towards online speech summarization. In *INTERSPEECH*.

Brian A Nosek, Emorie D Beck, Lorne Campbell, Jessica K Flake, Tom E Hardwicke, David T Mellor, Anna E van’t Veer, and Simine Vazire. 2019. Preregistration is hard, and worthwhile. *Trends in cognitive sciences*, 23(10):815–818.

C. Oertel, F. Cummins, Jens Edlund, P. Wagner, and N. Campbell. 2012. D64: a corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7:19–28.

Anne O'keeffe, Michael McCarthy, and Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Peter A. Raffensperger, Russ Webb, P. Bones, and A. McInnes. 2012. A simple metric for turn-taking in emergent communication. *Adaptive Behavior*, 20:104 – 116.

Z. Rahimi and D. Litman. 2020. Entrainment2vec: Embedding entrainment for multi-party dialogues. In *AAAI*.

Geetanjali Rakshit, Kevin K Bowden, Lena Reed, Amita Misra, and Marilyn Walker. 2018. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, volume 510, page 45. Springer.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

Paul Rayson, G. Leech, and Mary Hodges. 1997. Social differentiation in the use of english vocabulary: some analyses of the conversational component of the british national corpus. *International Journal of Corpus Linguistics*, 2:133–152.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

S. Renals, Thomas Hain, and H. Bourlard. 2007. Recognition and understanding of meetings the ami and amida projects. *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 238–247.

Matthieu Riou, Soufian Salim, and Nicolas Hernandez. 2015. Using discursive information to disentangle french language chat. In *2nd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC 2015)/Social Media at GSCL Conference 2015*, pages 23–27.

Alan Ritter, Colin Cherry, and Bill Dolan. 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, Los Angeles, California. Association for Computational Linguistics.

W. Roberts and Eidesstattliche Erklärung. 2012. Integrating syntax and semantics for word sense disambiguation.

Susan Robinson, Bilyana Martinovski, Saurabh Garg, Jens Stephan, and David Traum. 2004. Issues in corpus development for multi-party multi-modal task-oriented dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.

Anindya Roy, Camille Guinaudeau, Hervé Bredin, and Claude Barras. 2014. TVD: A reproducible and multiply aligned TV series dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 418–425, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, Lisbon, Portugal. Association for Computational Linguistics.

Isabella Seeber, Eva Bittner, Robert O Briggs, Triparna De Vreede, Gert-Jan De Vreede, Aaron Elkins, Ronald Maier, Alexander B Merz, Sarah Oeste-Reiß, Nils Randrup, et al. 2020. Machines as teammates: A research agenda on ai in team collaboration. *Information & management*, 57(2):103174.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Samira Shaikh, Tomek Strzalkowski, Aaron Broadwell, Jennifer Stromer-Galley, Sarah Taylor, and Nick Webb. 2010. MPC: A multi-party chat corpus for modeling social phenomena in discourse. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

C Shaoul and C Westbury. 2011. A usenet corpus (2005-2010). *Edmonton, AB: University of Alberta*.

Cyrus Shaoul and Chris Westbury. 2007. A usenet corpus (2005–2007). *University of Alberta, Edmonton, AB*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Penelope Sibun. 1997. Beyond dialogue: the six w's of multi-party interaction. In *Working Notes of AAAI97 Spring Symposium On Mixed-Initiative Interaction, Stanford, CA*, pages 145–150.

Rita C Simpson-Vlach and Sheryl Leicher. 2006. *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. University of Michigan Press ELT.

AB Stenström and Leiv Egil Breivik. 1993. The bergen corpus of london teenager language (colt). *ICAME journal*, 17:128.

Anna-Brita Stenström, Gisle Andersen, and Ingrid Kristine Hasund. 2002. *Trends in Teenage Talk: Corpus compilation, analysis and findings*, volume 8. John Benjamins Publishing.

Tomek Strzalkowski, Samira Shaikh, Ting Liu, George Aaron Broadwell, Jenny Stromer-Galley, Sarah Taylor, Umit Boz, Veena Ravishankar, and Xiaoai Ren. 2012. Modeling leadership and influence in multi-party online discourse. In *Proceedings of COLING 2012*, pages 2535–2552, Mumbai, India. The COLING 2012 Organizing Committee.

A. Stupakov, E. Hanusa, Deepak Vijaywargi, D. Fox, and J. Bilmes. 2012. The design and collection of cosine, a multi-microphone in situ speech corpus recorded in noisy environments. *Comput. Speech Lang.*, 26:52–66.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

D. Traum, S. Marsella, J. Gratch, J. Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *IVA*.

David C. Uthus and D. Aha. 2013. The ubuntu chat corpus for multiparticipant chat analysis. In *AAAI Spring Symposium: Analyzing Microtext*.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*, 4:60–68.

Lars Vilhuber. 2020. Reproducibility and replicability in economics. *Harvard Data Science Review*, 2(4).

A. Vinciarelli, Alfred Dielmann, S. Favre, and Hugues Salamin. 2009. Canal9: A database of political debates for analysis of social interactions. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–4.

Marilyn Walker, Grace Lin, and Jennifer Sawyer. 2012a. An annotated corpus of film dialogue for learning and characterizing character style. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1373–1378, Istanbul, Turkey. European Language Resources Association (ELRA).

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012b. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2748–2754, Portorož, Slovenia. European Language Resources Association (ELRA).

Rainer Winkler, Maya Lisa Neuweiler, Eva Bittner, and Matthias Söllner. 2019. Hey alexa, please help us solve this problem! how interactions with smart personal assistants improve group performance. In *ICIS International Conference of Information Systems*, Munich. ACM Digital.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2021. Qaconv: Question answering on informative conversations. *arXiv preprint arXiv:2105.06912*.

Richard Xiao and H. Tao. 2007. A corpus-based sociolinguistic study of amplifiers in british english. *Sociolinguistic Studies*, 1:241–273.

W. Xu, Charlie Hargood, Wen Tang, and F. Charles. 2018. Towards generating stylistic dialogues for narratives using data-driven approaches. In *ICIDS*.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021. Multi-document transformer for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14221–14229.

Zhaojun Yang, Baichuan Li, Yi Zhu, Irwin King, Gina-Anne Levow, and H. Meng. 2010. Collection of user judgments on spoken dialog system with crowdsourcing. *2010 IEEE Spoken Language Technology Workshop*, pages 277–282.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. End-to-end abstractive summarization for meetings. *ArXiv*, abs/2004.02016.