Characteristics of Common Experimental Dialogue Tasks: a Systematic Review & Taxonomy

Ella Cullen, Patrick Healey, Paraskevi Argyriou and Suyog Pipliwal

July 24, 2023

Characteristics of Common Experimental Dialogue Tasks: A Systematic Review & Taxonomy

Ella Cullen[1], Patrick Healey[1], Paraskevi Argyriou[2], and Suyog Pipliwal[1]

[1] School of Electronic Engineering and Computer Science, Queen Mary University

[2] School of Biological and Behavioural Sciences, Queen Mary University

Author Note

Correspondence should be addressed to phd candidate researcher Ella Cullen, Queen Mary University, email: e.l.a.cullen@qmul.ac.uk.

Abstract

Natural dialogue has a flexible, open-ended and collaborative character that makes controlled experiments difficult. One strategy for dealing with this is to use a dialogue task that reduces this complexity by limiting the content, format or structure of a dialogue. This paper introduces a systematic review of these tasks which aims to: i) provide an overview of the variety of dialogue tasks in the literature, ii) introduce a taxonomy for capturing the basic features of dialogue tasks, iii) introduce simple quantitative comparisons of existing corpora, and iv) identify potential gaps in the kinds of dialogue covered by current experimental work.

Keywords: Dialogue, task-oriented dialogue, domain-independent dialogue, face-to-face, taxonomy, natural language processing

Characteristics of Common Experimental Dialogue Tasks: A Systematic Review & Taxonomy

Dialogue tasks are a key practical tool for experimental investigations of human interaction; they bring useful experimental control to an inherently noisy and variable phenomenon (Sacks et al., 1974). However, these tasks also necessarily compromise some features of natural dialogue. To assess how well findings from these dialogue tasks generalise to other tasks and situations, it is important to understand how these task-oriented dialogues differ from each other and from natural dialogue. Section 1 of this paper introduces a survey of the literature and addresses the following questions:

1. What range of dialogue tasks have been used in the experimental literature?
2. What characteristics can we use to make meaningful, practical comparisons between dialogue tasks?
3. What aspects of natural dialogue are covered by the current literature? What gaps are there?

A further question that arises is; "should the variability and dynamism of natural dialogue be considered noise?". Section 2 explores this question. Comparison of quantitative dialogue measures across popular dialogue tasks will allow for this question to be answered, to see if the dynamism and variability of dialogue is significantly constrained by different task characteristics and how ecological validity is reduced.

Section 1

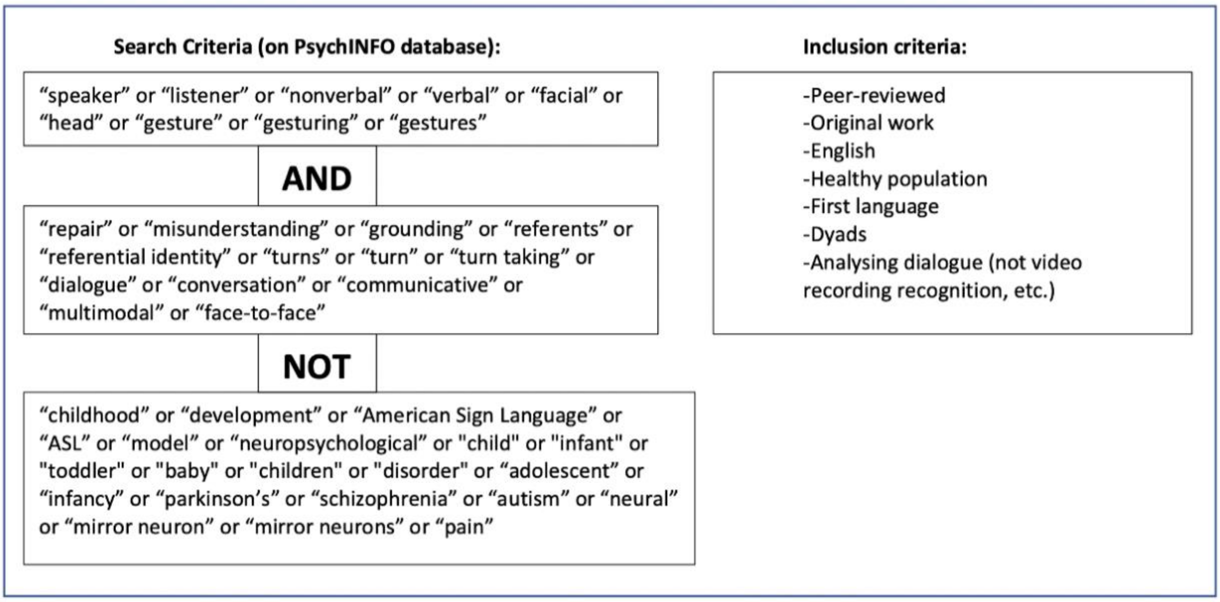
Methods

1. Systematic Review

With the research aim of systematically classifying existing dialogue tasks according to their characteristics, a systematic search of the literature was carried out on the PsychINFO database (See Figure 1) to answer the research question: "How can the methodological and structural variance between different dialogue task studies in the literature be characteristically organised?". Data extraction from this search involved the stages of inclusion illustrated in Table 1.

Figure 1.

Search strategy



Search Criteria (on PsychINFO database):

"speaker" or "listener" or "nonverbal" or "verbal" or "facial" or "head" or "gesture" or "gesturing" or "gestures"

AND

"repair" or "misunderstanding" or "grounding" or "referents" or "referential identity" or "turns" or "turn" or "turn taking" or "dialogue" or "conversation" or "communicative" or "multimodal" or "face-to-face"

NOT

"childhood" or "development" or "American Sign Language" or "ASL" or "model" or "neuropsychological" or "child" or "infant" or "toddler" or "baby" or "children" or "disorder" or "adolescent" or "infancy" or "parkinson's" or "schizophrenia" or "autism" or "neural" or "mirror neuron" or "mirror neurons" or "pain"

Inclusion criteria:

-Peer-reviewed
-Original work
-English
-Healthy population
-First language
-Dyads
-Analysing dialogue (not video recording recognition, etc.)

Table 1.

Summary of the stages of inclusion for systematic review.

Literature Review Stage	Number of Papers
Search of PsychINFO	8,931
Inclusion based on title	520
Inclusion based on abstract	242
Inclusion based on full text	78
Forward search (Google Scholar)	113

2. Taxonomy development

Taxonomy development followed Nickerson et al.’s (2013) guidelines. The purpose of this taxonomy was to classify existing dialogue tasks according to their characteristics. From collecting and coding the 113 dialogue task studies, dialogue task characteristics were synthesised and organised using a combination of the empirical-to-conceptual and conceptual-to-empirical approach through multiple iterations (Nickerson et al., 2013).

3. Application of taxonomy to tasks in the literature.

This taxonomy was then applied to the dialogue task studies collected in the systematic review to map out the space and distribution of task types in the experimental literature.

Results

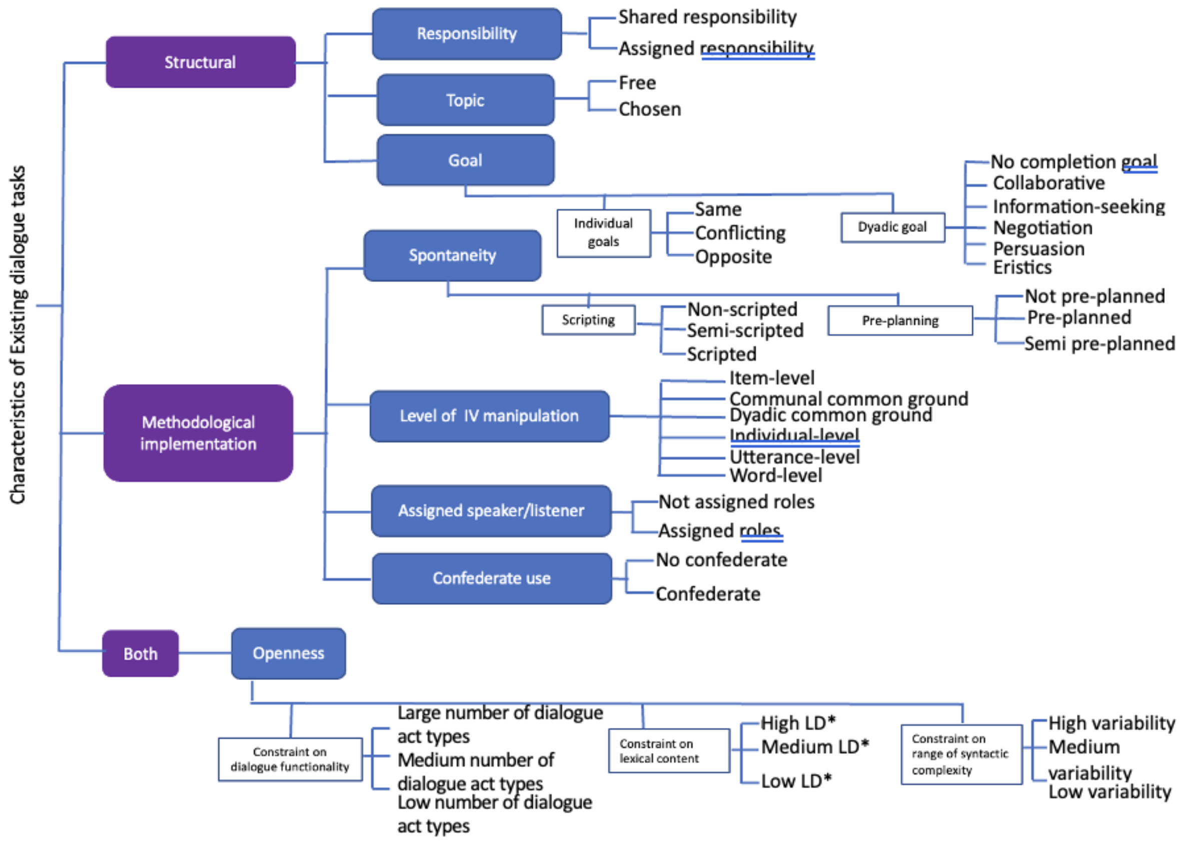
1. Systematic Review

The systematic search of the literature yielded 113 experimental dialogue task studies.

2. Proposed Taxonomy

Figure 2.

Taxonomy of the characteristics of existing dialogue tasks in the literature



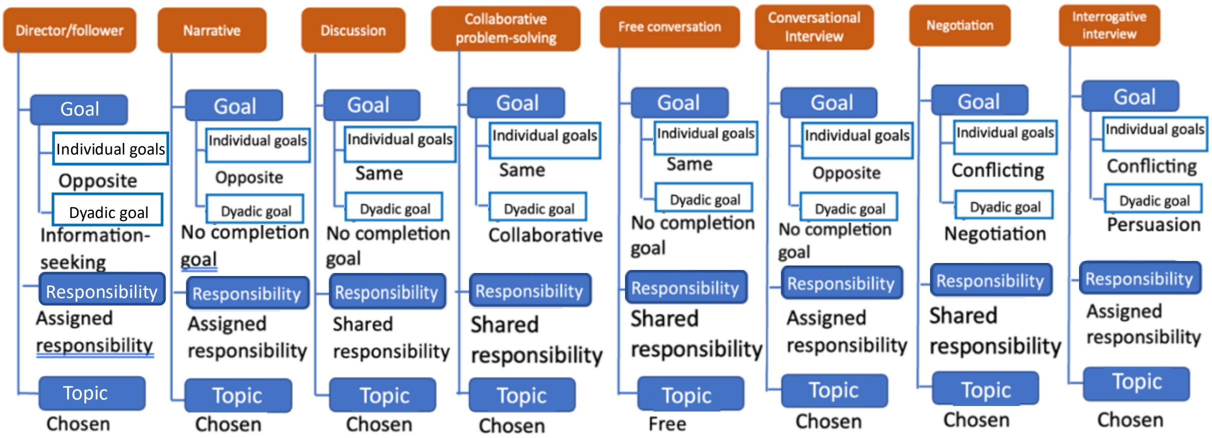
Note. The purple boxes are meta-dimensions, the blue boxes are dimensions, the white boxes are sub-dimensions and the non-boxed are characteristics of the sub-dimensions. A task can only have one characteristic of each sub-dimension.

*Lexical diversity (LD)

3. Grouping of dialogue tasks according to shared structural characteristics

Figure 3.

Dialogue task groups according to shared structural characteristics.



Note. The orange boxes illustrate resulting dialogue task groups. The following dimensions (blue boxes), sub-dimensions (white boxes) and characteristics under each task group highlight how tasks within a group homogenously meet the structural characteristics of their group.

4. Example classification of an existing dialogue task study with the proposed taxonomy

Table 2.

Example classification of a dialogue task

	Category / Dimension	Howarth & Anderson's (2007) Map Task
Structural	**Responsibility**	**Assigned:** Roles of 'instruction giver' and 'instruction follower'
	Topic	**Chosen:** Dialogue centres around instruction follower drawing a complete route that fits the instruction giver's description
	Individual goals	**Opposite:** One is to give instructions, one is to follow
	Dyadic goals	**Information-seeking:** asymmetric structure in which one individual consistently is seeking information from the other throughout the dialogue
Methodological	**Scripting**	**Non-scripted:** No part of either subjects' speech was directly scripted
	Confederate use	**No confederate:** Both were naïve participants whose behaviour was not directly instructed
	Level of IV manipulation	**Dyadic level:** Face-to-face vs. video-mediated
	Assigned speaker/listener	**Not assigned roles:** participants were instructed to freely speak about the task
	Pre-planned speech	**Not pre-planned:** describers and followers had to discuss the route given to them for that trial without previous examination of the route

Section 2

Methods

Comparative analysis of an assorted collection of transcripts of task-oriented and natural dialogues was carried out. Transcripts from the Spoken Demographic section of the British National Corpus (BNC) (Love et al., 2017) were used as a baseline measure for domain-independent, natural dialogue. Transcripts for dialogue tasks included the Map task (Anderson et al., 1991), Maze task (Garrod and Anderson, 1987), Tangram task (Clark & Wilkes-Gibbs, 1986), TRAINS corpus (Allen et al., 1995), DBOX corpus (Petukhova et al., 2014), SWBD (Godfrey et al., 1992).

Three measures were identified to enable direct, quantitative comparison of the effects of different dialogue tasks:

1. Lexical diversity

Lexical diversity was captured through the vocd-D index. A standard length of 200 words from each transcript was analysed.

2. Syntactic complexity

The Stanford parser (Klein & Manning, 2003) was used to parse each turn of the transcripts. Syntactic complexity was operationalised as the node count of the parse tree for each turn.

3. Dialogue act type range

Pre-annotated, ISO standard corpora were collected from DialogBank (Bunt et al., 2019). Available corpora included the SWBD, TRAINS, DBOX and HCRC Map task corpus. The dialogue act tags for a standard length of 100 turns per transcript were counted.

Results

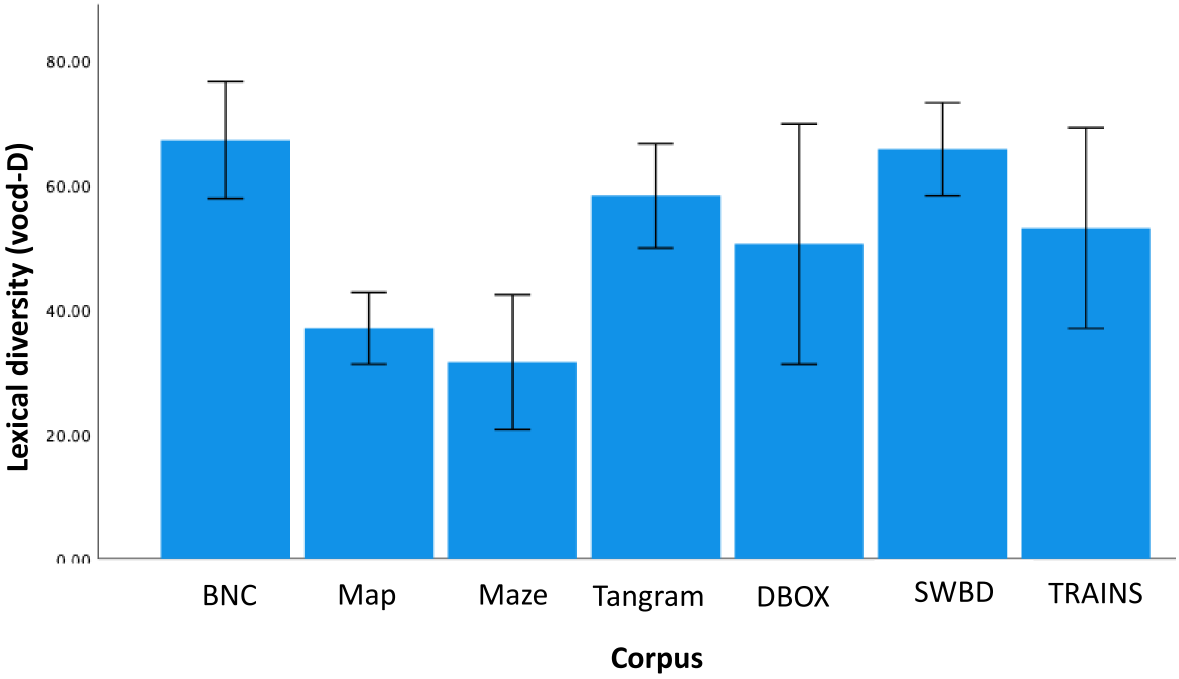
1. Lexical diversity

A one-way ANOVA comparing overall lexical diversity of dialogues between task transcripts and the BNC shows a statistically significant difference in lexical diversity between at least two groups [F(6, 38) = 12.29, p = <.001] (See figure 4).

The findings from post-hoc Tukey's HSD Test for multiple comparisons (Table 3) shows that the mean lexical diversity of the BNC was found to be significantly higher than the map and maze tasks. There was no significant difference found between the BNC and tangram, DBOX, SWBD, nor TRAINS task.

Figure 4.

Graph to illustrate the mean lexical diversity of different task corpora published from DialogBank



Note. Error bars show 95% confidence intervals.

Table 3.

Table of Tukey's HSD pairwise comparisons

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
BNC	Map	30.15167*	5.74174	<.001	12.2862	48.0171
	Maze	35.56167*	5.74174	<.001	17.6962	53.4271
	Tangram	8.90667	5.24147	0.621	-7.4022	25.2155
	DBOX	16.63667	6.41947	0.157	-3.3375	36.6108
	SWBD	1.46967	5.13557	1	-14.5097	17.449
	TRAINS	14.09667	6.41947	0.321	-5.8775	34.0708
Map	BNC	-30.15167*	5.74174	<.001	-48.0171	-12.2862
	Maze	5.41	5.74174	0.963	-12.4554	23.2754
	Tangram	-21.24500*	5.24147	0.004	-37.5538	-4.9362
	DBOX	-13.515	6.41947	0.37	-33.4892	6.4592
	SWBD	-28.68200*	5.13557	<.001	-44.6613	-12.7027
	TRAINS	-16.055	6.41947	0.188	-36.0292	3.9192
Maze	BNC	-35.56167*	5.74174	<.001	-53.4271	-17.6962
	Map	-5.41	5.74174	0.963	-23.2754	12.4554
	Tangram	-26.65500*	5.24147	<.001	-42.9638	-10.3462
	DBOX	-18.925	6.41947	0.073	-38.8992	1.0492
	SWBD	-34.09200*	5.13557	<.001	-50.0713	-18.1127
	TRAINS	-21.46500*	6.41947	0.028	-41.4392	-1.4908
Tangram	BNC	-8.90667	5.24147	0.621	-25.2155	7.4022
	Map	21.24500*	5.24147	0.004	4.9362	37.5538
	Maze	26.65500*	5.24147	<.001	10.3462	42.9638
	DBOX	7.73	5.9762	0.851	-10.8649	26.3249
	SWBD	-7.437	4.56941	0.666	-21.6547	6.7807
	TRAINS	5.19	5.9762	0.975	-13.4049	23.7849
DBOX	BNC	-16.63667	6.41947	0.157	-36.6108	3.3375
	Map	13.515	6.41947	0.37	-6.4592	33.4892
	Maze	18.925	6.41947	0.073	-1.0492	38.8992
	Tangram	-7.73	5.9762	0.851	-26.3249	10.8649
	SWBD	-15.167	5.88354	0.162	-33.4736	3.1396
	TRAINS	-2.54	7.03217	1	-24.4206	19.3406
SWBD	BNC	-1.46967	5.13557	1	-17.449	14.5097
	Map	28.68200*	5.13557	<.001	12.7027	44.6613
	Maze	34.09200*	5.13557	<.001	18.1127	50.0713
	Tangram	7.437	4.56941	0.666	-6.7807	21.6547
	DBOX	15.167	5.88354	0.162	-3.1396	33.4736
	TRAINS	12.627	5.88354	0.348	-5.6796	30.9336
TRAINS	BNC	-14.09667	6.41947	0.321	-34.0708	5.8775

(I) Task	(J) Task	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
	Map	16.055	6.41947	0.188	-3.9192	36.0292
	Maze	21.46500*	6.41947	0.028	1.4908	41.4392
	Tangram	-5.19	5.9762	0.975	-23.7849	13.4049
	DBOX	2.54	7.03217	1	-19.3406	24.4206
	SWBD	-12.627	5.88354	0.348	-30.9336	5.6796

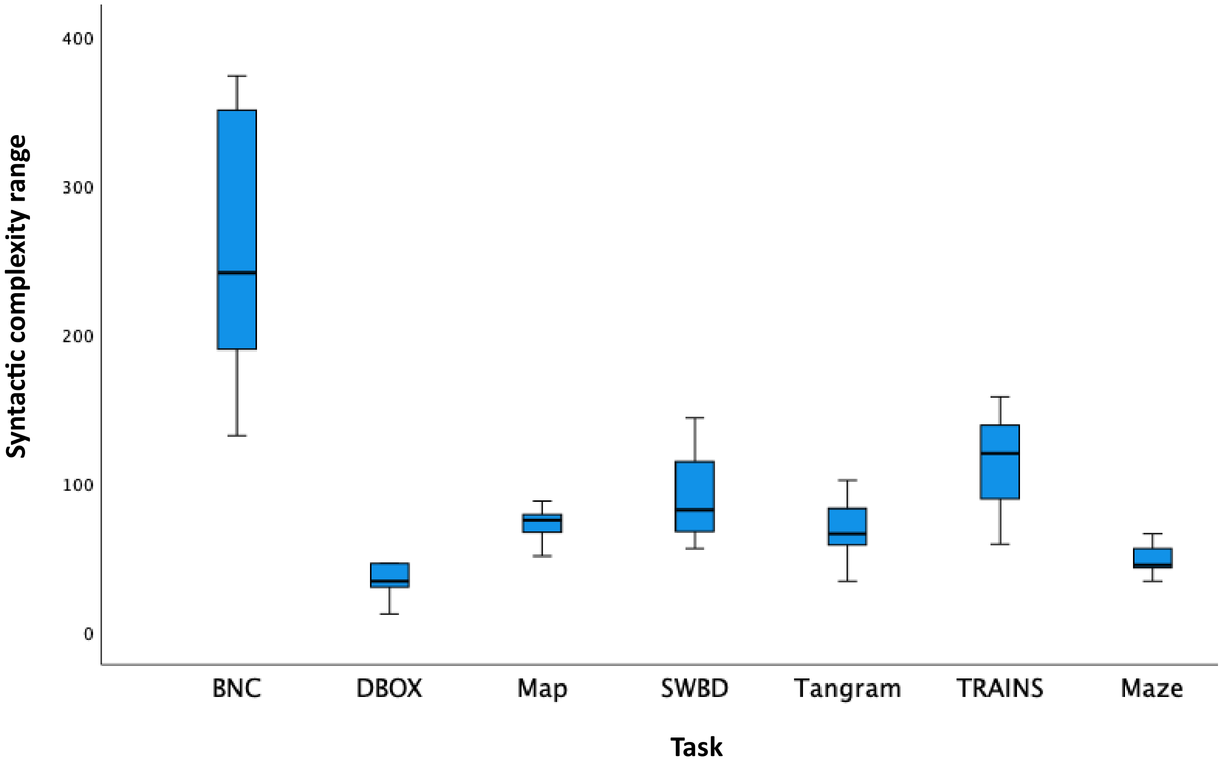
Note. * The mean difference is significant at the 0.05 level.

2. Median range of syntactic complexity

A Kruskal-Wallis test shows a reliable difference in the median range of syntactic complexity across tasks, χ2(6) = 27.96, p = <.001. See Figure 5. The median range of syntactic complexity was largest in the BNC (*Md* = 241.50), then TRAINS (*Md* = 120.00), then SWBD (*Md* = 82.00), then Map task (*Md* = 75.00), then Tangram task (*Md* = 66.00), then Maze task (*Md* = 45), then the smallest range was the DBOX corpus (*Md* = 34).

Figure 5.

Boxplot to illustrate the median range of syntactic complexity across tasks.



3. Range of ISO-standard dialogue act types

The mean percentage contribution of each ISO dialogue act in the transcripts of each corpus is illustrated in Table 6.

Table 6.

Relative distribution of dialogue act types across tasks (%)

ISO dialogue act	SWBD	TRAINS	DBOX	Map
inform	30.79%	16.19%	5.37%	6.43%
agreement	3.16%	0.54%	0.31%	1.47%
disagreement	-	-	-	-
correction	-	-	0.13%	0.43%
answer	3.13%	6.79%	0.56%	4.25%
confirm	0.69%	2.67%	0.96%	1.94%

ISO dialogue act	SWBD	TRAINS	DBOX	Map
disconfirm	-	-	0.49%	0.19%
question	-	-	-	-
set-question	1.81%	6.55%	7.07%	1.34%
propositional question	0.25%	1.14%	5.43%	3.27%
choice-question	-	0.34%	0.25%	0.52%
check-question	0.69%	2.49%	0.13%	4.13%
offer	-	1.41%	0.13%	-
address offer	-	-	-	-
accept offer	-	1.41%	0.13%	-
decline offer	-	-	-	-
promise	-	-	0.19%	-
request	-	0.27%	2.03%	0.27%
address request	-	-	0.63%	-
accept request	-	0.27%	0.89%	9.13%
decline request	-	-	-	-
suggest	-	0.54%	1.45%	0.40%
address suggest	-	-	-	-
accept suggest	-	-	-	-
decline suggest	-	-	-	0.13%
instruct	-	0.34%	0.37%	20.31%
setAnswer	-	-	5.47%	-
Propositional Answer	-	-	4.50%	-
Guess	-	-	1.70%	-
autopositive	7.93%	17.21%	11.65%	22.20%
autonegative	-	-	0.25%	0.49%
allopositive	0.25%	-	2.79%	4.00%
allonegative	-	-	0.13%	0.16%
feedbackelicitation	-	-	-	1.64%
stalling	30.05%	12.82%	16.73%	3.41%
pausing	-	3.27%	1.47%	0.35%
turn take	4.99%	4.68%	7.21%	4.14%
turn grab	0.74%	0.27%	-	1.06%
turn accept	-	6.03%	2.65%	0.13%
turn keep	8.42%	3.89%	11.62%	1.72%
turn give	-	-	-	0.39%
turn release	-	1.08%	-	0.13%
self-correction	4.61%	2.71%	2.25%	2.94%
self-error	-	0.54%	-	-
retraction	1.56%	-	-	-
completion	0.25%	0.34%	-	0.31%
correct misspeaking	-	0.27%	-	-

ISO dialogue act	SWBD	TRAINS	DBOX	Map
init-greeting	-	0.34%	-	-
return greeting	-	-	-	-
init-self-introduction	-	-	-	-
return-self-introduction	-	-	-	-
apology	-	-	0.50%	0.52%
accept apology	-	-	0.31%	-
thanking	-	-	0.30%	-
accept thanking	-	-	-	-
init-goodbye	-	-	-	-
return goodbye	-	-	-	-
opening	0.69%	1.41%	-	0.64%
turn assign	-	1.43%	-	0.51%
Congratulation	-	-	1.14%	-
Closing	-	-	1.78%	-
Contact indication	-	-	-	0.70%
Interaction structuring	-	2.72%	1.01%	0.31%

Discussion

The preceding analysis highlights some common themes in the dialogue tasks currently used in the literature. The most common task type is dyadic information-seeking dialogues. Eristic (debate/argument), persuasive and negotiative tasks have received much less attention. Experimental manipulations tend to be coarse-grained with utterance and word level manipulations relatively rare.

An important limitation of the quantitative results reported here is that relatively few dialogue tasks have published corpora in the public domain. DialogBank is currently the most comprehensive repository (Bunt et al., 2019). As a result direct comparisons were only possible for a small subset of tasks. However, this analysis highlights that organisation of dialogue task characteristics is important, as the quantitative comparisons indicate dialogue is clearly shaped by the task at hand, with even small samples of different tasks (with different characteristics) showing statistically significant differences in quantitative dialogue measures. This highlights the potential impact of task choice when using dialogue task findings and the caution required when making generalisations to domain-independent natural conversation.

References

Allen, J.F., Schubert, L.K., Ferguson, G., Heeman, P.A., Hwang, C.H., Kato, T., Light, M., Martin, N.G., Miller, B.W., Poesio, M., & Traum, D.R. (1994). The TRAINS project: a case study in building a conversational planning agent. *J. Exp. Theor. Artif. Intell., 7*, 7-48.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, *34*(4), 351-366. https://doi.org/10.1177/002383099103400404

Bunt, H., Petukhova, V. V., Chengyu Fang, A., Malchanau, A., & Wijnhoven, K. (2019). The DialogBank: Dialogues with Interoperable Annotations. *Language Resources and Evaluation*, *53*, 213-249. https://doi.org/10.1007/s10579-018-9436-9

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. https://doi.org/10.1016/0010-0277(87)90018-7

Godfrey, J.J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1,* 517-520 vol.1.

Howarth, B., & Anderson, A. H. (2007). Introducing objects in spoken dialogue: The influence of conversational setting and cognitive load on the articulation and use of referring expressions. *Language and Cognitive Processes*, *22*(2), 272-296.

Klein, D., & Manning, C.D. (2003). Accurate Unlexicalized Parsing. *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.3115/1075096.1075150

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. In *International Journal of Corpus Linguistics,* 22(3), pp. 319-344

Nickerson, R.C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*, 336-359. https://doi.org/10.1057/ejis.2012.26

Petukhova, V., Gropp, M., Klakow, D., Eigner, G., Topf, M., Srb, S., Motlícek, P., Potard, B., Dines, J., Deroo, O., Egeler, R., Meinz, U., Liersch, S., & Schmidt, A. (2014). The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. *International Conference on Language Resources and Evaluation*.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243