

Rethinking Conversation: Complexity and Evaluation in the Shifting Landscape of Dialogue Research

CLASP Seminar

Amandine Decker

June 2025



UNIVERSITÉ
DE LORRAINE



UNIVERSITY OF
GOTHENBURG

Introduction

Conversation is everywhere...

- Casual chats, work meetings, ...
- New modes of conversation;
- Modelling remains a challenge.

But it is hard to study.

Face-to-face: coordinate participants, physical space;

Naturalness: suffers in research settings;

Multimodal recording: costly, privacy-sensitive;

Ethical and technical barriers: consent, anonymisation, reproducibility.

In practice we use... whatever we can.

Reddit: large amount, various topics / communities – scraping, not shareable, asynchronous, text-based;

TV shows (e.g. Friends): variety, many situations – naturalness? accuracy of the transcripts?

Real data collection: adapted to task, more natural – slow, costly.

Research Questions

What kind of **data** do we use to study conversation — and for what kinds of **tasks**?

As we tackle **more complex tasks**, how do we **evaluate** them effectively?

What does it mean for **generalisability** when we work with “**degraded**” or **artificial** forms of conversation?

Table of contents

1. Introduction
2. Conversation Research Landscape (2024)
 - Tasks
 - Evaluation
3. A Few Parameters of Conversation
 - Structure & Control
 - Identity
 - Time & Context
 - Social & Interactional Dynamics
4. Discussion

Conversation Research Landscape (2024)

A Snapshot of Conversation Research (ACL Anthology 2024)

ACL Anthology API;

Queried paper titles with: “convers”, “dialog”, “discours”,
“discurs”;

Retrieved: data and data description, task type, evaluation metrics.

Venues

Venue	#	%
TACL	3 (4)	3.2 (4.2)
EMNLP	53 (59)	3.7 (4.1)
ACL	37 (40)	3.9 (4.2)
NAACL	32 (37)	4.4 (5.6)
EACL	14 (17)	5.0 (5.2)
JEP/TALN/RECITAL	9	6.7
ClinicalNLP	5	7.4
INLG	5 (6)	9.3 (11.1)
SemEval	32	11.4
CODI	5 (14)	17.6 (82.4)
SIGDIAL	44 (46)	65.2 (69.7)
Findings	(128)	(5.9)
LREC/COLING	~100 (123)	~6.4 (7.9)

Different Roles of Conversation in NLP Tasks.

Studying and Modelling conversation:

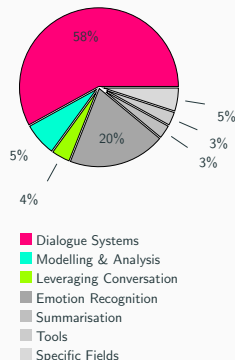
→ *discourse parsing, topic segmentation, turning point identification, breakdown prediction, empathetic alignment, feedback analysis;*

Building or evaluating dialogue systems:

→ *dialogue system, dialogue agent, dialogue state tracking, dialogue generation, QA;*

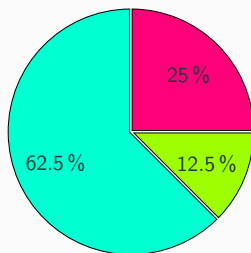
Leveraging conversation for other tasks:

→ *conversational search, conversational recommender system.*



< 7% of all the papers (16):

Dialogue structure: STAC & Molweni;
Others: All different (*Wikipedia*, *News*,
Movie subtitles, *Phone conversations*).



■ Dialogue Structure
■ Interactive features
■ Topics

> 50% of all the papers (dialogue systems, DST, QA):

- Lacking corpus descriptions;

- Almost exclusively task-oriented;

- Reference required;

- LLM-as-a-judge ($\sim 15\%$).

Dialogue Systems and LLMs in the Spotlight

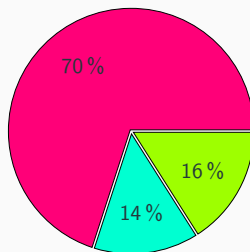
> 50% of all the papers (dialogue systems, DST, QA):

Lacking corpus descriptions;

Almost exclusively task-oriented;

Reference required;

LLM-as-a-judge ($\sim 15\%$).



■ Task-oriented
■ Open-ended
■ Topic Specific

Dialogue Systems and LLMs in the Spotlight

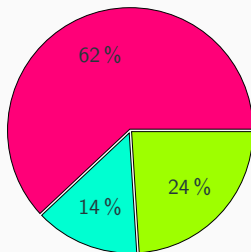
> 50% of all the papers (dialogue systems, DST, QA):

Lacking corpus descriptions;

Almost exclusively task-oriented;

Reference required;

LLM-as-a-judge ($\sim 15\%$).



■ Spontaneous
■ Scripted
■ Machine Generated

Dialogue Systems and LLMs in the Spotlight

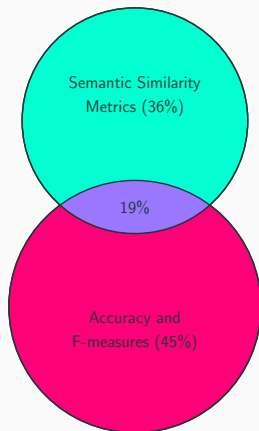
> 50% of all the papers (dialogue systems, DST, QA):

Lacking corpus descriptions;

Almost exclusively task-oriented;

Reference required;

LLM-as-a-judge ($\sim 15\%$).



Dialogue Systems and LLMs in the Spotlight

> 50% of all the papers (dialogue systems, DST, QA):

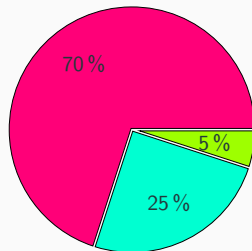
Lacking corpus descriptions;

Almost exclusively task-oriented;

Reference required;

LLM-as-a-judge (~ 15%).

Human Evaluation



■ With (14)
■ Without (5)
■ Bad (1)

Conversation as a facilitator...

Conversational Search (5);

Conversational Recommender System (4);

Conversational Task Solving (1).

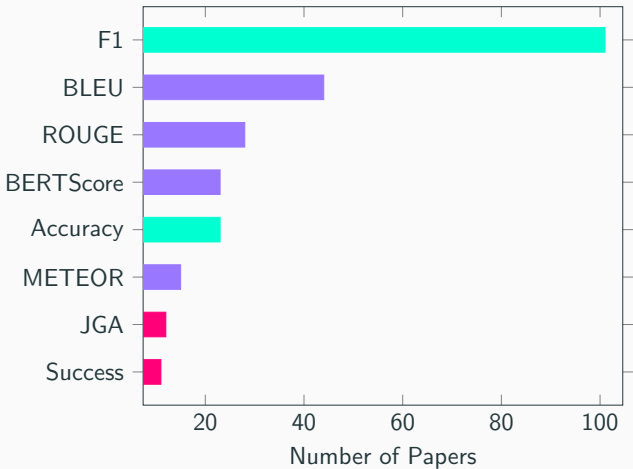
But the conversation is never evaluated:

Conversational Search → task only;

Conversational Recommender System → semantic diversity;

Conversational Task Solving → task only.

Dominant Metrics



BLEU, ROUGE, METEOR, BERTScore:

Initially designed for translation and summarisation;

Now dominate for text generation tasks;

Compute *n-gram overlap* / *semantic similarity*, not *interactional quality*;

Often used without additional human input ($\sim 1/3$).

Beyond-BLEU, Pseudo-Beyond-BLEU, Self-BLEU, BLEURT:

Better at semantic similarity, but still not about *conversation*;

→ Easy to compute but not very informative.

Beyond Metrics: What Do We Really (Want to) Evaluate?

Benchmarks: not always tested, probably in many models' training data;

Ranking: Not a quality assessment;

What do we expect from generative models?

→ What **features** should generative models really **retain**?

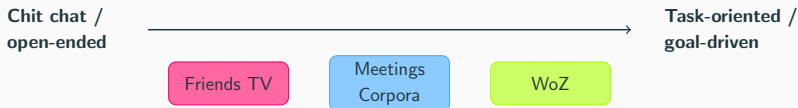
What features are **modified** in the different **corpora** we use?

A Few Parameters of Conversation

Structure & Control – Type and Genre

Overall communicative purpose and style:

→ Shapes lexical and syntactic complexity, politeness strategies, turn-taking norms, ...



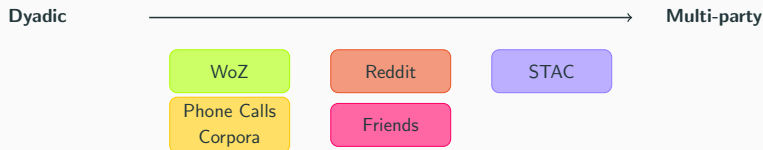
Structure & Control – Number of Speakers

Typical conversation: 2-4 (> 4: *dinner party problem* [5])

Multi-party: more chaotic (topics [10], turn-management, addressee resolution, long-distance attachment [7, 1]), tend to sub-divide [6];

→ Most current models assume dyadic setups.

Preprocessing: “removed meaningless text such as @someone”



Main speaker, addressee, overhearer, side participant, etc. [8, 4]:

Phrase utterances for target audience;

→ Corpus users / analysts = overhearer.

For direct
participants



External audi-
ence

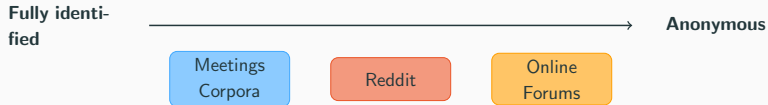
WoZ

English
CallFriend

Friends

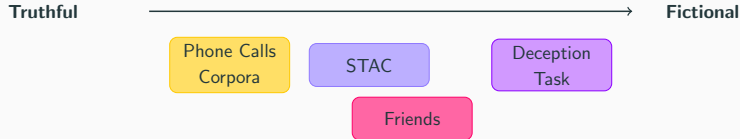
Degree to which participants know each other or are identified:

Influences tone, register, openness, politeness, aggression.



Identity – Layers of Interpretation

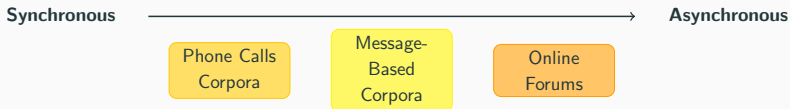
The extent to which participants are “being themselves” and speak about things the way they are [3].



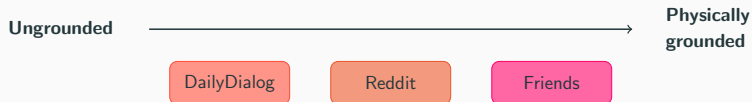
Time & Context – Simultaneity

Synchronous: live adaptation, self-correct and repairs, co-construction;

Asynchronous: more self-edits, no/less live feedback;
→ Strategies to relieve additional complexity.



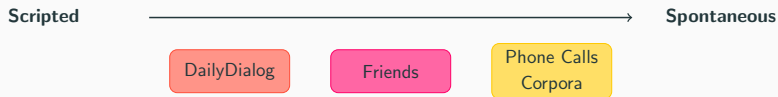
Pointing, real-world references, disturbances [9, 2].



Social & Interactional Dynamics – Spontaneity

Naturalness, disfluencies;

→ Many corpora are cleaned of spontaneity markers.



Social & Interactional Dynamics – Social Aspect of Conversation

The extent to which the conversation serves a practical goal vs. a social one:

Impacts structure;

Challenges usual coherence measures;

→ Approaching the limits of information state update models.

Task-
focused



Social

Booking
Dialogues

Reddit

?

Comparison of Some Corpora

Feature	Friends TV	Reddit Threads	DailyDialog	STAC	MultiWOZ
Type/Genre	Fictional, social	Mixed, informal	Synthetic, daily life	Strategy game chat	Task-oriented
# Speakers	Varies (1-6)	Varies	2	2+	2
Roles	Multiple	Poster, Participants, Readers	Simulated speakers	Game players	User/Agent
Anonymity	Known (scripted)	Anonymous/pseudonymous	Scripted	Pseudonymous	Anonymous
Simultaneity	Synchronous	Asynchronous	Synchronous-like	Synchronous	Synchronous
Spontaneity	Scripted	Natural/spontaneous	Scripted	Spontaneous	Wizard-style
Context Grounding	Physical scene	Limited/shared thread	Vague prompts	Gameboard state	One-sided
Interpretation Layers	High (role-play)	Variable	Minimal	Medium	One-sided
Social Aspect	High	Mixed	Low-Medium	Low(-Medium?)	Low

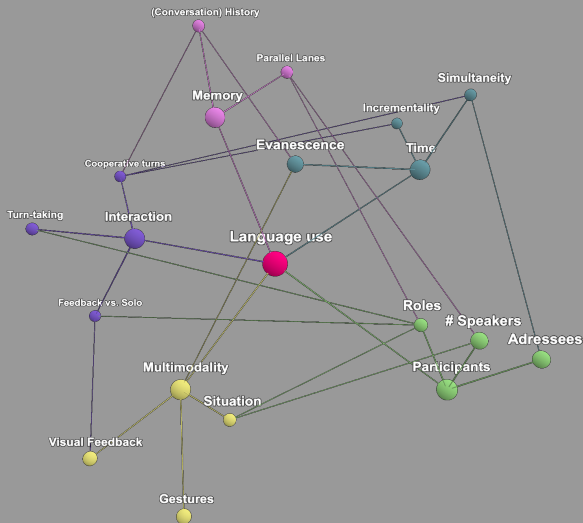
A Personal Example



Image: icon-icons.com

Discussion

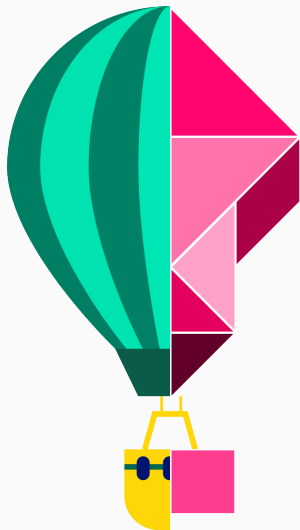
Changing one parameter can change everything.



Evaluation Challenges

- What do we really want to measure and evaluate?
- Reliability of Post-hoc Human Judgement;
- LLM-as-a-judge Trend.

Dialogue Collection Experiment



u2l.fr/montgolgram

Questions / Discussion?



Asher, Nicholas et al. (May 2016). **“Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus”**. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2721–2727.



Chevalier, Fabienne H.G. (2008). **“Unfinished turns in French conversation: How context matters”**. In: *Research on Language and Social Interaction* 41.1, pp. 1–30.



Clark, Herbert H. (1996). **Using language**. Cambridge university press.



Clark, Herbert H. and T. B. Carlson (1982). **“Hearers and speech acts”**. In: *Language* 58, pp. 332–373.



Dunbar, R. I. M., N. D. C. Duncan, and D. Nettle (1995). **“Size and Structure of Freely Forming Conversational Groups”**. In: *Human Nature* 6.1, pp. 67–78.



Fernández, Raquel et al. (June 2008). **“Modelling and Detecting Decisions in Multi-party Dialogue”**. In: *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Ed. by David Schlangen and Beth Ann Hockey. Columbus, Ohio: Association for Computational Linguistics, pp. 156–163.



Ginzburg, Jonathan and Raquel Fernández (June 2005). **“Scaling up from Dialogue to Multilogue: Some Principles and Benchmarks”**. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*. Ed. by Kevin Knight, Hwee Tou Ng, and Kemal Oflazer. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 231–238.



Goffman, Erving (1976). **“Replies and responses”**. In: *Language in Society* 5.3, pp. 257–313.



Stalnaker, Robert (2002). **“Common ground”**. In: *Linguistics and philosophy* 25.5/6, pp. 701–721.



Traum, David (2003). **“Issues in multiparty dialogues”**. In: *Workshop on Agent Communication Languages*. Springer, pp. 201–211.