

Coherence disruptions in human–chatbot interaction:
towards quantitative approach to conversation

Albert Maršík

Contents

| | |
|--|-----------|
| Introduction | 3 |
| 1 Theoretical foundations | 7 |
| 1.1 Textual dimension of conversation | 7 |
| 1.2 Interactional dimension of conversation | 11 |
| 1.3 Disruptions in conversation coherence | 13 |
| 1.3.1 Sources of incoherence in conversation | 14 |
| 1.3.2 What do people do about coherence disruptions? | 15 |
| 2 Experimental framework | 17 |
| 2.1 What are chatbots? | 17 |
| 2.2 Convform | 19 |
| 2.3 Conversation design in theory | 20 |
| 2.3.1 Rule-based approach | 20 |
| 2.3.2 Statistically driven approach | 21 |
| 2.4 Conversation design in practice | 23 |
| 2.4.1 Stimuli | 25 |
| 2.4.2 Ending the conversation | 28 |
| 3 Data | 29 |
| 3.1 Data collection | 29 |
| 3.1.1 Experimental designs | 29 |
| 3.1.2 Participants | 30 |
| 3.2 Data analysis | 30 |
| 3.3 Annotation | 31 |
| 3.3.1 Inter-annotator agreement | 32 |
| 4 Quantitative Analysis | 36 |
| 4.1 Expectations | 36 |
| 4.2 Result | 36 |
| 4.2.1 Ratings | 36 |
| 4.2.2 Stimulus and reaction | 37 |
| 4.2.3 Design rating | 37 |
| 4.2.4 Reaction to stimulus | 39 |
| 4.2.5 Comments | 40 |
| 4.2.6 Conversation style | 40 |
| 4.2.7 Zero anaphora | 41 |

| | | |
|----------|---|-----------|
| 5 | Qualitative observation | 44 |
| 5.1 | Pro-coherence perception | 44 |
| 5.1.1 | Remaps | 44 |
| 5.1.2 | Associative anaphora | 47 |
| 5.1.3 | Sequential coherence | 48 |
| 5.2 | Topical (in)coherence | 49 |
| 5.2.1 | Hard shift refusal | 50 |
| 5.2.2 | Incomplete control over conversation | 51 |
| 5.2.3 | Nontopicalized entities | 52 |
| 5.2.4 | Unexpected refusal | 53 |
| 5.2.5 | Unexpected acceptance | 55 |
| 5.2.6 | Nonfactual implication | 56 |
| 5.3 | Metacommunication | 56 |
| 5.4 | Conclusion | 58 |
| 6 | Final thoughts | 59 |
| 6.1 | Evaluation of anaphora categorization | 59 |
| 6.2 | Further steps | 59 |
| 6.2.1 | Experiment design improvement | 59 |
| 6.2.2 | Stimulus generation improvement | 60 |
| 6.2.3 | Hybrid approach | 60 |
| 6.3 | New opportunities | 60 |
| 6.4 | Conclusion | 61 |

Introduction

Recently, there has been a breakthrough in the way we interact with machines¹. We can now instruct a computer using natural language². Besides making existing technology an extra step accessible, new ways to use technology appear. Being able to simply talk to a machine and have it respond can help overcome longstanding challenges such as notably accessing a knowledge base via semantic search³. Until recently a knowledge base would usually be accessed via fulltext, meaning we would only be able to find information of which we knew a part of the formal encoding. Today, we can search for information simply by asking questions, including follow up ones all thanks to natural language computer interface.

The promise of much practical usage of the current wave of generative AI is ambitious and only brings its fruit slowly, perhaps slower, than was expected⁴. There is talk of a "plateau" in development of the technology powering the current cutting edge inventions⁵. That besides the current day issues of cutting edge AI like high electricity consumption⁶ and unpredictable and broad societal impact⁷

That being said, in context of conversation research, this development in technology promises to make things possible that previously were not. With a partial control of what happens in the conversation and a decent certainty, that our system will simulate human-human conversation to the user, new kind of conversational data is in reach – logs of the human-bot interaction, that could be categorized based on

- which researcher controlled stimulus and

¹D. Sharma et al. “Exploring The Evolution Of Chatgpt: From Origin To Revolutionary Influence”. In: *Educational Administration: Theory and Practice* 30.5 (2024).

²G. G. Hendrix. “Natural-language interface”. In: *American Journal of Computational Linguistics* 8.2 (1982), pp. 56–61.

³E. Makela. “Survey of semantic search research”. In: *Proceedings of the seminar on knowledge management on the semantic web*. Department of Computer Science, University of Helsinki, Helsinki. 2005.

⁴R. Metz et al. “OpenAI, Google, and Anthropic Are Struggling to Build More Advanced AI”. in: *Bloomberg News* (2024). Accessed: 2024-12-19. URL: <https://www.bloomberg.com/news/articles/2024-11-13/openai-google-and-anthropic-are-struggling-to-build-more-advanced-ai>; K. Hu and A. Tong. “OpenAI Rivals Seek New Path to Smarter AI as Current Methods Hit Limitations”. In: *Reuters* (2024). URL: <https://www.reuters.com/technology/artificial-intelligence/openai-rivals-seek-new-path-smarter-ai-current-methods-hit-limitations-2024-11-11>.

⁵G. Ritter and W. Lu. “The AI Plateau Is Real — How We Jump To The Next Breakthrough”. In: (Dec. 2024). URL: <https://www.emcap.com/thoughts/ai-s-curve-plateau-proprietary-business-data-breakthrough/>.

⁶H. Ritchie. “What’s the impact of artificial intelligence on energy demand?” In: *Sustainability by Numbers* (Nov. 2024). URL: <https://www.sustainabilitybynumbers.com/p/ai-energy-demand>.

⁷A Hagerty and I. Rubinov. “Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence”. In: *arXiv preprint arXiv:1907.07892* (2019); M. T. Baldassarre et al. “The social impact of generative ai: An analysis on chatgpt”. In: *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 2023, pp. 363–373.

- which participant reaction to given stimulus

they contain.

In the 1960s the relatively recent emergence and adoption of telephone technology allowed for recording and transcribing authentic conversational data. This advancement took place thanks to development in technology which is reminiscent of the current day situation. While human–bot conversational data is arguably less authentic than telephone conversation transcripts, experimental approach can be taken while the human element is present.

This paper's intention is to provide a debate on a metaresearch question – is using generative AI a viable methodology for conversation research? This is done by attempting to develop that very methodology. Proceeding we operate in a frontier, our first steps should be establishing data backed baseline knowledge and assessing possible lines of research.

Understanding what makes the unraveling text of a conversation a coherent one can be approached via obtaining conversational data containing coherence disruptions. This can be done using the discussed technology – it has the capability of conversing in a way that is found generally acceptable by humans and can drift away from the coherent interaction if appropriately instructed to do so.

The data this paper seeks to elicitate and analyse are actual human–chatbot exchanges containing moments which have the potential to be problematic for the human participant to process and follow up on. The boundary between what a coherent and an incoherent conversation is blurred. It is in no way a binary property of the text of the conversation⁸. The goal is therefore to touch on the gradual divide between them.

While chatbots are evaluated for how natural and error free their way of conversing is, human–human conversation is rarely flawless as errors happen and conversational coherence gets temporarily disrupted. In case of human–human communication, disruptions can however be cured easily⁹.

In human–bot communication, disruptions could derail a conversation completely, leaving the bot, who would only rely on surface level textual clues, in the dark.¹⁰ This has become rare with generative AI. Even though it brings a set of its own problems like frequently lacking factuality or the difficulty to handle data responsibly, the cutting edge technology powered conversation systems are

- better capable of understanding and producing relevant answers
- able to return to their conversational point of departure

Human–bot communication is often single–purpose. Companies and institutions deploy voice applications to interact with customers and clients, so there is usually a goal to be achieved. The coherence in each moment of such conversation can then be described based on whether the goal is currently coming closer to being achieved with ease. Another common frame for a human–bot interaction is an open–domain conversation, also

⁸T. Givón. *Coherence*. John Benjamins Publishing Company, 2020.

⁹Wolfram Bublitz and Uta Lenk. “Disturbed coherence: ‘Fill me in’”. In: *PRAGMATICS AND BEYOND NEW SERIES* (1999), pp. 153–174; M. Dingemanse and N. J. Enfield. “Interactive repair and the foundations of language”. In: *Trends in Cognitive Sciences* 28.1 (2024), pp. 30–42.

¹⁰M. F. McTear. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer International Publishing, 2020, pp. 43–70.

known as chit-chat or smalltalk¹¹. Since there is no global goal to achieve, the coherence of such interaction is dictated by a different set of factors¹²

Some factors that influence coherence in conversational texts, whether in human-human or human-bot exchanges, have been extensively studied.

Namely:

Politeness Brown and Levinson’s work on politeness strategies describes social alignment in smooth interactions¹³. Politeness strategies, such as using polite language, offering options, or softening potentially face-threatening comments, help to create a comfortable communicative environment. These strategies align with social norms, which people interpret as markers of respect, consideration, or even trust. A failure to employ these politeness strategies, or using them inconsistently, can disrupt conversational coherence. For example, blunt or overly direct responses may be perceived as abrupt or rude, diverting the conversation’s flow or causing discomfort. In such cases, the breakdown of polite norms can lead participants to question intent, hindering effective and smooth communication.

Speech acts Following Austin and Searle’s speech act theories, communication rely on expressing clear intentions and meanings that help build mutual understanding¹⁴¹⁵. When speakers convey intentions explicitly through statements, questions, requests, or assertions, it signals to listeners the purpose and direction of the conversation. Effective communication strategies help maintain coherence by ensuring each contribution builds logically on the last. On the other hand, unclear intentions or ambiguous phrasing can create misunderstandings, disrupting the conversation’s flow. Misalignment or mixed signals – such as using sarcasm without cues or making indirect requests without context—can leave listeners uncertain about how to respond, leading to off-track or irrelevant contributions and possibly creating need to address the communication to regain understanding.

Conversational Maxims Grice’s conversational maxims are fundamental to coherent dialogue¹⁶. They suggest that participants should:

- provide truthful information (Quality)
- neither too much nor too little (Quantity)
- remain on-topic (Relevance)
- communicate in an orderly, clear manner (Manner)

These maxims encourage effective exchange by setting a standard for contributions that are informative, truthful, relevant, and unambiguous. When violated, such as by

¹¹R. Ramnauth, D. Bršćić, and B. Scassellati. *More than Chit-Chat: Developing Robots for Small-Talk Interactions*. 2024. arXiv: 2412.18023 [cs.R0].

¹²Jun Xu et al. “Discovering Dialog Structure Graph for Coherent Dialog Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong et al. Association for Computational Linguistics, Aug. 2021, pp. 1726–1739.

¹³P. Brown and S.C. Levinson. *Politeness: Some Universals in Language Usage*. Politeness: Some Universals in Language Usage. Cambridge University Press, 1987.

¹⁴J. L. Austin. *How to do things with words*. Oxford University Press, 1962.

¹⁵J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

¹⁶H. P. Grice. “Logic and Conversation”. In: *Syntax and Semantics: Vol. 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. Academic Press, 1975, pp. 41–58.

giving excessive detail, omitting important context, or straying from the topic, coherence suffers. For instance, irrelevant tangents or over-detailed explanations may confuse the listener as to what is the main focal point of conversation in that moment. This misalignment can leave participants uncertain about the conversation's direction, ultimately diminishing coherence and the effectiveness of communication.

Sequence Structure The work of Schegloff and Sacks on sequence structure and turn-taking emphasizes that ordered interactions support predictability and continuity in dialogue¹⁷. Turn-taking conventions — where participants follow an implicit sequence of speaking and responding — help maintain the flow by structuring the conversation in a logical order. This sequence structure allows both parties to anticipate when to listen and when to speak, contributing to a well-paced, cohesive exchange. However, interruptions, abrupt changes in topic, or skipping expected responses can disrupt this sequence, introducing unpredictability that can confuse participants. These interruptions fragment coherence by shifting the conversation away from expected responses or structured flow, often leaving gaps in understanding or causing conversational breakdowns.

Message and Topic Interactional linguistics underscores that consistency in message and topic preserves continuity in conversation¹⁸. When speakers stick to a shared topic or make gradual, clear shifts, coherence is maintained because participants know what to expect. Frequent or abrupt topic shifts, however, or sending unclear or conflicting messages, can create disjointed exchanges. For instance, introducing a new topic without closure on the previous one can confuse listeners, leading to a scattered or fragmented interaction.

While all of the mentioned areas unveil much about the way conversation works, rarely do they concern themselves with the textual dimension of conversation. Most of the mentioned authors (with the notable exception of those operating within the interactional linguistics framework) could hardly be described as linguists, though their works significantly inform linguistics.

The lack of a true interpersonal dimension in human-chatbot communication allows to focus solely on the elements in conversational text, that make it cohesive and coherent or rather those that have the potential to prevent it from being that. The key concepts discussed in this paper are two closely related topics:

Coreference realized by anaphora and topic – what the text is about.

¹⁷E. A. Schegloff. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press, 2007.

¹⁸E. Couper-Kuhlen and M. Selting. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge University Press, 2017.

Chapter 1

Theoretical foundations

1.1 Textual dimension of conversation

The following concepts will be explored individually, in relation to one another and in relation to conversation: text, coherence, cohesion, coreference, anaphora, cataphora, endophora, exophora, topic, entity, and association. While the presented exploration draws on existing literature, it seeks to establish an independent and sustainable framework, rather than strictly adhering to established interpretations.

Text

Text, in its broadest sense, refers to any form of communication that conveys meaning through a combination of signs, symbols, or language¹². These semiotic structures can take various forms, including written, spoken, visual, or even non-verbal modes of expression³. A text can be as simple as a single sentence or as complex as a novel, and it can exist across different mediums, from books and articles to advertisements and digital content. What defines a text is its ability to convey a coherent message or idea, often intended for interpretation by an audience or an addressee. Texts can serve a wide range of purposes, including storytelling, instruction, persuasion, or simply recording information. Typically text is a structure that is linguistic, produced and perceived as intentional and coherent.

The text of a conversation is specific because it is multiproducer. Another example of a multiproducer text would be a sequence of commercial signs on a busy street. It is the spatial juxtaposition of the signs and temporal juxtaposition of utterances, that make them a text.

Another property of a conversation text is it is negotiated⁴. This is given by its multiproducer and temporal nature. Other types of text which are also negotiated are relatively rare. There are occurrences of debates which take place in written text, whether they are press columns or academic articles, which interact explicitly with each other, making them a negotiation. Such press discourse could however be considered a sequence of text units rather than a single temporarily juxtaposed text. This perspective could

¹J. Hrbáček. *Nárys textové syntaxe spisovné češtiny*. Praha: Trizonia, 1994, p. 7.

²L. Hjelmslev. *O základech teorie jazyka*. Trans. by F. Čermák. 2. doplněné a upravené vydání. Praha: Academia, 2016.

³R. Barthes and S. Heath. *Image, Music, Text*. A fontana original. Fontana Press, 1977, p. 13.

⁴H. Sacks. *Lectures on Conversation*. Ed. by G. Jefferson. Oxford: Basil Blackwell, 1992.

hardly be defended in regards to conversation, because its tight temporal coupling and cohesion, making conversation a unique phenomena.

Coherence

Coherence refers to the logical connections and consistent relationships that make a text easy to follow and possible to understand⁵⁶. It is achieved when the ideas, sentences, and paragraphs within a text are linked together in a meaningful way, allowing the reader to grasp the author's message without confusion. Coherence often depends on the use of transitions, the logical flow of arguments, and the proper sequencing of information. It ensures that each part of the text contributes to the overall meaning, creating a unified whole⁷. Incoherent text can be difficult or impossible to understand, even if the individual sentences are grammatically correct⁸. It is a property of the whole text, but textual elements can be pointed out that contribute to or diminish the given texts coherence. Those elements are however not referred to as 'coherence elements'.

Coherence is a cognitive phenomenon⁹ because it involves the mental processes of interpreting, organizing, and understanding information. When reading a text, coherence arises not only from the structure and linguistic cues provided by the author but also from the reader's ability to make connections between ideas based on prior knowledge, expectations, and context. This cognitive interaction between the text and the reader's mind is what makes the content understandable.

In conversation, coherence becomes even more complex, as multiple participants are simultaneously contributing to and interpreting the flow of information. Each individual brings their own perspective and understanding to the interaction, which requires constant negotiation to maintain coherence. Misunderstandings, different backgrounds, and interruptions can disrupt the coherence of a conversation, making it a more dynamic and fragile process compared to written text.

- whether a written text is coherent depends mostly on the reader¹⁰¹¹
- whether a conversation text is coherent depends on an ongoing negotiation

Coherence is a scalar property rather than a binary one¹². It is also tricky to measure. This paper seeks to explore one possible approach of declaring different levels of coherence disruptions and observing the acceptance rates in participants and correlation between them.

Cohesion

While coherence refers to the interpretative quality of a text, wherein the ideas form a logical and meaningful whole cohesion, focuses on the structural relations within a text,

⁵Givón, *Coherence*, p. 83.

⁶Hrbáček, *Nárys textové syntaxe spisovné češtiny*, p. 9.

⁷Hrbáček, *Nárys textové syntaxe spisovné češtiny*, p. 28.

⁸Hrbáček, *Nárys textové syntaxe spisovné češtiny*, p. 30.

⁹R. M. Roberts and R. J. Kreuz. "Nonstandard discourse and its coherence". In: *Discourse Processes* 16.4 (1993), pp. 451–464.

¹⁰Bublitz and Lenk, "Disturbed coherence: 'Fill me in'".

¹¹Roberts and Kreuz, "Nonstandard discourse and its coherence".

¹²Givón, *Coherence*.

achieved through grammatical and lexical links. It should be seen as an umbrella term covering specific relations within the structure of the text, where cohesive elements can be directly pointed out. While coherent text does not necessarily need to be cohesive, cohesive elements often support it. A coherent text tends to be at least somewhat cohesive.

Halliday and Hasan¹³ developed a detailed framework of cohesion, which includes endophoric references, relating parts of the text to each other, and exophoric references, which point outside the text¹⁴. Endophoric cohesion covers aspects like anaphoric references and cataphoric references¹⁵. Exophoric references, however, rely on shared context beyond the text itself, requiring readers to use prior knowledge. Following concepts can be considered cohesive elements.

Cataphore and Exophore

In Halliday and Hasan’s framework, cohesion in language is achieved through various devices that connect different parts of a text, forming a unified whole. They classify cohesive ties as references, substitutive forms, ellipsis, and connectors, with anaphoric references being one of the primary ways texts achieve cohesion¹⁶. When a text element cannot be mapped to a preceding referent, Halliday and Hasan suggest that cohesion is maintained through shared situational understanding, making the reference exophoric. Cataphoric references, though less common, involve elements that look forward in the text, showing intentionality by the author but contributing to cohesion primarily through the eventual resolution of the forward-pointing referent.

In conversation if a seemingly anaphoric text element is not successfully mapped to a preceding textual coreferent the reference can still be understood, because shared context. Such element reaches out of the text with its reference, making it an exophoric one. Cataphore is a related phenomena – a reference which points forward in the text. Such occurrence is relatively rare in written text and even more so in conversation. In fact it is somehow pointless to account for cataphore in a multi-producer text. A cataphore denotes an authors intention to reveal the nature of a referent explicitly after first mentioning them. In conversation, where multiple contributors cocreate given text, and mutual understanding and agreement is the measure of how coherent the produced text is, later realisation of a vague reference does not contribute to how coherent it is. Regardless, in case of a cataphore, only the referent is a cohesive element, not the cataphore, as it ties back to the previous text, creating bonds across large textual units.

Anaphora, Endophore and Coreference

A common cohesive text element is an anaphora¹⁷. It is a reference inside the text pointing back to a previously mentioned entity. Often it is realised via personal pronouns. Though

¹³M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, 1976.

¹⁴Halliday and Hasan, *Cohesion in English*, p. 31.

¹⁵E. Hajičová, J. Havelka, and P. Sgall. “Discourse Semantics and the Salience of Referents”. In: *Journal of Slavic Linguistics* 11.1 (2003), pp. 127–140; S. Loaiciga, S. Dobnik, and D. Schlangen. “Anaphoric Phenomena in Situated dialog: A First Round of Annotations”. In: *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Oct. 2022, pp. 31–37. URL: <https://aclanthology.org/2022.crac-1.4/>.

¹⁶Halliday and Hasan, *Cohesion in English*, p. 68.

¹⁷A. Nedoluzhko. “Rozšířená textová koreference a asociační anafora (koncepte anotace českých dat v pražském závislostním korpusu)”. PhD thesis. Prague: Univerzita Karlova, Filozofická fakulta, Ústav českého jazyka a teorie komunikace, 2010.

there are other ways for anaphora to realise. In Czech, anaphoric references often rely on grammatical gender and number, making participial endings essential for identifying the referent. For instance, when a grammatically masculine entity is mentioned, later references might use a participle in the masculine form, such as *šel* ("he went"), connecting back to it without repeating the noun or using a demonstrative. Demonstratives, such as *ten* ("that") or *tento* ("this one"), also frequently serve anaphoric functions, guiding the reader to a previously mentioned subject. Temporal and locative adverbs, such as *tam* ("there") and *tehdy* ("then"), also contribute cohesion by indirectly referencing time and place details introduced earlier in the text. These anaphoric elements strengthen textual coherence by reducing redundancy and maintaining flow. The reader identifies coreferential links through these markers, following the cohesive threads without needing explicit repetitions.

An anaphoric element is by definition also endophoric. It points inside the text it appears in. By definition an anaphoric element has a referent, which occurs earlier in the text. These two elements are then coreferent. As such they also share an identical exophoric reference – they point outside of the text.

In conversation, many aspects of which are subject to negotiation, also specific coreference relations can be questioned¹⁸. The reference realised by one communication participant may be unclear to the other resulting in a repair request coming from another participant. In conversation analysis, Sacks's concept of repair traditionally addresses misunderstandings related to intentions and actions, loosely drawing on frameworks like Austin's and Searle's speech act theories. From this perspective, repairs often target interpretative gaps about what a speaker intends to do with their utterance. However, viewed from a broader, more abstract level, what is called repair triggers can extend beyond intentions alone, encompassing issues on the textual level as well. For instance, an nonassignable anaphora — a reference that lacks a clear antecedent — may lead to a repair request, thereby showing how textual ambiguities prompt interactional responses. This approach expands the causes of repair in conversation, integrating elements of reference and interactional misalignment, where a structural aspect of the language itself can become a repairable issue in the communicative exchange.

Topic

Topic is what a text is about. That makes topic very complicated to define. Among others, some issues with topic and annotating it in text are:

- A text can and typically does cover multiple topics
- Different framing will produce different topic annotations of text
- The span of a topic section can be impossible to delimit within text.
- Topic annotation is by its nature always more text, so even it can be annotated for topic. making topic annotations recursive. One cannot therefore achieve a definite topic description of a text.

Despite all these complications, topic cannot be skipped in conversation research as it is deeply intertwined with the concepts mentioned above. Topic progressions across

¹⁸S. Loaiciga, Simon D., and D. Schlangen. "Reference and coreference in situated dialogue". In: *Proceedings of the Second Workshop on Advances in Language and Vision Research*. 2021, pp. 39–44.

text are realised via anaphora and association and tightly interact with coherence. An sufficient amount of time has to be spent on a given topic unit, enough information has to be said about a given topic in order for it to be possible to move on or add another one in the conversation. Closure has to be provided in order for a topic to be done. Transitioning from one topic to another has a potential to disrupt coherence, if the association between the topics is too distant. A divergence in topic has to be justified.

Association

Association is a textual realization of an isotopic relation¹⁹. By their exophoric properties, referents exist in a semantic web of relationships. Similarly to coherence, associative relationships are a cognitive phenomena. They come to exist when they are perceived. While association is a cohesive element it is difficult to formalize the way it can and has been done with anaphoric text relations. It is however a major factor in a coherence of text as in some cases a text can only rely on association in its coherence.

Entity

An entity is an exophoric referent, descriptions of objects, people, events etc²⁰. Words or text elements which can be referred to by an anaphora will be called entities. Since a phrase containing an anaphora typically adds more information about the referent the new information must be semantically compatible, in other words association has to be possible between the referent and the added information. Entity also has to do with topic. In text topic can be represented by a single or multiple entities. Coreferent words will be regarded as a single entity. It can serve to partially map a topic distance in the texts chronology.

1.2 Interactional dimension of conversation

A conversation text is produced by multiple producers. This complicates things:

- Conversation is an interactive process, distinct from static text, which is created collaboratively.
- Conversational content is continuously negotiated by participants, who continuously adapt one another.

Due to its temporal and cooperative nature, conversation allows for:

- Overlaps in speech,
- Swift corrections of minor errors,
- Multiple layers of perspective, including:
 - Each participant's personal viewpoint,

¹⁹T. Koblížek. *Interpretační sémantika. Úvod do textové teorie Françoise Rastiera*. Univerzita Karlova v Praze, Filozofická fakulta, 2015.

²⁰D. Ayuso. "Discourse entities in Janus". In: *27th Annual Meeting of the Association for Computational Linguistics*. 1989, pp. 243–250.

- Each participant’s perception of others’ viewpoints,
- Each participant’s understanding of the shared conversation as it’s being co-created.

each of these perspectives can desynchronize resulting in misunderstandings. Humans however are excellent at correcting misunderstandings this is because under regular circumstances, people cooperate.

The Cooperation Principle, introduced by philosopher H.P. Grice, suggests that participants in a conversation typically work together to achieve effective communication. Grice proposed that, to ensure this cooperation, speakers follow four conversational maxims. In practice, people may not always follow these maxims but they do so in ways that still rely on shared expectations of cooperation. Even when misunderstandings arise, humans naturally engage in conversational repair, using their social intuition and mutual cooperativity to clarify intention and realign perspectives.

Contemporary conversation research can be understood to draw from conversation analysis. Modern conversation research traces its roots to conversation analysis, a field pioneered by sociologists Harvey Sacks and Emanuel Schegloff in the 1960s. They sought to understand the structure and social rules of everyday interactions, focusing on the patterns and norms that govern turn-taking and response. Thanks to recordings of phonecalls, transcripts could be qualitatively analyzed. This research has led to coining new terminology.

Adjacency pair

Adjacency pairs describe sequences of two related utterances by different speakers²¹. These pairs are characterized by their predictable and reciprocal nature, where the first part sets up the expectation for a specific type of response. Common examples include greetings ('Hi' → 'Hello'), questions and answers ('What time is it?' → '3 PM'), or offers and acceptances/declines ('Would you like some coffee?' → 'Yes, please' or 'No, thank you').

Sequence structure

Sequence structure refers to the organization of conversational turns into coherent patterns or sequences. It describes how interactions are shaped by predictable structures, such as adjacency pairs. These sequences provide order and meaning to conversations, guiding participants in understanding when and how to respond. Schegloff²² emphasized that sequence structure is central to the social organization of talk, as it allows participants to manage and negotiate interaction effectively.

Topic shading

Topic shading, as discussed by Sacks²³, refers to the subtle way in which a conversation naturally shifts from one topic to another while maintaining coherence. Instead of abruptly changing the subject, speakers introduce a related idea or concept, gradually

²¹Sacks, *Lectures on Conversation*, p. 188.

²²E. A. Schegloff. "On the organization of sequences as a source of 'coherence' in talk-in-interaction". In: *Conversational Organization and its Development*. Ed. by B Dorval. Ablex, 1990, pp. 51–77.

²³E. A. Schegloff and H. Sacks. "Opening up Closings". In: *Semiotica* 8.4 (1973), pp. 289–327.

steering the discussion in a new direction. This process allows for smooth transitions in dialogue, helping participants maintain engagement and avoid confusion.

Dis/preferred answers

Preferred answers, according to Sacks²⁴, are responses in conversations that align with social norms and expectations, making interactions smoother and more cooperative. In conversation analysis, preferred answers typically follow the format or intent of the preceding question or statement. They contrast with "dispreferred" answers, which might include refusals or disagreements and often require additional explanation or mitigation to maintain social harmony.

Conversational repair

Conversational repairs refer to how participants address and resolve problems in understanding, hearing, or speaking during interactions²⁵. These issues, can occur at any point in a conversation. Repairs are classified into self-repair, where the speaker corrects their own error, and other-repair, where a different participant addresses the issue. They can further be classified into self-initiated repair and other-initiated repair.

As a descendant of conversation analysis interaction linguistics has emerged, building on its insights to examine language use in social contexts. It broadens the focus to study not only verbal exchanges but also multimodal cues like gestures, gaze, and intonation, analyzing how these elements contribute to meaning. Interaction linguistics aims to understand the dynamic aspects of conversations, such as how topics shift and how sequences of speech acts unfold, reflecting the fluid nature of human communication.

1.3 Disruptions in conversation coherence

While the question of what makes for a coherent text is too broad, the answer to what makes for a coherent conversation can be somewhat easier to answer. Because conversation participants negotiate understanding, it is up to them, when a conversation is and is not coherent to describe what a coherent conversation is, it is worth pursuing the moments, when the conversation stops flowing with ease. Such moments can be called coherence disruptions. A coherence disruption is a complex phenomena as

- it penetrates through some or all of mentioned perspectives on an on-going conversation
- it can't be evaluated in a binary fashion

There are different degrees to which a conversation coherence can be disrupted:

- if a participant suddenly starts speaking in an a way that can hardly be considered interaction due to its irrelevance or

²⁴Sacks, *Lectures on Conversation*, p. 410.

²⁵E. A. Schegloff, G. Jefferson, and H. Sacks. "The Preference for Self-Correction in the Organization of Repair in Conversation". In: *Language* 53.2 (1977), pp. 361–382.

- if the utterance simply is not grammatical or understandable, while the conversation has been compromised and becomes incoherent, it has more to do with incoherent written text, because the incoherence is encapsulated on the level of a single utterance

Roberts²⁶ Discusses various types of incoherent text. He exemplifies so called giberish as incoherent text that is absent of structural relations. On the other hand he discusses experimental theater or literature as a type of text which is assumed to be coherent in the sense that there is an intention behind it but contains little to no structural relations. Lastly he mentions a so called "schizophrenic discourse" as a speech that is not assumed to be coherent even if it has structural relations to it. In any case Roberts definitively states that coherence is assumed and is therefore a receptive phenomena. The incoherence Roberts discusses is considerably different from when the source of incoherence stems from the structure of the conversational text or relationship between different utterances – this is when another participant assesses, they are simply speaking leading a different conversation perhaps with a differing intention or that they are conversing under differing set of circumstances which manifests formally in the linguistic fabric of the conversation – its text. All that despite everyone included being cooperative.

1.3.1 Sources of incoherence in conversation

Schegloff shows how incoherence arises when people interpret sequence structure differently, namely in terms of which turn is seen as an answer to which previously occurring turn. In his example, the participants misread each other's intentions, leading to confusion about how their turns fit together. They each project different expectations for how the conversation should unfold, which causes misaligned sequence structure interpretations. When this happens, they turn to brief metacommunication — comments about the conversation itself to try to clarify and re-align their understanding. Schegloff illustrates how these efforts to "repair" the misalignment are central to managing and resolving incoherent moments in conversation.

Coherence disruptions are also discussed in literature. Hrbáček's approach to coherence and cohesion in text distinguishes the two concepts, noting how they often interact but can also be independent. He highlights that while cohesion involves grammatical or lexical links that make sentences flow together, coherence relies on the logical and meaningful progression of ideas. This means that a text could be cohesive – using connectives, repetitions, and consistent lexical choices – yet lack coherence if the sequence of ideas doesn't make logical sense or follow a clear progression. Conversely, a text may be coherent in its narrative flow without relying heavily on cohesive devices. In Czech linguistics, the distinction between *téma* (theme) and *réma* (rheme), as used by Daneš, underlines the role of topic progression. Hrbáček illustrates this by discussing examples where a story progresses logically from one point to the next while being incoherent despite being clear about its topic structure due to never coming back to a previously mentioned topic.

Two kinds of phenomena are at hand when it comes to ways in which conversation coherence can be disrupted – topic shifts and nonassignable anaphora²⁷. While not unique to conversation both take on specific forms in it worth looking at.

²⁶Roberts and Kreuz, "Nonstandard discourse and its coherence".

²⁷Bublitz and Lenk, "Disturbed coherence: 'Fill me in'".

Topic shifts

When conversations shift abruptly from one topic to another, it can create confusion for the conversation partner. They might find themselves trying to reconnect to the previous discussion or wondering how the new subject relates. This can lead to misunderstandings as the transition can feel jarring.

One interesting question is, how do we determine when a topic has run its course? What common traits do conversations share when a subject is truly exhausted? Perhaps observing transcripts could reveal repeating patterns in topic progression or sequence structure.

Moreover, what makes for a smooth transition between topics? Is it related to the cues participants give each other, or perhaps the context of the discussion? How do we navigate the flow of conversation and what indicates a natural shift versus a disruptive one?

Nonassignable anaphora

nonassignable anaphora is closely tied to topic progression. Currently established topic or topics help assigning anaphora and determining between an anaphora and an exophore. Even if an anaphoric device is not assignable, and the reference is presumably an exophoric one, The reason for employing this reference must be relevant to an established topic. In conversation meaning of demonstratives is to be negotiated. If an anaphoras assignability causes confusion, chances are it is caused by one of the following

- there are no relevant assignment candidates
this situation can be understood as a vague or unjustified exophore
- there are multiple equally relevant candidates
- candidate has occurred in the conversation text too long ago
can be understood as an abrupt return to previously established topic

1.3.2 What do people do about coherence disruptions?

In conversation, coherence disruptions often prompt participants to employ strategies to maintain understanding and flow. Schegloff suggests that people manage these disruptions through interactive repair or inference. Interactive repair often involves explicitly addressing misunderstandings or clarifying intentions, often by rephrasing or asking questions. Interactive repair refers to immediate, collaborative corrections within dialogue, where one speaker might correct the other or themselves to enhance clarity. Inference and pragmatic reasoning, the most seamless methods, allow participants to fill gaps based on context and social cues, helping conversations continue smoothly without explicit repair.

Dingemanse and Enfield²⁸ echoes this from a cognitive perspective, highlighting how inference and pragmatic reasoning are particularly effective. Participants rely on shared understanding and contextual knowledge to interpret ambiguous statements. Together, they use both explicit (metacommunication and repair) and implicit (inference and reasoning) methods work to restore coherence.

²⁸Dingemanse and Enfield, “Interactive repair and the foundations of language”.

It needs to be noted however that both interactive repair and reasoning are deployed in a number of other contexts other than conversation coherence disruption. Inference takes place constantly²⁹. Each of those moments could be hardly considered a coherence disruption. There is however always potential for it, particularly via unclear or nonassignable anaphora or abrupt unjustified topic shifts. Repair and metacommunication also takes place in a mutually informed and synchronized interaction. It is for example deployed when it is revealed that the interaction participants intentions or opinions differ.

These uses of interaction management are however hardly possible to analyse on a textual level since they do not cooccur with coherence disruptions. What can be observed are – as mentioned above – troublesome anaforic references and topic progressions.

²⁹H. Garfinkel. “Studies in ethnomethodology”. In: Routledge, 2023, pp. 209–261.

Chapter 2

Experimental framework

2.1 What are chatbots?

A chatbot is a conversation simulation application. An attempt to make a machine converse with a human user requires mimicking human speech. The intention is to have a user interact with a chatbot that would communicate so well that the user would be convinced this is another human they are talking to. Whether just that has been achieved would be measured by a so called turing test proposed by Alan Turing in 1950¹

Initial attempts at making a computer converse were rule-based². The content of the chatbot utterances would be predetermined and there would be a decision tree that would decide what to say next. In the early days as well as often times in modern day systems string matching³ would be used to analyse user input.

The chatbot ELIZA⁴ is regarded as a milestone in conversation AI. It pretended to be a therapist using general phrases and questions. This led to a relatively believable conversation.

As long as interaction frame is strictly defined and the robot has some level of authority like in case of ELIZA the rule-based approach can work. In modern day this principle is still being used in large enterprise systems. In practice this approach works as long as there is surveillance and maintenance⁵.

Machine learning moved the possibilities of conversational AI forward. Multidimensional semantic space based classifiers like the ones used in IBM Watson⁶ allow for understanding declared categories while keeping the structure rule based.

The recent breakthrough of large language models using the transformer architecture seemingly solved the conversational AI problem altogether. It is now possible to generate near natural speech this gives the possibility to just let the conversation be taken over by one answer generator.

¹A. M. Turing. “Computing Machinery and Intelligence”. In: *Mind* 59.236 (1950), pp. 433–460.

²McTear, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*, p. 43.

³C. Moy. “Knuth–Morris–Pratt illustrated”. In: *Journal of Functional Programming* 34 (Jan. 2024).

⁴J. Weizenbaum. “ELIZA a Computer Program for the Study of Natural Language Communication Between Man and Machine”. In: *Commun. ACM* 9.1 (Jan. 1966), pp. 36–45.

⁵Š. Kološová. “Konverzační design: principy designu hlasového robota pro přirozenou komunikaci s lidmi”. Diplomová práce. Praha: Univerzita Karlova, Filozofická fakulta, Ústav informačních studií - studia nových médií, 2022.

⁶D. Ferrucci et al. “Building Watson: An Overview of the DeepQA Project”. In: *AI Magazine* 31.3 (July 2010), pp. 59–79. URL: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2303>.

Chatbot implementations of large language models are very useful for some usecases like open domain conversation or accessing knowledge base. Moreover in 2024 this technology is now closer to beating the Turing test than any other model or approach before it⁷ by having 54% of participants thinking they are talking to a human. While Eliza convinced 22% of participants, actual humans only convince 67% of participants.

The downside at hand is however loss of control. To give a dialog system power to make changes in adjacent systems, its behavior needs to be programmable.

Turn taking in chatbot interactions

Even if the Turing test is passed, a true simulation of conversation can only be achieved if the temporal aspect of conversation is simulated⁸.

As established in previous chapter, turn taking is a crucial aspect of conversation. The way participants determine who is to talk explains the difference between a structure of the text of conversation and a single-producer text. The mechanism of turn taking differs between actual human conversation and an interaction between a chatbot and a user.

Interaction between chatbot and user typically take place in a strict fashion where both participants, human and virtual, have unlimited time to come up with the next answer. While the chatbot should be optimized to answer as fast as possible, the user has as much time as they need until a technical fallback.

“Research in sociolinguistics, psycholinguistics, and conversational analysis has revealed that turn-taking is a mixed-initiative, locally coordinated process, in which a variety of verbal and nonverbal cues such as eye gaze, body pose, head movements, hand gestures, intonation, hesitations, and filled pauses play a very important role. We continuously produce and monitor each other for these signals and can coordinate seamlessly at the scale of hundreds of milliseconds across these different channels with multiple actors.”⁹

People are capable of producing and picking up clues that indicate opportunities for turn taking easily. Research is going figuring out the correct time to start speech¹⁰¹¹ or actually creating a system that will be able to produce such behavior¹²¹³. This research field has the potential to push conversation technology closer to true conversation simulation.

A truly flexible turn taking has not however been attempted in this paper. In fact, data consisting of conversation transcripts with no overlap are helpful for textual analysis that is conducted here.

⁷C. R. Jones and B. K. Bergen. *People cannot distinguish GPT-4 from a human in a Turing test*. 2024. arXiv: 2405.08007 [cs.HC]. URL: <https://arxiv.org/abs/2405.08007>.

⁸A. Raux and M. Eskenazi. “Optimizing the turn-taking behavior of task-oriented spoken dialog systems”. In: *ACM Trans. Speech Lang. Process.* 9.1 (May 2012).

⁹S. Andrist et al. “Turn-Taking and Coordination in Human-Machine Interaction”. In: *Ai Magazine* 37 (Dec. 2016), pp. 5–6.

¹⁰G. Skantze. “Turn-taking in Conversational Systems and Human-Robot Interaction: A Review”. In: *Computer Speech & Language* 67 (May 2021), pp. 101–178.

¹¹A. Gravano and J. Hirschberg. “Turn-taking cues in task-oriented dialogue”. In: *Computer Speech & Language* 25.3 (2011), pp. 601–634.

¹²G. Jonsdottir and K. Thórisson. “A Distributed Architecture for Real-Time Dialogue and On-Task Learning of Efficient Co-Operative Turn-Taking”. In: Oct. 2013, pp. 293–323.

¹³F. Gervits et al. “It’s About Time: Turn-Entry Timing for Situated Human-Robot Dialogue”. In: *Proceedings of the Special Interest Group on Discourse and Dialogue*. 2020. URL: <https://hrilab.tufts.edu/publications/Gervits2020Sigdial.pdf>.

2.2 Convform

An exploration has been carried out using a custom tool called Convform¹⁴.

At its core Convform is a computer program which accepts a configuration, user input and context and determines next chatbot answer. Other than that it offers a collection of utilities to help design and run chatbots.

Participant facing chat interface

In order to handle the inputs, convform provides a chatting environment for the participants to interact with a chatbot. The convform environment differs from a usual chat log because it does not display the entire history the conversation. In attempt to simulate spoken conversation it only displays the last chatbot response. This way the participant has to rely on their memory in taking part in the conversation like they would in spoken interaction. Other than that the participant may enter their next response and send it. They are also instructed to end to conversation by a red button if the chatbot behaviour is "unnatural" (nepřirozené) After the conversation is over whether it has been ended by the user or the chatbot, there is a questionnaire which asks the participants to rate how "natural" the conversation was and mark and comment on utterances in the now fully displayed conversation.

Conversation design tool

Convform lets admin user create chatbots and define their behavior. The behavior can be defined by string matching rules or prompts. It is capable of working as a statemachine or a stateless chatbot. It provides a level of control over references within the design.

Testing and debugging of various conversation contexts

While designing chatbots it is necessary to be able to simulate various situations to fine tune various possible scenarios that might occur in the conversation. To achieve this, there must be a way to encode required context to convform. The convform chatbots use a custom conversation status object to represent their current understanding of the conversation. It contains information about the history of the conversation which in conjunction with the configuration file and user input helps determine the next response. The configuration file is static and conversation status is updated automatically. User input comes from the user. This conversation status can simulate any possible conversation context from the chatbots perspective. For testing and debugging specific contexts, convform allows admin user to tweak the conversation status.

Accessing the conversation data

Lastly convform naturally includes a convenient way to read user interactions and browse associated conversation status objects

¹⁴A. Maršík. *Convform*. <https://github.com/almarsk/convform>. 2024.

2.3 Conversation design in theory

Designing the behavior of a dialog system is referred to as conversation design¹⁵¹⁶¹⁷. It is not the course of any one conversation that is being designed here but rather as many possible ways any conversation could go for a given use case. Conversation design as a profession is deeply connected with the rule-based approach that has been used since ELIZA times. A conversation designers task is typically maintaining all the possible utterances and rules under which they would be uttered in enterprise dialog systems.

A conversation designer operates between the business logic and use case of the dialog system the clients, customers or users interacting with the system and the developers maintaining the system.

2.3.1 Rule-based approach

In order to be able to design a rule-based dialog system, one needs to be able to encode the following:

- The possible utterances, that the dialog system can produce
- Rules under which the next utterance will be chosen

If the conversation is supposed to be a state machine e.g. it needs be able to use different sets of rules under different contexts in the conversation. This way a dialog system can be context aware to a degree. A conversation design of this sort can be displayed as a diagram. Then a way to maintain context of conversation is also necessary. This context needs to encode rules to choose an immediate ruleset which helps determine the next utterance. This principle is a simplification of how people decide what they will say next in conversation.

Pros and cons

This approach to designing a dialog system has been the standard for decades. It offers a granular control over how a dialog system should behave. In case of the state machine variant it allows to guide the user through a relatively complex process. It however suffers from how unpredictable the user can be. It is up to the conversation designer to cover all the possible ways of answering which not only is hardly possible but also poses a necessity to segment the spectrum of possible answers which can generate conflict when a user input semantically spans across multiple determined categories. This issue is even stronger while using the string matching approach, because there the string literal can decide about the following dialog system answer as if meanings and their speech representations would map one-to-one, which they are not. Even if a certain meaning is included in a ruleset, the system might not grasp the meaning and react in an incoherent way. With the state machine the distribution of various rules across various rulesets requires big effort. Extending the capabilities of a rule-based dialog system hardly scale

¹⁵Kološová, “Konverzační design: principy designu hlasového robota pro přirozenou komunikaci s lidmi”.

¹⁶McTear, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*.

¹⁷H. Brabra et al. “Dialogue Management in Conversational Systems: A Review of Approaches, Challenges, and Opportunities”. In: *IEEE Transactions on Cognitive and Developmental Systems* 14.3 (2022), pp. 783–798. DOI: 10.1109/TCDS.2021.3086565.

and tend to have regressions. In case of dialog systems relying on user input by speech transcription the text input processed by the system is not guaranteed to represent what the user actually said. In conclusion rule-based approach to conversation design provides control over the dialog system behavior but tends to be inflexible and unreliable.

2.3.2 Statistically driven approach

Some of the issues tied to rule-based systems are resolved incorporating machine learning techniques into the dialog management of the dialog system. As mentioned above the recent breakthroughs in the field of speech generation have been significant allowing for letting the dialog system play a bigger role in what is being said next. In its simplest form, it is possible to just let the answer be generated "end-to-end". The user input is sent to a model which generates an answer. While not perfect¹⁸ this technology is capable of staying on topic¹⁹, mirroring²⁰ and other things that make for a coherent conversation.

Large language models

The main component that is responsible for this way of simulating conversation at this level of flexibility are so called large language models²¹. These models, powered by advanced neural networks, have revolutionized the field of natural language processing. Among the most influential architectures are transformers, which enable these models to handle vast amounts of text data and capture complex patterns of meaning, context, and grammar. In simple terms they use their training data to generate the next most probable token.

These systems are trained on immense datasets, allowing them to generate coherent and contextually relevant responses across various topics. This flexibility has made them increasingly mainstream, being integrated into tools for writing, education, customer service, and more.

Prompt engineering techniques

The rapid advancement of LLM technology has outpaced research into optimal interaction strategies. Understanding how to engage effectively with these systems has been a developing area²², which illustrates both their power and their novel nature. The foundational idea is: an LLM performance improves significantly if it is prompted correctly.

Over time, researchers and practitioners have developed techniques for crafting effective prompts to optimize outputs. The simplest approach is known as "zero-shot" prompting, where a user poses a direct question or request without additional context

¹⁸J. Tie et al. "LLMs are Imperfect, Then What? An Empirical Study on LLM Failures in Software Engineering". In: *arXiv preprint arXiv:2411.09916* (2024).

¹⁹M. N. Sreedhar et al. *CantTalkAboutThis: Aligning Language Models to Stay on Topic in Dialogues*. 2024. arXiv: 2404.03820 [cs.CL].

²⁰J. Ivey et al. "Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Dialogue". In: *arXiv preprint arXiv:2409.08330* (2024).

²¹Humza H. Naveed et al. "A comprehensive overview of large language models". In: *arXiv preprint arXiv:2307.06435* (2023).

²²P. Sahoo et al. "A systematic survey of prompt engineering in large language models: Techniques and applications". In: *arXiv preprint arXiv:2402.07927* (2024).

or examples^{23,24}. However, zero-shot prompting may not always yield the desired depth or accuracy. It is common for the model to "misunderstand" the assignment and generate tokens so that it will "confuse" itself and lead to generated answer in a completely irrelevant direction²⁵.

More sophisticated strategies include "few-shot" prompting²⁶, where examples are provided to guide the model's response style or focus. This way there is a reference for the structure of the answer and there is a protection to the answer leading somewhere it is not meant to. Framing the task as an analogy can help improve the output²⁷.

Another very prevalent way that has proven to improve the performance of LLMs is a so called chain-of-thought prompting²⁸. It encourages the model to articulate its reasoning step-by-step, enhancing logical accuracy. There are many ways to achieve this, but the primary one is a few-shot approach where a description of the logic is explicitly described. The model is then prompted to produce a similar chain of thought and end the answer with the sought after information.

This principle can be further improved by chaining several LLM calls and having one evaluate the previous one. Such strategy has proven to surpass other models in available metrics²⁹.

Pros and cons

Using large language models as core component of dialog systems brings resolution to many issues rule-based systems introduce.

An LLM powered dialog system is flexible in understanding the user input. The user input is processed in a much more sophisticated than the string-matching or approach. While the classifier approach is a lot more capable to understand, it is still forced to choose a predefined answer, whereas an LLM has the capability to tailor an answer for every input. It can do this in a way that would be very hard to come up with especially in advance with the help of a conversation designer, leveraging the fact that LLM is primarily a text generator and only functions as a component in a dialog system. It can be relatively well controlled as it can accept complex instructions as to how to behave and these instructions can be tuned in runtime.

Systems of this sort however introduce their own set of problems. A big issue with factuality is called hallucinating³⁰. Factuality is a challenge for LLMs overall. It is

²³A. Kong et al. "Better zero-shot reasoning with role-play prompting". In: *arXiv preprint arXiv:2308.07702* (2023).

²⁴Yinheng Li. "A practical survey on zero-shot prompt design for in-context learning". In: *arXiv preprint arXiv:2309.13205* (2023).

²⁵Y. Hwang et al. "LLMs can be easily Confused by Instructional Distractions". In: *arXiv preprint arXiv:2502.04362* (2025).

²⁶B. Yao et al. "More samples or more prompts? exploring effective in-context sampling for llm few-shot prompt engineering". In: *arXiv preprint arXiv:2311.09782* (2023).

²⁷C. E. Stevenson et al. "Can Large Language Models generalize analogy solving like people can?" In: *arXiv preprint arXiv:2411.02348* (2024); C. Qin et al. "Relevant or Random: Can LLMs Truly Perform Analogical Reasoning?" In: *arXiv preprint arXiv:2404.12728* (2024).

²⁸J. Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

²⁹S. Wu et al. "A Comparative Study on Reasoning Patterns of OpenAI's o1 Model". In: *arXiv preprint arXiv:2410.13639* (2024).

³⁰A. Bruno et al. "Insights into Classifying and Mitigating LLMs' Hallucinations". In: *arXiv preprint arXiv:2311.08117* (2023); G. Perkovic, A. Drobnjak, and I. Boticki. "Hallucinations in llms: Understanding and addressing challenges". In: *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*.

possible to ask questions, that do not have a correct answer³¹. It has been shown, that LLMs have an issue knowing that they do not know something³². Recognizing that is the case requires an extra level of reasoning that is an object of research as of recently³³. They are also required to be ethically aligned with humans which is an ongoing research³⁴.

Even if all the programming and training is done to the most benefit of humans, information technology is susceptible to be broken by malicious action or so called jail breaks³⁵. LLMs being trained on vast amount of data, they hold knowledge that can be illegal or unethical to spread like steps to create explosives for example. The typical examples of jail breaks are ways to manipulate the LLM to give out this information which under regular circumstances it would not give.

This is why for dialog systems that are supposed to achieve anything else on top of the conversation itself, if they are meant to be powered by LLMs, a regulating structure needs to be placed on top of the LLM.

2.4 Conversation design in practice

Conversation design in Convform attempts to combine elements of rule-based design with text generation. It allows creating a purely rule-based chatbots which analyse the user input based on string matching and say exactly what they are prescribed to. On the other hand it also allows to make the chatbot understand the user input by adding it in a prompt and answer using a generated response. Both these approaches can be combined in various ways. Other than that, convform also allows to predetermine the chatbot personality for the entire conversation. The building blocks of a convform chatbot are states and intents which represent the utterance and the understanding.

State

A state is an object which carries several pieces of information bundled together. At its core it contains the utterance of the chatbot whether it is a hardcoded one or a prompt component which is to be called. A state however also contains information about which intents to listen to in the next user input, which states to add automatically to the next response and other navigation instructions like this one. Each convform status associated with a response can contain multiple states. This is to make convform generate more complex answers which can react in a flexible way. However it also comes with a challenge to order these responses correctly and make sure that they are not contradicting each other content-wise.

IEEE. 2024, pp. 2084–2088.

³¹A. Payandeh et al. “How susceptible are llms to logical fallacies?” In: *arXiv preprint arXiv:2308.09853* (2023).

³²Z. Yin et al. “Do large language models know what they don’t know?” In: *arXiv preprint arXiv:2305.18153* (2023).

³³C. Olah. *Mechanistic Interpretability – July 2024 Update*. Accessed: 2025-03-29. July 2024. URL: <https://transformer-circuits.pub/2024/july-update/index.html#dark-matter>.

³⁴Zhichao Z. Wang et al. “A comprehensive survey of LLM alignment techniques: RLHF, RLAIIF, PPO, DPO and more”. In: *arXiv preprint arXiv:2407.16216* (2024).

³⁵Y. Zhou et al. “Don’t Say No: Jailbreaking LLM by Suppressing Refusal”. In: *arXiv preprint arXiv:2404.16369* (2024); A. Wei, N. Haghtalab, and J. Steinhardt. “Jailbroken: How does llm safety training fail?” In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 80079–80110.

Intent

An intent is an object representing a category of a user response. It contains the information to determine whether user input fits in given category and the state or states to respond with next. Just like in state, the information about whether the user input corresponds with the intent can be encoded via string-matching patterns or a prompt. As mentioned intent is a problematic concept, because it forces an outside logic and categorization on user input, which might not be able to fit well in the framework declared by the current intent set. It is however the only way for a conversation designer to peek into what is going on in the conversation and to direct the dialog system in the correct way.

Using a combination of a states and intents a convform chatbot can be created, that will be instructed to lead from one state to another make decisions based on intents while being able to use any combination of hardcoded responses and intent patterns and descriptions of responses or user inputs used in intents. Detailed description of how convform works can be found in the wiki of its github repository ³⁶

Coherence

With support of LLM powered responses convform can be used to simulate an open-domain conversation with a user and simultaneously using a combination of intents and prompting a convform chatbot can be created that will act incoherently under a predefined set of conditions allowing to create experimental stimuli. First however, regular conversation needs to be achieved using convform.

Conversation style

To simulate conversation, it is useful to simulate a persona. The persona can then have a simulated motive to converse which can interest the user enough to engage in interaction with the dialog system. For rule-based systems, persona can be defined ahead of time and it can manifest itself via the specific writing of the hardcoded responses that the system is able to give. With generated responses, the persona of the chatbot has to be included in the prompt. The personality of LLMs and conversation technology more broadly is being discussed³⁷. The general characteristics of a machine talking to a human are typically friendliness and helpfulness. For conversation research with convform, the goal is to achieve just that. The chatbot persona needs to be friendly, polite and curious. It needs to be able to keep the conversation going but not change topic too often. It needs to be able to add a little bit of its own perspective.

The conversation style e.g. the amount of participation and initiative in conversation is something people adapt in to their conversation counter-part. Since developing a system that would imitate this behavior requires additional effort and expands scope beyond the coherence research this paper focuses on, this approach to conversation design has not been taken here. Instead two versions of conversation style have been developed and distributed evenly between participants.

³⁶<https://github.com/almarsk/convform/wiki>

³⁷A. Deshpande et al. *Toxicity in ChatGPT: Analyzing Persona-assigned Language Models*. 2023. arXiv: 2304.05335 [cs.CL]. URL: <https://arxiv.org/abs/2304.05335>; G. Laban et al. “Robo-Identity: Exploring Artificial Identity and Emotion via Speech Interactions”. In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2022, pp. 1265–1268.

The initial conversation style used in the experiments represents a curious and friendly chatbot who is instructed via prompt to ask lots of follow up questions. This tends to result in a conversation that moves forward in its topical structure in way deemed incoherent by Hrbáček³⁸. It depends on the participants impression whether it would be perceived as curious and initiative or shallow and dismissive.

A second version of conversation style has been introduced to get some insights on participants acceptance and the course of conversation itself. This one would interleave topical questions with remarks on the topic The intention behind this would be to slow down conversation tempo and give the participant the opportunity to bring their own initiative.

Prompting

Entity recognition In order to track entities that could be referred to a few-shot prompt was deployed that would help keep track of which entities have been mentioned.³⁹ Since GPT4o, the model used in the experiment, tended to consider too many things an entity, most examples are negative and do not capture an entity. It also contains some repetition as a result of fine-tuning the best wording.

Anaphorization To be able to create conversation designs which contain various types of anaphora, convform first needs to be able to give a response that has an anaphoric reference to an entity from the previous conversation in it.

The GPT-4o model used for this use case does not tend to generate sentences with anaphoras in them. Instead it will rather mirror the entity phrase.

The anaphorization prompt therefore tasks the model to modify a generated response so that the mirrored entity is replaced with an anaphoric device. For this another few-shot prompt was used to modify a just generated response found in Appendix under *anaphorization*. The approach here is opposite of automated annotation of anaphora in text⁴⁰ Instead we are generating it. This prompt has been tuned to catch as many tricky cases as possible.

2.4.1 Stimuli

With these tools multiple chatbots were created that would generate conversation situations which serve the role of experimental stimuli. Participant reactions to these stimuli can be then compared. This way conversational experimental designs can be created. There are three types of stimuli created for the purpose of this paper. They are shallow anaphora, deep anaphora and nonassignable anaphora.

Shallow anaphora

A shallow anaphora is a kind of anaphora where the referent of the anaphoric device should be relatively easy to map as opposed to a deep anaphora. The referent will always occur in the preceding utterance of the participant. This type of stimuli is relatively

³⁸Hrbáček, *Nárys textové syntaxe spisovné češtiny*, p. 30.

³⁹S. Loaigca, A. Beyer, and D. Schlangen. “New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities”. In: *International Conference on Computational Linguistics*. 2022.

⁴⁰Loaigca, Dobnik, and Schlangen, “Anaphoric Phenomena in Situated dialog: A First Round of Annotations”.

simple to achieve in convform generating a response and using the anaphorization prompt on it afterwards.

Shallow or surface anaphora⁴¹ is common in regular conversation and should not pose a problem for participant to understand in a conversation with chatbot. It should therefore not have an impact on the user acceptance of the chatbot and should generally go unnoticed. It is regardless worth using as stimuli for a reference unproblematic case that still requires the same kind of processing as other more interesting stimuli.

Example:

Participant I love coffee

Chatbot (not anaphorized) What kind of coffee do you like best?

Chatbot (anaphorized) What kind of it do you like best?

Deep anaphora

A deep anaphora is a situation where the referent of the anaphoric device occurs several utterances ago. The depth is not measured by number of occurrences but by number of new entities that occur since the referent which the anaphora refers to. Measuring depth of anaphora by the number of utterances does not capture the dynamic nature of topic progression in the text of the conversation. The number of utterances does not map on how many topics have been visited. While the number of new entities does not map exactly either, it is closer to the topic progression and close to what is being sought after here – how far in the conversation is an entity still acceptable or even available to speakers. An entity is can represent a topic but can also be one of several entities to represent a topic or can cover several topics at once all depending on which way the conversation goes.

As stated earlier, both topic and entity are difficult to define and their annotations tend to be recursive. A close-enough approach has been adopted in this paper. While runtime topic annotation by an LLM is not necessary for generating deep anaphora and has therefore not been attempted in this paper, entity tracking is made possible by entity recognition prompt.

This prompt runs in parallel with the next response generation and writes down its results in the conversation status. A chatbot that contains the deep anaphora stimuli chooses a participant mentioned entity relatively early on in the conversation (though not at the very beginning) and then tracks new mentioned entities. When there have been 4 new entities mentioned, the next response generation prompt will be modified. The modification lies in that the context of the conversation that has so far taken place will be cut so that the chatbot only has access to the conversation until the point of the mention of the entity. Given the response generation prompt the response will contain a question about the mentioned entity. Then the only thing that needs to be done is modifying the response via the anaphorization prompt.

The trick here the participant and the chatbot differ in their perspectives on what the conversation currently is. The chatbot refers to something that from the perspective of the participant has been mentioned a while ago.

⁴¹H. Hoji. “Surface and deep anaphora, sloppy identity, and experiments in syntax”. In: *Anaphora: A reference guide* (2003), pp. 172–236.

This approach is relatively imprecise and relies on luck to a certain degree. Compared to the shallow anaphora it is expected to be somewhat more problematic and perhaps cause the participant to request clarification.

There has been one issue that has arisen while developing this stimuli that has proven to alter the character of the data in an unwanted way. Since the chatbot has no access to the conversation that happens between the occurrence of the referred to entity and the participants present moment, chances are the chatbots question will be on a piece of information that has been mentioned in the meantime. Whenever that happens the degree of participant acceptance decreases significantly due to a topical incoherence rather than due to struggling to mapping a deep anaphora. This has been dealt with via providing the chatbot with the rest of the conversation in another component of the prompt with the instruction to avoid any of the topics mentioned there. LLMs are known to handle negative instruction with less success than positive ones⁴² but this measure seems to have mitigated the problem as can be seen in the data attached in the appendix.

Nonassignable anaphora

The last type of stimuli used in this paper has been called nonassignable anaphora. It is a device that the participant will tend to interpret as an anaphoric device, typically a personal or demonstrative pronoun, but one such that the participant will not be able to map to any of the candidate entities in the previous conversation text. This stimuli is expected to lower the participant acceptance by the greatest amount.

To make a chatbot contain this stimuli entities are tracked to make sure there are candidates to be considered in case an anaphora occurs. Once there is a sufficient number of entities recognized in the conversation a hardcoded response is returned instead of an LLM generated one. The response contains a pronoun that to make sense of participant needs to interpret as an anaphora.

Since the response containing the stimuli is hardcoded, there is no guarantee that is actually is incoherent with the previous conversation and that there is no candidate to map the anaphora on to. Though odds are high enough every conversation that is supposed to contain this stimuli will have to be manually checked to confirm the required stimuli is present. This will be the case for all the conversations regardless because presence of stimuli is not guaranteed for shallow and deep anaphora either.

An approach not explored in this paper is achieving a nonassignable anaphora is also possible via generating a response using a prompt that instructs an LLM to come up with a question containing an unrelated entity avoiding all mentioned entities and anaphorize it before showing it to the participant. Although LLMs tend to perform worse with negative instructions⁴³ this could be achieved using a chain of prompts. The notion of nonassignable anaphora brings into scope the question of what makes an anaphora assignable. It is the semantic compatibility of the words around the anaphora that determine which of the candidates the anaphora is referring back to. The generation of the lexical surroundings of the anaphora needs to be handled carefully when coming up with an LLM based response.

⁴²J. Jang, S. Ye, and M. Seo. “Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts”. In: *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*. Ed. by Alon Albalak et al. Vol. 203. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 52–62.

⁴³Jang, Ye, and Seo, “Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts”.

2.4.2 Ending the conversation

While recognizing when the conversation is ending or especially when it should not end people rely on a set of clues similarly to knowing when to take turns speaking⁴⁴⁴⁵⁴⁶. In open-domain conversation like the one a convform chatbot holds with experiment participants, the main challenge is to recognize when there is a topic at hand that interests the participant. Another discipline in the realm of ending the conversation is recognizing it is a good time to end the conversation due to the participants lack of interest or exhaustiveness of all topics at hand. Conversation designs made for this paper do not take much of this into account. The main goal for a convform chatbot here is to present the participant with a stimuli. Once they manage that, if the participant is willing to continue the conversation continues for a hardcoded number of responses. This leads to participants sometimes noticing the conversation ending abruptly and mentioning they would like to continue in the questionnaire or even at the very end of the conversation itself. This can potentially have effect on the score given by the participant and therefore brings noise into this parameter.

Since each conversation has to be manually checked, an assessment of how to deal with this noise can be made while and after processing the data. A runtime topic annotation and other prompting techniques could potentially help make the convform chatbot converse in such a way that would be more aware of the general course of the conversation perhaps giving hints about ending the conversation ahead of time or reacting to and handling the participants hints of the same type.

⁴⁴Adam M. Mastroianni et al. “Do conversations end when people want them to?” In: *Proceedings of the National Academy of Sciences* 118.10 (2021), e2011809118.

⁴⁵J. A. Y. Davidson. *Ending structures in conversation*. University of California, Irvine, 1975.

⁴⁶L. Coppock. “Politeness strategies in conversation closings”. In: *Unpublished manuscript: Stanford University* (2005).

Chapter 3

Data

The data collected using a convform chatbot is a transcript of the conversation between a participant and the chatbot. Depending on the conversation design of the given chatbot the conversation may contain a record of the participant being exposed to a specific situation and their reaction to it. Other than that the collected data contains an information about whether the participant quit the conversation, the participants rating of the conversation and their comment on it.

Unfortunately it cannot be guaranteed the required stimuli actually occurs in the conversation. Though the probability is relatively high, the LLM technology responsible for most answers is nondeterministic and the participants tend to be unpredictable. On many levels conversation can take an unintended direction which can spoil the stimuli. Whether thematic, textual or interactional, anything can go wrong. That is why as mentioned earlier, each conversation needs to be visited manually to confirm required stimuli is present. That does not present too much of a problem since a qualitative analysis of the data has the potential to bring lots of corrections to experiment design, relevant observations or inspiration for further steps. It would however pose a complication was one interested in scaling up this approach that would have to be dealt with.

3.1 Data collection

3.1.1 Experimental designs

Data collection took place with the use of six chatbots. They each represent a possible combination of a 3x2 design. The two design variables are anaphora type and conversation style.

The anaphora types are as discussed above a shallow, deep and unassignable. The conversation styles were two. One, inquisitive, where the instruction part of the prompt would always instruct the LLM to pose a question. The other, relaxed, would interleave this question instruction with an instruction to comment on the conversation in a nonquestion way. The conversation style variable was deployed for purely explorative purposes.

All the chatbots had the same persona instruction except their names. They would communicate in Czech with participants who were czech or slovak native speakers. The first utterance in the conversation would be the chatbots and would be hardcoded to contain greeting and self introduction.

They would be instructed to chat for several turns and then the stimulus would come.

After the stimulus the chatbot would continue for a couple more turns and then say goodbye. The questionnaire would then appear to the participant. That is unless the participant aborted earlier, in which case the questionnaire would appear immediately. In some cases participants simply left the user interface leaving no extra information. Other than that, participants were asked to grade the conversation on scale 1 to 5 where 1 was most acceptable and 5 was least acceptable.

Timing of the prompt would differ based on the anaphora type. For shallow type the chatbot simply converses for two turns and thereafter it starts tracking entities. At first recognized entity the prompt fires off prompting an utterance about the extracted entity and anaphorizing it. With deep anaphora similar thing happens, but the chatbot waits several more entity occurrences before also prompting an utterance about the extracted entity and anaphorizing it. Lastly the nonassignable anaphora has the same timing method as the shallow anaphora, but the prompt itself is hardcoded to contain an anaphoric reference which is very unlikely to be assignable.

In the experiment the participant would be instructed to end the conversation, if the chatbots communication wasnt natural. The concrete Czech word used was "přirozená komunikace".

3.1.2 Participants

The data has been collected in two waves. The pilot wave generated 50 conversations the participants for which were recruited from the networks of the papers author. In the attached full data the identifiers of these conversations have the letter p for pilot appended at the end. The second wave generated 325 conversations the participants for which were recruited from the students of the Charles University. They were instructed to take two conversations with a week pause in the middle. However since for this experiment explorative in nature contrast and timing were not critical they were free to revisit the chatbot however they liked. There is no attempt made to indicate unique participants in the conversation as this would not bring any observable results. If such approach should be taken in the future, more control has to be gained over the course of the conversation as discussed below. No changes were made to the design for the second wave other than changing the structure of the questionnaire. While in the pilot wave was tasked to make a comment on the conversation or explain why they aborted it if they did so, the second wave was shown the entire conversation and was asked to mark and comment on specific chatbot responses they found odd or interesting.

3.2 Data analysis

Since the collected data is relatively complex multiple layers of analysis need to be performed on it. The steps of analysis will be following:

- Conversation design stimuli annotation
- Participant reaction quantitative analysis
- General qualitative analysis

In order to assess the datapoints for quantitative analysis each conversation needs to be visited to confirm or deny the presence of expected stimuli. This preliminary step of

qualitative analysis provides feedback on whether the conversation design that is meant to facilitate the experimental design was succesful in doing so. Further analysis leading up to a quantitative assessment should only be done on conversations where convform succesfully exposed the participant to the stimuli. It has been observed in collected data that sometimes an unintended stimuli takes place instead of the intended one. Then simply changing the label of the conversation is appropriate. It is also necessary to filter the collected data for noncooperative participants who make it impossible for convform to perform the stimuli in the first place. The participant needs to take on a role of a friendly conversator if the experiment is to work. If they for example attempt to take advantage of the LLMs obedience and give it an unrelated task that involves any sort of text generation, they are pronounced noncooperative.

The qualitative analysis of the participant reaction to stimuli is mostly of explorative nature. The convform environment lets the user to input any text which even under the condition that they are cooperative can be unexpected and can derail the conversation. While these incidents are typically anecdotal and cannot be used to make generalizations they need to be taken into account as a possible participant behavior. From the perspective of dialog system development such cases would be considered to be edge cases and dealing with them on the conversation design level would probably be considered lower priority. However the conversation analysis perspective will be very interested in all the potential paths that a certain conversation situation can go in. This is ultimately why any weakness to the conversation design does not matter too much as all conversation are be visited anyways in search for unique situations and effort to understand full consequences of the conversation design.

In conclusion the main reason for using an experiment environment solution like convform can be summed up in two points:

- Partial control over what happens in the conversation
- The ability to expose multiple people to a comparable conversational stimuli

The first point has been described in detail above. As far as the second point goes, this approach is a step towards a quantitative analysis of conversation. Though conversation is multifactored and various types of situation always come with a different set of circumstances if a quantitative analysis of a set of conversations containing a certain reoccurring pattern proves contrastive in some of the quantifiable parameters to a set of conversations containing a different reoccurring pattern case could be made this is caused by the observed patterns regardless of all the noise naturally present in a conversation text.

3.3 Annotation

All 375 conversations were visited to confirm whether the stimulus expected based on the conversation design. Other than that the participant reaction was annotated to be either continuation, metacommunication or aborting of the conversation. These three types of participant reaction to stimulus represent the amount of understanding the reference. In simple terms, continuation means the stimulus went under the participants radar or has been accepted. Metacommunication, an utterance which somehow addresses the ongoing communication, shows the participant not understanding, but perhaps attributing

the misunderstanding to themselves or believing the issue can be resolved. Lastly the conversation is aborted when the participant loses trust in the ongoing conversation being coherent now or in the future. Then the illusion of social interaction falls and the conversation the participant is having dissolves into typing text into a computer program and getting text in return. The immersiveness necessary for simulating a conversation is here no longer achieved.

3.3.1 Inter-annotator agreement

To make sure that annotating the data is possible, inter-annotator agreement has been measured on 37 conversations which gives 10Three annotators annotated each conversation for anaphora type and whether the participant reaction to stimulus is metacommunicative. Metacommunication is the only reaction type that needs manual annotation since aborting the conversation is observable in the transcript and marked in the data automatically and any reaction that is not aborting and metacommunication has to be a continuation.

Stimulus annotation

For anaphora type Cohen's Kappa was calculated for each of the three annotator pairs with following results:

Cohen's Kappa for Annotator 1 and Annotator 2: 0.89

Cohen's Kappa for Annotator 1 and Annotator 3: 0.88

Cohen's Kappa for Annotator 2 and Annotator 3: 0.92

Coming to the average Pairwise Cohen's Kappa of 0.90 the inter-annotator agreement appears to be high enough that this annotation categorization to be used and trusted.

Each of the conversations that caused disagreement is interesting in its own right:

conversation id: 145

...

b: Jaké dárky plánuješ koupit?

Which gifts do you intend to buy?

h: ještě nevím, potřebuji pro tatku nějaký dárek a pro mamku

don't know yet, need for dad a gift and for mom

b: To může být výzva, ale určitě najdeš něco skvělého.

That can be a challenge, but surely you'll find something great.

h: jo, taky si myslím

yeah, i also think

b: Jaké koníčky má tvůj táta a mamka?

Which hobbies does your dad a mom have?

h: nevím, neznám je

don't know, don't know them

b: To je škoda, můžeš se jich zeptat.
That is a shame, you can ask them.

h: bydlí moc daleko
they live too far

b: Bot: To je těžké, ale co si zjistit jejich zájmy?
That is tough, but how about finding out what their interests are?

h: taky nevímš
also don't know

b: Možná bys mohl zjistit, co mají rádi.
maybe you could look into, what they like

h: ok
ok

b: To je dobrý nápad, jak to uděláš?
That's a good idea, how will you do it?

h: zeptám se tety, ta určitě bude vědět
I'll ask my aunt, she'll surely know

b: Jaké mají oni koníčky?
Which hobbies do they have?

h: kdo?
who?

...

This conversation potentially has a deep anaphora in it, since the first occurrence of the mappable entity *dad and mom* is deep in the conversations history at the stimulus time. There are however other entities that are mappable that are mentioned in the meantime like *aunt* or the general group of *close ones* mentioned before the snippet starts. This points at the potentially missing category of *ambivalent anaphora*. Another issue with this stimulus is also that it asks a question that is already being discussed. This can be seen as a both topical and sequential coherence disruption. Why would the same question with the same reference in it be posed? That also makes the anaphora that much more vague and leads to a metacommunication *who?*.

conversation id: 150

...

h: Běhám 400m překážek a dělám sedmiboj
I run 400m with obstacles and do heptathlon

b: To je náročné! Jak dlouho už se tomu věnuješ?
That's demanding! How long have you been doing this?

... 7 turns (14 utterances) ...

h: Jsem po zánětu v patě, tak snad se to nevrátil, hodně to uvolňuji a cvičím na to
I had a heel inflammation, hopefully it won't return, I release and exercise it a lot

b: Jaké disciplíny tvoří on? Jakou máš mezi nimi nejoblíbenější?
Which disciplines does it consist of? Which is your favorite among them?

h: Kdo?
Who?

...

Here the mappable entity *heptathlon* is so far deep in the conversation that the annotator who would not agree with calling a deep anaphora stimulus perhaps did not even notice it or insisted that this is just too deep. While that is technically against the annotators instruction based on the above description of stimuli but it points at the interesting problematic of anaphora depth.

The other two conversations that had an annotation disagreement about them were IDs 151 and 153. Both of them have an imperfect generation in them that excludes them from the stimuli containing data.

Metacommunication annotation

Inter-annotator agreement regarding the metacommunication has produced following results:

Cohen's Kappa for Annotator 1 and Annotator 2: 0.94

Cohen's Kappa for Annotator 1 and Annotator 3: 0.88

Cohen's Kappa for Annotator 2 and Annotator 3: 0.94

With the average Pairwise Cohen's Kappa coming to 0.92, it can be considered reliable to annotate metacommunication in the collected data.

The two conversations that had disagreement about them in terms of presence of metacommunication in the participant reaction to stimulus are both interesting in their own right.

conversation id: 164

...

b: Jakou písničku z ní máš nejraději?
Which song from here do you like the best?

h: z rádia?
from the radio?

...

This reaction can be seen to contain an ellided metacommunicative device like *do you mean*. Since it is not present explicitly and the participant reaction is topically anchored in the conversation, the case is ambivalent.

conversation id: 173

...

b: Tak to by mě zajímalo, jak na to pak ona zareagovala.
Then I wonder, how she reaction to that.

h: Ano, pokazilo
Yes, it went wrong

...

It is unclear where the participant is coming from in their response. For this reason, it is also unclear whether the participant reaction is a metacommunication. There is a good chance the participant is somehow commenting on the chatbot breaking, in which case such utterance would indeed be considered to address the ongoing communication.

In conclusion the annotator agreement is strong enough for both stimulus type and metacommunication reaction and quantitative analysis can be performed on the measured metrics.

Chapter 4

Quantitative Analysis

4.1 Expectations

In the performed quantitative analysis, following result were expected to be found.

For user rating the rating was expected to be ordered like so from best to worse: shallow, deep, nonassignable.

As far as aborting the conversation goes, The amount of aborted conversations were expected to be ordered from least to most as: shallow, deep, nonassignable.

The reactions were expected to map onto the stimuli like so:

| | |
|---------------|----------|
| Shallow | Continue |
| Deep | Meta |
| Nonassignable | Abort |

As far as conversation style, the inquisitive variant was expected to have a worse score since every next question shifts the topic forwards, which should generate a topical progression that sooner or later becomes incoherent as Hrbáček demonstrates¹.

4.2 Result

4.2.1 Ratings

The collected ratings are of the following distribution:

| rating | times |
|---------------|--------------|
| Rating 1 | 22 |
| Rating 2 | 62 |
| Rating 3 | 138 |
| Rating 4 | 37 |
| Rating 5 | 6 |
| No rating | 19 |

¹Hrbáček, *Nárys textové syntaxe spisovné češtiny*, p. 30.

The average rating of the collected conversations which have been rated by participants is 2.77. Since the missing ratings are participants who clicked off of the conversation the absence of their rating can be interpreted as the worst rating. Then the average rating of all the conversations is 2.97.

The instruction to the participants was to rate the conversation using the same system as the grading in czech public school system. That may explain the majority of ratings being a 2 or a 3. The ratings in the collected data can only be compared against each other and for them to give a better picture of the stimulus effect, a more complex experiment design would need to be deployed containing a reference conversation and its rating for each participant.

4.2.2 Stimulus and reaction

The annotation result is as following:

| anaphora type | times |
|----------------------|--------------|
| Shallow | 104 |
| Deep | 64 |
| Nonassignable | 116 |
| Other | 85 |

Ignoring the conversation that lack a rating the result is following:

| anaphora type | times |
|----------------------|--------------|
| Shallow | 102 |
| Deep | 58 |
| Nonassignable | 105 |
| Other | 71 |

There is less conversation with the deep anaphora stimulus. This is because the deep anaphora stimulus requires the most difficult operation by the dialog system and is therefore the most likely to fail. The annotation result also confirms its necessity given almost a fifth of the conversations were marked as *other* and thereby excluded from the stimulus reactions quantitative analysis.

4.2.3 Design rating

The rating for the various conversation types came out to be following:

| anaphora type | rating |
|----------------------|---------------|
| Shallow | 2.75 |
| Deep | 2.53 |
| Nonassignable | 2.95 |

Opposed to expectations shallow anaphora stimulus type has worse score than deep anaphora. P-value from Kruskal-Wallis test being 0.02 is sufficiently low, but when values stimulus types are compares post-hoc via Dunn test, result come to:

| anaphora types | | p-value |
|----------------|---------------|---------|
| Deep | Nonassignable | 0.005 |
| Deep | Shallow | 0.158 |
| Nonassignable | Shallow | 0.136 |

The only sufficiently low p-value is between deep and nonassignable anaphora. The results only partially correspond with the expectations. Number rating is however noisy as other things happen in conversation that affect it. The chatbot contributions to the conversation outside of the stimulus has to be under control and produce natural responses in order to really be able to view the ratings as reflecting the stimuli hence the high p-values.

Design rating with abandonment interpretation

Some conversations do not have a rating because participant clicked off of the experiment webpage. Since the participants were instructed to leave the conversation if it is not *natural*, these conversations can be interpreted as worse possible reaction - 5.

We can further split these into a situation where the participant clicked off of the experiment right after the stimulus. This only happened in case of nonassignable anaphora design. Its new average score comes to 3.06.

If the participant left the conversation at any time after the stimulus, this can be because the stimulus disturbed the conversation coherence and while it was not serious enough at the stimulus time for the participant to end the conversation, the coherence never recovered. If we interpret conversations abandoned at any point after stimulus as rated with the worse possible rating, we get following results:

| anaphora type | rating |
|---------------|--------|
| Shallow | 2.80 |
| Deep | 2.77 |
| Nonassignable | 3.15 |

Shallow anaphora that is expected to have the best result still gets a slightly worse rating than the deep one. The differences between these ratings and the ones where abandoned conversations are excluded are:

| anaphora type | difference |
|---------------|------------|
| Shallow | -0.04 |
| Deep | -0.23 |
| Nonassignable | -0.19 |

The difference shows that the shallow anaphora had by far the least number of abandoned conversations. Updated significance check provides a slightly different picture with Kruskal-Wallis test p-value being a similar value, but the post-hoc Dunn test showing more significant differences:

| anaphora types | | p-value |
|----------------|---------------|---------|
| Deep | Nonassignable | 0.016 |
| Deep | Shallow | 0.653 |
| Nonassignable | Shallow | 0.019 |

Here the results are more significant but still not enough for deep vs shallow anaphora. In conclusion the shallow anaphora was not rated as expected in relation to the other stimuli, but the difference does not seem to be significant and in accordance with the expectations has the lowest number of abandoned conversations. There are many factors in the conversation data and stimulus does not seem to have a strong enough effect on the rating. If better more seamless conversation could be simulated, perhaps stronger effect of the stimulus on rating could be observed. The small difference can also be attributed to the fact that participants tend to avoid extremes in their grading.

4.2.4 Reaction to stimulus

A quantitative metric that is expected to show better the acceptability and understandability of the various stimulus designs is the direct conversational reaction to them. The annotated data came to the following result:

| anaphora type | continuation | meta | abort |
|---------------|--------------|------|-------|
| Shallow | 91 | 8 | 5 |
| Deep | 34 | 26 | 4 |
| Nonassignable | 13 | 57 | 46 |

Which percentage wise gives the following:

| anaphora type | continuation | meta | abort |
|---------------|--------------|--------|--------|
| Shallow | 87.50% | 7.69% | 4.81% |
| Deep | 53.12% | 40.62% | 6.25% |
| Nonassignable | 11.21% | 49.14% | 39.66% |

With the extremely low p-value coming to $P < 10^{-28}$ there is no doubt that the different stimuli have a clear effect on the participant reaction and that the results will be replicable. The expectations are met with the continuation being highest for shallow anaphora and lowest for nonassignable anaphora while the abort reaction proves to have the opposite tendency.

One unpredicted feature in the results is deep anaphora having a relatively close percentage of continuation and meta reactions and nonassignable anaphora sharing similar values for meta reactions and abort reactions. This suggests there is a inner division in the data that can be described and simulated more closely.

Stimulus x Reaction specific ratings

Looking at ratings for specific groups of stimuli, results are following:

| anaphora type | continuation | meta | abort |
|----------------------|---------------------|-------------|--------------|
| Shallow | 2.73 | 3.12 | 2.60 |
| Deep | 2.43 | 2.62 | 2.75 |
| Nonassignable | 2.45 | 3.04 | 2.98 |

Interestingly for shallow and nonassignable anaphora, the ratings do not follow the expected course. The unfulfilled expectation lies in the meta reaction having worse ratings than the abort reactions. The only anaphora type the expectation were met for is the deep one. Yet the significance of this result is low and more control over the conversation outside of the stimulus needs to be acquired to be able to rely on these results. To make conclusions about the meta reaction being indicative of a less acceptable experience in participant would be hurried.

4.2.5 Comments

While annotating, participants comments were collected to gain an extra metric on the reflection of the stimuli. While the shallow anaphora was mentioned once due to a generation issue and generally went by unnoticed, the deep anaphora got 11 mentions, which makes for slightly less than a fifth of conversations with this stimulus. Finally the nonassignable anaphora was mentioned 69 times in the participant comments. This makes over half of its occurrences were commented upon. All the comments expressed confusion about mapping the anaphora to a possible preceding referent. This result follows the expectations.

4.2.6 Conversation style

As mentioned above, there was an attempt to expose different participants to different conversation styles. This came out of the necessity to create a default way for the chatbot to communicate. First style is the inquisitive, where the chatbot is prompted to ask questions. An elaboration on it is a style of conversation, where the chatbot is at one time prompted to ask questions and at another to simply comment on the topic at hand. There were 183 conversations collected with the chatbot communicating in an inquisitive style and 128 conversations where the chatbot communicated in the second, more relaxed style. The relaxed style has been expected to have a better score as was confirmed with following result:

| conversation style | rating |
|---------------------------|---------------|
| Inquisitive | 2.63 |
| Relaxed | 2.45 |

However the Kruskal-Wallis test p-value of 0.10 suggests too small effect to consider this result replicable. To find statistically significant effect on conversation styles the conversation needs to be controlled in a more nuanced way by the dialog system.

The relaxed style however better rated on average also suffered from some issues in terms of generating natural answers. Sometimes the chatbot would simply repeat the participants response without adding anything new to it. For example in following conversation lack of topical input lead to aborting the conversation:

conversation id: 164

...

b: To chápu, co bys chtěl v histologii zlepšit?
I understand, what would you like to improve about histology

h: chtela bych ji nemit prostě
I would just like to not have it

b: To chápu, chtěla bys ji prostě nemít."
I understand, you would just like to not have it.

aborted

4.2.7 Zero anaphora

Since Czech is a prodrop language², it is expected that where the stimulus anaphora stands in subject position, there will be tendency of it being dropped. A response with zero anaphora³ would then be expected to be more acceptable than an equivalent response with an explicit one.

During annotation there have been found to be only 16 conversations with zero anaphora stimulus. This is partially due to the prompting of the stimulus, the few-shot prompt contained no example of a zero anaphora and partially because a lot of the stimuli contained an object positioned anaphora.

Coincidentally the stimuli containing zero anaphora were split equally between shallow and deep anaphora stimuli. The nonassignable anaphora stimuli mostly consisted of a hardcoded response that already contained the anaphora therefore it is not a part of the zero anaphora collection of conversations.

| reaction type | # occurrences |
|---------------|---------------|
| Continuation | 14 |
| Meta | 2 |
| Abort | 0 |

This result very clearly shows that the zero-anaphora stimuli were mostly going by unnoticed and have a higher level of acceptability. The same goes for rating with following result:

| anaphora type | rating |
|---------------|--------|
| Zero | 2.19 |
| Nonzero | 2.73 |

²A. Pešková. "Slavic and Romance pro-drop in contrast: Evidence from Czech and Spanish". In: *Languages in contrast* 19.2 (2019), pp. 310–333.

³Ning Yang et al. "A Study of Zero Anaphora Resolution in Chinese Discourse: From the Perspective of Psycholinguistics". In: *Frontiers in Psychology* 12 (2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.663168. URL: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.663168>.

Only shallow and deep anaphora type stimuli were used for this calculation. The zero anaphora score is a lot better considering the range of averages that the data has produced so far. This shows future designs need to take zero-anaphora into account more in order to achieve stronger contrast between stimuli.

The data contains conversations where stimulus was inappropriately generated with a nonzero anaphora like here:

conversation id: 347

...

h: ...hráli jsme je přes aplikaci ale já mu k vánocům koupila pravou šachovnici, takže konečně můžeme hrát normálně.

... we would play via an app but I bought him a chessboard for christmas so we can finally play normally. ...

b: Jakou máte ji, klasickou nebo něco speciálního?

Which do you have her, classic or something special?

h: jak to myslíš?

what do you mean?

...

The meta reaction is likely caused by the depth of the anaphora, but the bot response would have been more likely had the anaphora not been a part of it.

In some cases zero anaphora can also be inappropriate like here:

conversation id: 357

...

h: Nooo teď jsme jeli kolem břevnovského kláštera, nikdy jsem tam nebyla, tak bych zajela tam.

Well we just drove past Břevnov monastery, I never visited, so I would go there.

b: To zní zajímavě! Co tě na Břevnovském klášteře láká nejvíce?

That sounds interesting! What about Břevnov monastery is the most attractive for you?

h: To nevím právě, překvapilo me to jak je to velký.

I don't know, it surprised me how big it is.

b: Máš rád objevování nových míst a památek?

Do you like discovering new places and monuments?

h: Ráda bych na to měla více prostoru.

I would like to have more space to do it.

b: To chápu! Jaké místo bys chtěla objevit jako první?

I get that! Which place would you like to discover first?

h: No nevím ale třeba bys mi mohl doporučit nějaký hezký místo v Česku spjato i s nějakou novou moderní architekturou, možná nějaká vyhlídka.
I don't know but maybe you could recommend a nice place in Czechia connected with some new modern architecture, maybe a lookout tower.

b: Slyšela jsi o historii a zajímavostech?
Have you heard about history and trivia?
aborted

The participant commented after the conversation that the question did not fit in with the context which is why they aborted. That is reasonable as the response was generated based on the monastery topic from deeper in the conversation. Perhaps if there was an explicit anaphora here, the mapping would have been easier as dropping in a nonsubject position is not preferred in Czech. Regardless, the hard topical shift that disrupts the coherence caused by the chatbot would remain in the conversation.

Chapter 5

Qualitative observation

Visiting each conversation during annotation gave a detailed insight on the nature of the data, success of generating the correct stimuli and creating the sought after situations. The following observations about the experiment design, the various stimuli and conversation coherence in general were made.

5.1 Pro-coherence perception

Most noticeably the data shows how participants communicate with a cooperativity in mind and assume a lot in order to maintain the conversation flow. The quality with which the chatbot could generate responses was high enough, that the illusion of social interaction would hold up even if minor issues did come up.

5.1.1 Remaps

In following conversation, the participant maps an anaphora on a possible candidate even though the hardcoded response was meant to be nonassignable:

conversation id: 100

...

h: mám jednu mourovatou doma

I have a tabby one at home

b: Jak se jmenuje tvá mourovatá kočka?

What is the name of your tabby cat?

h: to nevím

I don't know

b: Jak dlouho už ji máš doma?

How long have you had it at home?

h: nemám vlastně

I don't actually

b: Jak ses k ní tedy dostal?

So how did you get it?

h: ráda pečů
I like to bake

b: Jaké oblíbené dobroty obvykle pečeš?
Which treats do you usually bake?

h: dorty
cakes

b: Tak to by mě zajímalo, jak na to ona potom zareagovala.
Well I'd like to know how did she react to it afterwards.

h: myslíš tu kočku?
You mean the cat?

...

This conversation along with several others is the reason annotation was necessary for this type of stimulus. The tolerance on participant side was sometimes very high, even if the anaphora was incongruent with the assumed mapping or even if some of the reference contained in the hardcoded stimulus does not map perfectly like in following conversation:

conversation id: 257

...

b: Stále je co zlepšovat, přemýšlel jsi o trenérovi?
There is always room to improve, did you consider a trainer?

h: na dvou hodinách už jsem byl s trenérem
I already did two lessons with a trainer

b: Tak to by mě zajímalo, jak na to ona potom zareagovala.
Well I'd like to know how did she react to it afterwards.

h: že musím lépe používat nohy, pohyb vychází z kyčle
that I have to use my legs better, the movement start in the hip

...

On this regard, a general third person seems to be acceptable, though from the annotation point of view it should be nonassignable as can be seen here:

conversation id: 36

...

b: Co tě nejvíce zaujalo v této knížce?
What did you find the most interesting in this book?

h: Nádherné fotografie minerálů pod mikroskopem.
Beautiful photos of minerals under a microscope.

b: Jaké další knihy o minerálech ti doporučili?
What other books about mineral did they recommend?

h: Většinou mi nikdo nic nedoporučuje, hledám si knihy sám.
Usually nobody recommends me anything, I look for my books alone.

...

In some cases the nonassignable anaphora stimulus has the potential to be interpreted as a third person reference to the participant. Such cases seem to have been tolerated by the participants as well though in human-human conversation these would probably cause metacommunication.

conversation id: 276

...

b: Co tě dneska tak unavilo?
What exhausted you today?

h: konverzace s klíčícím pohlavím
conversations with the male gender

b: To zní náročně, o čem jste si povídali?
That sounds demanding, what did you talk about?

h: o škole
about school

b: Tak to by mě zajímalo, jak na to ona potom zareagovala.
Well I'd like to know how did she react to it afterwards.

h: docela dobře
pretty well

...

In case of the deep anaphora stimulus, one of the big reasons annotation was necessary was also the risk of a remap. Generating the deep anaphora response, the LLM is only shown the appropriate section of the conversation. This exposes the generated and anaphorized response to the risk of another mapping candidate to have been mentioned in the meantime like in the following conversation:

conversation id: 259

...

h: ano, mám - zaklínač 3, kingdomcome, nebo dota 2
yes I do - Wither 3, kingdomcome or dota 2

b: Zaklínač 3 má nádherně propracovaný svět a silný příběh.
Withcer 3 has an elaborate world and a strong story.

h: přesně tak, je to úplně jiný svět než ten náš, takže je to skvělá relaxace.
exactly, it is a completely different world from ours, so it is a great relaxation.

b: Do které části herního světa Zaklínače 3 se nejraději vracíš?
To which part of the Witcher 3 world do you like to return the most?

h: Jako které místo se mi tam líbilo? Novigrad je moc hezký
As in which place there I liked? Novigrad is very pretty.

b: Kterého hrdinu v ní hraješ nejraději?
Which hero do you play most often in her? (dota 2 is feminine in Czech)

h: No tak tam není moc široký výběr hrdinů, Geralta
Well there is not a broad selection of heroes, Geralt

...

Since multiple videogames were mentioned, the participant interpreted the question to be about the most recently mentioned one, *Witcher 3* which has also been topicalized by the chatbot. The chatbot was prompted to ask about the *dota 2* where also the question makes more sense.

Due to stimulus generation issues, sometimes the reference is misunderstood. In following conversation the stimulus reference the the *african continent* but the participant answers in a more broad way. They even mention this misunderstanding later in a comment. This means during the conversation, the anaphora is nonassignable and general enough, that the participant continues the conversation even if it entails a topic shift.

conversation id: 243

...

b: Dnes jsem se dozvěděla, že existuje více než 7 000 jazyků.
Today I learned there is over 7000 languages.

h: to je hodně zajímavé, už jsem to taky slyšela. Mám pocit, že třetinou z nich se mluví na africkém kontinentu
that is very interesting, I also heard about this. I think that third of them are spoken on the african continent.

b: Máš nějakou oblíbenou zemi nebo jazyk na něm?
So you have a favorite country or language on him? (continent is masculine in Czech)

h: mám docela ráda anglii a angličtinu, co ty?
I like england and english, what about you?

...

5.1.2 Associative anaphora

hardcoded phrase with intended nonassignable anaphora is sometimes assignable by association

An anaphora which lacks an explicit candidate mapping does not necessarily need to be nonassignable. Such associative anaphora¹ also occurs in the collected data:

conversation id: 248

¹M. Charolles. “Associative anaphora and its interpretation”. In: *Journal of pragmatics* 31.3 (1999), pp. 311–326.

...

h: Byla jsem v nemocnici za pacienty
I went to the hospital to see the patients

b: To muselo být zajímavé. Jaká byla tvoje nejčennější zkušenost?
That must have been interesting. What was your most valuable experience?

h: Je jich hodně
There was a lot

b: To zní dobrodružně, co tě tam dneska konkrétně zaujalo?
That sounds adventurous, what specifically interested you today?

h: Zvětšena slezina
enlarged spleen

b: Znamenalo to pro něj nějaké specifické komplikace?
Did it for him entail some complications?

h: No... je to příznak lymfomu
Well... it is a symptom of lymphoma

...

Here the chatbot response contains an anaphora which cannot be mapped onto a specific mentioned entity in the previous text but is understandable as there is clearly a person at hand in the situation the participant describes. The participant has then no issues continuing the conversation and moving the topic forward by answering the chatbot's question.

A spectrum could be shown of anaphoric devices that stems from an associative anaphora which is clearly related to a mentioned entity though lacks an explicit antecedent and a newly mentioned entity which then has the potential to retrospectively be pronounced a cataphora if it is lexically realized later on in the conversation.

5.1.3 Sequential coherence

Without directly provoking it via experiment design, some conversations show a pro-coherence perception on the participant side in terms of the sequential structure () of the conversation:

conversation id: 123

...

h: Boj s rakovinnými buňkami znáš?
Do you know the fight with cancer cells?

b: Jak se rozsivky konkrétně podílejí na boji s ní?
How diatoms specifically participate in the fight against it?

h: Ano.
Yes.

...

Here the participant answer hints at a misinterpretation of the chatbot response. While the chatbot is instructed to ask open-ended questions to keep the conversation going, the participant seems to understand the question in the chatbot response as a request for confirmation which is also given. This is likely because the chatbot question adds new topical content with its question which allows for interpreting the question as demonstrating a level of knowledge about the topic which the participant inquires about in the first question.

5.2 Topical (in)coherence

The deep anaphora stimulus type had the highest error rate. The annotation showed that only 64 conversations of this type were collected. That is approximately half of how many conversations were collected of the shallow and nonassignable anaphora. This relative failure has mainly been caused by the remaps discussed above. In its ideal form it looked something like this:

conversation id: 253

...

h: Začínám se připravovat na zkouškové období na vejšce, protože už se kvapem blíží

I am starting to prepare for the exam period in my university, because it is coming up quickly.

b: Jaké předměty ti dají nejvíc zabrat?

Which subjects are the hardest?

h: Asi demografická analýza

I guess demographical analysis.

b: Co tě na demografické analýze nejvíc baví, nebo naopak stresuje?

What is the most fun aspect or the most stressful about demographical analysis?

h: Nejvíce mě stresuje, že je tam opravdu hodně učiva. Baví mě tabulky v excelu

What stresses me the most is that there is really a lot of material. I like the excel tables.

b: Na co se pak těšíš nejvíc, až zkouškové skončí?

What do you look forward to the most, when the exam period is over?

h: Na ten klid a žádný stres

The peace and no stress

b: Na co ještě jiného, kromě klidu, se těšíš?

Is there anything else other than peace you are looking forward to?

h: Na to, že budu moct zase trávit víc času se svými přáteli

To being able to spend more time with my friends again

b: Jak dlouho už ji studuješ?
How long have you been studying it?

h: druhým rokem
second year

...

The chatbot question goes back to an entity that has only been mentioned several turns ago and the conversation topic has somewhat drifted apart since then. The participant would in this case be able to map the anaphora with the help of its lexical environment. Especially the verb of the question containing the anaphora makes or breaks the understanding. This is clear in following conversation, where the verb is repeated:

conversation id: 30

...

h: včera me bolela hlava
I had a headache yesterday

b: Jak dlouho už tě to trápí?
How long has that bothered you?

h: krk jen dnes a hlava dva dny
throat only today and head two days

b: Bolí tě i něco jiného, nebo jen hlava a krk?
Does anything else hurt other than head and throat?

h: nene
no no

b: Co si myslíš, že by ti mohlo pomoci?
What do you think could help you?

h: snad léky z lékárny
Perhaps medicine from the pharmacy

b: Co jsi dělal, když tě začala ona bolet?
What did you do when she started hurting? (note: head is feminine in Czech)

h: vzala si paralen
I took a paralen (note: paralen is a pill)

...

This mapping is so obvious due to the repetition that the inappropriate nonzero anaphora does not distract the participant.

5.2.1 Hard shift refusal

Various reactions to the deep anaphora stimulus type were recorded. The data shows sometimes mapping the anaphora does not warrant acceptance of the response.

conversation id: 114

...

b: To je těžké říct, ale miluju Gaudího. Co ty?
That is hard to say, but I love Gaudí. What about you?

h: Já také, co se ti na Gaudím líbí?
Me too, what do you like about Gaudí?

b: Líbí se mi jeho kreativita a přírodní inspirace. Co tě na něm oslovuje?
I love his creativity and inspiration in nature. What talk to you about him?

... 2 turns (4 utterances) ...

h: Vinylových desek má až moc :(nenapadá tě něco jiného?
He has way too many vinyls :(can't you think of something else?

b: Jeho organické tvary a barvy mě fascinují. Jak se ti líbí ta?
His organic shapes and color fascinate me. How to you like her?

h: Cože? O čem to mluvíš?
What? What are you talking about?

...

In this conversation the stimulus suffers from a generation issue and causes a meta-communicative reaction and later is commented upon:

comment: Chatbot nepochopil změnu tématu, pravděpodobně reagoval na romantismus, ale já se s ním chtěla bavit o dárcích. Navíc, nedokázal strukturovat otázku.

Chatbot misunderstood the topic shift, probably reacted to romanticism, but I want to talk about gifts. Besides he couldn't structure the question.

Other than the ideal case where deep anaphora supports the conversation coherence by following up on a previously mentioned topic, deep anaphora as it was programmed here would sometimes abruptly shift to an abandoned topic.

5.2.2 Incomplete control over conversation

One other frequent reason conversations that were meant to contain a deep anaphora stimulus would fail was insufficient topical distance from the anaphorized entity. Since convform does not relate tracked entities to each other conversations like following could not be considered to contain a deep anaphora because the anaphorized entity would last occur too recently:

conversation id: 192

...

b: Pravěk je zajímavý! Co tě na něm nejvíc fascinovalo?
Prehistoric era is fascinating! What fascinated you most about it?

h: Mamuti
Mammoths

b: Mamuti jsou fascinující! Četl jsi o jejich vyhynutí?
Mammoths are fascinating! Did you read about their extinction?

h: Ještě jsem je viděl
I saw them still

b: Opravdu? Kde jsi je viděl a jaký to byl zážitek?
Really? Where did you see them and how was that experience?

... 4 turns (8 utterances) ...

b: Mamuti a jejich adaptace na ledové podmínky mě fascinují.
Mammoths and their adaptation to icy conditions fascinate me.

h: V těch knihovnách?
In those libraries?

b: Odkud sis vybral materiály k jejich výzkumu?
Where did you pick the materials for their research?

...

5.2.3 Nontopicalized entities

One of the requirements for an entity to be able to be mapped onto by a deep anaphora might be that it has been topicalized. One such case is found in the data:

conversation id: 162

...

b: Jaké hry máš nejraději hrát s kamarády?
Which games do you like to play with your friends the most?

h: Asi Sea of Thieves, Minecraft, nebo střílečky
Sea of Thieves, Minecraft or shooters I guess

b: Jakou hudbu nejraději posloucháš při hraní her?
Which music do you prefer to listen to while playing these games?

h: Ne vždycky u toho poslouchám hudbu, ale když ano, tak buď rap, nebo doporučení od dvou kanálů na Youtube
I don't always listen to music, but when I do it's either rap or recommendations of two Youtube channels

b: Jaké kanály na YouTube ti dělají hudební doporučení?
Which channels on Youtube give you recommendations?

h: Asi je neznáš, ale xkito a MrSuicideSheep
You probably don't know them but xkito and MrSuicideSheep

b: Jaké stavební projekty jsi v něm nejvíc dokončil?
Which building projects did you finish the most in him (note: Minecraft is masculine in Czech)

aborted

The participant aborts the conversation after a question is asked that contains an anaphora which due to its lexical environment can only be mapped to *minecraft* from a previous participant response as it is the only entity that has to do with *building projects*. This entity however has only been in the rhema position and mentioned long ago from the perspective of topical progression. Other issues contributing to the participant aborting the conversation are lack of the topic shift annotation and lack of closure of the interrupted topic. There has however also been an annotator disagreement on this conversation as one of the annotators considered this conversation to contain a nonassignable anaphora. While that was slightly against the annotator manual, because formally there is indeed a mapping candidate present in the conversation text, this supports the participants decision to stop the conversation. In terms of Daneš² the conversation topic progression is as follows:

topical progression of conversation id: 162

...

b: Which T1?

h: R1a, R1b , or R1c.

b: Which T2?

h: Either R2a or R2b.

b: Which T3(<-R2b)?

h: T3a and T3b.

b: Which T4(<-R1b)?

aborted

Here T4 is tied to R1b which never stood in the topic position. More conversations of this type need to be gathered to understand the role of topicalization in conversation coherence.

5.2.4 Unexpected refusal

Sometimes stimuli that were considered to be mappable and coherent cause metacommunicative reactions - the participants misunderstood.

In the following conversation the stimulus was mappable:

conversation id: 96

...

b: Co by tě na tom předmětu víc zaujalo?

What would interested you the most in that course?

h: kdyby ho učil pes

If it was given by a dog

b: Jaké bys od psa čekal vyučovací metody?

Which teaching methods would you expect from a dog?

²F. Daneš. "Functional sentence perspective and text connectedness". In: *Text and discourse connectedness* (1989), pp. 23–31.

h: štěkání v morseovce
barking morse code

b: A co bys dál od psa očekával v hodině?
And what else would you expect from a dog in the lesson?

h: mazlení. co ty?
cuddling. what about you?

b: Jakého bys chtěl mít jako učitele?
Which would you like to have as a teacher?

h: jak to myslíš?
how do you mean?

...

The question itself was however ambivalent in some way as the participant comment suggests:

comment: nenapsal jaké plemeno
didn't write which race

In following conversation, zero anaphora referring to the topic at hand caused meta-communication:

conversation id: 312

...

h: Mám, jmenuje se: She is the man
I do, it is called: She is the man

b: To je skvělý film! Co tě na něm nejvíc baví?
That is a great movie! What do you like about it the most?

h: Je to romantická komedie, takže mě baví to, že je to spíše na oddech.
It is a romcom so I like it more to relax.

b: Jaké typy máš nejraději?
Which types do you like best?

h: Typy čeho?
Types of what?

b: Typy filmů nebo seriálů, jaký žánr preferuješ?
Types of movies or series, which genre do you prefer?

...

Referring to a the current hypertopic via a zero anaphora has the potential to cause a spectrum of reaction.

5.2.5 Unexpected acceptance

Sometimes participants would accept a hardly or not at all assignable anaphora.

conversation id: 279

...

b: Kterou koledou začínáte zkoušky?
Which carol do you start your rehearsals with?

h: Každou zkoušku začínáme jinak
We start each rehearsal differently

b: Tak to by mě zajímalo, jak na to pak ona zareagovala.
Then I wonder, how she reaction to that.

h: To mě taky
Me too

b: Z jakého důvodu měníte pořadí hned po každém zkouškovém dni?
For which reason do you change the order after every rehearsal day?
aborted

This conversation has a gramatically feminine mapping candidate and somehow the chatbot remark though hardcoded makes sense to the participant.

In the following conversation the participant seems to want to continue in the conversation regardless of the incoherent response from the chatbot and so they simply change the topic:

conversation id: 142

...

h: a co jsi dělal ty?
and what did you do?

b: Dělal jsem si radost četbou a učením nových věcí. Ty máš nějaké oblíbené předměty?
I was enjoying myself by reading and learning new things. Do you have any favorite subjects?

h: biofyzikální chemii
biophysical chemistry

b: To zní zajímavě! Co tě na ní nejvíc baví?
That sounds interesting! What do you like most about it?

h: jako baví mě farmakologie
like I enjoy pharmacology

b: Tak to by mě zajímalo, jak na to pak ona zareagovala.
Then I wonder, how she reaction to that.

h: a ty studuješB
and you studyB

...

5.2.6 Nonfactual implication

The chatbot would sometimes hallucinate information that has not been said. Here the chatbot implies that the participant mentioned going to work that day:

conversation id: 264p

b: Ahoj, já jsem Vladimatic.
Hi, I am Vladimatic.

h: Ahojky
Hello

b: Jaký jsi měl den?
How was your day?

h: Docela to šlo. Co ty?
It was ok. What about you?

b: Taky se mám dobře, díky za optání.
I am also well, thanks for asking.

h: To je dobře
That's good.

b: Taky ráda něco podnikám po práci, co tě dneska bavilo?
I also like to do things after work, what did you enjoy today?

h: Já jsem dneska nebyla v práci a navíc mě dnes nic nebavilo...
I wasn't at work today and besides I enjoyed nothing today...

...

Implying things that were not mentioned leads to a special type of metacommunicative reaction which can contain negated repetition of the incorrect statement.

5.3 Metacommunication

One of the stimulus reaction types is metacommunication. This is a response which somehow addresses the course of the interaction itself. It is not trivial and was therefore annotated. In its simplest form metacommunicative reactions would be a question that contains a repetition of the element causing confusion like here:

conversation id: 317

...

b: Tak to by mě zajímalo, jak na to pak ona zareagovala.
Then I wonder, how she reaction to that.

b: kdo ona?
who she?

...

In other cases a more general metacommunicative formulas are deployed which express general confusion. This happened in following case for example since there were multiple unmapped anaphoras preventing a simple metacommunication.

270 double anaphora -> meta "

conversation id: 270

...

b: To je skvělý plán, co nejraději v posilovně cvičíš?
That is a great plan, what do you like to excersise in the gym?

h: nejraději cvičím záda. myslím, že budu dělat cviky na záda zrovna dneska
I like the back the most. I think I will do back excersise today

b: Které konkrétně na něj mají vliv?
Which exactly have influence on him?

h: jak to myslíš, napiš mi prosím otázku líp
what do you mean, write the question better please

...

Metacommunication signals a conversation coherence disruption but they give a chance to fix the issue at hand. The clarification or correction may not suffice like in the following case:

conversation id: 101

...

b: To jsi naprosto pravdu! Jaké knižní překvapení tě naposledy potkalo?
You are completely right! Which book surprise did you get last?

h: Asi brandon sanders, i když sci-fi vlastně vůbec nečtu
I guess brandon sanders even though I don't actually read sci-fi at all.

b: Tak to by mě zajímalo, jak na to pak ona zareagovala.
Then I wonder, how she reaction to that.

h: Kdo?
Who

b: Omlouvám se, myslel jsem na postavy z jeho knih. Co se ti na Sandersonovi líbilo?
Sorry, I meant the characters from his books. What did you like about Sander-son?

abort

5.4 Conclusion

The inter-annotator agreement shows that the anaphora stimulus categories are valid. However a deeper look into the data on both quantitative and qualitative levels shows that there are further subgroups of coherence disruptions and anaphora types which the categorization did not cover. This was to a degree intentional. Using a simplified model to understand the studied area via describing the models weaknesses indeed brought an insight and managed to illuminate further steps.

The general research question has been to confirm that the coherence would be increasingly more disrupted with each anaphora type - shallow, deep and nonassignable. In most measurable ways this proved to be true. This was not a surprising result as rather than answering the question the goal of the research has been to gather concrete data supporting this claim and confirming the usefulness of the conversation simulation methodology in the process. Participant reactions to the different types of stimuli differed greatly, the more disruptive stimuli had more participants abandoning the conversation and the participant comments show that while the shallow anaphora stimulus went by practically unnoticed, the deep and nonassignable ones had increasingly more participants make a remark on them. The only metric that did not quite succeed showing the predicted outcome was global rating of the conversation. The results seem to go in the expected direction, but are not always statistically significant and in some cases the deep anaphora conversations have a better average score than the shallow anaphora ones. This is generally because the rating covers the entire conversation the participants had and many other things happen in the conversation other than the stimulus. For future research it is necessary to get the general course of the conversation more under control.

On a similar token, the general conversation style variable did not show any significant results. Same cause can be attributed to this. More control over the chatbot has to be gained to program such that might generate better or worse participant ratings.

One aspect where the participant rating give an interesting result is the zero anaphora. Conversation containing it have much better average rating than ones where it is explicit. The data confirms an already known fact that Czech is a pro-drop language. This must be used in the future to increase the contrast in the observed phenomena by decreasing noise from this nonacceptability source.

The data shows a strong pro-coherence perception. The participants are often willing to map the anaphora to whichever candidate is at hand even if it is only implicit and associative.

The deep anaphora stimulus type has been an interesting look into the inner workings of conversation. The chatbot meant to elicit this type of data was relatively unsuccessful. Improvements to be made in that regard mainly lie in more thorough real-time topic annotation and letting the post stimulus prompt access the intended reference. That way the acceptance of the explanation in case of metacommunicative reaction can also be observed.

In general it has been observed that the deep anaphora acceptance depends on several aspects. The topic shift needs to be annotated in the discourse, there needs to be enough information to perform mapping, there needs to be no critical ambivalence to perform mapping and an entity that has not been previously topicalized is less likely to be mapped onto.

Chapter 6

Final thoughts

6.1 Evaluation of anaphora categorization

The categorization used in this research has been simplified for exploration. The immediate observation has been that both deep and nonassignable anaphora have two dominant participant reactions. The deep anaphora was had a similar number of acceptance and metacommunicative reaction. This begs a question whether it is possible to program and prompt two different chatbots, that will be capable of sharing these two types of reactions. In other words the question is whether it is possible to simulate an immediately acceptable and a confusing deep anaphora. The same goes for the nonassignable anaphora where the split was on the metacommunication and abroting the conversation. Here the task to split these two types up among two different chatbots is even trickier, because the criteria of not understanding and directly ending the conversation might depend on individual circumstances. That way the textual aspect of the conversation would have a lesser effect.

That qualitative analysis shows also that at least one disruptive anaphora type has been elided from the categorization. In many cases the participant confusion stemmed from the ambivalence of the anaphora - there has been multiple mapping candidates. There is a hint of a more granular categorization of disruptive anaphora.

Finally on anaphora and conversation coherence in general, it is not easy to just break the anaphora, something always comes along with it as the various factors at hand in conversation are intertwined. The data certainly showed, that everytime there is a anaphora related coherence disruption in conversation, this impacts the topical structure of the unfolding text of conversation which is where the incoherence ends up being perceived.

6.2 Further steps

6.2.1 Experiment design improvement

For future research better control over the conversation needs to be gained. That way the amount of noise can be decreased as there would be less sources of confusion or nonacceptability in the conversation text other than the stimulus. Participants could also be exposed to a sequence of conversations with different stimulus each time. That way the stimuli could be ordered a certain way and a rating development metric would be gained. Lastly more granular approach to the participant reaction to stimulus could be taken. At least one type has been ommited or rather fused with the metacommunication

reaction type - the late aborting of the conversation. After the stimulus if the participant reacts in a metacommunicative way, the chatbot has a chance to fix the disruption. If it does not succeed in that, the coherence remains disrupted and the social interaction illusion falls due to the stimulus but not directly after the stimulus. For this the chatbot also needs to have access to which entity the anaphora has been referring to.

6.2.2 Stimulus generation improvement

Some of the challenges met during the course of the research have to do with the stimulus generation. One improvement that has been mentioned above is the chatbot being aware of what it is referring to with its anaphoras. In this version of convform, this information has not been stored anywhere. It should be present in the prompt in case the participant is confused and asking about it. The secondary reaction might show which types of referents will be acceptable or less so.

The conversation quality needs to improve. The zero anaphora has to be deployed where it is appropriate to decrease the nonacceptability noise. Better topic progression awareness needs to be developed in convform. The chatbot sometimes ignores questions and asks its own which is topically and sequentially inappropriate.

Lastly the nonassignable anaphora stimulus could be improved to generate stimulus responses via prompt. That would require generating an irrelevant entity and anaphorically referring to it. This challenge could provide more insight on associative anaphora if tackled successfully.

6.2.3 Hybrid approach

All of the aforementioned improvements are hard to achieve with the current state of convform. At the time of this research it functions as a hybrid tree structure where the response is either hardcoded or a prompt. To gain more flexibility, the dialog management needs to be more prompt oriented. However to keep the control, the prompt needs to consist of components that self regulate based on logic and analysis prompts. Essentially with each turn metadata like topical and sequential status of the conversation should be collected. The components should then be allowed to assemble the prompt based on this data while following the conversation designer instruction. This process should allow for creating a dialog system that is aware of what is going on and can conditionally produces prompts and generate responses that are defined by the conversation designer while maintaining the flexibility with which state-of-the-art systems interact.

6.3 New opportunities

With a dialog system of this nature intentionally creating the stimuli hinted at in the analysis would become possible:

- Understanding the bimodality of deep and nonassignable anaphora
- Deep anaphora referring to a (non)topicalized entity
- Closer look at zero reference and associative anaphora
- Observing participant reactions to nonfactual implications

- Comparing different anaphora depths
- Topic shift with or without discourse markers
- Secondary participant reactions to repair attempts

6.4 Conclusion

The methodology used in present research has been experimental on two levels. One was to see whether it can be useful for conversation research. While there are many improvements to be made on both the experimental framework and the stimuli, it has proven to be able to generate data that would be otherwise impossible to get. This comes at the price of annotation. Every conversation has to be visited to confirm whether the required stimulus is actually present. With improving conversation design, this might become a smaller issue and with a certain degree of precision and a certain scale, annotations might become obsolete. For now and near future, the conversations will be visited. Seeing the interactions with one's own eyes has been the goal of the research regardless, so annotating can also be perceived as a reward rather than a price.

The other level was to expose participants to a chatbot that would subvert the course of the conversation. This has shown to be successful to a decent degree although here also it is a lot of improvement to be made in both practical and theoretical aspect. The result itself has been more or less unsurprising and the experiment served mainly to generate data that would support this uncontroversial claim. From here, multiple research paths can be taken in both understanding the inner workings of conversation and attempting to simulate various situations using the dialog system.

Bibliography

- Andrist, S. et al. “Turn-Taking and Coordination in Human-Machine Interaction”. In: *Ai Magazine* 37 (Dec. 2016), pp. 5–6.
- Austin, J. L. *How to do things with words*. Oxford University Press, 1962.
- Ayuso, D. “Discourse entities in Janus”. In: *27th Annual Meeting of the Association for Computational Linguistics*. 1989, pp. 243–250.
- Baldassarre, M. T. et al. “The social impact of generative ai: An analysis on chatgpt”. In: *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 2023, pp. 363–373.
- Barthes, R. and S. Heath. *Image, Music, Text*. A fontana original. Fontana Press, 1977.
- Brabra, H. et al. “Dialogue Management in Conversational Systems: A Review of Approaches, Challenges, and Opportunities”. In: *IEEE Transactions on Cognitive and Developmental Systems* 14.3 (2022), pp. 783–798. DOI: 10.1109/TCDS.2021.3086565.
- Brown, P. and S.C. Levinson. *Politeness: Some Universals in Language Usage*. Politeness: Some Universals in Language Usage. Cambridge University Press, 1987.
- Bruno, A. et al. “Insights into Classifying and Mitigating LLMs’ Hallucinations”. In: *arXiv preprint arXiv:2311.08117* (2023).
- Bublitz, Wolfram and Uta Lenk. “Disturbed coherence: ‘Fill me in’”. In: *PRAGMATICS AND BEYOND NEW SERIES* (1999), pp. 153–174.
- Charolles, M. “Associative anaphora and its interpretation”. In: *Journal of pragmatics* 31.3 (1999), pp. 311–326.
- Coppock, L. “Politeness strategies in conversation closings”. In: *Unpublished manuscript: Stanford University* (2005).
- Couper-Kuhlen, E. and M. Selting. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge University Press, 2017.
- Daneš, F. “Functional sentence perspective and text connectedness”. In: *Text and discourse connectedness* (1989), pp. 23–31.
- Davidson, J. A. Y. *Ending structures in conversation*. University of California, Irvine, 1975.
- Deshpande, A. et al. *Toxicity in ChatGPT: Analyzing Persona-assigned Language Models*. 2023. arXiv: 2304.05335 [cs.CL]. URL: <https://arxiv.org/abs/2304.05335>.
- Dingemanse, M. and N. J. Enfield. “Interactive repair and the foundations of language”. In: *Trends in Cognitive Sciences* 28.1 (2024), pp. 30–42.
- Ferrucci, D. et al. “Building Watson: An Overview of the DeepQA Project”. In: *AI Magazine* 31.3 (July 2010), pp. 59–79. URL: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2303>.
- Garfinkel, H. “Studies in ethnomethodology”. In: Routledge, 2023, pp. 209–261.
- Gervits, F. et al. “It’s About Time: Turn-Entry Timing for Situated Human-Robot Dialogue”. In: *Proceedings of the Special Interest Group on Discourse and Dialogue*. 2020. URL: <https://hrilab.tufts.edu/publications/Gervits2020Sigdial.pdf>.

- Givón, T. *Coherence*. John Benjamins Publishing Company, 2020.
- Gravano, A. and J. Hirschberg. “Turn-taking cues in task-oriented dialogue”. In: *Computer Speech & Language* 25.3 (2011), pp. 601–634.
- Grice, H. P. “Logic and Conversation”. In: *Syntax and Semantics: Vol. 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. Academic Press, 1975, pp. 41–58.
- H. Naveed, Humza et al. “A comprehensive overview of large language models”. In: *arXiv preprint arXiv:2307.06435* (2023).
- Hagerty, A and I. Rubinov. “Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence”. In: *arXiv preprint arXiv:1907.07892* (2019).
- Hajičová, E., J. Havelka, and P. Sgall. “Discourse Semantics and the Salience of Referents”. In: *Journal of Slavic Linguistics* 11.1 (2003), pp. 127–140.
- Halliday, M. A. K. and R. Hasan. *Cohesion in English*. Longman, 1976.
- Hendrix, G. G. “Natural-language interface”. In: *American Journal of Computational Linguistics* 8.2 (1982), pp. 56–61.
- Hjelmlev, L. *O základech teorie jazyka*. Trans. by F. Čermák. 2. doplněné a upravené vydání. Praha: Academia, 2016.
- Hoji, H. “Surface and deep anaphora, sloppy identity, and experiments in syntax”. In: *Anaphora: A reference guide* (2003), pp. 172–236.
- Hrbáček, J. *Nárys textové syntaxe spisovné češtiny*. Praha: Trizonia, 1994.
- Hu, K. and A. Tong. “OpenAI Rivals Seek New Path to Smarter AI as Current Methods Hit Limitations”. In: *Reuters* (2024). URL: <https://www.reuters.com/technology/artificial-intelligence/openai-rivals-seek-new-path-smarter-ai-current-methods-hit-limitations-2024-11-11>.
- Hwang, Y. et al. “LLMs can be easily Confused by Instructional Distractions”. In: *arXiv preprint arXiv:2502.04362* (2025).
- Ivey, J. et al. “Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Dialogue”. In: *arXiv preprint arXiv:2409.08330* (2024).
- Jang, J., S. Ye, and M. Seo. “Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts”. In: *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*. Ed. by Alon Albalak et al. Vol. 203. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 52–62.
- Jones, C. R. and B. K. Bergen. *People cannot distinguish GPT-4 from a human in a Turing test*. 2024. arXiv: 2405.08007 [cs.HC]. URL: <https://arxiv.org/abs/2405.08007>.
- Jonsdottir, G. and K. Thórisson. “A Distributed Architecture for Real-Time Dialogue and On-Task Learning of Efficient Co-Operative Turn-Taking”. In: Oct. 2013, pp. 293–323.
- Koblížek, T. *Interpretační sémantika. Úvod do textové teorie Francoise Rastiera*. Univerzita Karlova v Praze, Filozofická fakulta, 2015.
- Kološová, Š. “Konverzační design: principy designu hlasového robota pro přirozenou komunikaci s lidmi”. Diplomová práce. Praha: Univerzita Karlova, Filozofická fakulta, Ústav informačních studií - studia nových médií, 2022.
- Kong, A. et al. “Better zero-shot reasoning with role-play prompting”. In: *arXiv preprint arXiv:2308.07702* (2023).
- Laban, G. et al. “Robo-Identity: Exploring Artificial Identity and Emotion via Speech Interactions”. In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2022, pp. 1265–1268.
- Li, Yinheng. “A practical survey on zero-shot prompt design for in-context learning”. In: *arXiv preprint arXiv:2309.13205* (2023).

- Loaiciga, S., A. Beyer, and D. Schlangen. “New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities”. In: *International Conference on Computational Linguistics*. 2022.
- Loaiciga, S., Simon D., and D. Schlangen. “Reference and coreference in situated dialogue”. In: *Proceedings of the Second Workshop on Advances in Language and Vision Research*. 2021, pp. 39–44.
- Loaiciga, S., S. Dobnik, and D. Schlangen. “Anaphoric Phenomena in Situated dialog: A First Round of Annotations”. In: *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Oct. 2022, pp. 31–37. URL: <https://aclanthology.org/2022.crac-1.4/>.
- Makela, E. “Survey of semantic search research”. In: *Proceedings of the seminar on knowledge management on the semantic web*. Department of Computer Science, University of Helsinki, Helsinki. 2005.
- Maršík, A. *Convform*. <https://github.com/almarsk/convform>. 2024.
- Mastroianni, Adam M. et al. “Do conversations end when people want them to?” In: *Proceedings of the National Academy of Sciences* 118.10 (2021), e2011809118.
- McTear, M. F. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer International Publishing, 2020, pp. 43–70.
- Metz, R. et al. “OpenAI, Google, and Anthropic Are Struggling to Build More Advanced AI”. In: *Bloomberg News* (2024). Accessed: 2024-12-19. URL: <https://www.bloomberg.com/news/articles/2024-11-13/openai-google-and-anthropic-are-struggling-to-build-more-advanced-ai>.
- Moy, C. “Knuth–Morris–Pratt illustrated”. In: *Journal of Functional Programming* 34 (Jan. 2024).
- Nedoluzhko, A. “Rozšířená textová koreference a asociační anafora (koncepte anotace českých dat v pražském závislostním korpusu)”. PhD thesis. Prague: Univerzita Karlova, Filozofická fakulta, Ústav českého jazyka a teorie komunikace, 2010.
- Olah, C. *Mechanistic Interpretability – July 2024 Update*. Accessed: 2025-03-29. July 2024. URL: <https://transformer-circuits.pub/2024/july-update/index.html#dark-matter>.
- Payandeh, A. et al. “How susceptible are llms to logical fallacies?” In: *arXiv preprint arXiv:2308.09853* (2023).
- Perkovic, G., A. Drobnjak, and I. Boticki. “Hallucinations in llms: Understanding and addressing challenges”. In: *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE. 2024, pp. 2084–2088.
- Pešková, A. “Slavic and Romance pro-drop in contrast: Evidence from Czech and Spanish”. In: *Languages in contrast* 19.2 (2019), pp. 310–333.
- Qin, C. et al. “Relevant or Random: Can LLMs Truly Perform Analogical Reasoning?” In: *arXiv preprint arXiv:2404.12728* (2024).
- Ramnauth, R., D. Brščić, and B. Scassellati. *More than Chat-Chat: Developing Robots for Small-Talk Interactions*. 2024. arXiv: 2412.18023 [cs.R0].
- Raux, A. and M. Eskenazi. “Optimizing the turn-taking behavior of task-oriented spoken dialog systems”. In: *ACM Trans. Speech Lang. Process.* 9.1 (May 2012).
- Ritchie, H. “What’s the impact of artificial intelligence on energy demand?” In: *Sustainability by Numbers* (Nov. 2024). URL: <https://www.sustainabilitybynumbers.com/p/ai-energy-demand>.

- Ritter, G. and W. Lu. “The AI Plateau Is Real — How We Jump To The Next Breakthrough”. In: (Dec. 2024). URL: <https://www.emcap.com/thoughts/ai-s-curve-plateau-proprietary-business-data-breakthrough/>.
- Roberts, R. M. and R. J. Kreuz. “Nonstandard discourse and its coherence”. In: *Discourse Processes* 16.4 (1993), pp. 451–464.
- Sacks, H. *Lectures on Conversation*. Ed. by G. Jefferson. Oxford: Basil Blackwell, 1992.
- Sahoo, P. et al. “A systematic survey of prompt engineering in large language models: Techniques and applications”. In: *arXiv preprint arXiv:2402.07927* (2024).
- Schegloff, E. A. “On the organization of sequences as a source of ‘coherence’ in talk-in-interaction”. In: *Conversational Organization and its Development*. Ed. by B. Dorval. Ablex, 1990, pp. 51–77.
- *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press, 2007.
- Schegloff, E. A., G. Jefferson, and H. Sacks. “The Preference for Self-Correction in the Organization of Repair in Conversation”. In: *Language* 53.2 (1977), pp. 361–382.
- Schegloff, E. A. and H. Sacks. “Opening up Closings”. In: *Semiotica* 8.4 (1973), pp. 289–327.
- Searle, J. R. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.
- Sharma, D. et al. “Exploring The Evolution Of Chatgpt: From Origin To Revolutionary Influence”. In: *Educational Administration: Theory and Practice* 30.5 (2024).
- Skantze, G. “Turn-taking in Conversational Systems and Human-Robot Interaction: A Review”. In: *Computer Speech & Language* 67 (May 2021), pp. 101–178.
- Sreedhar, M. N. et al. *CantTalkAboutThis: Aligning Language Models to Stay on Topic in Dialogues*. 2024. arXiv: 2404.03820 [cs.CL].
- Stevenson, C. E. et al. “Can Large Language Models generalize analogy solving like people can?” In: *arXiv preprint arXiv:2411.02348* (2024).
- Tie, J. et al. “LLMs are Imperfect, Then What? An Empirical Study on LLM Failures in Software Engineering”. In: *arXiv preprint arXiv:2411.09916* (2024).
- Turing, A. M. “Computing Machinery and Intelligence”. In: *Mind* 59.236 (1950), pp. 433–460.
- Wei, A., N. Haghtalab, and J. Steinhardt. “Jailbroken: How does llm safety training fail?” In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 80079–80110.
- Wei, J. et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.
- Weizenbaum, J. “ELIZA a Computer Program for the Study of Natural Language Communication Between Man and Machine”. In: *Commun. ACM* 9.1 (Jan. 1966), pp. 36–45.
- Wu, S. et al. “A Comparative Study on Reasoning Patterns of OpenAI’s o1 Model”. In: *arXiv preprint arXiv:2410.13639* (2024).
- Xu, Jun et al. “Discovering Dialog Structure Graph for Coherent Dialog Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong et al. Association for Computational Linguistics, Aug. 2021, pp. 1726–1739.
- Yang, Ning et al. “A Study of Zero Anaphora Resolution in Chinese Discourse: From the Perspective of Psycholinguistics”. In: *Frontiers in Psychology* 12 (2021). ISSN: 1664-

1078. DOI: 10.3389/fpsyg.2021.663168. URL: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.663168>.
- Yao, B. et al. “More samples or more prompts? exploring effective in-context sampling for llm few-shot prompt engineering”. In: *arXiv preprint arXiv:2311.09782* (2023).
- Yin, Z. et al. “Do large language models know what they don’t know?” In: *arXiv preprint arXiv:2305.18153* (2023).
- Z. Wang, Zhichao et al. “A comprehensive survey of LLM alignment techniques: RLHF, RLAIIF, PPO, DPO and more”. In: *arXiv preprint arXiv:2407.16216* (2024).
- Zhou, Y. et al. “Don’t Say No: Jailbreaking LLM by Suppressing Refusal”. In: *arXiv preprint arXiv:2404.16369* (2024).