

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

Albert Maršík

Contents

1	Introduction and Theory	2
2	Experimental framework	3
2.1	What are chatbots?	3
2.2	Convform	5
2.3	Conversation design in theory	6
2.3.1	Rule-based approach	6
2.3.2	Statistically driven approach	7
2.4	Conversation design in practice	9
2.4.1	Stimuli	11
2.4.2	Ending the conversation	13
3	Data	14
3.1	Data collection	14
3.1.1	Experimental designs	14
3.1.2	Participants	15
3.2	Data analysis	15
3.3	Annotation	16
3.3.1	Inter-annotator agreement	17

Chapter 1

Introduction and Theory

..

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

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

²M. F. McTear. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer International Publishing, 2020, pp. 43–70, 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>.

by one answer generator.

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. Brabrá 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. It has been convincingly shown that this technology has the capability of reacting in a flexible way to much of what is being thrown at it (). While not perfect () this technology is capable of staying on topic (), mirroring () 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 models (). 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 area (), 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 or examples (). 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 direction ().

More sophisticated strategies include "few-shot" prompting, 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. Since analogy is a task LLMs are doing really well in () framing the task as an analogy can help improve the output significantly ().

Another very prevalent way that has proven to improve the performance of LLMs is a so called chain-of-thought prompting (). 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 metrics ().

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 hallucinating (). Factuality is a challenge for LLMs overall. It is possible to ask questions, that do not have a correct answer (). It has been shown, that LLMs have an issue knowing that they do not know something (). Recognizing that is the case requires an extra level of reasoning that is an object of research as of recently ()[dark matter of ai]

They are also required to be ethically aligned with humans. This is an ongoing research field which has been making some troublesome observations recently with the most intelligent models available ().

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 breaks (). 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.

With all this in mind, how much of a responsibility do we want to allow LLMS to have? LLM powered technology is expected to be making decisions that will and do affect people on a daily basis. Hopefully that will not happen before the challenges of alignment and hallucinations are solved.

As far as low-stakes open-domain conversation simulation goes, LLM powered chat-bots are relatively safe though there are cases of dangerous or tragic situations especially for vulnerable individuals ().

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.

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 ¹⁸

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.

¹⁸<https://github.com/almarisk/convform/wiki>

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 discussed^{19,20}. 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.

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 Hrbáček²¹. 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. () New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities 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 anaphore, convform first needs to be able to give a response that has an anaphoric reference to an entity from the previous conversation in it.

¹⁹A. Deshpande et al. *Toxicity in ChatGPT: Analyzing Persona-assigned Language Models*. 2023. arXiv: 2304.05335 [cs.CL]. URL: <https://arxiv.org/abs/2304.05335>.

²⁰G. 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.

²¹J. Hrbáček. *Nárys textové syntaxe spisovné češtiny*. Praha: Trizonia, 1994, p. 30.

The GPT-4o model used for this use case does not tend to generate sentences with anaphores 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 text () Annotating anaphoric phenomena in situated dialogue. 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 anaphore, deep anaphore and nonassignable anaphore.

Shallow anaphore

A shallow anaphore is a kind of anaphore where the referent of the anaphoric device should be relatively easy to map as opposed to a deep anaphore. The referent will always occur in the preceding utterance of the participant. This type of stimuli is relatively simple to achieve in convform generating a response and using the anaphorization prompt on it afterwards.

Shallow anaphore 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 anaphore

A deep anaphore 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 anaphore refers to. Measuring depth of anaphore 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 covnersation 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 anaphore 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 anaphore 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.

This approach is relatively imprecise and relies on luck to a certain degree. Compared to the shallow anaphore 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 anaphore. 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 ones () but this measure seems to have mitigated the problem as can be seen in the data attached in the annex.

Nonassignable anaphore

The last type of stimuli used in this paper has been called nonassignable anaphore. 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 anaphore 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 anaphore.

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 anaphore 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 anaphore either.

An approach not explored in this paper is achieving a nonassignable anaphore 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 instructions () this could be achieved using a chain of prompts. The notion of nonassignable anaphore brings into scope the question of what makes an anaphora assignable. It is the semantic compatibility of the words around the anaphore that determine which of the candidates the anaphore is referring back to. The generation of the lexical surroundings of the anaphore needs to be handled carefully when coming up with an LLM based response.

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 speaking (). 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.

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 successful in doing so. Further analysis leading up to a quantitative assessment should only be done on conversations where convform successfully 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 conversations are 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.

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.
- Bublitz, Wolfram and Uta Lenk. “Disturbed coherence: ‘Fill me in’”. In: *PRAGMATICS AND BEYOND NEW SERIES* (1999), pp. 153–174.
- Couper-Kuhlen, E. and M. Selting. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge University Press, 2017.
- 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.
- 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.
- Hjeldslev, L. *O základech teorie jazyka*. Trans. by F. Čermák. 2. doplněné a upravené vydání. Praha: Academia, 2016.
- 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>.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Turing, A. M. “Computing Machinery and Intelligence”. In: *Mind* 59.236 (1950), pp. 433–460.
- 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.
- 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>.