MSc Business Information Techonology

# Exploring Lexical Alignment Influences in Price Bargain Chatbot

Amy Zhao
z.zhao-2@student.utwente.nl

Supervisor: Maurice van Keulen m.vankeulen@utwente.nl
Supervisor: Mariet Theune m.theune@utwente.nl
Supervisor: Daniel Braun d.braun@utwente.nl
Supervisor: Sumit Srivastava s.srivastava-1@utwente.nl

February, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

**Abstract**

This study investigates the integration of lexical alignment into text-based negotiation chatbots and its impact on user perception. Lexical alignment, where participants in a conversation adopt similar language, is examined for its potential to enhance chatbot user experiences. This study proposed a chatbot architecture designed for price bargaining, incorporating lexical alignment. The architecture consists of seven components: user interface, intent classifier, price extractor, product extractor, dialogue management, response generation, and database. Key among these are the intent classifier, which categorizes user inputs; the price and product extractors, which identify prices and products mentioned in the conversation; dialogue management, which follows predefined rules for chatbot actions; and response generation, which crafts appropriate responses using OpenAI's API and prompt engineering techniques. This response generation can include lexical alignment features or not. To evaluate the effects of lexical alignment, a user experiment was conducted. While the results did not show statistical significance, they suggest that lexical alignment might positively influence user satisfaction. This finding indicates a potential direction for enhancing user interaction with chatbots in the future.



*Keywords*: chatbot, lexical alignment, chatGPT

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The field of Artificial Intelligence (AI) has witnessed rapid advancements in recent years, reaching new heights with the introduction of ChatGPT [37], a state-of-the-art conversational agent. Chatbots, a subset of these agents, are increasingly prevalent across various industries including healthcare [9], education [24], and BFSI (Banking, Financial Services, and Insurance), offering cost-effective, efficient, and customizable services [2]. While commonly used by many businesses in customer service to increase customer interaction and improve consumer service [2], there's a growing trend to employ chatbots in more complex tasks like negotiation. Companies like Pactum are providing chatbots to automate the negotiation process to reduce operational costs with clients like Walmart and Maersk [4].

Despite widespread implementation, there is still a preference among customers for more human-like interactions, especially in problem-solving scenarios [67]. Enhancing user experience and trust in chatbots is crucial for customer adoption [64, 7], and linguistic alignment is emerging as a key factor in this regard. Linguistic alignment is a phenomenon where individuals adapt their language usage to align with their conversation partners during interaction [44]. It occurs at various levels, encompassing not just the mirroring of word choices, sentence structures, and phrases, but also aligning in understanding word meanings and higher-level concepts [65]. Research has shown that speakers intentionally use similar linguistic patterns to enhance understanding and foster positive feelings [12, 26, 43, 17] toward each other [73].

Despite a large number of research on linguistic alignment in human-human interaction, its application in human-computer interaction has only begun to gather attention. Most studies in this area have relied on the Wizard-of-Oz technique or have been limited to simple task-oriented chatbots, leading to short-term interactions and an inability to capture complex content [70]. This gap limits our understanding of linguistic alignment in more complex, real-life scenarios.

## 1.1 Research Questions

This study aims to explore the effect of lexical alignment in complex scenarios, specifically focusing on price bargaining tasks. Lexical alignment has demonstrated benefits in dialogue systems, including enhancing likability, trustworthiness, and satisfaction, and improving task success rates [59, 60, 53]. These aspects are critical in negotiations, which involve considerations like user emotions and cooperation [40]. Additionally, the increasing demand for negotiation chatbots presents unique challenges, such as analyzing customer behavior and controlling negotiation strategies [11, 46].

The main research question is formulated as follows:

- *How does lexical alignment in text-based negotiation chatbots influence user perceptions?*

This is supported by two sub-questions:

- *How can a chatbot be developed to perform lexical alignment and price bargaining with users?*
- *How to evaluate user perceptions and negotiation outcomes influenced by lexical alignment in chatbots?*

## 1.2 Research Structure

This study is organized into several chapters to address the research questions outlined above. Chapter 2 reviews relevant literature to establish a foundational understanding of chatbots and linguistic alignment. Chapter 3 examines current methodologies in developing chatbots for linguistic alignment and negotiation. Chapter 4 details the research methodology, including the design and development of a negotiation chatbot with lexical alignment features. Chapter 5 describes the user testing procedures and analyzes the results. Chapter 6 discusses these results, addressing the study's limitations. Chapter 7 concludes the thesis, answering the research questions and proposing future research directions.

# Chapter 2

# Background

This chapter presents an overview of chatbots, including an introduction to the various technologies they utilize and the growing interest in chatbot research within academia and their implementation in the industry. It then delves into the specialized area of negotiation chatbots and the dynamics of customer adoption of chatbots. This Chapter also introduces the phenomenon of linguistic alignment and its underlying mechanisms in human-to-human interactions. It then extends to the human-computer interaction area, highlighting how linguistic alignment enhances communication between humans and chatbots.

## 2.1 Chatbot

Conversational agents are software systems engineered to replicate human interaction, engaging through text, speech, and even gestures as outlined by Khatri et al. [52]. Chatbots, which represent a subset of these agents, have evolved significantly from their earliest simple version, ELIZA [75], which showed basic conversational capability. Presently, chatbots have diversified to provide assistance in various aspects of daily life. There are informative chatbots that serve to provide information, task-based chatbots that assist with specific tasks, and conversational chatbots that aim for more natural human-like interactions [6].

Current chatbots utilize different advanced technologies that can be classified into six main categories: template-based, corpus-based, intent-based, RNN (Recurrent Neural Network)-based, RL (Reinforcement Learning)-based, and those with hybrid models [63]. Template-based chatbots, like ALICE [74], use predefined response patterns, making them suitable for simple applications due to their straightforward development and deployment [63]. Corpus-based chatbots, in contrast, utilize a database to store and retrieve information, offering more efficient and scalable solutions [66]. Intent-based chatbots, widely used for task-oriented systems, exemplified by Rasa [16], leverage Natural Language Understanding (NLU) for multi-turn dialogues [16]. RNN-based chatbots, also known as generative chatbots, employ deep learning models like Seq2Seq (Sentence to Sentence), LSTM (Long Short-Term Memory), and GPT (Generative Pre-trained Transformer) to produce responses dynamically [6]. However, the unpredictable nature of deep learning models limits these chatbots' use in informative or task-based applications [63]. RL-based chatbots rely on reinforcement learning to generate responses but require comprehensive pre-defined dialogues for training [63]. Hybrid approaches aim to enhance chatbot performance by combining different techniques. For instance, certain intent-based chatbots utilize NLG (Natural Language Generation) to create responses aligned with dialogue actions [23, 76].

### 2.1.1 Chatbots in Academia and Industry

The interest in chatbots spans both academic research and industry applications. Figure 2.1 shows the number of publications related to Chatbot from 2013 to 2023, sourced from Scopus. This data was gathered using search terms: ("chatbot" OR "conversational agents" OR "dialogue system"). The search suggests the advancements in Artificial intelligence (AI) and Natural Language Processing (NLP) have been the gasoline for developing more advanced chatbots.

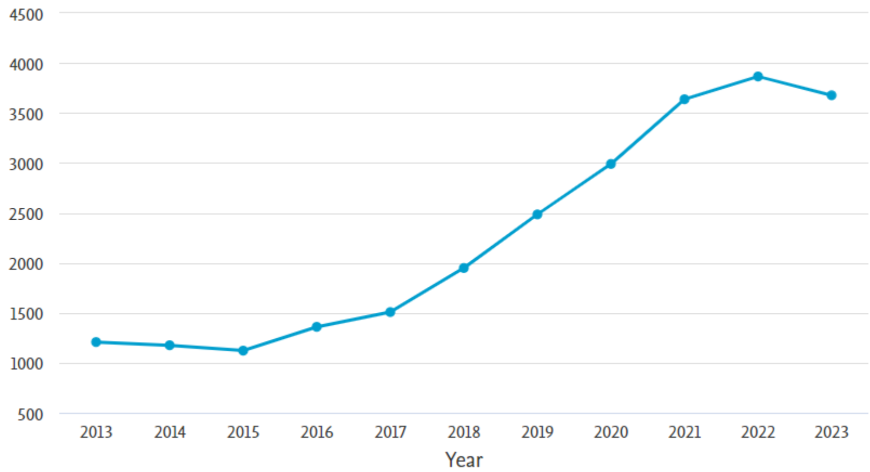



FIGURE 2.1: Publications on Scopus From 2013 to 2023

The implementation of chatbots across industries has seen a marked increase, caused by the 24/7 availability and cost-efficiency they offer. The global chatbot market, valued at approximately $5.13 billion in 2022, is expected to expand at an annual growth rate of 23.3% until 2030 [2]. North America is expected to see a steep increase in this market, as illustrated in Figure 2.2.

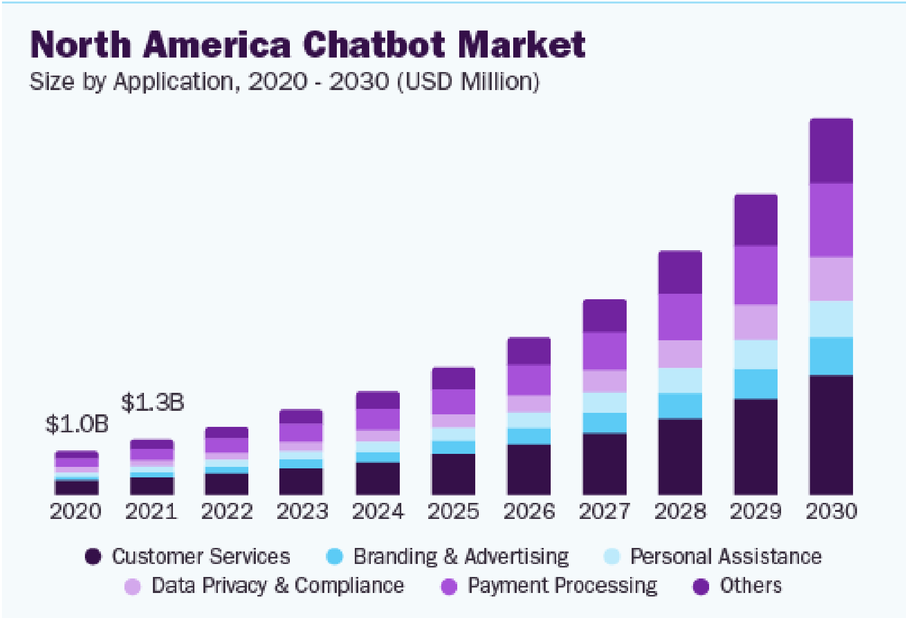



FIGURE 2.2: North America Chatbot Market[2]

In 2022, the retail and e-commerce sectors held significant shares of the chatbot market at 30.34%, respectively [2]. Moreover, diverse sectors like healthcare, banking, financial services, insurance, media, travel, and others are also adopting chatbots at a significant rate, as highlighted in Figure 2.3. Chatbots in e-commerce enhance customer interaction and support the shopping process. For instance, Lego's e-commerce chatbot "Ralph" successfully decreased the cost per conversion by 71% while accounting for a 25% increase in social media sales [1]. In healthcare, chatbots like Wysa addresses mental health issues by offering 24/7 support [5]. Banks and financial institutions utilize AI assistants like IBM's Watson to improve customer service and drive digital transformation [3].

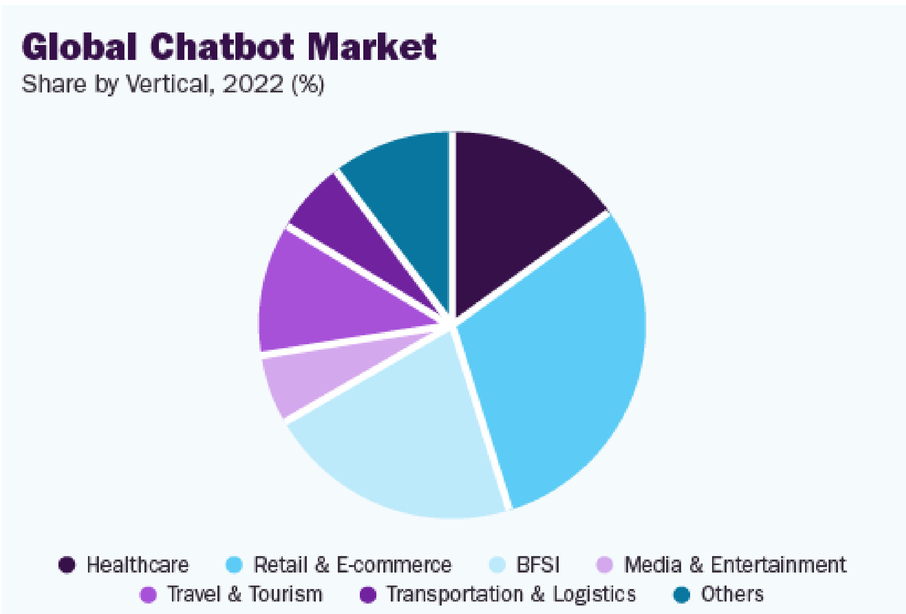



FIGURE 2.3: Global Chatbot Market[2]

### 2.1.2 Negotiation Chatbots

Negotiation chatbots represent an emerging area of application, facilitating negotiations in personal and business contexts. Negotiation, as defined by Fisher et al. pp. xxvii [39], is a "back-and-forth communication designed to reach an agreement when you and the other side have some interests that are shared and others that are opposed". Researchers have examined various factors that influence negotiation outcomes. One crucial factor is emotion during the negotiation process. Several studies suggest that positive feelings have a beneficial impact on negotiation by increasing confidence in judgments and promoting cooperative behaviors [33]. Furthermore, positive emotions contribute to building relationships among bargainers and facilitating social bonding, leading to more cooperative processes and expected outcomes [33, 38]. Additionally, the foundation of negotiation, the information processing, is influenced by emotions as well [33]. Another critical factor influencing negotiation outcomes is the level of trust among negotiators. Defining trust and distrust precisely remains an ongoing discussion [61]. However, one explanation is that trust (or distrust) is the response to positive (or negative) expectations regarding the actions of the other negotiators [61]. The higher the levels of trust, the greater the willingness to cooperate [61]. Moreover, politeness can also play a role in the negotiation process. Research shows that nonpolite conversations can damage the relationship among negotiators and the outcomes of deals [55]. However, excessive politeness may be interpreted as a lack of trustworthiness or result in a loss of leverage [55].

The implementation of negotiation chatbots in real business scenarios presents significant challenges due to the complex nature of negotiation processes and the "black box" nature of deep learning models. One major issue is ensuring the accuracy of information provided by these chatbots [27]. For example, a negotiation chatbot developed by Facebook turned out to lie to achieve better deals [27]. Moreover, controlling negotiation strategies is another challenge [46], such as easy to compromise [57]. Despite these challenges, the potential benefits of employing negotiation chatbots are substantial. For instance, the negotiation chatbot developed by Pactum and used by Walmart for supplier negotiations resulted in a 5.7% increase in savings across winning supplier bids. This success story exemplifies the significant advantages that can be realized when negotiation chatbots are effectively implemented.

### 2.1.3 Customer Adoption of Chatbots

Despite widespread implementation, customer resistance to chatbot usage persists. Several factors have been investigated by researchers. One notable impact is on customer experience, which encompasses usage convenience and perceived usefulness. Usage convenience refers to how easily and effortlessly customers can adopt the use of conversational agents [64]. Perceived usefulness, on the other hand, refers to the extent to which the consumers believe that using chatbots will enhance their performance [64]. Kasilingam developed a customized Facebook e-commerce chatbot that assists users in information searches, product recommendations, and order processing [51]. Participants were then asked to engage with the chatbot and complete a questionnaire. the results indicated that both usage convenience and perceived usefulness positively influence consumers' attitudes towards shopping chatbots [51]. Moreover, trust plays a significant role in chatbot adoption. Surprisingly, privacy does not have a substantial impact on trust [31]. In a study conducted by Arcand et al., an online survey was distributed to 375 participants who used mobile platforms for banking activities. The findings demonstrated trust significantly and positively impacts customer adoption and customer satisfaction [7].

## 2.2 Linguistic Alignment

Linguistic alignment is a phenomenon where participants in a conversation adjust their language use, encompassing both style and content, to resemble that of their counterparts. This adjustment can be observed at various linguistic layers: lexical, syntactic, and semantic [65]. At the lexical layer, alignment manifests among other things when individuals use similar referential expressions [25]. Syntactically, it involves matching sentence structures, exemplified by either using a Prepositional Object (e.g., "I gave her an apple") or a Double Object structure (e.g., "I gave an apple to her") [18]. Semantic alignment pertains to the convergence in understanding word meanings, expressions, and overarching concepts [18]. The Interactive Alignment Model (IAM) posits that the linguistic alignment between interlocutors leads to more effective communication [65].

### 2.2.1 Underlying Mechanisms in Human-Human Interaction

Several psychological theories have been proposed to elucidate linguistic alignment. The Interactive Alignment model, proposed by Pickering and Garrod [65], suggests that this alignment is multi-layered, commencing from superficial levels, such as word usage, and ascending to deeper aspects like linguistic style [65]. Notably, they argue that this alignment is predominantly subconscious, with participants often unaware of their linguistic mirroring behaviors [65, 19, 41]. Children's innate alignment tendencies during language learning further underscore this subconscious nature [45].

Conversely, Allan [12] propounds that speakers deliberately align their expressions to enhance comprehension—a term for "audience design" [12]. Further, empirical evidence by Clark and Fussell [26, 43] highlights that speakers accommodate their linguistic choices based on cultural knowledge. Similarly, when interacting with non-native speakers, native individuals strategically adapt their vocabulary to facilitate comprehension [17]. Thus, apart from subconscious mechanisms, conscious strategic considerations are also at play.

Moreover, social motivations may influence alignment. Using a shared lexicon can foster positive sentiments between speakers, leading to beneficial outcomes [73]. Overall, linguistic alignment evidently underpins successful human-to-human communication.

### 2.2.2 Linguistic Alignment in Human-Computer Interaction

While the beneficial implications of linguistic alignment in human-human interactions are well-established, its dynamics in human-computer interactions (HCI) remain an emerging research area. Many research endeavors aim to confirm whether the positive effects of linguistic alignment observed in human-human dialogues translate equivalently to HCI contexts. However, a limited number of studies have actually implemented conversational agents for experimental purposes. The majority of other experiments, inspired by Branigan et al. [20], utilized the Wizard-of-Oz setup and involved a picture-naming and matching task. This method involves a process where participants believe that they are interacting with an autonomous system, but instead, the system is actually being operated by a human. It is a common and valuable tool in human-computer interaction research, providing flexibility and feasibility. In these experiments, the presented pictures had a preferred name and an acceptable but less favored name. For instance, a favored name could be 'bus' while the alternative name could be 'coach'. Choosing the same words achieved alignment. After the experiments, users' feedback was gathered using questionnaires. This approach provides invaluable insights into HCI.

Regarding the lexical alignment of HCI, multiple studies affirm its occurrence [29, 14, 49, 53, 59, 60, 70, 71, 48, 13]. This alignment appears to be reciprocal, affecting both the human and computer entities [13, 53, 70]. Interestingly, in the Wizard-of-Oz experiment, participants aligned more with the opposite partner when they thought they were communicating with a computer compared to when they thought they were communicating with a human [14, 48], although contrasting findings exist [29].

The benefits of lexical alignment in HCI are manifold. It bolsters likability [59, 60], trustworthiness [60], and satisfaction [60] in spoken dialog system. Task success rates also improved, likely due to reduced communication ambiguities [53, 62]. Moreover, lexical alignment in HCI has been linked to improved information recall and comprehension [71] and reduced perceived effort and frustration [70].

## 2.3 Conclusion

This chapter has established the foundation for understanding chatbot implementation in the industry and the role of linguistic alignment in both human-human and human-computer interactions. It has identified current trends in chatbot usage, particularly in the area of negotiation chatbots, and has revealed the effects of linguistic alignment. This understanding is crucial for formulating the hypotheses of this study. It sets the stage for further exploration into these areas. The insights gained in this chapter will help guide the subsequent parts of this study, especially in how lexical alignment can affect negotiation chabot to enhance the user experience and achieve better negotiation outcomes.

# Chapter 3

# Related work

This chapter delves into research on linguistic alignment in text-based dialogue systems and innovative approaches in negotiation dialogue systems, highlighting key contributions and a comparative analysis of various methodologies.

## 3.1 Linguistic Alignment Dialogue System

Linguistic alignment in text-based conversational agents has been the focus of a few previous studies. Spillner et al. [70] introduced a method to enhance chatbot responses by incorporating both lexical and syntactic alignments. For the lexical alignment, they utilized spaCy's 300-dimensional GloVe word vectors to embed words and calculate their similarity, substituting the template answer terms with user-provided terms. For instance, while the default term for a movie's evaluation might be 'score', their chatbot could recognize and substitute with alternatives like 'rating'. For syntactic alignment, they align responses by transforming user queries using grammatical rules. This involved recognizing the components of English's structured sentence form, such as subject, verb, and object. Based on the type of question, responses are generated by inserting relevant information and adjusting linguistic elements. For non-standard or unrecognized inputs, the system resorts to predefined rules, offering greetings or fallback responses to maintain engagement. Lopes et al. [62] introduced a similar approach to substitute terms, employing a two-stage algorithm to select the best "primes". The first stage, called Long-Term entrainment, determines the initial prime that the system uses based on past conversations. The second stage, called Short-Term entrainment, selected the primes from users' inputs and substituted the initial prime.

The more dynamic approach proposed by Hu [47] is known as Personage Primed Architecture which is designed for the pedestrian direction giving domain. This dynamic architecture consists of three main stages: INPUT, Sentence Planning, and OUTPUT. In the INPUT stage, user utterances are processed and divided into various prime values based on a text plan and set of entrainment target values. For instance, an utterance like "Okay, now I'm at the corner of Cedar Street and Elm, so should I head toward the clock tower from here?" is deconstructed into primed values capturing prepositions (at, toward, from; Noun), Noun (I, corner, here), Verbs (am, head), etc. The Sentence Planning phase has three essential transformations. Firstly, the template that best matches the prime values is chosen, and these values are then combined with it to create the current utterance. Secondly, there is the pragmatic markers insertion stage. Pragmatic markers, such as "okay, so, now" in the previous example, do not contribute to semantic content but occur often in the natural flow of speech. In Personage Primed, if prime values corresponding to these markers are found in user input and align with the template, they are incorporated. The third stage involves lexical choice. This includes synonym selection, where synonymous verbs and prepositions in the user's prime values replace those in the current utterance. Referring expression selection replaces each destination noun in the current utterance against the prime values. Lastly, the tense transformation and modal insertion adjust the current utterance based on the prime values' tenses and modals. Similarly, Buschmeier [21] introduced the SPUD prime approach, which extends the microplanning system known as SPUD lite, for natural language generation (NLG), that paraphrases utterances based on breaking-down elements of user input. SPUD (sentence planning using description) lite is a system to generate utterances considering input grammar, knowledge base, and communicative goals [72]. Based on this, SPUD prime calculates values associated with the frequency of use of various linguistic structures and chooses the higher calculated values, thus facilitating alignment with previous language use.

Furthermore, Dubuisson Duplessis et al. [34] proposed a method combining the natural language generation process with a selection model that evaluates lexical alignment. The process consists of two modules. The first module, overgeneration, is responsible for generating alternative utterances. The second module ranks them based on their lexical alignment scores, selecting the top scorer as the ideal aligned response. To illustrate, if a user inputs the question, "Can you tell me something about the character, the white rabbit?" A basic response generated might be, "The rabbit is being chased by Alice." However, the alternative utterance module then refines this basic response. As a result of this refinement and ranking alignment scores, the response, "The white rabbit is being chased by Alice" is considered lexically aligned. Conversely, "The bunny is being chased by Alice" is perceived as not being lexically aligned. Similarly, Dušek et al. [35] introduced a related method employing a sequence-to-sequence (seq2seq) generation model and n-gram ranking module. The generation model has three variants. The first is a basic seq2seq model, used as a benchmark. Usually, seq2seq model has two main components, an encoder and a decoder. The encoder processes the input sequence, captures the most important features, and transforms it into a vector. Then the decoder transforms the vector into the output sequence. The second variation is the prepending context model. Unlike the basic version that only passes the input sequence into the encoder, this variant also incorporates the prior context into the encoder. The third version is the context encoder model. This version introduces an additional, separate encoder specifically for context utterances. So, the model consists of three components, one encoder for the current input, another separate encoder for the previous context, and a decoder to produce the output. After generating a set of possible responses (k responses), they are re-evaluated based on how well they align with the input's context. The one that aligns best becomes the chosen response.

## 3.2 Negotiation Dialogue System

Negotiated dialogue systems require advanced communication and strategic skills, and as a result, a variety of innovative approaches have emerged. Lewis et al. [57] proposed an end-to-end model demonstrating the possibility of combining linguistic ability and strategic reasoning in negotiation tasks. They collected 5808 human-human dialogues of distributing available items between two users from Amazon Mechanical Turk. Then they employed the training approach that combines supervised learning to produce the complete dialogue and then used reinforcement learning to fine-tune the agent's ability to achieve higher rewards within a dialogue context, simulating future parts of the conversation to evaluate different responses. This model employs a two-phase training approach. Initially, it uses supervised learning to generate complete dialogues. Subsequently, reinforcement learning is applied to refine the agent's ability to secure higher rewards within a dialogue. This is achieved by simulating future conversations to evaluate potential responses. Furthermore, He et al. [46] introduced a realistic dataset named Craigslistbargain, also via Amazon Mechanical Turk, for a buyer-seller negotiation setup. This dataset included complex negotiation behaviors such as casual conversation and side offers. They used a training approach similar to Lewis et al., combining supervised learning with reinforcement learning to train the Craigslistbargain dataset. This approach has shown effectiveness in generating responses that are not only meaningful and human-like but also diverse, decoupling various negotiation strategies. However, these models sometimes produced less desirable negotiation outcomes due to limited consideration of negotiation strategies. They occasionally generated aggressive responses or were too insistent on the offer price, aiming to maximize the deal value. This led to scenarios where humans were more likely to become upset or offended, potentially quitting the negotiation during the interaction.

Zhou et al. [79] proposed a framework for evolving strategies in negotiation that utilized a combination of machine learning classifiers and rule-based detection on the Craigslistbargain dataset. The paper categorizes negotiation strategies into two main types: integrative and distributive. Integrative strategies focus on mutual gain and collaboration, while distributive strategies are more about asserting one's position and interests. For instance, integrative strategies might include tactics like addressing buyers' concerns or negotiating side offers. Distributive strategies might involve expressing negative sentiment or using certainty words. The framework is integrated into future dialog systems, potentially enhancing their effectiveness.

Further advancements were made by Zhou et al [80] and Joshi et al. [50], who integrated these identified tactics into their negotiation dialog systems, outperforming previous models. Instead of only utilizing deep learning models, Zhou et al [80] utilized finite state transducers (FSTs) and recurrent neural networks. The combination can better track the entire dialogue history and make it easier for humans to interpret model decisions. However, FST-based approaches may lack expressivity [50]. Thus, Joshi et al. [50] adopted Graph Attention Networks (GAT), which effectively modeled complex negotiation strategies while maintaining interpretability through intermediate structures. This approach not only enhanced expressivity but also achieved a higher sale price ratio compared to the FST-based model.

To address the shortcomings of previous supervised learning and reinforcement learning methods, namely their inadequacy in long-term planning and neglect of opponent reactions, Yang et al. [77] introduced a model incorporating the theory of mind (ToM) into negotiation dialogue systems. Theory of mind refers to the cognitive ability to understand not only one's own thoughts but also the thoughts of others [42]. In the proposed model, this concept was applied to predict how a dialogue act or an utterance produced by the system would influence the opponent's response. This model exhibited a 20% improvement in dialogue agreement rates compared to its predecessors.

Finally, Batra et al. [10] proposed a model focusing specifically on pricing strategies in one-on-one customer negotiations. This model breaks down the negotiation process into phases like intent identification, price extraction, and response generation using predefined texts. The approach involves processing buyer messages to determine their intent and price, followed by generating strategic responses and counteroffers. These responses are guided by a structured decision tree and a set of established rules. It also incorporates sentiment analysis to adjust counteroffers according to the buyer's sentiment towards the product. This bot demonstrates effective argumentation and smooth communication, indicating its practical application in real-world negotiation contexts.

## 3.3 Conclusion

This chapter has provided a comprehensive overview of existing research on linguistic alignment in text-based dialogue systems and various approaches in negotiation dialogue systems. It highlighted key studies that have developed methods for featuring lexical alignment in chatbot responses through substituting lexical choices and employing Seq2Seq models. The chapter also explored different methodologies in developing negotiation dialogue systems, from incorporating strategic reasoning to employing theory of mind concepts. These insights inform the development of a negotiation dialogue system by breaking down the workflow into several components and integrating lexical alignment features into the components.

# Chapter 4

# Architecture

This chapter discusses the structural design of the chatbot, its deployment, and the preliminary testing phases. The architecture draws inspiration from Batra [10]. This design was selected due to its ability to manage the bargaining strategy deftly. By controlling strategy, the focus remains primarily on the investigation of lexical alignment effects.

## 4.1 System Architecture

The chatbot's architecture consists of seven fundamental components, illustrated in Figure 4.1:

- **User Interface:** Enable straightforward user-chatbot interactions.
- **Intent Classifier:** Interprets and categorizes user input, thereby producing user intent variables.
- **Price Extractor:** Extracts the price the user suggests, creating related variables.
- **Product Extractor:** Identifies the product entity in user input.
- **Dialog Management:** Employs predefined rules, utilizing user intention, suggested price, and product specifics to derive the chatbot's intent and counter-proposal.
- **Response Generation:** Constructs tailored responses through the OpenAI API based on the chatbot's intent and the proposed counteroffer.
- **Database:** Records conversation logs and questionnaire outcomes.

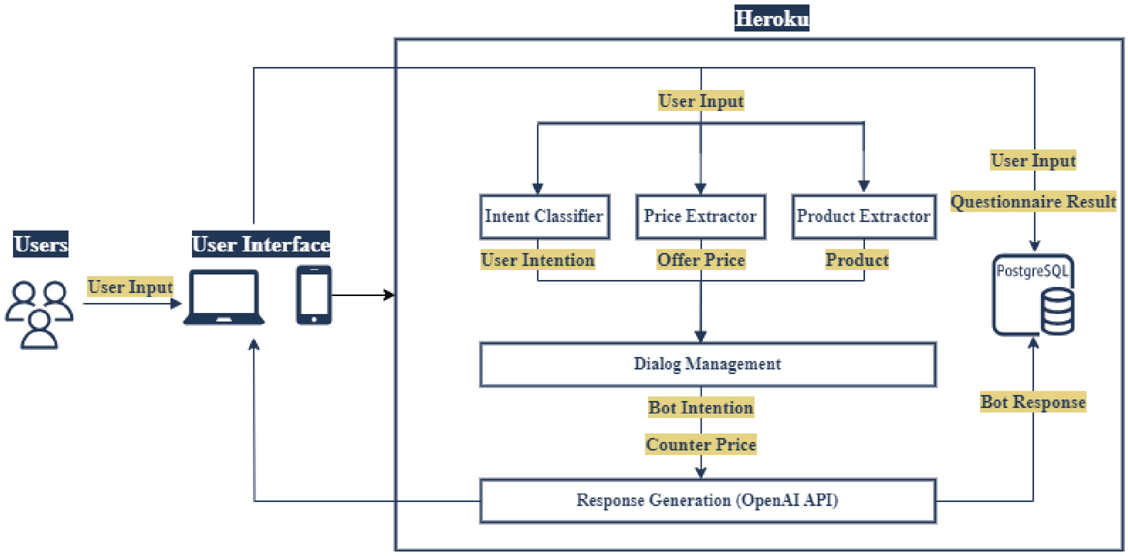



FIGURE 4.1: Overview

The User Interface is designed using HTML, CSS, and Javascript. The chatbot's workflow was developed in Python. Flask, a lightweight web application framework, serves as the bridge connecting the front-end and back-end. The entire system is hosted on Heroku. Heroku is a cloud platform as a service that enables developers to build, run, and operate applications entirely in the cloud. It was chosen because of its ease of use, multi-language support, and diverse add-ons, eliminating the need for infrastructure management.

For data storage, SQLite was utilized in the initial local developing phase due to its lightweight nature and quick operations. However, during the deployment on Heroku, PostgreSQL was adopted to overcome SQLite's retention limitations as SQLite is a file-based database all the data stored in it would be lost at least once every 24 hours. Also, Heroku's PostgreSQL add-on provided a convenient solution, seamlessly integrating without having to manually install, configure, or manage the database server.

To simplify the building process while maintaining relatively a rich user experience, the chatbot is designed to negotiate sales for four distinct products: the Switch OLED game console Nespresso Lattissima One coffee machine, Roland FP-30 digital piano, and the Fujifilm X-T5 camera. These products were chosen for two reasons. First, each product is widely recognized and populated by students. Second, the products selected are diverse, ranging from game consoles to cameras, with prices ranging from \$200 to \$800. This diversity ensures their appeal to different people. In-depth descriptions for each chatbot component are explored in the following sections of this chapter.

## 4.2 User Interface

To thoroughly assess user experiment results and mitigate user effects, the entire experiment procedure is executed online via a website. This website was structured into three sequential web pages, each to guide the participant through a specific phase of the experiment. Considering the likelihood of participants accessing the website via mobile devices, the website was crafted with CSS styling adjustments to accommodate mobile screen size, ensuring a mobile-friendly experience.

### 4.2.1 Introduction and Consent

On this website homepage, participants are introduced to the project, followed by a consent form. Given that the majority of the participants are either friends or acquaintances of friends, all descriptions are written in an informal, friendly, and casual way. This is designed to enhance their interest in participating in the testing. This webpage ensures participants are well-informed and voluntarily participating. After consent, clicking the "Start now" button will direct them to the primary webpage interacting with the chatbot, as shown in Figure 4.2.

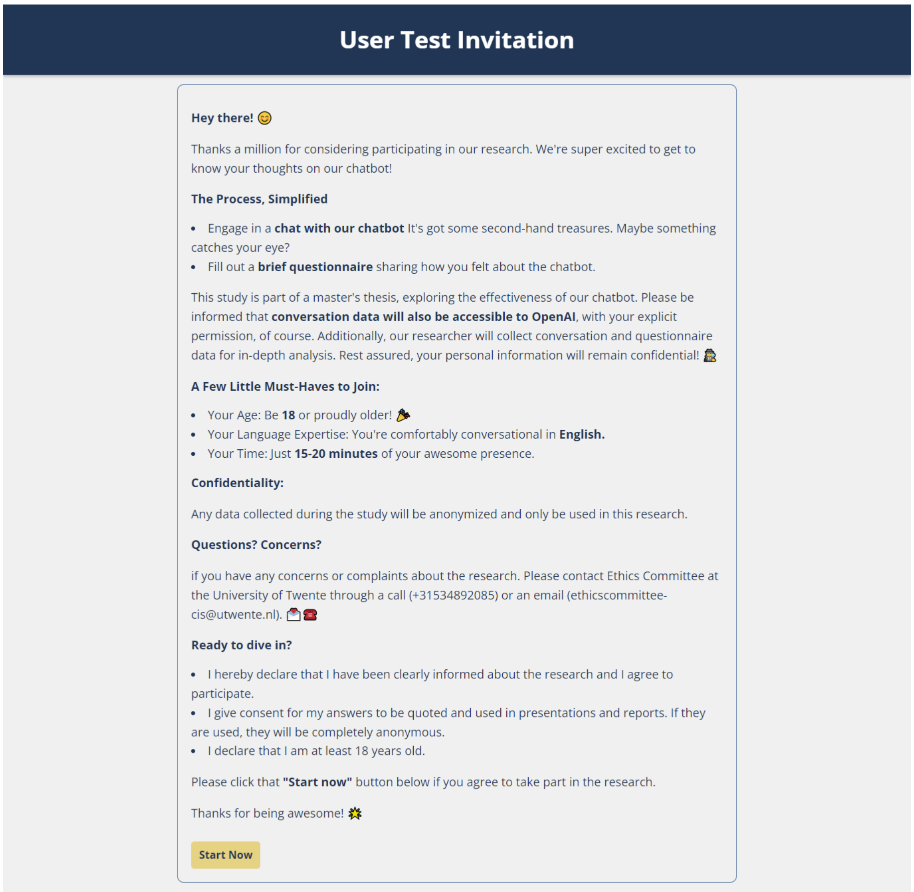



FIGURE 4.2: Homepage of website

### 4.2.2 Interaction with the Chatbot

Illustrated in Figure 4.4, this page offers guidelines for purchasing products for participants. The primary goal of this study is to investigate lexical alignment effects. The longer the user's utterance, the more content for the chatbot to mimic user's lexical choice. Participants are, therefore, encouraged to use complete sentences throughout the interaction.

To improve user experience, especially considering the bot's prolonged response time, a "thinking" message appears as soon as users input their queries. This serves as a placeholder, subsequently replaced by the chatbot's actual response. This mechanism, illustrated in Figure 4.3, ensures users are consistently engaged and minimizes potential frustration.

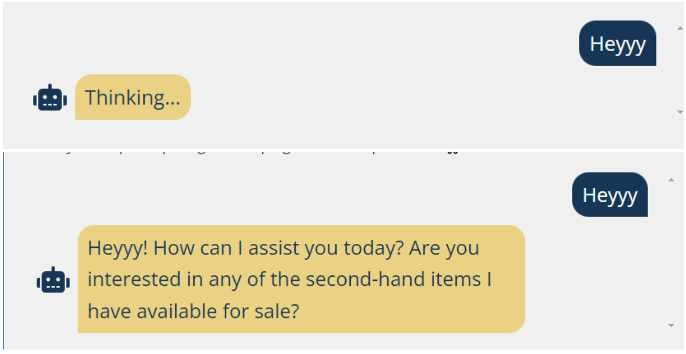



FIGURE 4.3: "Thinking" Message Followed by actual response

Participants are instructed to click the "Go to questionnaire" button after they attempt to purchase one or more products. This action will redirect them to the questionnaire webpage, as illustrated in Figure 4.4.

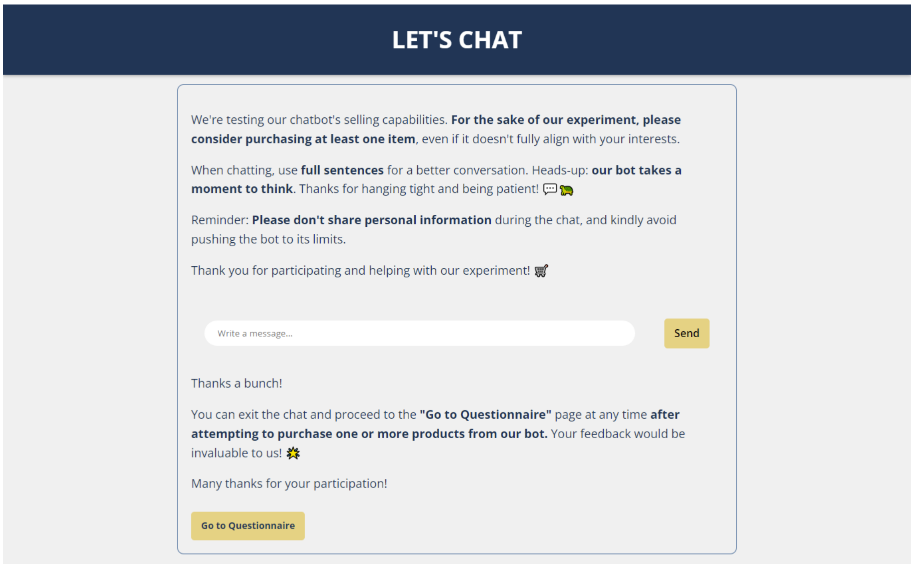



FIGURE 4.4: Chatbot Interaction page

### 4.2.3 Questionnaire and Feedback

The questionnaire is spread over three pages to optimize user experience, shown in Figure 4.5. This questionnaire comprises 28 multiple-choice questions and 1 open-ended question. For the purpose of future results analysis, progression to subsequent pages or final submission requires completion of all questions, except the open-ended question. Further explanation of the questionnaire is in Chapter 5 Evaluation. Upon submission, the feedback is stored, and participants are thanked via a message, signifying the experiment's conclusion.

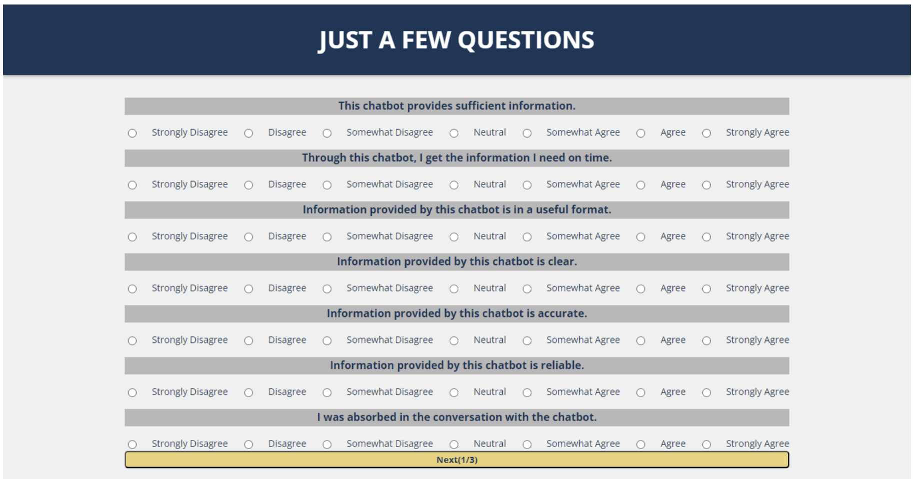



FIGURE 4.5: Questionnaire Page

## 4.3 Chatbot Workflow

This section discusses the chatbot's workflow, highlighting five main components: intent classifier, price extractor, product extractor, dialog management, and response generation.

### 4.3.1 Intent Classifier

The chatbot, upon receiving user input, leverages the intent classifier for categorization. The primary basis for classes derives from the Craigslistbargain dataset proposed by He [46]. The dataset divided user inputs into nine classes: disagree, agree, insist, inquire, intro, propose, vague-price, counter, and inform, based on rule-based phrases as illustrated in Table 4.1. Comprehensive statistics on the buyer's intent distribution are provided in Table 4.2.

| Intent | Matching Patterns |
|---|---|
| disagree | no, not, n't, nothing, dont |
| agree | not disagree and ok, okay, great, perfect, deal, that works, i can do that |
| insist | the same offer as the previous one is detected |
| inquire | starts with an interrogative word (e.g., what, when, where) or particle (e.g., do, are) |
| intro | hi, hello, hey, hiya, howdy or how are you, interested |
| init-price | first price mention |
| vague-price | no price mention and come down, highest, lowest, go higher/lower, too high/low |
| counter | new price detected |
| inform | previous utterance dialogue act was inquire |

TABLE 4.1: Craigslistbargain Intention Classes [46]

| Intent | Number of Utterance |
|---|---|
| unknown | 3578 |
| inquiry | 3571 |
| counter-price | 3412 |
| init-price | 2477 |
| intro | 2117 |
| agree | 758 |
| disagree | 308 |
| inform | 288 |
| vague-price | 214 |
| insist | 194 |

TABLE 4.2: Craigslistbargain Intention Number

In the original dataset, nine classes were designated for user inputs. However, this study has refined these into eight primary categories, as detailed in Table 4.3. The greet class mirrors the intro function from the original dataset. Similarly, the inquiry, agree and disagree classes are retained as they were. Due to the overlap in classes like init-price, counter-price, vague-price, and insist, this study consolidates them into a single counter-price intent. To enhance lexical alignment through increasing conversation turns, the chatbot initially withholds product information. This led to the introduction of a new ask-list intent, recognizing users to inquire about the products on offer. Additionally, a goodbye intent has been added to recognize user inputs indicating to end of the conversation. The introduction of the "chitchat" class utilizes OpenAI capabilities, fostering extended and more natural dialogues, thereby enhancing the user experience. Given the original dataset using a rule-based parser that leverages regular expression matching and if-then rules, a significant number of unidentified intents persist. Supervised learning might not yield optimal results, given its reliance on precise label accuracy.

Given the importance of intent classification in the chatbot workflow and its direct impact on the chatbot's performance, this research chose to employ the OpenAI API for this function. When user inputs are received, the OpenAI API is employed to analyze these inputs and categorize them into predefined intent classes. This classification is crucial for directing the chatbot's responses appropriately. In situations where the inputs are either unrecognized or fall outside the standard categories, an "error" method is implemented to indicate the users need to re-input.

| Intent | Description |
|---|---|
| greet = Intro | User greets to bot, such as hello, good morning. |
| New ask-list | User asks the bot what the bot is selling. |
| inquiry | User asks the product more detailed information or shows interest in one product. |
| counter-price | User offers the price for products or wants to negotiate the price of the product. |
| agree | Users accepts the offer. |
| disagree | User rejects the offer. |
| New goodbye | User goodbye to the bot. |
| New chitchat | chitchat. |

TABLE 4.3: Intent of User Input

### 4.3.2 Price Extraction

Price extraction is conducted on every user input to determine if the user has provided an offer for a product. Recognizing that users might just type a numerical value as an indicative price, the system employs a predefined rule to identify the offered price: it captures all numbers from the user's input and selects the smallest one as the offer. However, there is a need to differentiate between quantity and price. For instance, in inputs like "I want to buy 2 switches, do you have them?", the number "2" signifies quantity, not price. Given that all products being sold are priced above 200 euros, the system selects the lowest number from the input as long as it is above 20, considering it as the offered price.

### 4.3.3 Product Extraction

Given the product's scope is limited to four products, string matching is favored over named entity recognition as it is more accurate with limited data. TheFuzz package, previously known as FuzzyWuzzy, is utilized with its fuzzy string matching capabilities. It employs the Levenshtein Distance to calculate the similarity between instances. this method allows for partial matches, accommodating user typos. For instance, even if "switch" is misspelled as "swithc", the match can still be recognized. The primary terms for matching include "switch", "coffee machine", "piano", and "camera". Anticipating users might refer to brand names, additional terms like "roland", "nintendo", "fujifilm", and "nespresso" were incorporated. The degree of similarity between strings is determined using a partial ratio.

However, during the testing phase, setting an optimal matching score proved challenging. TheFuzz calculates similarity scores by assessing how closely two text strings match each other. For shorter strings, the similarity between the user's input and the targeted product name tends to be quite high. As a result, a higher threshold, typically above 80, is necessary to ensure precise product extraction. On the other hand, with longer strings, the similarity score between the user input and the product name often appears lower. Therefore, a lower threshold, around 60, is more suitable.

To enhance accuracy, this component was refined to first dissect the input, focusing on extracting noun phrases. This was accomplished using the spaCy library, an advanced tool for Natural Language Processing. For instance, given the input "I would like to know about the switch", the extracted noun phrases would be "I" and "the switch". Following this extraction, TheFuzz would perform the string matching. If the matching score is higher than 80, the terms are extracted as product variables.

### 4.3.4 Dialog Management

Once user intent, proposed price, and product are identified, the bot manages the dialogue accordingly. This is based on a rule-based system that is designed to emulate a realistic negotiation experience and allow for closer monitoring and control over the negotiation strategy. This ensures a focus on investigating lexical alignment effects. An overview of these rules is presented in Figure 4.6 and the detailed description is listed below. The prompt of each intention will be generated and passed to the response generation component.

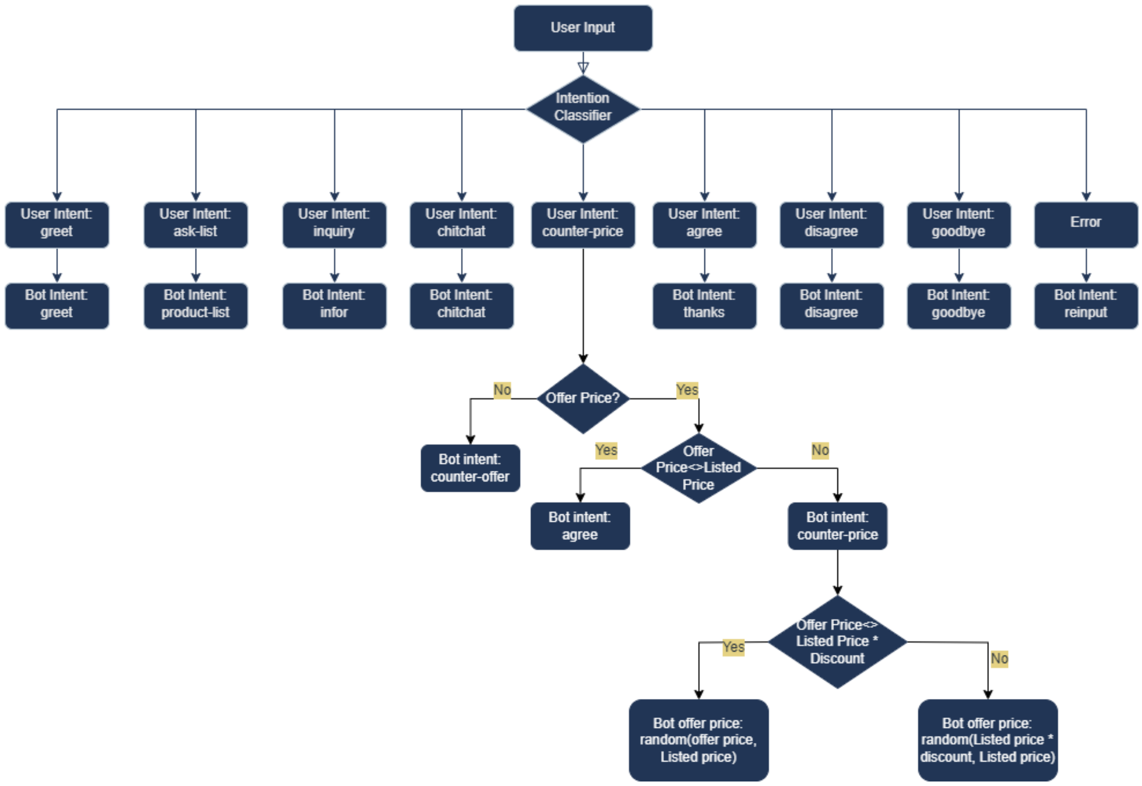



Figure 4.6: Dialog Management

**Bot-Intent: greet**

- **User-Intent:** greet
- **Description:** The chatbot acknowledges the greeting and responds in kind.
- **Prompt:** Extend a warm welcome to the user.

**Bot-Intent: product-list**

- **User-Intent:** ask-list
- **Description:** The bot lists all product types and their prices upon the user's request but avoids revealing detailed product information immediately to prolong the conversation.
- **Prompt:** List only the product Type and Price, then inquire which one the user wishes to purchase.

**Bot-Intent: infor**

- **User-Intent:** inquiry
- **Description:** Provides comprehensive product details, such as color, model, age, condition and side offers. Follows up by inquiring about user preferences, for instance, their interest in gaming if they ask about a gaming console.

- **Prompt:** When the user asks about a specific product, respond with a product description and then ask a relevant question based on their choice. For example, if they ask about the Switch, provide the description and then ask if they like games. If they ask about the coffee machine, provide the description and then ask if they enjoy coffee. if they ask about the camera, provide the description and then ask if they like photography. If they ask about the piano, provide the description and then ask if they enjoy playing piano.

**Bot-Intent: chitchat**

- **User-Intent:** chitchat
- **Description:** If the user engages in casual conversation, like mentioning their love for games, the bot continues the chat and subtly redirects towards product negotiation "That's great to hear! The Switch OLED version is a fantastic console for gaming. Would you like to proceed with the purchase?"
- **Prompt:** Craft a reply referencing the prior conversation and guide the conversation to sell product.

**Bot-Intent: counter-offer**

- **User-Intent:** counter-price
- **Description:** If the user doesn't specify a price, the bot prompts the user to suggest one.
- **Prompt:** Prompt the user to suggest a price when they haven't provided one.

**Bot-Intent: counter-price**

- **User-Intent:** counter-price
- **Description:**
  - First Counter Attempt:
    * If the user's offer equals or is higher than the default price, the bot finalizes the deal.
    * If the user's offer price is between 95% and the default price, the bot randomly generates a counteroffer between the user's offer and the default price.
    * If the user offer price is below 95% of the default price, the bot randomly generates a counteroffer between 95% of the default price and the default price.
  - Second Counter Attempt:
    * If the user's offer equals or is higher than 95% of the default price, the bot finalizes the deal.
    * If the user offer price is between 88% and 95% of the default price, the bot randomly generates a counteroffer between the user's offer and 95% of the default price.
    * If the user offer price is below 88% of the default price, the bot randomly generates a counteroffer between 88% and 95% of the default price.

- Third Counter Attempt:
  * If the user's offer equals or is higher than 88% of the default price, the bot finalizes the deal.
  * If the user offer price is between 80% and 88% of the default price, the bot randomly generates a counteroffer between the user's offer and 88% of the default price.
  * If the user offer price is below 80% of the default price, the bot randomly generates a counteroffer between 80% and 88% of the default price.
- Fourth Counter Attempt:
  * If the user's offer equals or is higher than 80% of the default price, the bot finalizes the deal.
  * If the user offer price is between 75% and 80% of the default price, the bot randomly generates a counteroffer between the user's offer and 80% of the default price.
  * If the user offer price is below 75% of the default price, the bot randomly generates a counteroffer between 75% and 80% of the default price.
- Fifth Counter Attempt:
  * If the user's offer equals or is higher than 75% of the default price, the bot finalizes the deal.
  * If the user offer price is below 75% of the default price, the bot declines the offer and closes the negotiation.

- **Prompt:** Politely decline the user's proposal and present a counteroffer at counter price.

**Bot-Intent: thanks**

- **User-Intent:** agree
- **Description:** If a deal is agreed upon, the bot expresses gratitude.
- **Prompt:** Express gratitude to the user for finalizing the agreement and hope they have an excellent day.

**Bot-Intent: disagree**

- **User-Intent:** disagree
- **Description:** If the user indicates no room for negotiation, the bot displays disappointment and suggests other available products to keep the conversation ongoing.
- **Prompt:** Politely decline the user's offer and suggest alternative products along with their prices.

**Bot-Intent: goodbye**

- **User-Intent:** goodbye
- **Description:** A simple farewell message from the bot.
- **Prompt:** Goodbye to the user and wish them a wonderful day.

**Bot-Intent: re-input**

- **Error**
- **Description:** When the user's input isn't recognized or doesn't fit any intention category, the bot requests a re-input.
- **Prompt:** Apologize for the oversight of the product and kindly request the user to specify it again.

**Special scenarios**

- If a user inquires about multiple products simultaneously, the bot will apologize, stating it can handle one product at a time, and ask for a reinput.
- If the bot is unsure about which product the user refers to, especially during price negotiations, it will request product name clarification.

### 4.3.5 Response Generation

In this research, the OpenAI API was employed to produce responses. Specifically, the "ChatCompletion.create" function was used. The function mainly takes a list of message objects as input, where each object specifies serves a role ("system", "user", or "assistant"). The "system" role helps to set up the assistant's behavior and persona, essentially providing high-level instructions for the whole conversation. The "user" role provides requests for the "assistant" role to respond to, as shown in Figure 4.7.

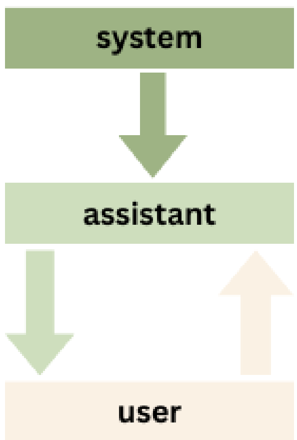



FIGURE 4.7: API Call

Through the Dialog Management component, the preset prompts, along with user input, are passed into message history, as shown below. Then the system retrieves the generated response and returns it to the user.

```
{"role": "user", "content": "{user_input}"}
{"role": "assistant", "content": "{prompt}"}
{"role": "assistant", "content": "{responses}"}
```

{"role": "user", "content": "{user_input}"}: This line represents a message from the user. The "{user_input}" part is a placeholder, which indicates that whatever the user inputs will be placed there. {"role": "assistant", "content": "{prompt}"}: This line is an instruction to the assistant. It directs the assistant to provide a response based on a given prompt ("{prompt}" is another placeholder indicating that the prompt will be placed there). {"role": "assistant", "content": "{responses}"}: This line is the response generated by the model, as "{prompt}" is also a placeholder indicating that the generated response will be placed there.

## 4.4 Prompt Testing

The prompt testing aimed to evaluate the efficacy of various prompts regarding lexical alignment in simulated conversations. It compares the lexical alignment results of different prompts and selects the two prompts for maximum difference between the alignment and unalignment versions. This section details the setup, prompt notifications, lexical alignment measurements, and results of the prompt testing. In the end, it presents the outcomes of the prompt testing and the finalized prompt.

### 4.4.1 Prompt Modification

Prompt modifications were used to influence the assistant's responses, targeting two distinct scenarios: maximizing and minimizing lexical alignment. Inspired by Clavie [28], the different prompt modifications are summarized in Table 4.4, and each description is elaborated as well.

| Short Name | Description |
|---|---|
| Baseline | No specific lexical alignment instructions. |
| Zero-shot | Lexical alignment/unalignment without examples. |
| Few-shot | Lexical alignment/unalignment with two examples. |
| Rawinst | Instructions in the user message. |
| Sysinst | Instructions in the system message. |
| Mock | Instructions using a simulated discussion. |
| Reit | Reinforced Lexical alignment/unalignment instructions. |

TABLE 4.4: Overview of Prompt Modification

**Baseline:**
The baseline prompt is simply set up without any detailed instructions to guide model behavior related to lexical alignment. This setup serves as the reference point against which other prompts are evaluated for their impact on lexical alignment. This helps determine if changes in the experiment have any significant effects regarding lexical alignment.

```
{"role": "user", "content": "{user_input}"}
{"role": "assistant", "content": "You respond in a short, within three
sentences based on instruction: "{prompts}"}
```

As mentioned above, `The "{user_input}" part is a placeholder, which indicates that whatever the user inputs will be placed there.` `"{prompt}"` directs the assistant to provide a response, which is another placeholder indicating that the prompt will be placed there.

**Zero-shot prompt:**
Zero-shot prompts do not provide explicit examples for the model. The prompt specifically instructed the chatbot to pay attention to particular lexical elements such as prepositions, nouns, and verbs. In the alignment version of the chatbot, the prompt is asked to mirror these lexical elements in its responses. Conversely, in the unaligned version, the prompt is asked to use different words for these elements, thereby varying its responses from the user's input.

**Lexical Alignment**

```
{"content": "Your primary objective is to mimic the user's closely
choice of words in your responses.\
Specifically, mirror their prepositions, nouns, tenses, modals, verbs,
product names, and hedges."}
```

**Lexical Unalignment**

```
{"content": "Your primary objective is to use different words from users
in your responses.\
Specifically, substitute their prepositions, nouns, tenses, modals,
verbs, product names, and hedges.\"}
```

**Few-shot prompt**
Few-shot prompts offer a few examples for the model. In addition to the instructions given in the zero-shot prompt, this version elaborates with specific examples for better understanding and performance.

**Lexical Alignment**

```
{"content": "Your primary objective is to mimic the user's closely
choice of words in your responses.\
Specifically, mirror their prepositions, nouns, tenses, modals, verbs,
product names, and hedges.\
For instance, if the user uses verb buy, you should use verb buy too;
if the user uses noun switch, you should use noun switch too."}
```

**Lexical Unalignment**

```
{"content": "Your primary objective is to use different words from users
in your responses.\
Specifically, substitute their prepositions, nouns, tenses, modals,
verbs, product names, and hedges.\
For instance, if the user uses verb buy, you should use verb purchase
instead; if the user uses noun switch, you should use noun Nintendo
instead."}
```

**Instruction of user or system message**

The content in the Zero-shot and Few-shot prompts is defined, but their roles aren't specified. So two modifications were employed for passing instructions to the model: using a "user message" or using a "system message". It is stated that the "system message" should have a greater influence on the chatbot's behavior compared to the "user message." However, the exact impact on performance and lexical alignment remains unclear. Consider this as an example of a zero-shot message below.

**Zero shot for Lexical Alignment**

```
{"role": "user", "content": "Your primary objective is to mimic the
user's close choice of words in your responses.\
Specifically, mirror their prepositions, nouns, tenses, modals, verbs,
product names, and hedges."}
```

**Zero shot for Lexical Alignment**

```
{"role": "system", "content": "Your primary objective is to mimic the
user's close choice of words in your responses.\
Specifically, mirror their prepositions, nouns, tenses, modals, verbs,
product names, and hedges."}
```

**Mock-exchange instructions**

Clavie [28] used the fine-tuned ability of LLM to follow a conversational format to get better performance. This study employed same step to process instruction as a mock conversation. The model is queried to confirm understanding.

```
{"role": "user", "content": "[...] Do you understand?"}
{"role": "assistant", "content": "Yes, I understand."}
```

**Re-iterating instructions**

This modification involves reinforcing the instructions by appending a message that repeats its main objective to align lexically with the user or not.

**Lexical Alignment**

```
{"role": "assistant", "content": "I will try to use the same words
as user's."}
```

**Lexical Unalignment**

```
{"role": "assistant", "content": "I will try to use the different words
as user's."}
```

### 4.4.2 Lexical Alignment Measurement

Several methodologies of measuring lexical alignment have been introduced in existing literature. Campano et al. [22] adopt an approach to quantify lexical alignment, taking into account not only the shared vocabulary between interlocutors but also the self-repetition within each speaker's dialogue. This approach allows examining whether the user tends to share vocabulary with themselves rather than others. The final alignment score is computed as a mean value. Dubuisson Duplessis et al. [34] examined shared expressions (sequences of tokens) and self-repetition lexicons. A similar approach is implemented by Sinclair et al. [68]. The methodology involves identifying shared expressions in conversations and calculating five metrics, namely Expression Lexicon, Expression Variety, Expression Repetition speaker, Initiated Expression, and Vocabulary Overlap. Additionally, Spillner et al. [70] calculate the local alignment score for each utterance as the same terms that also appear in the preceding utterance, based on the assumption that a speaker is more likely to repeat terms from the preceding utterance.

For this pilot test under discussion, the method of Spillner et al. [70] is selected for its simplicity in accessing prompt efficacy. Given that the sentences during price bargain negotiation are generally short, lexical alignment is calculated based on the ratio of tokens appearing in the current bot response and all previous user responses. This score is then averaged over the entire conversation to get a mean alignment score. The formula is detailly described in detail below.

Given the following variables:

- $B_i$: Set of tokens in the bot's $i^{th}$ response.
- $U_{1:i}$: Set of tokens in all user responses from the start to the $i^{th}$ turn.
- $n$: Total number of bot's responses during the conversation.

The alignment score for the $i^{th}$ bot's response can be represented as:

$$A_i = \frac{|B_i \cap U_{1:i}|}{|U_{1:i}|} \tag{4.1}$$

where $|B_i \cap U_{1:i}|$ is the number of tokens that are common in both the bot's $i^{th}$ response and all user responses up to the $i^{th}$ turn, and $|U_{1:i}|$ is the total number of tokens in all user responses up to the $i^{th}$ turn.

The mean alignment score for the entire conversation is then:

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i \tag{4.2}$$

### 4.4.3 Tests and Results

To thoroughly examine prompts performance regarding lexical alignment, simulated user inputs were created, drawing from the Craigslistbargain dataset that covers all user intents mentioned above. A total of 10 dialogues were created, each comprising an average of 5 user inputs, to simulate realistic negotiation scenarios. These dialogues were then processed using lexical alignment prompts and unaligned prompts accordingly. Subsequently, the average lexical alignment scores for the chatbot's responses were calculated accordingly and compared.

| | Alignment Score |
|---|---|
| Baseline | 0.293 |
| Zero-shot Rawinst | 0.402 |
| Zero-shot Sysinst | 0.397 |
| Few-shot Rawinst | 0.321 |
| Few-shot Sysinst | 0.383 |
| Zero-shot+Rawinst+Mock+Reit | 0.411 |
| Zero-shot+Sysinst+Mock+Reit | 0.365 |

Table 4.5: Alignment Scores of Alignment Prompt Modification

The experiment started with testing each prompt individually. After calculating the alignment score, the highest score prompt was retained, and an additional prompt modification was introduced. This approach ensured that the testing could evaluate the combined effects of multiple prompts. For example, after the "Zero-shot" and "Few-shot" prompt was tested in both the "Rawinst" and "Sysinst" modifications, the more effective version was selected. Next, the "Mock" and "Reit" prompt was introduced and was tested again in both "Rawinst" and "Sysinst" notification.

As shown in Table 4.5, the optimal prompt for maximizing lexical alignment was identified based on the results. The zero-shot instructions are passed as the "user" message to generate responses. Additionally, the prompt simulates a conversation to reinforce the instructions. Notably, this prompt resulted in a mean lexical alignment score of 0.411, which was the highest average score across all experiments.

**Final Prompt for Alignment**

```
{"role": "user", "content": "Your primary objective is to closely mimic
user's choice of words in your responses.\
Specifically, mirror their prepositions, nouns, tenses, modals, verbs,
product names, and hedges.\
Do you understand?"}
{"role": "assistant", "content": "Yes, I understand and I will try to
use the same words as user's."}
```

```
{"role": "user", "content": f"{user_input}"}
{"role": "assistant", "content": "You respond in a short, within three
sentences based on instruction: "f"{instruction}"}
```

| | Alignment Score |
|---|---|
| Baseline | 0.293 |
| Zero-shot Rawinst | 0.299 |
| Zero-shot Sysinst | 0.351 |
| Few-shot Rawinst | 0.358 |
| Few-shot Sysinst | 0.299 |
| Few-shot+Rawinst+Mock+Reit | 0.259 |
| Few-shot+Sysinst+Mock+Reit | 0.269 |

TABLE 4.6: Alignment Scores of Unalignment Prompt Modification

The same approach was adopted to minimize the alignment score. As indicated by Table 4.6, the most effective prompt for minimizing lexical alignment was identified. In this experiment, a few-shot prompt is employed as the "user" message. Also, it incorporates a simulated conversation aiming at reinforcing the instructions. The final result of a mean lexical alignment score was 0.259, the lowest average score across all experiments.

**Final Prompt for Unalignment**

```
{"role": "user", "content": "Your primary objective is to use different
words from users in your responses.\
Specifically, substitute their prepositions, nouns, tenses, modals,
verbs, product names, and hedges.\
For instance, if the user uses verb buy, you should use verb purchase
instead; if the user uses noun switch, you should use noun Nintendo
instead.\
Do you understand?"}
{"role": "assistant", "content": "Yes, I understand and I will try to
use the different words as user's."}
{"role": "user", "content": f"{user_input}"}
{"role": "assistant", "content": "You respond in a short, within three
sentences based on instruction: "f"{instruction}"}
```

## 4.5 Preliminary Testing

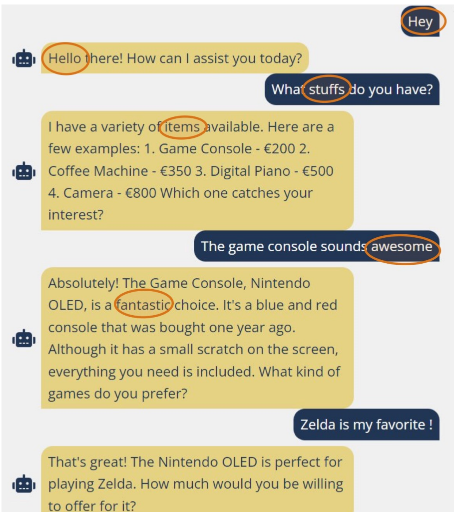
Before rolling out the full user testing, a preliminary test was executed. Feedback was garnered from my supervisors and a group of six friends who interacted with the website. To enhance the user experience, several modifications were incorporated:

- To address instances during the preliminary phase where the bot did not respond due to backend system errors, error-catching mechanisms were implemented. These mechanisms address OpenAI API, database integration, and internal server errors.

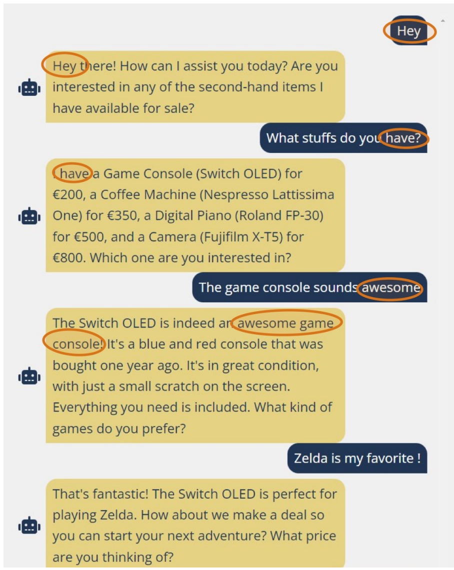
When such errors arise, users are now guided to restart the experiment from the homepage or refresh the page.

- Given the constraints on the OpenAI API for GPT-3.5, which allows 3500 calls and processes 10,000 tokens per minute, an exponential backoff retry system was established to handle rate limit errors that users might encounter. Exponential backoff is a strategy used to manage retry attempts. Tracity library was used so that if an API call failed due to rate limit, it should be retried with a random wait time between 1 and 10 seconds before each attempt and cease retrying after a total delay of 22 seconds.
- For an improved user interface, users can now simply press "Enter" to send messages instead of manually clicking the "Send" button. To prevent potential database integrity errors due to continual user input, users are now prevented from sending messages while awaiting the bot's reply.

In conclusion, below are sample finalized conversations of both the lexical aligned and unaligned versions.




(A) Unalinged




(B) Aligned

Figure 4.8: Example Conversations

## 4.6 Conclusion

This chapter details the structural design and preliminary testing of the chatbot. It outlines the chatbot's components, such as the User Interface, Intent Classifier, Price and Product Extractors, Dialog Management, and Response Generation, and explains their functionalities. The chapter also discusses the methodologies used for prompt testing and measuring lexical alignment. Through these detailed descriptions, the chapter provides a comprehensive view of the technical and functional aspects of the chatbot, announcing its effective implementation in evaluating lexical alignment during the user experiment phase.

# Chapter 5

# Evaluation

This chapter presents an evaluation of the chatbot. First, the process of recruiting participants and the experiment procedure are outlined. This is followed by a detailed introduction to the measurement matrices employed in the study. Subsequently, the chapter delves into presenting and analyzing the statistical results obtained from these experiments.

## 5.1 Methods

### 5.1.1 Participants

This study recruited participants primarily through personal networks, including friends and acquaintances. The target sample size of 40 was calculated using Python to attain a statistical power of 0.8, considering a large effect size (0.8) and the standard alpha error probability of 0.05 for an independent t-test. Ultimately, 52 individuals participated in the experiment. However, data from 21 participants were excluded based on several criteria: 1) failure to complete the questionnaire, 2) insufficient conversational turns (fewer than four), or insufficient conversation utterances (using numerical responses rather than sentences), considering the study's focus on lexical effects, and 3) failure to follow the instructions in terms of purchasing one or two predefined products, as the study also examined negotiation outcomes. Consequently, a total of 31 participants were included in the final dataset: 13 in the unaligned group and 18 in the aligned group.

### 5.1.2 Procedure

The experiment was conducted online. Participants were invited via a message containing a link to the experiment website and a brief introduction. To avoid bias, details about the study's purpose on lexical alignment were withheld. A between-subjects design was used. Participants were divided into two groups: one interacted with a chatbot with the lexical alignment feature, while the other group engaged with the chatbot without the lexical alignment feature. Upon accessing the website, participants were automatically assigned a unique user ID. Those with even user IDs interact with a chatbot featuring lexical alignment, while participants with odd IDs experienced a non-aligned version.

### 5.1.3 Measurements

This study employs a combination of objective and subjective measures to evaluate lexical alignment effects from the user's viewpoint.

### Objective

The objective metrics encompass four specific measures, as outlined in Table 5.1. The deal price refers to the ratio of the final agreed price to the original price. It reflects the financial outcome of the negotiation. The deal rate indicates the proportion of participants who successfully reached an agreement by the end of the experiment. It reflects the outcome of the negotiation. In order to assess user engagement or interest, the dialogue turns and the user utterance length will be measured as well [78]. Specifically, the dialogue turns counts the number of turns taken by chatbot and participant during conversation. On the other hand, the user utterance length measures the word count of each user's input during these interactions.

| Objective metrics | Description |
|---|---|
| Deal Price | The ratio of original price |
| Deal Rate | Successful agreement rate |
| Dialogue Turns | Total number of turns |
| User Utterance Length | Total number of words user input |

Table 5.1: Objective metrics

Previous research indicates that positive emotions and trustworthiness influence the outcomes of negotiations negotiation by increasing confidence in judgments and promoting cooperative behaviors [33, 38, 61]. Studies have shown that lexical alignment can enhance trustworthiness and evoke positive emotions, such as likeability [59, 60]. In addition, the stronger lexical alignment results in the higher user engagement [69]. Therefore, the following hypotheses are proposed:

- **H1a:** Lexical alignment will lead to an increase in the final deal price during negotiations with the chatbot.
- **H1b:** Lexical alignment will lead to a higher deal rate with the chatbot.
- **H2a:** Lexical alignment will result in a higher number of dialogue turns with the chatbot.
- **H2b:** Lexical alignment will increase the length of user utterances during interactions with the chatbot.

H1a and H1b are based on the assumption that improved trust and positive emotion will result in more favorable outcomes in the negotiation. H2a and H2b suggest that lexical alignment fosters user engagement, promoting more detailed and extended interactions.

### Subjective

For a comprehensive evaluation of user satisfaction, this study employs the model proposed by Ashfaq et al. [8]. This model was chosen because this framework combined the ECM [15], ISS model [32], and TAM [30] to create a simplified yet robust model. for assessing user satisfaction. Unlike models that measure satisfaction solely, this approach evaluates both the components of user satisfaction and overall satisfaction.

The model defines five dimensions primarily for accessing user satisfaction: Information quality, service quality, perceived enjoyment, perceived usefulness, and perceived ease of use.

- Information quality can be defined as “the accuracy, format, completeness, and currency of information produced by digital technologies” [32].
- Service quality emphasizes timely responses and personalized attention, both of which can significantly enhance user satisfaction [32].
- Perceived enjoyment is described as users who do experience intrinsic enjoyment and find a system enjoyable during their use [30].
- Perceived usefulness measures the degree users believe how effectively the technology aids in performing specific tasks [30].
- Perceived ease of use reflects the user’s expectation of a seamless experience with the system [30].
- Satisfaction describes the overall user satisfaction with the chatbot [30].

In this study, the questionnaire focused on three dimensions: Information Quality, Service Quality, and Perceived Enjoyment. This choice was made because perceived ease of use was found not to have a statistically significant effect on satisfaction in the validation phrase [30] and was thus excluded. Furthermore, since the negotiation process with the chatbot in this study was more similar to a one-time interaction, the "perceived usefulness" dimension related to customer service scenarios was considered unnecessary.

Trustworthiness in this study is considered as the reliability, honesty, and credibility of the agent [56]. To measure trustworthiness, survey instruments from [56] were employed. To assess overall user satisfaction with the chatbot, instruments from [58] were utilized.

For consistency, all questionnaire items were rephrased in past tense and the term "conversational agent" was replaced with "chatbot." The details of these rephrased questions can be found in Table 5.2. The questionnaire adopted a 7-point Likert scale format.

| Dimension | Subjective Questions |
|---|---|
| Information quality | This chatbot provided sufficient information. |
| Information quality | Through this chatbot, I got the information I need on time. |
| Information quality | Information provided by this chatbot was in a useful format. |
| Information quality | Information provided by this chatbot was clear. |
| Information quality | Information provided by this chatbot was accurate. |
| Information quality | Information provided by this chatbot was reliable. |
| Perceived enjoyment | I was absorbed in the conversation with the chatbot. |
| Perceived enjoyment | I enjoyed a conversation with the chatbot |
| Perceived enjoyment | It was fun and pleasant to share a conversation with the chatbot. |
| Perceived enjoyment | The conversation with the chatbot was exciting. |
| Service quality | The chatbot has a modern-looking interface. |
| Service quality | The chatbot gave me a prompt response. |
| Service quality | The chatbot has visually appealing materials. |
| Service quality | The chatbot gave me individual attention. |
| Trustworthiness | I have faith in what the chatbot is telling me. |
| Trustworthiness | The chatbot is honest. |
| Trustworthiness | The chatbot is trustworthy |
| Trustworthiness | I can trust the information provided by the chatbot |
| Satisfaction | My overall experience of chatbot was very satisfactory. |
| Satisfaction | My overall experience of chatbot was very pleased |
| Satisfaction | My overall experience of chatbot was very contented |
| Satisfaction | My overall experience of chatbot was absolutely delightful. |

Table 5.2: Subjective Questions

Given the evidence that lexical alignment enhances trustworthiness and user satisfaction [59, 60]., the following hypotheses were proposed:

- **H3:** Lexical alignment will lead to increased user satisfaction with the chatbot.
- **H3a:** Lexical alignment will lead to improved Information Quality in interactions with the chatbot.
- **H3b:** Lexical alignment will enhance Perceived Enjoyment during interactions with the chatbot.
- **H3c:** Lexical alignment will improve Service Quality in interactions with the chatbot.
- **H4:** Lexical alignment will increase the perceived trustworthiness of the chatbot.

H3 is based on the assumption that lexical alignment increases user satisfaction. H3a, H3b, and H3c disaggregate satisfaction by hypothesizing that lexical alignment increases information quality, perceived enjoyment and service quality, which then leads to increased satisfaction. H4 shows that lexical alignment increases trustworthiness.

## 5.2 Results

Before conducting statistical analyses, the lexical alignment score of the chatbot was calculated using the methodology outlined in Chapter 4. This preliminary step was crucial to confirm the chatbot's effectiveness in lexical alignment during the experiment process. The results revealed a difference in alignment scores between the two groups: the alignment score of the unaligned group is 0.315 (SD=0.105), while the aligned group scored 0.393 (SD=0.144). This variance in scores lays the groundwork for a more in-depth analysis of how these differences impacted the outcome of the users' interactions with the chatbots.

### 5.2.1 Data Distribution

In this section, data distribution of both objective and subjective measures were tested through Kolmogorov-Smirnov and Shapiro-Wilk tests. Data distribution is crucial for selecting appropriate statistical methods for further analysis. Since the participants were divided into two groups: Unaligned and Aligned, the testing was applied separately.

#### Objective

Table 5.3 presents the results of various objective measures including Price, Deal Rate, Dialogue Turns, and User Utterance Length. For Deal Price and User Utterance Length, the significance value of both tests is greater than 0.05 for both Unalinged and Aligned groups, which indicates the data is normally distributed. However, for Dialogue Turns, the significance value of the Shapiro-Wilk test is slightly lower than 0.05 for the Unaligned group, which suggests the data may not follow the normal distribution. As the Deal Rate is binominal data for two groups, there is no distribution among those data.

| | | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | | Statistic | df | Sig. | Statistic | df | Sig. |
| Deal Price | Unaligned | 0.229 | 6 | 0.200 | 0.939 | 6 | 0.651 |
| | Aligned | 0.165 | 12 | 0.200 | 0.911 | 12 | 0.218 |
| Deal Rate | Unaligned | - | 6 | - | - | 6 | - |
| | Aligned | - | 12 | - | - | 12 | - |
| Dialogue Turns | Unaligned | 0.275 | 6 | 0.176 | 0.785 | 6 | 0.043 |
| | Aligned | 0.168 | 12 | 0.200 | 0.914 | 12 | 0.242 |
| User Utterance Length | Unaligned | 0.244 | 6 | 200 | 0.895 | 6 | 0.346 |
| | Aligned | 0.114 | 12 | 0.200 | 0.967 | 12 | 0.874 |

Table 5.3: Data Distribution Test for Objective Metrics

#### Subjective

Table 5.4 provides the results for subjective measures including Information Quality, Service Quality, Perceived Enjoyment, Trustworthiness, and Satisfaction. For Service quality, Perceived Enjoyment, Trustworthiness, and Satisfaction, the significance value of both tests is greater than 0.05 for both the Unalinged and Aligned groups, which indicates the data is normally distributed. However, for Information quality, the significance value of the Shapiro-Wilk test is higher than 0.05 for the Aligned group, which suggests the data may not follow the normal distribution.

| | | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|---|
| | | Statistic | df | Sig. | Statistic | df | Sig. |
| Information quality | Unaligned | 0.224 | 6 | 0.200* | 0.927 | 6 | 0.557 |
| | Aligned | 0.266 | 12 | 0.200* | 0.814 | 12 | 0.014 |
| Service quality | Unaligned | 0.176 | 6 | 0.200* | 0.955 | 6 | 0.783 |
| | Aligned | 0.204 | 12 | 0.180 | 0.909 | 12 | 0.208 |
| Perceived enjoyment | Unaligned | 0.176 | 6 | 0.200* | 0.944 | 6 | 0.690 |
| | Aligned | 0.184 | 12 | 0.200* | 0.959 | 12 | 0.764 |
| Trustworthiness | Unaligned | 0.207 | 6 | 0.200* | 0.892 | 6 | 0.331 |
| | Aligned | 0.149 | 12 | 0.200* | 0.923 | 12 | 0.309 |
| Satisfaction | Unaligned | 0.195 | 6 | 0.200* | 0.965 | 6 | 0.861 |
| | Aligned | 0.268 | 12 | 0.017 | 0.879 | 12 | 0.084 |

TABLE 5.4: Data Distribution Test for Subjective Metrics

The results from both tables indicate that most of the objective and subjective measures do follow the normal distribution, with few exceptions. This suggests that for most variables, parametric tests can be applied for further analysis.

### 5.2.2 Two-Sample t-Test

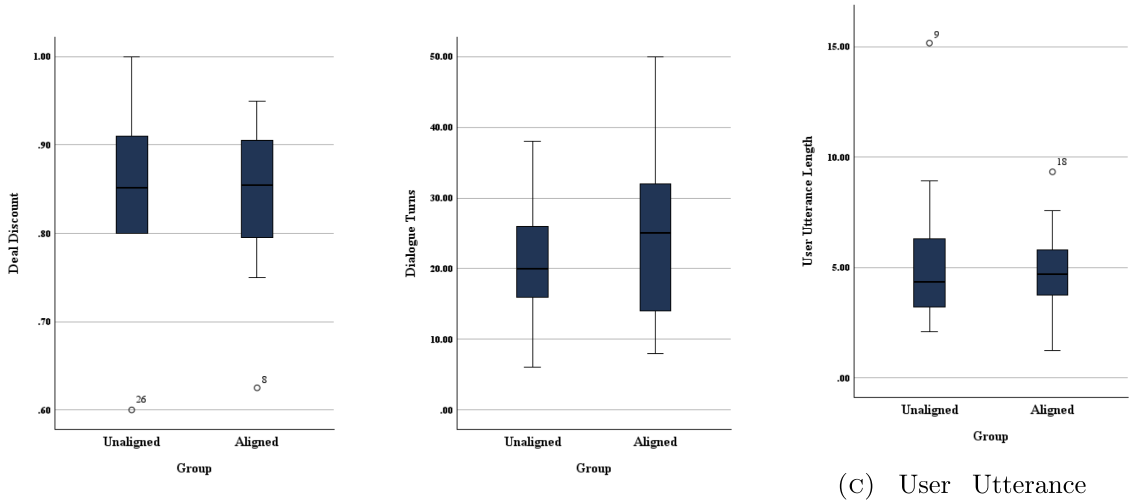
Given that most variables are normally distributed and the experimental design is between-subject, a two-sample t-test was conducted to evaluate the hypotheses.

#### Objective

The distributions of the deal price, the dialogue turns, and user utterance length are illustrated in Figure 5.1. An overview of the analysis results for these objective measures is presented in Table 5.5. The mean values of the deal price are closely matched (Unaligned = 0.83, Aligned = 0.84), with a t-value of -0.074 and a p-value of 0.471, indicating no statistically significant difference between the groups. Similarly, for dialogue turns and user utterance length, the means show minor differences (Unaligned = 20.46, Aligned = 24.00 for dialogue turns; Unaligned = 5.47, Aligned = 4.74 for user utterance length) with p-values of 0.176 and 0.230, respectively, signifying no significant differences. Regarding the deal rate, the aligned group (0.67) is slightly higher than the unaligned group (0.46).

| Objective metrics | Mean | Standard Deviation | T-Value | P-Value | Effect Size |
|---|---|---|---|---|---|
| Deal Price | Unaligned=0.83<br>Aligned=0.84 | Unaligned=0.14<br>Aligned=0.09 | -0.074 | 0.471 | 0.11 |
| Deal Rate | Unaligned=0.46<br>Aligned=0.67 | / | / | / | / |
| Dialogue Turns | Unaligned=20.46<br>Aligned=24.00 | Unaligned=8.29<br>Aligned=11.48 | -0.946 | 0.176 | 10.28 |
| User Utterance Length | Unaligned=5.47<br>Aligned=4.74 | Unaligned=3.49<br>Aligned=1.97 | 0.749 | 0.230 | 2.70 |

TABLE 5.5: Two Sample t-Text results for objective metrics




(A) Deal Price (B) Dialogue Turns (C) User Utterance Length

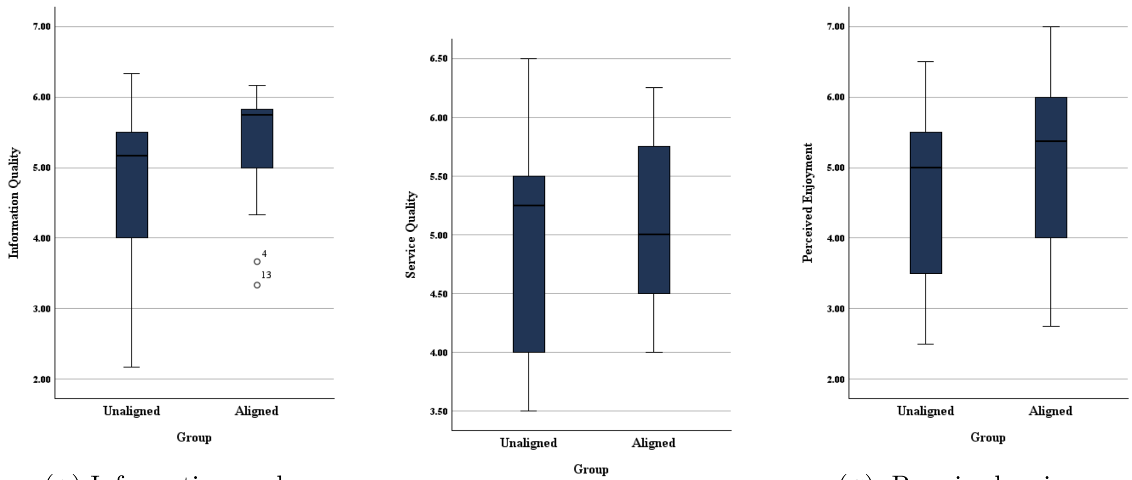
FIGURE 5.1: The Deal Price, Dialogue Turns, User Utterance Length distribution

### Subjective

The distributions of information quality, service quality, and perceived enjoyment are depicted in Figure 5.2. Table 5.6 provides a summary of the subjective measures' analysis results. Information quality shows a marginal difference (Unaligned = 4.79, Aligned = 5.36) with a t-value of -1.542 and a p-value of 0.067, which is slightly above the conventional alpha level of 0.05, suggesting a trend towards significance. Service quality means are similar (Unaligned = 5.00, Aligned = 5.10), with a non-significant t-value (-0.334) and p-value (0.370). Perceived enjoyment also shows some differences (Unaligned = 4.65, Aligned = 5.04); however, the t-value and p-value indicate these differences are not statistically significant.

| Subjective metrics | Mean | Standard Deviation | T-Value | P-Value | Effect Size |
|---|---|---|---|---|---|
| Information quality | Unaligned=4.79<br>Aligned=5.36 | Unaligned=1.22<br>Aligned=0.83 | -1.542 | 0.067 | 1.00 |
| Service quality | Unaligned=5.00<br>Aligned=5.10 | Unaligned=0.97<br>Aligned=0.65 | -0.334 | 0.370 | 0.80 |
| Perceived enjoyment | Unaligned=4.65<br>Aligned=5.04 | Unaligned=1.24<br>Aligned=1.15 | -0.898 | 0.188 | 1.19 |
| Satisfaction | Unaligned=4.52<br>Aligned=5.28 | Unaligned=1.42<br>Aligned=1.09 | -1.689 | 0.051 | 1.234 |
| Trustworthiness | Unaligned=4.75<br>Aligned=5.15 | Unaligned=1.12<br>Aligned=1.13 | -0.981 | 0.167 | 1.130 |

TABLE 5.6: Two Sample t-Text results for subjective metrics



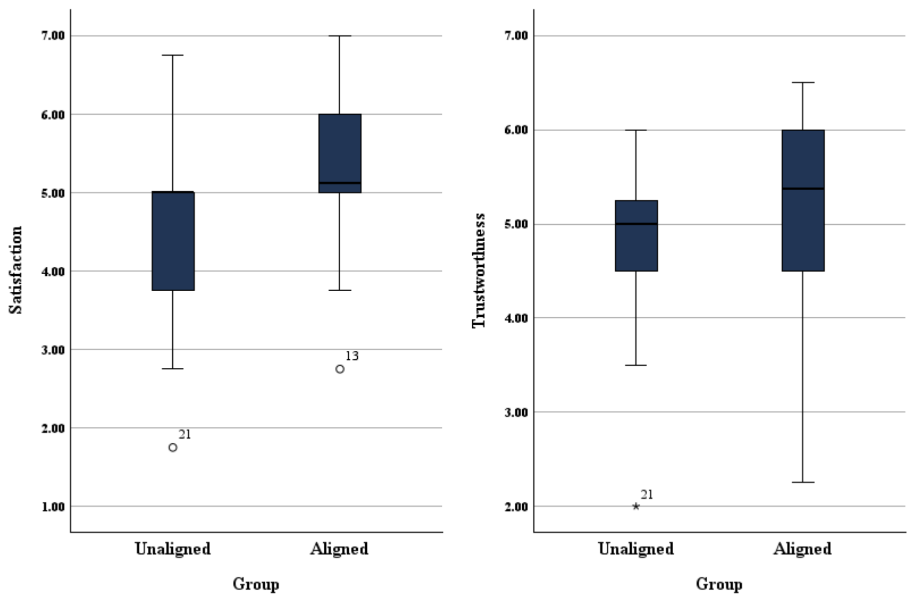

(A) Information quality

(B) Service quality

(C) Perceived enjoyment

FIGURE 5.2: The Information quality, Service quality, Perceived enjoyment distribution

Satisfaction levels differ (Unaligned = 4.52, Aligned = 5.28), but the t-value (-1.689) and p-value (0.051) do not indicate statistically significant differences. However, the distribution plot of satisfaction in Figure 5.3 highlights the differences between the two groups. Trustworthiness means are close (Unaligned = 4.75, Aligned = 5.15), with a non-significant t-value (-0.981) and p-value (0.167).




(A) Satisfaction

(B) Trustworthiness

FIGURE 5.3: The Satisfaction, Trustworthiness distribution

### Mann-Whitney U-test

Although most variables appeared suitable for parametric tests, the distributions of dialogue turn and information quality data suggested potential deviations from normality. To verify the results, the non-parametric Mann-Whitney U test was applied to these two variables. The results, presented in Table 5.7, indicate no statistically significant differences between the groups for these variables.

| | Dialog Turn | Information Quality |
|---|---|---|
| Mann-Whitney U | 138.500 | 150.000 |
| Z | 0.864 | 1.328 |
| Asymptotic Significance (1-sided test) | 0.194 | 0.092 |
| Effect Size | 0.16 | 0.24 |

Table 5.7: The Results of Mann-Whitney U-test

## 5.3 Conclusion

This chapter outlines the design of a user experiment conducted to test hypotheses derived from prior research. Utilizing a between-subjects design, the experiment was executed online, where participants were split into two groups: one interacted with a chatbot featuring lexical alignment, and the other with a chatbot lacking this feature. Given that linguistic alignment often occurs subconsciously [65, 19, 41], participants were not briefed on the experiment's purpose. Instead, they were asked to evaluate the overall performance of the chatbot by completing a questionnaire. This study employed both objective and subjective measures to evaluate the effects of lexical alignment. The objective measures focused on the outcomes of the negotiation, while the subjective measures user perceptions of the chatbot.

The analysis of these measures was conducted using a two-sample t-test, which compared the means of various variables between the groups with and without lexical alignment. The results indicated that the means for most variables were slightly higher in the aligned group compared to the unaligned group. However, these differences did not reach statistical significance. This pattern was consistent across objective measures, including transaction price, number of dialogues, and successful transaction rate, as well as subjective metrics like information quality, service quality, perceived enjoyment, trustworthiness, and satisfaction. The lack of statistically significant differences indicates that, although there were observable variations in the experiences between the two groups, these were not substantial enough to validate the previously proposed hypotheses. Therefore, the results suggest that the presence or absence of lexical alignment in the chatbot did not lead to major differences in the measured outcomes.

# Chapter 6

# Discussion

This chapter delves into additional analyses conducted on the collected dataset, providing a thorough discussion of the findings from the results. It also explores the limitations of the study, offering a critical perspective on the research methodology and results.

## 6.1 Further Analysis

Despite previous research indicating positive effects of lexical alignment on user satisfaction and trustworthiness, the two-sample t-test results showed only slightly higher means in various objective measures (transaction price, number of dialogues, successful transaction rate) and subjective measures (information quality, service quality, perceived enjoyment, trustworthiness, satisfaction) for the aligned group compared to the unaligned group, without statistical significance. To delve deeper into the relationship between lexical alignment and these variables, lexical alignment scores were calculated for each participant, reflecting the degree of chatbot alignment. Subsequent statistical analyses were conducted to identify potential patterns and associations.

#### Correlation Analysis

Appendix B presents the results of the correlation analysis. While a positive correlation was found between satisfaction and factors such as information quality, perceived enjoyment, service quality, and trustworthiness, indicating that higher levels of these factors lead to increased satisfaction, satisfaction exhibited a negative correlation with alignment scores. This is contrary to our initial assumptions. One potential explanation for this could be a lack of authenticity, the increasing alignment scores indicate the repeating teams that may be perceived as lacking intelligence or authenticity, thereby reducing satisfaction. Moreover, significant correlations among information quality, perceived enjoyment, service quality, and trustworthiness were observed, as expected. However, the alignment score's negative correlation with user utterance tokens and dialogue turns suggests that alignment may not necessarily equate to more engaged or longer conversations. This counterintuitive result needs further investigation to understand the nature of user interaction with chatbots and how alignment plays a role in perceived satisfaction.

#### Chatbot and User Lexical Alignment

Prior studies [69, 54] have confirmed lexical alignment happens mutually, suggesting a direct correlation between user and chatbot alignment. In this study, Appendix B indicated a trend where higher chatbot alignment corresponded with greater user alignment, although the correlation was not statistically significant. In contrast to previous findings, where user and chatbot alignment were in a similar rate, this study observed significantly higher alignment levels in the chatbot compared to the user. A possible explanation for this might be attributed to the method of calculating lexical alignment scores. The scores are determined by the percentage of overlapping words in total utterances. Given that the chatbot's responses are generally lengthier than user utterances, this could result in disproportionally lower alignment scores for users. This finding highlights the need to consider response length in alignment score calculations to better understand the dynamics of user-chatbot interactions.

#### PLS Analysis

To more effectively assess the relationship between lexical alignment scores and other variables, Partial Least Squares Structural Equation Modeling (PLS-SEM) was chosen considering the complex relationships among different variables, the non-normal distribution of alignment scores, and the relatively small sample size. A structural model (see Appendix D.2) was proposed based on hypotheses mentioned in Chapter 5 and the reliability of each construct was evaluated using factor loading, Cronbach's alpha ($\alpha$), convergent validity, and discriminant validity. Based on results shown in Appendix D.2, Perceived Enjoyment and Engagement had Cronbach's  values of 0.317 and 0.034, respectively, which are below acceptable levels, indicating reliability issues with these constructs. So those two were excluded from subsequent analysis. As the Composite Reliability (CR) and the Average Variance Extracted (AVE) for each construct (except PE and EG) exceeded the thresholds of 0.7 and 0.5 and the Variance Inflation Factor (VIF) were mostly below 10, other factors were included for subsequent analysis.

After refining and validating the measurement model, we tested the hypotheses by examining the structural model. A bootstrapping procedure was used to evaluate the significance of path coefficients and standard errors in the model. These results are detailed in Appendix D.2, with a summary in Appendix D.2, including path coefficients, standard deviations, T-values, and significance levels. The findings indicated a negative correlation between lexical alignment and information quality and between alignment and service quality, with no significant correlation between alignment and trustworthiness. While there was a suggested correlation between alignment and satisfaction and between alignment and deal price, these effects were not conclusively significant.

## 6.2 Discussion

Prior research has indicated that lexical alignment can positively influence user satisfaction and trust in human-computer interactions [60, 59]. These studies, however, primarily involved simple dialogue tasks such as Wizard of Oz picture-matching experiments or straightforward information retrieval, limiting the complexity and depth of the conversations. This study uses a more complex conversational scenario—negotiation—to evaluate the effects of lexical alignment from both an objective way (negotiation outcomes) and a subjective way (user perception of the chatbot). Although the two-sample t-test indicated differences between the aligned and unaligned groups, they weren't statistically significant. However, additional analyses, including correlation analysis and PLS-SEM, suggested a negative impact of alignment on these variables. The lack of support for the original hypothesis could be attributed to several factors:

- Initially, the study assumed a large effect size to calculate the targeted sample size for two-sample t-tests, expecting significant differences between the two groups. However, past experiments typically show medium to small effect sizes, indicating a need for a larger sample size. Additionally, the exclusion of 21 participants further reduced the sample size, potentially affecting the significance of observed differences.

- Participants' feedback highlighted issues with the chatbot often providing repetitive information, possibly leading to failure to access the relationship between lexical alignment and satisfaction. Firstly, while linguistic alignment typically occurs in conversations allowing equal information access, the chatbot's redundant responses might lead users to merely scan for key information, like prices, rather than fully engage with the content. Secondly, the minimal variation in lexical alignment scores across interactions suggests that the impact of lexical alignment on user experience might be marginal. Consequently, these factors challenge the evaluation of a direct relationship between lexical alignment and satisfaction and may lead to contrasting findings. An illustrative example of this issue is the chatbot's repetitive descriptions of a coffee machine, which resulted in user frustration and reduced satisfaction.

```
U: What are you providing
B: I have a variety of second-hand items available for sale.
Here are a few examples:
1. Game Console - Switch OLED, 200
2. Coffee Machine - Nespresso Lattissima One, 350
3. Digital Piano - Roland FP-30, 500
4. Camera - Fujifilm X-T5, 800
Which one are you interested in?
U: Describe coffee machine
B: The Nespresso Lattissima One is a white coffee machine that was
bought two years ago and is in perfect condition.
It comes with some capsules to get you started. What type of
coffee do you like?
U: Americano
B: The Nespresso Lattissima One is a white coffee machine that was
bought two years ago and is in perfect condition. It comes with some
capsules to get you started. What type of coffee do you like? like
an Americano or something else?
U: Any further description? Like how it compares with other machines
B: The Nespresso Lattissima One is a white coffee machine that was
bought two years ago and is in perfect condition. It offers the
convenience of making delicious Americano coffee with just a touch
of a button. What other features or specifications are you looking
for in a coffee machine?
```

- The difference in lexical alignment rates between chatbots and users can be attributed to several factors. Firstly, the bot's utterances are typically longer than the user's utterances, contributing to a redundancy issue with ChatGPT [36]. The user lexical alignment score was calculated in the same way as the bot alignment score, so the longer the bot utterances, the lower the user lexical alignment score. Secondly, the bot's tendency to provide all information in a single response may lead users to merely scan and extract the necessary information, overlooking words used by the chatbot. Lastly, the simplicity and consistency of user language during the negotiation process. The examples below suggest that users often employ straightforward, repetitive language when negotiating prices, regardless of the chatbot's language or sentence structure. This tendency for users to stick to familiar, simple sentences in negotiations can result in lower user alignment scores.

```
U: can you offer me a lower price?
U: I would go for 100
U: how about 300?
U: I can offer 100 euros for it
U: Maybe 170?
```

## 6.3 Limitations

The study faced several limitations, categorized into design and analysis aspects:

**Design Aspects**

- The study employed the GPT-3.5 model over the more advanced GPT-4.0 due to API call limitations, as the 3.5 model accommodates more calls per minute. This decision could have impacted the quality of the chatbot interactions. GPT-4.0 performs better in producing more human-like, coherent, and less redundant responses, suggesting that using the 3.5 model might have limited the effectiveness of the interactions, potentially influencing the study's outcomes. Moreover, during the user testing phase, OpenAI launched a new model, which may affect the results as well.
- To encourage longer dialogues, the study provided intentionally vague instructions to participants. This strategy, aiming to allow for more natural conversations, led to some participants not following the experimental process, such as not purchasing predetermined items. This resulted in the exclusion of their data, leading to a reduced sample size, which could have affected the study's statistical power.

**Analysis Aspects**

- The study's approach to calculating lexical alignment scores, based purely on token repetition, presents a significant limitation. This simplistic and straightforward method fails to capture the other aspects of effective communication, such as context and semantics. Consequently, a higher alignment score does not necessarily indicate a stronger or more meaningful alignment. It overlooks the subtle dynamics of human-computer interaction and can lead to misleading interpretations that chatbots' repetitive but context-independent replies increase alignment scores without actually improving user satisfaction or interaction effectiveness. This limitation highlights the need for more advanced methods in future studies that can accurately assess lexical alignment in complex conversational scenarios.

# Chapter 7

# Conclusion

This chapter provided a comprehensive reflection on this study by addressing the research questions that were established in the beginning accordingly. Then conclude the chapter by outlining recommendations for future research in this domain.

## 7.1 Conclusion

The primary objective of this thesis is to explore the effects of lexical alignment on text-based negotiation chatbots and how they influence user perceptions. Although the results are not statistically significant, the findings still indicate that lexical alignment may have a potential influence on user experiences during negotiations with chatbots, paving the way for future research in this area.

This study aims to answer the main research questions:

*How does lexical alignment in text-based negotiation chatbots influence user perceptions?*

By dividing the research question into two sub-questions, the study was summarized by answering these sub-questions:

- *How can a chatbot be developed to perform lexical alignment and price bargaining with users?*

Through exploring previous methodologies of developing the price bargain chatbot and the chatbot featuring lexical alignment, the architecture was proposed for developing the chatbot to perform lexical alignment and price bargain with users. The architecture consists of seven fundamental components: namely user interface, intent classifier, price extractor, product extractor, dialog management, response generation, and database to provide smooth negotiation conversation with users for user experiments. The chatbot mainly relied on five main components to perform, namely intent classifier, price extractor, product extractor, dialog management, and response generation. The intent classifier categorized user input into different intent classes, the price extractor extracts the price the user suggests, and the product extractor identifies the product the user mentioned. Based on identified intent, price, product, and dialogue management utilized to trigger the chatbot action through predefined rules. Then based on the chatbot actions, response generation generates the appropriate responses. The response management utilizes OpenAI API to generate responses with or without lexical alignment through prompt engineering.

- *How to evaluate user perceptions and negotiation outcomes influenced by lexical alignment in chatbots?*

In order to evaluate user perceptions and negotiation outcomes, a user experiment is conducted online. In total 31 participants are included in the final analysis, 13 in the unaligned group and 18 in the aligned group. A combination of both objective and subject measures is used to evaluate user perceptions and negotiation outcomes. The objective matrix mainly measures the outcomes of negotiation, including deal price, deal rate, dialogue turns, and user utterance length, focused on assessing whether the lexical alignment brings out better negation results and enhances user engagement. The subjective matrix evaluates user satisfaction and trustworthiness through a questionnaire filled out by participants. User satisfaction is measured with four dimensions: information quality, perceived enjoyment, service quality, and satisfaction. The two-sample t-test is used to assess the results of the experiment. The results indicate that the means for most variables are slightly higher in the aligned group compared with the unaligned group. However, these differences are not statistically significant. This pattern is consistent across various measures, including transaction price, number of dialogues, and length of user utterances, as well as subjective metrics such as information quality, service quality, trustworthiness, and satisfaction. The absence of statistically significant differences suggests that while there are observable distinctions in the experiences of the two groups, they are not substantial enough to be deemed statistically meaningful to support the proposed hypothesis.

## 7.2 Future Work

Previous research on lexical alignment in human-computer interaction often involved Wizard of Witch picture-matching experiments or simple dialogues for retrieving the information. The conversations involved in these experiments are often brief and straightforward. Even though this study explored the effects of lexical alignment in more complex scenarios, involving richer conversations, the time of interaction and the conversation turns are still relatively low and the user input during the negotiation process is not very complex, prefer using simple, straightforward, and same words and sentence. The impact of lexical alignment in longer and more intimate interactions, like longer relationships as friends or partners, remains unknown and presents an interesting for further study.

Also, most existing studies compare different groups, one is chatbots with the linguistic alignment feature and another one without that feature, to assess the impact of lexical alignment. Future research could explore the effects of varying degrees of lexical alignment in human-computer interaction. For instance, determining if stronger lexical alignment leads to increased user likeability. And whether there will be a breakpoint where excessive alignment is considered and recognized as mimicry, potentially leading to a decrease in user satisfaction and likeability.

## Appendix A

# Use of AI

During the preparation of this work the author(s) used OpenAI API in order to build a chatbot with lexical alignment feature for further analysis. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

During the preparation of this work the author(s) used ChatGPT in order to proofread the writing language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

# Appendix B

# Correlation Analysis

| | Turns | IQ | PE | SQ | TR | SA | Utokens | Score | Price |
|---|---|---|---|---|---|---|---|---|---|
| IQ | .130 | 1 | .588** | .484** | .526** | .762** | -.044 | -.307* | -.050 |
| | (.243) | | (<.001) | (.003) | (.001) | (<.001) | (.408) | (.047) | (.423) |
| PE | .040 | .588** | 1 | .251 | .472** | .649** | .011 | -.338* | -.097 |
| | (.416) | (<.001) | | (.086) | (.004) | (<.001) | (.478) | (.032) | (.350) |
| SQ | .021 | .484** | .251 | 1 | .407* | .648** | .170 | -.272 | -.118 |
| | (.455) | (.003) | (.086) | | (.011) | (<.001) | (.180) | (.069) | (.318) |
| TR | -.145 | .526** | .472** | .407* | 1 | .645** | .008 | -.048 | -.326 |
| | (.219) | (.001) | (.004) | (.011) | | (<.001) | (.484) | (.398) | (.093) |
| SA | .149 | .762** | .649** | .648** | .645** | 1 | .125 | -.302* | -.185 |
| | (.213) | (<.001) | (<.001) | (<.001) | (<.001) | | (.253) | (.045) | (.251) |
| Turns | 1 | .130 | .040 | .021 | -.145 | .149 | .017 | -.490** | .086 |
| | | (.243) | (.416) | (.455) | (.219) | (.213) | (.463) | (.003) | (.367) |
| Utokens | .017 | -.044 | .011 | .170 | .008 | .125 | 1 | -.371* | .470* |
| | (.463) | (.408) | (.478) | (.180) | (.484) | (.253) | | (.020) | (.026) |
| Score | -.490** | -.307* | -.338* | -.272 | -.048 | -.302* | -.371* | 1 | -.182 |
| | (.003) | (.047) | (.032) | (.069) | (.398) | (.045) | (.020) | | (.235) |
| Price | .086 | -.050 | -.097 | -.118 | -.326 | -.185 | .470* | -.182 | 1 |
| | (.367) | (.423) | (.350) | (.318) | (.093) | (.251) | (.026) | (.235) | |

Table B.1: Result of Correlation Analysis

# Appendix C

# Lexical Alignment

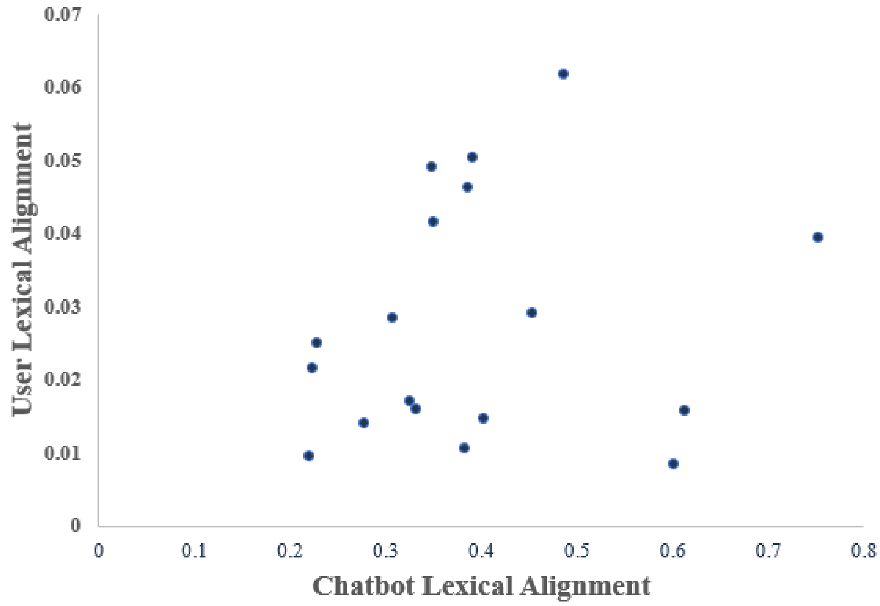



Figure C.1: Correlation between chatbot alignment and user alignment

# Appendix D

# PLS Analysis

## D.1 Basic structural model

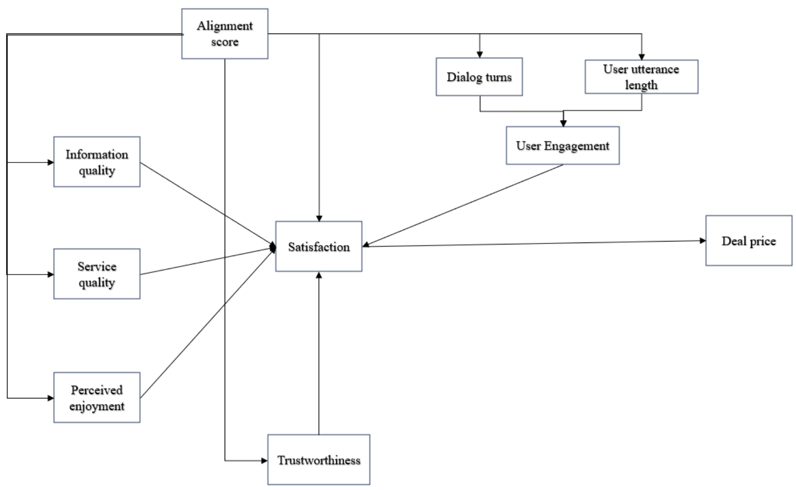



Figure D.1: Basic structural model

## D.2 Measurement Model

| Construct | Loading | $\alpha$ | CR | AVE | VIF |
|---|---|---|---|---|---|
| IQ1 ← IQ | 0.831 | 0.877 | 0.906 | 0.616 | 3.563 |
| IQ2 ← IQ | 0.845 | | | | 3.395 |
| IQ3 ← IQ | 0.780 | | | | 5.298 |
| IQ4 ← IQ | 0.768 | | | | 4.637 |
| IQ5 ← IQ | 0.772 | | | | 2.196 |
| IQ6 ← IQ | 0.706 | | | | 1.736 |
| PE1 ← PE | 0.561 | 0.317 | 0.645 | 0.318 | 1.369 |
| PE2 ← PE | 0.519 | | | | 1.066 |
| PE3 ← PE | 0.452 | | | | 1.367 |
| PE4 ← PE | 0.694 | | | | 1.030 |
| SA1 ← SA | 0.937 | 0.957 | 0.969 | 0.886 | 5.715 |
| SA2 ← SA | 0.928 | | | | 4.197 |
| SA3 ← SA | 0.969 | | | | 10.051 |
| SA4 ← SA | 0.930 | | | | 5.298 |
| SQ1 ← SQ | 0.670 | 0.866 | 0.910 | 0.719 | 1.573 |
| SQ2 ← SQ | 0.897 | | | | 2.877 |
| SQ3 ← SQ | 0.888 | | | | 3.188 |
| SQ4 ← SQ | 0.913 | | | | 3.310 |
| TR1 ← TR | 0.857 | 0.864 | 0.904 | 0.705 | 2.222 |
| TR2 ← TR | 0.693 | | | | 1.615 |
| TR3 ← TR | 0.899 | | | | 2.819 |
| TR4 ← TR | 0.893 | | | | 2.333 |
| EG1 ← EG | 0.801 | 0.034 | 0.670 | 0.508 | 1.000 |
| EG2 ← EG | 0.612 | | | | 1.000 |

Table D.1: The reliability and validity of the measurement

## D.3 Structural Model

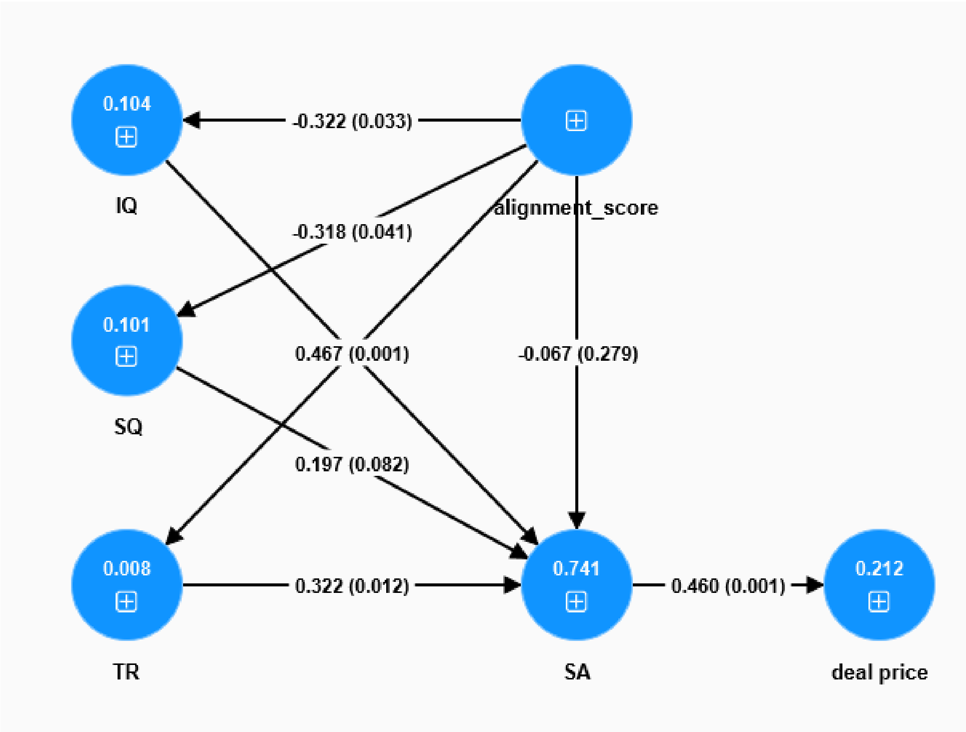



FIGURE D.2: Result of structural model

| | Original sample | Standard deviation | T statistics | P values |
|---|---|---|---|---|
| Path | | | | |
| IQ -> SA | 0.467 | 0.153 | 3.047 | 0.001 |
| SA -> deal price | 0.456 | 0.147 | 3.136 | 0.001 |
| SQ -> SA | 0.197 | 0.142 | 1.389 | 0.082 |
| TR -> SA | 0.322 | 0.143 | 2.252 | 0.012 |
| score -> IQ | -0.322 | 0.176 | 1.834 | 0.033 |
| score -> SA | -0.067 | 0.114 | 0.587 | 0.279 |
| score -> SQ | -0.318 | 0.183 | 1.738 | 0.041 |
| score -> TR | -0.089 | 0.148 | 0.603 | 0.273 |
| Total indirect effects | | | | |
| score -> SA | -0.242 | 0.156 | 1.545 | 0.061 |
| score -> deal price | -0.142 | 0.100 | 1.423 | 0.077 |
| Specific indirect effects | | | | |
| score -> IQ -> SA -> deal price | -0.069 | 0.056 | 1.227 | 0.110 |
| score -> SA -> deal price | -0.031 | 0.056 | 0.551 | 0.291 |
| score -> SQ -> SA -> deal price | -0.029 | 0.030 | 0.946 | 0.172 |
| score -> TR -> SA -> deal price | -0.013 | 0.030 | 0.434 | 0.332 |
| Total effects | | | | |
| score -> SA | -0.309 | 0.166 | 1.856 | 0.032 |
| score -> deal price | -0.142 | 0.100 | 1.423 | 0.077 |

TABLE D.2: The results of direct, indirect and total effects

# Bibliography


[1] Ecommerce chatbot: 9 best chatbots for ecommerce from successful brands, 2022. URL: https://gettalkative.com/info/ecommerce-chatbot.

[2] Chatbot market size, share, trends & growth report, 2023. Accessed: (2023-12-15). URL: https://www.grandviewresearch.com/industry-analysis/chatbot-market.

[3] IBM watson to watsonx. https://www.ibm.com/watson, 2023. Accessed: (2023-12-15).

[4] Pactum clients case studies. https://pactum.com/clients/, 2023. Accessed: (2023-12-15).

[5] Wysa case studies reports. https://www.wysa.com/case-studies-and-reports, 2023. Accessed: (2023-12-15).

[6] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*, pages 373–383. Springer, 2020.

[7] Manon Arcand, Sandrine PromTep, Isabelle Brun, and Lova Rajaobelina. Mobile banking service quality and customer relationships. *International Journal of Bank Marketing*, 35(7):1068–1089, 2017.

[8] Muhammad Ashfaq, Jiang Yun, Shubin Yu, and Sandra Maria Correia Loureiro. I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of ai-powered service agents. *Telematics and Informatics*, 54:101473, 2020.

[9] Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. Chatbot for healthcare system using artificial intelligence. In *2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*, pages 619–622. IEEE, 2020.

[10] Kabir Batra, Niraj Nair, Atharva Chaudhary, and Dipti Jadhav. Intelligent negotiation bot using machine learning techniques. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–5. IEEE, 2022.

[11] Rodrigo Bavaresco, Diórgenes Silveira, Eduardo Reis, Jorge Barbosa, Rodrigo Righi, Cristiano Costa, Rodolfo Antunes, Marcio Gomes, Clauter Gatti, Mariangela Vanzin, et al. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review*, 36:100239, 2020.

[12] Allan Bell. Language style as audience design. *Language in society*, 13(2):145–204, 1984.

[13] Stefan Benus. Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6:802–813, 2014. URL: https://api.semanticscholar.org/CorpusID:17127960.

[14] Kirsten Bergmann, Holly Branigan, and Stefan Kopp. Exploring the alignment space – lexical and gestural alignment with real and virtual humans. *Frontiers in Human-Media Interaction*, 04 2015. doi:10.3389/fict.2015.00007.

[15] Anol Bhattacherjee. Understanding information systems continuance: An expectation-confirmation model. *MIS quarterly*, pages 351–370, 2001.

[16] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[17] Heather Bortfeld and Susan E Brennan. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147, 1997.

[18] Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25, 2000.

[19] Holly P Branigan, Martin J Pickering, Janet F McLean, and Alexandra A Cleland. Syntactic alignment and participant role in dialogue. *Cognition*, 104(2):163–197, 2007.

[20] Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, and Ash Brown. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57, 2011.

[21] Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. *Modelling and Evaluation of Lexical and Syntactic Alignment with a Priming-Based Microplanner*, page 85–104. 01 2010. doi:10.1007/978-3-642-15573-4_5.

[22] Sabrina Campano, Jessica Durand, and Chloé Clavel. Comparative analysis of verbal alignment in human-human and human-agent interactions. In *International Conference on Language Resources and Evaluation*, 2014. URL: https://api.semanticscholar.org/CorpusID:10697066.

[23] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

[24] Fabio Clarizia, Francesco Colace, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. Chatbot: An education support system for student. In *Cyberspace Safety and Security: 10th International Symposium, CSS 2018, Amalfi, Italy, October 29–31, 2018, Proceedings 10*, pages 291–302. Springer, 2018.

[25] Eve V Clark. Conceptual perspective and lexical choice in acquisition. *Cognition*, 64(1):1–37, 1997.

[26] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[27] L Clark. Facebook teaches bots how to negotiate. they learn to lie instead. wired, june 15, 2017.

[28] Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell. Large language models in the workplace: A case study on prompt engineering for job type classification. In *International Conference on Applications of Natural Language to Information Systems*, pages 3–17. Springer, 2023.

[29] Benjamin Cowan and Holly Branigan. Does voice anthropomorphism affect lexical alignment in speech-based human-computer dialogue? pages 155–159, 09 2015. doi:10.21437/Interspeech.2015-75.

[30] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[31] Massilva Dekkal, Manon Arcand, Sandrine Prom Tep, Lova Rajaobelina, and Line Ricard. Factors affecting user trust and intention in adopting chatbots: the moderating role of technology anxiety in insurtech. *Journal of Financial Services Marketing*, pages 1–30, 2023.

[32] William H DeLone and Ephraim R McLean. The delone and mclean model of information systems success: a ten-year update. *Journal of management information systems*, 19(4):9–30, 2003.

[33] Daniel Druckman and Mara Olekalns. Emotions in negotiation. *Group decision and negotiation*, 17:1–11, 2008.

[34] Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frederic Landragin. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55, 06 2021. doi:10.1007/s10579-021-09532-w.

[35] Ondřej Dušek and Filip Jurcicek. A context-aware natural language generator for dialogue systems. pages 185–190, 01 2016. doi:10.18653/v1/W16-3622.

[36] Adrian Egli. Chatgpt, gpt-4, and other large language models: The next revolution for clinical microbiology? *Clinical Infectious Diseases*, 77(9):1322–1328, 2023.

[37] Gunther Eysenbach et al. The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers. *JMIR Medical Education*, 9(1):e46885, 2023.

[38] Michael Filzmoser, Rudolf Vetschera, and Sabine T Koeszegi. Negotiation processes: Empirical insights. *Handbook of Group Decision and Negotiation*, pages 61–79, 2021.

[39] Roger Fisher, William L Ury, and Bruce Patton. *Getting to yes: Negotiating agreement without giving in.* Penguin, 2011.

[40] Joseph P Forgas. On feeling good and getting your way: Mood effects on negotiator cognition and bargaining strategies. *Journal of personality and social psychology*, 74(3):565, 1998.

[41] Carol A Fowler, Julie M Brown, Laura Sabadini, and Jeffrey Weihing. Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of memory and language*, 49(3):396–413, 2003.

[42] Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

[43] Susan R Fussell and Robert M Krauss. Coordination of knowledge in communication: effects of speakers' assumptions about what others know. *Journal of personality and Social Psychology*, 62(3):378, 1992.

[44] Simon Garrod and Anthony Anderson. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218, 1987.

[45] Simon Garrod and Aileen Clark. The development of dialogue co-ordination skills in schoolchildren. *Language and Cognitive Processes*, 8(1):101–126, 1993.

[46] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. *arXiv preprint arXiv:1808.09637*, 2018.

[47] Zhichao Hu, Gabrielle Halberg, Carolynn Jimenez, and Marilyn Walker. *Entrainment in Pedestrian Direction Giving: How Many Kinds of Entrainment?*, pages 151–164. 04 2016. doi:10.1007/978-3-319-21834-2_14.

[48] Shen Huiyang and Min Wang. Effects of social skills on lexical alignment in human-human interaction and human-computer interaction. *Computers in Human Behavior*, 143, 02 2023. doi:10.1016/j.chb.2023.107718.

[49] Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Katsunori Shimohara, Mitsunori Miki, and Norihiro Hagita. Lexical entrainment in human robot interaction. *International Journal of Social Robotics*, 7, 04 2015. doi:10.1007/s12369-014-0255-x.

[50] Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, and Yulia Tsvetkov. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. *arXiv preprint arXiv:2106.00920*, 2021.

[51] Dharun Lingam Kasilingam. Understanding the attitude and intention to use smartphone chatbots for shopping. *Technology in Society*, 62:101280, 2020.

[52] Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Raefer Gabriel, Ashwin Ram, and Rohit Prasad. Alexa prize—state of the art in conversational ai. *AI Magazine*, 39(3):40–55, 2018.

[53] Theodora Koulouri, Stanislao Lauria, and Robert Macredie. Do (and say) as i say: Linguistic adaptation in human-computer dialogs. *Human–Computer Interaction*, 31:1–79, 12 2014. doi:10.1080/07370024.2014.934180.

[54] Theodora Koulouri, Stanislao Lauria, and Robert D Macredie. Do (and say) as i say: Linguistic adaptation in human–computer dialogs. *Human–Computer Interaction*, 31(1):59–95, 2016.

[55] Alice J Lee, Malia F Mason, and Claire S Malcomb. Beyond cheap talk accounts: A theory of politeness in negotiations. *Research in Organizational Behavior*, 41:100154, 2021.

[56] SeoYoung Lee and Junho Choi. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies*, 103:95–105, 2017.

[57] Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

[58] Chechen Liao, Jain-Liang Chen, and David C Yen. Theory of planning behavior (tpb) and customer satisfaction in the continued use of e-service: An integrated model. *Computers in human behavior*, 23(6):2804–2822, 2007.

[59] Gesa Linnemann and Regina Jucks. As in the question, so in the answer? language style of human and machine speakers affects interlocutors convergence on wordings. *Journal of Language and Social Psychology*, 35, 01 2016. doi:10.1177/0261927X15625444.

[60] Gesa Linnemann and Regina Jucks. 'can i trust the spoken dialogue system because it uses the same words as i do?'—influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction. *Interacting with Computers*, 03 2018. doi:10.1093/iwc/iwy005.

[61] Meina Liu and Chongwei Wang. Explaining the influence of anger and compassion on negotiators' interaction goals: An assessment of trust and distrust as two distinct mediators. *Communication Research*, 37(4):443–472, 2010.

[62] José Lopes, Maxine Eskenazi, and Isabel Trancoso. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech Language*, 31, 05 2015. doi:10.1016/j.csl.2014.11.007.

[63] Bei Luo, Raymond YK Lau, Chunping Li, and Yain-Whar Si. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):e1434, 2022.

[64] Marcello M Mariani, Novin Hashemi, and Jochen Wirtz. Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. *Journal of Business Research*, 161:113838, 2023.

[65] Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190, 2004.

[66] Karen Pudner, Keeley A Crockett, and Zuhair Bandar. An intelligent conversational agent approach to extracting queries from natural language. In *World Congress on Engineering*, volume 1, page 305, 2007.

[67] Tom Puthiyamadam and José Reyes. Experience is everything: Here's how to get it right, 2018.

[68] Arabella Sinclair, Katherine McCurdy, Christopher Lucas, Adam Lopez, and Dragan Gasevic. Tutorbot corpus: Evidence of human-agent verbal alignment in second language learner dialogues. In Collin F Lynch, Agathe Merceron, Michel Desmarais, and Roger Nkambou, editors, *Proceedings of the 12th International Conference on Educational Data Mining*, pages 414–419, July 2019. Educational Data Mining 2019, EDM 2019 ; Conference date: 02-07-2019 Through 05-07-2019. URL: http://educationaldatamining.org/edm2019/.

[69] Laura Spillner and Nina Wenig. Talk to me on my level–linguistic alignment for chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, pages 1–12, 2021.

[70] Laura Spillner and Nina Wenig. Talk to me on my level – linguistic alignment for chatbots. pages 1–12, 09 2021. doi:10.1145/3447526.3472050.

[71] Sumit Srivastava, Mariet Theune, and Alejandro Catala. The role of lexical alignment in human understanding of explanations by conversational agents. pages 423–435, 03 2023. doi:10.1145/3581641.3584086.

[72] Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. Microplanning with communicative intentions: The spud system. *Computational Intelligence*, 19, 05 2001. doi:10.1111/1467-8640.00221.

[73] Rick B Van Baaren, Rob W Holland, Bregje Steenaert, and Ad Van Knippenberg. Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, 39(4):393–398, 2003.

[74] Richard Wallace. The elements of aiml style. *Alice AI Foundation*, 139, 2003.

[75] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[76] Jason D Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016.

[77] Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. Improving dialog systems for negotiation with personality modeling. *arXiv preprint arXiv:2010.09954*, 2020.

[78] Ran Zhao, Oscar J Romero, and Alex Rudnicky. Sogo: a social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on intelligent virtual agents*, pages 239–246, 2018.

[79] Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. A dynamic strategy coach for effective negotiation. *arXiv preprint arXiv:1909.13426*, 2019.

[80] Yiheng Zhou, Yulia Tsvetkov, Alan W Black, and Zhou Yu. Augmenting non-collaborative dialog systems with explicit semantic and strategic dialog history. *arXiv preprint arXiv:1909.13425*, 2019.