Talk to Me on My Level – Linguistic Alignment for Chatbots

Laura Spillner
Digital Media Lab, TZI, University of Bremen
Bremen, Germany
laura.spillner@uni-bremen.de

Nina Wenig
Digital Media Lab, TZI, University of Bremen
Bremen, Germany
nwenig@uni-bremen.de

ABSTRACT

Digital companions and conversational agents are becoming increasingly popular in our everyday lives. Natural language interfaces play an important role in ubiquitous computing: voice assistants are used to control smart home devices and smartphone applications; chatbots serve as an interface to solve tasks and acquire information easily. However, misunderstandings due to non-standard language, expressions that serve social functions without conveying information, or a lack of situational awareness still pose problems for these interfaces. Humans are able to prevent or repair communication failures by imitating their conversation partner's lexical choices, sentence structures, and overall language style; a mechanism known as Linguistic Alignment. In this paper we present different strategies to easily integrate an alignment effect in natural language interfaces. We implemented a chatbot that imitates alignment and tested it in an online user study with 75 participants. Our results show that alignment helps to decrease user frustration and perceived task workload.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *User studies*; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

conversational agents, chatbots, linguistic alignment, lexical entrainment, adaptation

ACM Reference Format:
Laura Spillner and Nina Wenig. 2021. Talk to Me on My Level – Linguistic Alignment for Chatbots. In *23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '21), September 27-October 1, 2021, Toulouse & Virtual, France.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3447526.3472050

1 INTRODUCTION

For a fluent and easy interaction with mobile and ubiquitous devices, natural language plays an emerging role. Communication via chat apps is arguably among the most common uses of mobile devices: in 2019, 50% of the time users spent on mobile devices was in social and communication apps[1]. For several years, the most popular apps have been on Messenger applications [39]. WhatsApp alone had 2 billion monthly users in January 2021[2]. At the same time, voice input and free text input for other apps are becoming more and more common, especially in voice assistants, chatbots and new search interfaces. One major breakthrough for voice assistants was the invention of Siri in 2011 on the iPhone. In spite of the renewed popularity and the considerable amount of recent research in Artificial Intelligence (AI) based Natural Language Understanding (NLU) and Generation (NLG), these interfaces still face some difficult challenges.

Communication comes with informal and often non-standard language use and new types of expressions (such as emojis) that may hinder successful human-computer communication. Users tend to react socially to computer agents [45], especially so if these agents also employ natural language or imitate social and affective cues [1, 28]. For conversational agents, this often leads to misunderstandings or communication breakdowns, and currently available solutions (e.g. voice assistants or chatbots) still fail quite often in real-world usage [40]. Therefore, it is important to find methods to support successful human-computer communication.

In the field of psycho-linguistics, many researchers have studied how two interlocutors, i.e. the people engaged in a conversation, are able to achieve mutual understanding and prevent communicative failures. The social and collaborative nature of human conversation is well established: Theories such as *Communication accommodation theory* [26] and *Interactive alignment theory* [52] describe how successful communication between humans depends on both participants' ability to adapt to the language of their conversation partner. Speakers imitate each other's language in many ways; by repeating word choices, phrasings, and sentence structure, as well as becoming more similar to one another in their overall language style. In this way, they are able to achieve a more similar understanding of the situation under discussion, thus paving the way for successful communication.

We suggest that considering the psychological model of human communication can be essential for further improvements in conversational agents. Chatbots in particular are intrinsically tied to the linguistic modality of chatting; which is more social, more mobile, and therefore less adherent to classic dialog structure and formal grammar or spelling rules.

In this paper we investigate how users perceive and evaluate a chatbot that mimics human behavior through alignment, either by using similar terms as the user or by using both similar terms and similar sentence structures. Thus we implement an alignment effect in a chatbot using two different approaches. The chatbot is then tested in a user study in order to investigate the perceived task workload by the user during a task as well as user engagement.

[1] https://www.appannie.com/en/go/state-of-mobile-2020/

[2] https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/, updated 2021-04-18

MobileHCI '21, September 27-October 1, 2021, Toulouse & Virtual, France

In human interaction, linguistic alignment plays an essential role in assuring successful communication. Higher alignment between conversational partners can impact how positively they are perceived. This is even more important when it comes to solving tasks: higher alignment between two people improves task success [53], as well as workload and engagement in information-seeking conversations [58]. Solving tasks and acquiring information are key use cases of chatbots and other conversational agents. Since previous studies have shown that users linguistically align to computers in several ways [12], we reason that alignment can similarly improve communication between humans and conversational agents. If alignment from the chatbot can increase alignment between human and agent, this capability could play an important role in improving user interactions with chatbots and reducing the perceived workload when solving tasks. Therefore, we hypothesize (1) that users will align more strongly to a chatbot that also exhibits alignment; and (2) that increased alignment will impact both workload and user engagement.

This paper makes two primary contributions: we show that aligning the chatbot automatically to the users does in fact lead to stronger alignment from users, and that this helped to lower the workload perceived by the user and increase user engagement. Therefore, we propose that these strategies should be further investigated in research on human-computer communication, and can be utilized in future implementations of chatbots and voice assistants.

2 RELATED WORK

A number of connected theories in psycho-linguistics describe how two interlocutors (participants in a conversation) can achieve mutual convergence by striving to accommodate each other's linguistic choices [5, 17, 24, 37, 52]. This work is based mainly on Pickering and Garrod [52]'s Interactive Alignment Theory, and we will use the term *alignment* throughout to refer to this phenomenon. Pickering and Garrod [52]'s theory posits that interlocutors will over the course of their conversation align on different levels of linguistic representation (e.g. lexical choices, syntactic structure, pronunciation, language style). They argue that this is what ultimately leads to alignment of the speakers' understandings of the situation they are discussing, and thus makes successful communication possible.

It has been shown before that interlocutors will imitate one another on non-linguistic levels, such as facial expressions and gestures [2]. The same phenomenon can be observed across linguistic levels, such as phonetic realizations & prosody [48], lexical choices [15, 24, 42] and syntactic sentence structures [9, 32, 51]. There is strong evidence that how someone is perceived is influenced positively both by their non-verbal alignment [17, 35, 41] as well as by their linguistic alignment [7, 59]. In our study, we focus on the effects of lexical and syntactic/structural alignment. While the most commonly used term in the literature is syntactic alignment, 'structural priming' is the terminology preferred by Pickering and Ferreira [51]. Since our work differs from lab studies investigating specific syntactic choices (e.g. [11, 13, 19]), we will use the term *structural alignment*.

It is well established in Human-Computer Interaction (HCI) that humans will apply social scripts to interaction with computers [45], especially so when it comes to inherently social activities such as natural language dialog [1, 28, 44]. A number of studies have been able to show that human speakers will align to computers in the same way in which they align to other humans [12, 33, 56], including aspects such as prosody [3, 57], lexical [4, 49] and structural choices [13, 14]. Many of these studies are based on Wizard of Oz experiments. There has to our knowledge not been a considerable amount of research on how alignment might be implemented in a conversational agent, and whether or not that would influence user interaction.

Previous studies in HCI have however shown that social characteristics (sometimes referred to as social cues, anthropomorphic features, or human-like behavior) impact interaction with conversational agents [1, 27, 38]. Implementing social characteristics can positively influence how conversational agents are perceived, and can lead to a stronger social response from users: Lee et al. [36] evaluated the influence of different self-disclosure techniques on the users' self-disclosure. Therefore, they tested the chatbots over a period of three weeks with small talk and sensitive questions. They found out that high self-disclosure of the chatbot fosters users' self-disclosures regarding sensitive questions.

In spite of this, several researchers have noted that conversational agents still struggle with some aspects of understanding and generating natural dialog which are inherent to human communication, such as situatedness and context awareness, phatic responses (expressions that serve social functions instead of conveying information), and non-standard language (sociolects, dialects, register, informal language) [6, 16, 18, 21]. Additionally, some researchers have shown cases in which social cues in fact lead to adverse effects instead of influencing interaction positively [8, 25, 60]. One possible explanation for this discrepancy might be that when employing social characteristics in conversation, it is more important to align to the conversational partner than to employ the 'generally correct' level of such a characteristic.

Notably, Thomas et al. [58] found this to be the case in their study on language style. They analyzed information-seeking conversations between humans, the kind of dialogs that are of central importance for closed-domain conversational agents. They found that there was no single 'best style' or 'generally good' style that lead to the lowest effort or highest engagement for the user. Instead, workload was lower and engagement was higher when both participants were more similar to one another's style.

However, there is only a small number of studies that we are aware of that attempt to implement an alignment effect in a conversational agent. Based on Thomas et al. [58]'s study, Hoegen et al. [34] developed a voice-based conversational agent that was capable of matching their language style of its conversational partner, and found that users rated the agent better and as more trustworthy when it aligned to their personal style. This agent is capable of 'chit-chat' conversation, and adapts to its user across linguistic levels (word choice and syntax as well as prosody). In contrast, earlier works were concerned with prosodic and phonetic alignment: Suzuki and Katagiri [57] analyzed how changes in loudness and response latency can be used to induce prosodic alignment in HCI. Nishimura et al. [46] also analyzed prosodic alignment in conversations with dialog systems, and proposed a model that could actively change the system's prosody in order to imitate that of the user.

We hypothesize that linguistic alignment should also positively impact communication between humans and computers. We think that this will be the case not just for overall language style, but also for other linguistic levels such as lexical and structural alignment. Instead of there being an ideal 'level of formality' for a chatbot, we hypothesize that a chatbot that can align to the user's word choices and sentence structures should be perceived better, and lead to lower workload and higher engagement.

3 KONRAD - AN AI-BASED CHATBOT WITH DIFFERENT ALIGNMENT LEVELS

In order to test our hypothesis, we developed a chatbot for a simple closed-domain information-seeking task, and compared three variants of that chatbot in a user study. The three variants are 1) a baseline variant with no intentional alignment from the chatbot; 2) a variant that aims to create lexical alignment by substituting default terms with those preferred by the user; and 3) a variant that aims to create both lexical and structural alignment by using grammatical transformations to create the chatbot's reply based on the user's query. Our goal in the study is, firstly, to ascertain whether we can create measurable alignment and how that in turn impacts user alignment to the chatbot; and secondly, to analyze the influence of alignment on the user's interaction with the chatbot in terms of task workload and user engagement.

Konrad is able to answer questions about movies running in a fictional local cinema - it can inform users about which movies are running when, what they are about, who stars in them or directed them. The goal in choosing this subject domain was to have a relatively common use case, which would not require sophisticated methods of knowledge representation, and at the same time call for language and terminology that allowed enough opportunities for simple lexical alignment (whereas a bot about an 'expert' topic might require specific, fixed terminology).

The chatbot classifies user queries as one of a number of predefined intents (e.g. questions asking about when a movie starts), retrieves the necessary information, and constructs an answer based on the identified intent. In current real-world chatbot applications, intent recognition is often achieved with neural networks, but answers are commonly generated without the use of AI. For example, Google's Dialogflow[3] is a popular platform for the creation of conversational agents and interfaces, which provides deep learning NLU methods but only allows for static answers. Although current state-of-the-art transformer models (such as GPT-3[4]) are capable of producing very good sentences and context-dependent answers, using these deep learning models for NLG has a number of drawbacks (such as size, training speed, and susceptibility to biased training data).

We decided to build Konrad using template and rule-based methods for creating an answer. In the core implementation, depending on the intent that was recognized, Konrad replies either with one of a number of static answers, or uses an answer template into which the requested information is inserted. For example, the template for a question about the start time of a specific movie (e.g. "When does Batman begin?") is "[movie title] starts at [time]" (so the chatbot answers "Batman starts at 8:00"). This has the advantage that it allows for a directed addition of alignment at specific points in the chatbot's answers, as we can entirely control how answers are formed (which would not be the case using e.g. Transformer models).

3.1 Implementation

We implemented the chatbot with the help of spaCy[5], an open-source library for Natural Language Processing (NLP), and one of the most common libraries for text analysis offering many state-of-the-art algorithms. SpaCy provides an English language model, can integrate deep learning models, and can be used for tasks such as tokenization, part-of-speech tagging, named entity recognition, and dependency parsing.

Konrad can differentiate between 21 different intents, which include a number of general intents (greeting, request for help, and thanks), a few general questions about the cinema (requesting a list of all available movies, the name of the cinema, or the ticket cost), as well 13 specific intents that users can ask about the movies (the movie title, date, start time, end time, when the movie plays (time and date), duration, genre, plot, year of release, director, actors, actor's roles, and critical rating). Additionally, Konrad is able to remember which movies or subsets of available movies were talked about last, and what the most recent intent was. We collected 14 movies in total, covering a variety of different genres and eras. The information about the movies was collected from IMDB[6] and Rotten Tomatoes[7].

Notably, the goal in creating Konrad was not to develop a particularly knowledgeable chatbot, or one that could significantly improve upon the traditional user interaction with a cinema's website. Rather, we want to utilize Konrad as a simplified example of a possible real-world application, in order to use it to investigate the effect of alignment on user interaction.

During the iterative development process, the chatbot underwent a number of pre-tests with potential users, which served to test the chatbot's intent recognition and later the different methods of language generation. Based on the first successful pre-tests, we learned that this limited number of intents and the amount of memory are sufficient for users to find a fitting movie. Additionally, these conversations with the chatbot supplemented the training data used to train a model for intent recognition.

We realized intent recognition with a feed-forward neural network (FFNN) implemented in Python using TensorFlow[8]. An FFNN is the simplest kind of neural network, in which information can only flow through the network in one direction (it is acyclic). We trained the model on data that was partially created manually, and partially collected from the pre-tests and manually annotated with the correct intents. We implemented three variants, one as a baseline with no user-dependent results, one where we substitute specific words, and one where we additionally transform the sentences based on the user input. The intent recognition and the information that the user receives do not differ between the three variants of the chatbot - the only difference between them is how Konrad's

[3] https://dialogflow.cloud.google.com/cx/

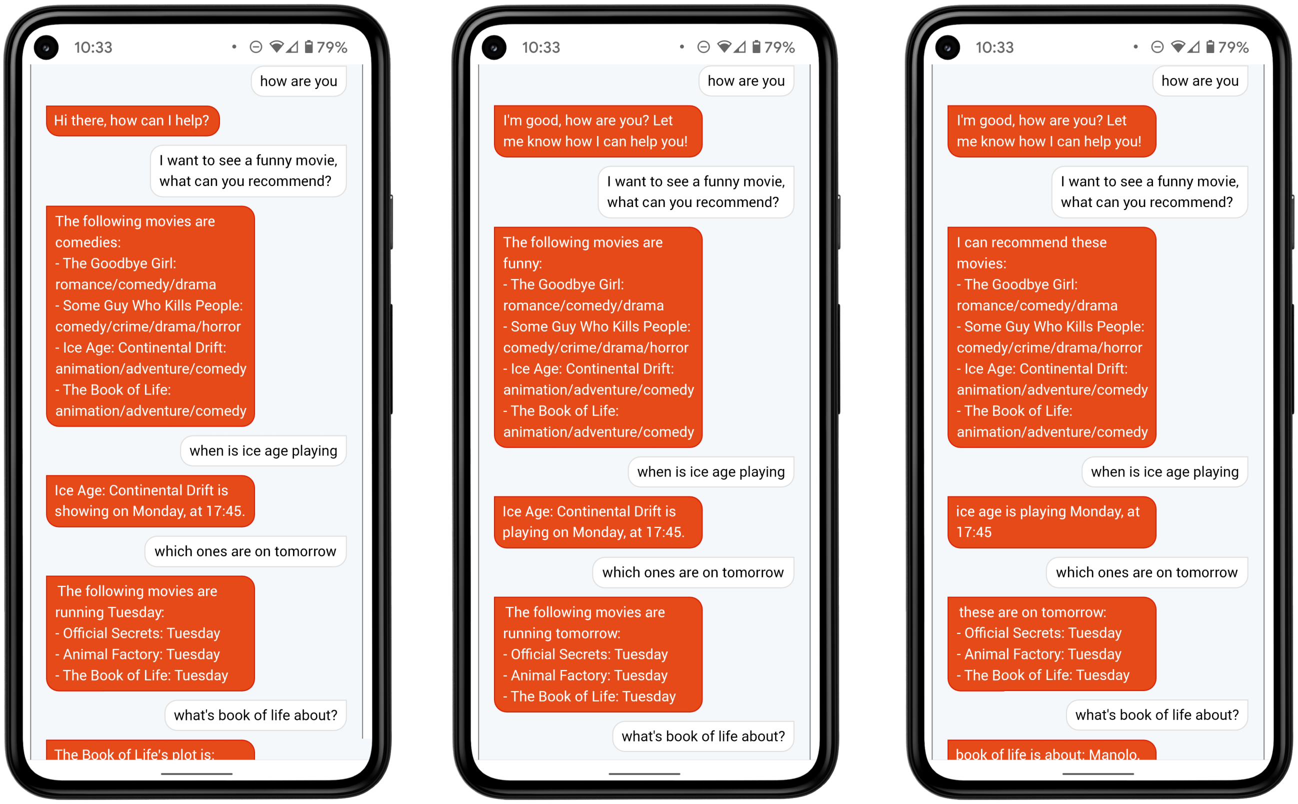
[4] https://openai.com/

[5] https://spacy.io/

[6] https://www.imdb.com/

[7] https://www.rottentomatoes.com/

[8] https://www.tensorflow.org/



how are you

Hi there, how can I help?

I want to see a funny movie, what can you recommend?

The following movies are comedies:
- The Goodbye Girl: romance/comedy/drama
- Some Guy Who Kills People: comedy/crime/drama/horror
- Ice Age: Continental Drift: animation/adventure/comedy
- The Book of Life: animation/adventure/comedy

when is ice age playing

Ice Age: Continental Drift is showing on Monday, at 17:45.

which ones are on tomorrow

The following movies are running Tuesday:
- Official Secrets: Tuesday
- Animal Factory: Tuesday
- The Book of Life: Tuesday

what's book of life about?

The Book of Life's plot is:

(a) Baseline - no additional alignment

how are you

I'm good, how are you? Let me know how I can help you!

I want to see a funny movie, what can you recommend?

The following movies are funny:
- The Goodbye Girl: romance/comedy/drama
- Some Guy Who Kills People: comedy/crime/drama/horror
- Ice Age: Continental Drift: animation/adventure/comedy
- The Book of Life: animation/adventure/comedy

when is ice age playing

Ice Age: Continental Drift is playing on Monday, at 17:45.

which ones are on tomorrow

The following movies are running tomorrow:
- Official Secrets: Tuesday
- Animal Factory: Tuesday
- The Book of Life: Tuesday

what's book of life about?

(b) Substitution

how are you

I'm good, how are you? Let me know how I can help you!

I want to see a funny movie, what can you recommend?

I can recommend these movies:
- The Goodbye Girl: romance/comedy/drama
- Some Guy Who Kills People: comedy/crime/drama/horror
- Ice Age: Continental Drift: animation/adventure/comedy
- The Book of Life: animation/adventure/comedy

when is ice age playing

ice age is playing Monday, at 17:45

which ones are on tomorrow

these are on tomorrow:
- Official Secrets: Tuesday
- Animal Factory: Tuesday
- The Book of Life: Tuesday

what's book of life about?

book of life is about: Manolo,

(c) Transformation

Figure 1: Example Conversations with all three variants of Konrad

replies are generated. Figure 1 illustrates example conversations for all three variants.

3.2 Baseline Variant

In the baseline variant, language generation is achieved purely through the pre-defined templates. There is no additional alignment effect, therefore, the chatbot's lexical choices, sentence structure, and overall language style do not depend on those of the user at all. Any measured alignment in this condition happens either randomly, or results from the alignment of the user.

3.3 Lexical Alignment through Substitution

In this variant, we aim to imitate a basic lexical alignment effect by substituting terms in the template answers with terms applied by the user. E.g. when the user uses one term such as *story*, the chatbot also uses that term instead of the default expression *plot*. Research on linguistic alignment shows a *priming* effect in regards to the lexical choice [10, 29, 54]: People are likely to repeat the terms that their conversational partner has used to refer to objects or concepts in the preceding utterance, even if these terms are not their preferred terms. Based on this, the chatbot we implemented attempts lexical alignment by scanning the user's preceding utterance for terms that are synonymous with other terms that the chatbot usually uses in its replies.

Research has shown that there is a large amount of variability of terms that users will intuitively choose when interacting with a computer program, e.g. when using commands [22]. Therefore, we did not want to rely on a limited set of pre-defined synonyms, and instead utilized *word embeddings* to test for semantic similarity. Word embeddings (or word vectors) represent words as vectors in space. State-of-the-art methods are able to generate word embeddings that capture semantic relationships, i.e. words that are more similar in meaning have representations that are closer to each other in vector space and represent analogies between words Mikolov et al. [43]. We utilized spaCy's 300-dimensional GloVe word vectors [50] and similarity testing in order to substitute terms in this variant of Konrad.

Using a small number of manually defined reference terms in combination with semantic similarity testing, the chatbot is able to recognize a variety of synonyms for terms, and replace terminology used in the template answer with the terms seen in the user's priming sentence. For example, Konrad's default term for a movie's critical score is 'score', but the chatbot is able to recognize alternate terms such as 'rating' and use these instead.

We expect that in analyzing user dialogs, we should be able to measure higher lexical alignment from this variant of the chatbot. Furthermore, we hypothesize that this should lead to increased alignment from the user, which in turn should have a positive influence on the workload for a task and user engagement as it does in human dialog [58].

3.4 Lexical and Structural Alignment through Transformation

In order to achieve lexical alignment as well as alignment of sentence structures, we implemented another method of answer generation, in which Konrad's answers to user queries are created by transforming the user queries based on grammatical rules.

The generation of the chatbot's answer from a well-formed, grammatically correct question is straightforward, since English sentence structure is quite rigid. We identify subject, finite verb, and object through part-of-speech tagging and dependency parsing. Then, depending on whether the user query is formed as a yes-no question or a *wh*-question[9], additional information may be inserted at the appropriate place (e.g. "Does the movie run on Wednesday?" is rearranged into "Yes, the movie runs on Wednesday", whereas "When does the movie start?" is answered with "The movie starts at 20:00"). Additional transformations include reversing pronouns ('you' vs. 'I' and 'we'), and adjusting the verb tense and noun form if necessary. Finally, some of the stylistic markers of informal conversation are also applied to imitate the user sentence, such as the use of emoji and punctuation as well as the capitalization of the first word of the sentence.

For user messages that are not phrased as questions (or not recognized correctly), some additional rules were created manually. These include appropriate greetings (e.g. if the user greets the chatbot with 'How are you?', it answers 'I'm good, how are you?'), as well as fall-back answers in case that the query was phrased as a statement (e.g. 'I want to see a movie on Sunday') or if the subject or finite verb could not be identified in the user query. The latter happens either because the user did not form a full sentence (e.g. 'And batman?'), because the dependency parsing failed, or because spelling or grammar mistakes led to the language model misidentifying the part-of-speech of some of the tokens.

4 STUDY DESIGN

We compared the three variants of Konrad in an online user study. The study was shared primarily among students and researchers at the University of Bremen, and thus includes mostly participants who are native German speakers, some native English speakers, as well as some native speakers of other languages. Since previous work has shown that the alignment effect persists in non-native speakers as well as for language learners [55], we do not further distinguish participants based on their native language.

On our website, we presented the participants with one task that they should solve with the help of Konrad: finding a movie that they would want to watch. The study participants were not aware that we were investigating linguistic alignment, instead, they were under the impression that we were testing the quality of the chatbot. We designed the study as a between-group design because we saw a considerable training effect during the pre-tests, which we wanted to avoid. After finishing the task, each participant was asked to complete a questionnaire about their perception of the interaction.

84 persons completed the questionnaire, however, nine of those were excluded (four because they did not complete the task or their conversation with Konrad was less than eight turns long, and five in order to balance the number of participants per condition to that of the smallest group). Thus, we included 25 participants per group, to a total of 75 participants (21 female, 47 male, 7 no answer or other), with an average age of 24.93 ($SD = 6.69$). Ten participants are native English speakers, 63 are native German speakers, and two are native speakers of other languages.

Out of all 75 participants, only 15 took part in the study on a mobile device, while 60 tested the chatbot on a PC. However, 71 participants indicated that they were regular users of chat apps such as WhatsApp, and 51 regularly use intelligent assistants (e.g. Siri by Apple or Amazon Alexa). Additionally, 45 of the participants answered that they used an 'AI chatbot' before.

4.1 Measuring Linguistic Alignment

We analyzed the participants' conversations with the chatbot in order to quantify both lexical (word choice) and structural (syntactic sentence structure) alignment both from the chatbot and from the user. We hypothesize that users will react to the aligned variants of Konrad by aligning more strongly with the chatbot as well, and we hypothesize that this will positively impact their workload and engagement during the task.

In order to quantify lexical alignment, we consider the priming effect: Due to lexical priming, a speaker should be more likely to repeat terms used in the preceding, priming sentence instead of using different, semantically equivalent terms [11, 23, 49]. Doyle and Frank [20] define a local, word-wise *conditional* alignment measure as the proportion of terms that are repeated in the sentence following the priming utterance. Based on this, we calculate a local alignment score for each message as the ratio of tokens in that message that also appear in the priming message. We then average that measure over the conversation to get one mean alignment score for the chatbot and the user each.

In analyzing structural or syntactic alignment, most previous studies have not measured the alignment in natural dialogs. Instead, they use a laboratory study set-up that constrains what types of syntactic structures participants are likely to use, by having them describe specific pictures. In these experiments, researchers usually compare two alternative syntactic choices, such as double object vs. prepositional object [11], active voice vs. passive voice, prenominal adjective vs. relative clause [19]. In one of the few studies that instead calculates syntactic alignment for natural language conversations, [53] identify which syntactic rules sentences are built on based on their grammatical structure. We take a similar, due to the shorter conversations in our study simplified, approach: Utilizing the dependency tree generated from the language model, we differentiate between sentences that consist only of subject and verb; those that include objects (direct and indirect); adverbials (complements and modifiers); nominal complements; or at least

[9] *Wh*-questions are those questions starting with what, when, where, who, whom, which, whose, why and how

one subordinate clause. Utterances that are missing either subject or verb (e.g. commands that a user might try with a chatbot such as 'Help') are not counted. An alignment score is calculated as the proportion of sentences that repeat the syntactic structure of the preceding, priming utterance.

Both of these measures quantify alignment as the average similarity of pairs of consecutive utterances, either in terms of word choices or in terms of sentence structures. Because of this, we expect that even if there is no actual, psychological alignment from an interlocutor (the chatbot or the user), the calculated alignment score is likely to be higher than zero due to random repetitions. The alignment score is influenced by several factors (e.g. length of the sentences). Therefore, it is only important how much the alignment scores differ for the chatbot variants.

4.2 Measuring User Perception

We used the NASA Task Load Index (NASA-TLX) [31] to measure user workload during the task. The original NASA-TLX consists of six subscales (mental demand, physical demand, temporal demand, performance, effort, and frustration) which are rated on a 100-point range divided into five-point steps; as well as a separate part that uses pairwise comparisons to establish how the individual subscales are weighted when calculating the mean perceived workload. However, many researchers have forgone these pairwise comparisons, and instead calculate a 'Raw TLX'[30] as the mean between the subscale ratings. When doing so, it is also common to exclude subscales that are irrelevant to the task at hand [30]. Therefore, we have excluded the subscale on physical demand.

Furthermore, we used the User Engagement Scale (UES) short form [47] to measure user engagement with the chatbot. The short form of the UES consists of four subscales (focused attention, perceived usability, aesthetic appeal, and reward), which are measured with twelve statements in a randomized order on a five-point Likert Scale (1='Strongly disagree', 5='Strongly agree'). A general engagement score is calculated as the mean across all four measurements[47].

Lastly, we added nine additional questions regarding the participants' opinion on the interaction with the chatbot (whether it was e.g. friendly, polite, natural) as well as the extent of the chatbot's domain knowledge, which were also rated on a 5-point Likert scale.

5 USER STUDY RESULTS

The online study took place over four weeks; the participants talked to Konrad over an average of 17 turns during a conversation (the minimum number of user turns was eight, the maximum 48).

As expected, in the substitution condition, Konrad was able to recognize a variety of different user-preferred terms. That includes for example terms that describe a movie shown at the cinema (e.g. 'running', 'playing', 'airing'), that refer to concepts such as a movie's critical score (e.g. 'rating', 'reviews', 'critiques' and even 'critics' [sic]) or a movie's plot (e.g. 'story', 'summary', 'synopsis', 'abstract'), as well as a variety of descriptions of movie genres (e.g. 'happy', 'funny', 'sad', 'scary'). Since many of our testers were not native English speakers, we noted that the chatbot was also able to recognize terms that are not generally used in English in this way, such as 'cards' instead of 'tickets'. In only a few cases, Konrad recognized a wrong term as a synonym ('[movie] is seeing' instead of 'showing').

In the transformation condition, the chatbot was able to directly transform user queries into answers as intended, although some user expressions lead to grammatical mistakes from the chatbot.

The number of user queries that the chatbot either could not answer (because it was not programmed to answer that kind of query) or did not answer correctly averaged 22.8%. This number does not differ significantly between the three variants (baseline: $M = .26$, $SD = .18$; substitution: $M = .21$, $SD = .11$; transformation: $M = .22$, $SD = .16$).

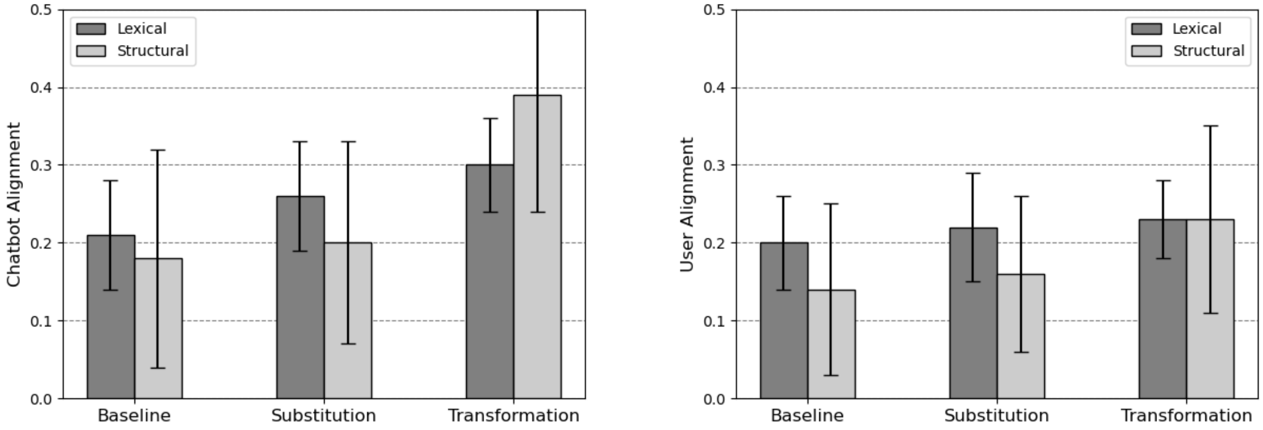
5.1 Linguistic Alignment

The analysis of the conversations using an ANOVA showed that there was a significant effect of chatbot variant on **chatbot** lexical alignment ($F_{2,72} = 11.171, p < .001$) and on chatbot structural alignment ($F_{2,72} = 16.346, p < .001$). The lowest lexical alignment was measured in the baseline variant, compared to higher scores in the substitution and transformation variants (baseline: $M = .21$, $SD = .07$; substitution: $M = .26$, $SD = .07$; transformation: $M = .30$, $SD = .06$). A Tukey-corrected post-hoc T-test showed that this was a significant difference between baseline and substitution variants ($p < .05$, $d = 0.749$) as well as between baseline and transformation variants ($p < .001$, $d = 1.321$). As expected, structural alignment is very close in the baseline and the substitution variant (baseline: $M = .18$, $SD = .14$; substitution: $M = .20$, $SD = .13$), and is significantly higher in the transformation variant (transformation: $M = .39$, $SD = .15$) compared to both the baseline ($p < .001$, $d = 1.400$) and the substitution variant ($p < .001$, $d = 1.358$).

User lexical alignment is also lowest in the baseline condition (baseline: $M = .20$, $SD = .06$), but is similar in both conditions with added alignment (substitution: $M = .22$, $SD = .07$; transformation: $M = .23$, $SD = .05$), see Figure 2b; there was not a significant difference. However, there is a significant effect of chatbot variant on user structural alignment ($F_{2,72} = 4.771, p < .05$). Similar to chatbot alignment, the users' structural alignment is only slightly higher in the substitution condition than in the baseline one, and highest in the transformation condition (baseline: $M = .14$, $SD = .11$; substitution: $M = .16$, $SD = .10$; transformation: $M = .23$, $SD = .12$) The difference between baseline and substitution variant is significant ($p < .05$, $d = 0.813$).

Correlation analysis shows that user lexical alignment is indeed correlated with chatbot lexical alignment ($r = .401$), Figure 3a, and user structural alignment is correlated both with chatbot structural alignment ($r = .741$), see Figure 3b and chatbot lexical alignment ($r = .462$).

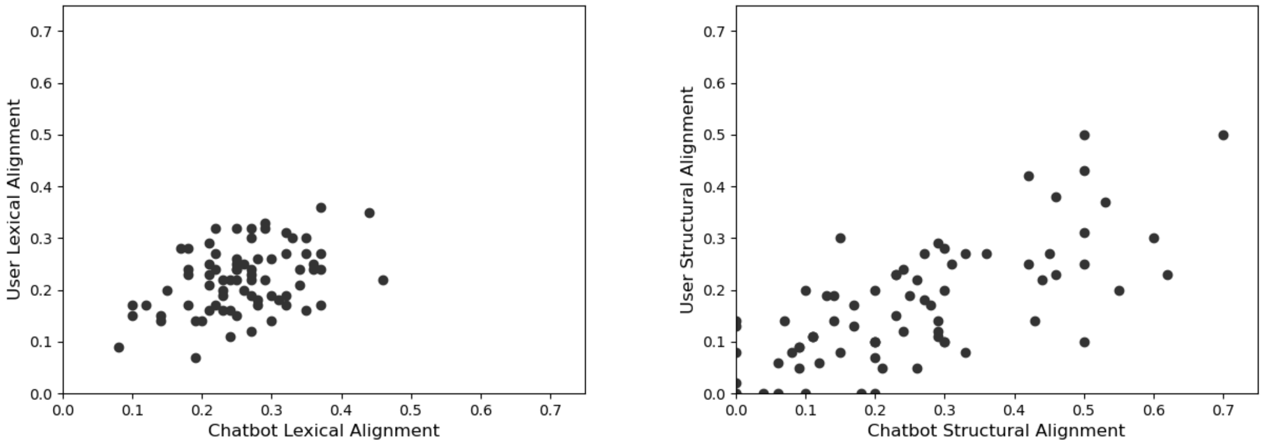
Additionally, using Student's T-test, we compared the alignment measured from users between those that interacted with the chatbot on a mobile device and those that did so on a desktop PC. However, the device used for the study did not have a significant effect on users' lexical alignment or their structural alignment.



(a) Chatbot Alignment

(b) User Alignment

Figure 2: Alignment per chatbot variant for Konrad and the user



(a) Lexical Alignment

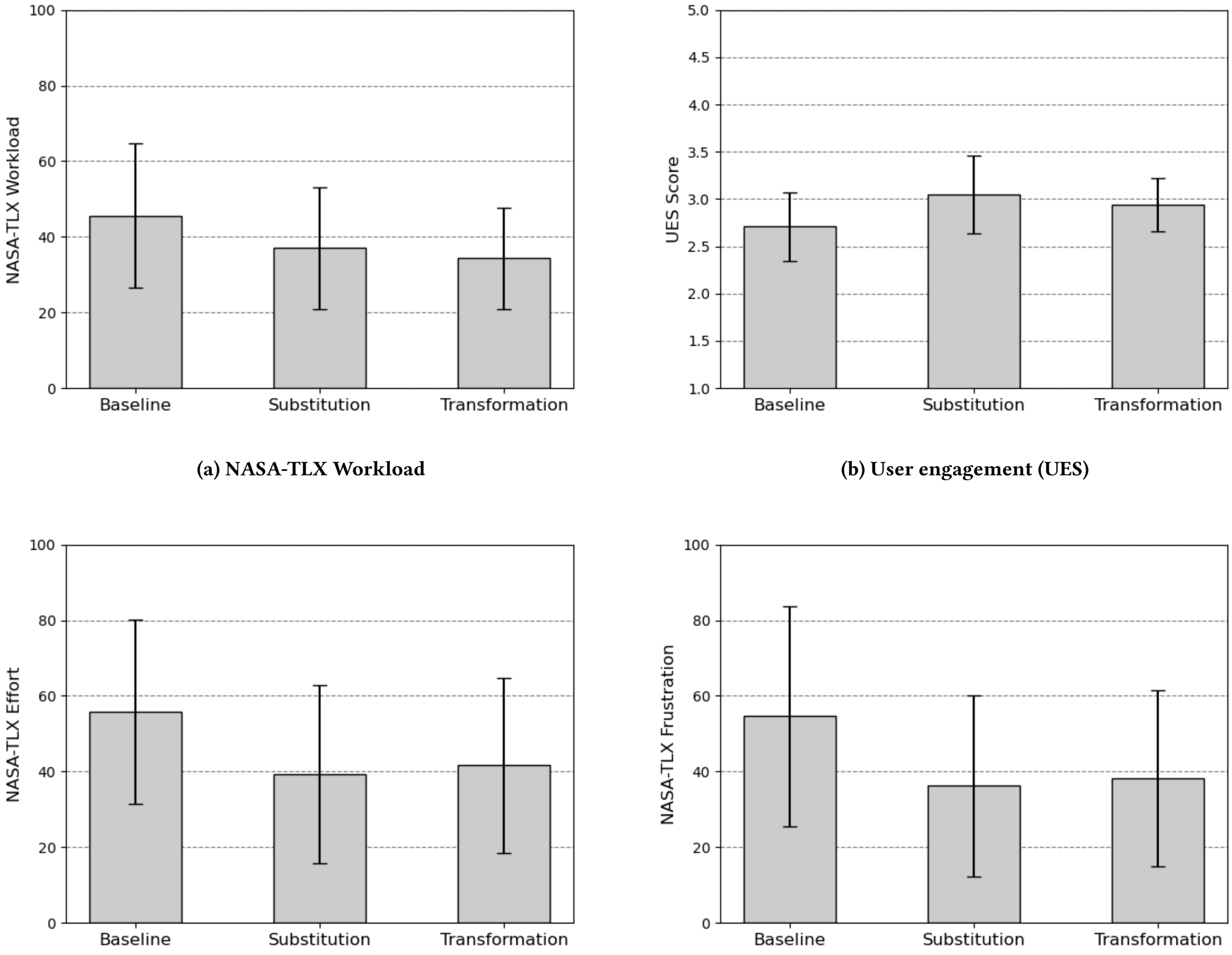
(b) Structural Alignment

Figure 3: Correlation between user alignment and chatbot alignment

5.2 User Perception

We measured the highest mean workload with the NASA-TLX for the baseline condition ($M = 45.64$, $SD = 19.00$), with a decrease both in the substitution alignment condition ($M = 37.04$, $SD = 16.13$) and the transformation alignment condition ($M = 34.36$, $SD = 13.43$), see Figure 4a. A one-way ANOVA revealed that the difference between the three groups is significant ($F_{2,72} = 3.250, p < .05$) A Tukey post-hoc analysis showed a significant difference in the pairwise comparisons between baseline and transformation variants ($p < .05, d = 0.686$), but no difference between baseline and substitution.

For the subscale Effort, a one-way ANOVA also showed a significant difference between the three chatbot variants ($F_{2,72} = 3.617, p < .05$). The baseline condition scored highest and alignment through substitution scored lowest (baseline: $M = 55.80$, $SD = 24.31$; substitution: $M = 39.20$, $SD = 23.35$; transformation: $M = 41.60$, $SD = 23.08$), see Figure 4c. The Tukey post-hoc analysis revealed a significant difference between the baseline and substitution variants ($p < .05, d = 0.696$). Thus, user effort was significantly lower when the chatbot used alignment through substitution than when it did not exhibit any alignment in the baseline variant. Similarly, which chatbot variant was used also significantly affected the Frustration subscale ($F_{2,72} = 3.903, p < .05$). Here,



(a) NASA-TLX Workload

(b) User engagement (UES)

(c) NASA-TLX Effort

(d) NASA-TLX Frustration

Figure 4: Comparison of NASA-TLX and UES results between the three variants of Konrad

the baseline condition again scored highest and substitution scored lowest (baseline: $M = 54.60$, $SD = 29.05$; substitution: $M = 36.20$, $SD = 23.86$; transformation: $M = 38.20$, $SD = 23.36$), see Figure 4d. A Tukey post-hoc analysis showed that user frustration is significantly lower in the substitution alignment variant than in the baseline ($p < .05$, $d = 0.692$), while there is no significant difference between baseline and alignment through transformation.

In analyzing user engagement as measured by the UES score, the Shapiro-Wilk test for the baseline alignment group indicated a non-normal distribution. Since the assumption of normality is not given, we used the non-parametric Kruskal-Wallis test, which reveals a significant difference in UES scores between the three conditions ($p < .05$). The alignment through substitution condition has the best UES engagement score ($M = 3.04$, $SD = .41$), with transformation alignment ($M = 2.94$, $SD = .28$) scoring slightly lower user engagement, and the baseline variant scoring lowest ($M = 2.71$, $SD = .36$) (see Figure 4b). The post-hoc test revealed that the difference between substitution and baseline variants was significant ($p < .005$, $d = 0.860$). There were however no significant differences for the individual subscales of the UES (focused attention, perceived usability, aesthetic appeal, and reward).

Furthermore, there were no notable differences between the three conditions either in terms of users' subjective opinions about the chatbot, or their perception of its domain knowledge.

A correlation analysis was performed to investigate whether there was a direct link between alignment and user workload (mean raw NASA-TLX score) or user engagement (mean UES score). This analysis showed that the users' perceived workload is indeed negatively correlated with the lexical chatbot alignment ($r = -.324$) and the structural chatbot alignment ($r = -.286$). However, we did not find a direct correlation between the measured alignment and user engagement as measured by the UES.

Additionally, we compared the results between mobile users and PC users, however, there was no significant difference.

6 DISCUSSION

With our implementation of Konrad, we wanted to show that we could generate an alignment effect from the chatbot. We hypothesized that a higher amount of repetition of terms and syntactic structures by the chatbot should in turn lead to higher user alignment.

In the substitution variant, we measured higher lexical alignment from the chatbot than in the baseline condition, while it did not notably impact the structural alignment measure. This means that in the substitution variant, the chatbot showed a higher-than-random amount of repetition of lexical choices, while the repetition of sentence structures was not increased compared to the random repetitions in the baseline variant. In contrast, by transforming user messages to generate chatbot answers, we are able to achieve both a stronger lexical alignment effect, and significantly higher structural alignment.

Additionally, users responded to the chatbot exhibiting alignment, supporting our hypothesis: While lexical alignment from users increases only slightly with both chatbot alignment conditions, user structural alignment increases much more strongly in the transformation condition. Moreover, the correlation analysis showed that user alignment is directly correlated with alignment displayed by the chatbot, more strongly so in case of structural alignment.

Figure 3a and Figure 3b show user lexical alignment depending on chatbot lexical alignment and user structural alignment depending on chatbot structural alignment, respectively. The lexical alignment measure appears to be noticeably lower than the structural alignment; however, that is to be expected: while the highest possible score is theoretically 1, in lexical alignment, that would mean that all of that speaker's utterances consist entirely of terms repeated from their respective priming utterances (which would most likely not result in meaningful conversation). For example, the utterance pair "Which movies run on Thursday?"; "These movies run on Thursday:" is scored as only 0.8, and the highest average (over one conversation) lexical alignment score that was achieved is 0.46. In contrast, in structural alignment, an average score of 1 would mean that all of their utterances use the same sentence structure as the priming utterance, which is theoretically possible especially in shorter dialogs. The highest average structural alignment score was 0.7, i.e. 70% of their utterances used the same sentence structure as the priming sentence. Therefore, these values are not directly comparable between the two measures.

More interestingly, we notice that user lexical alignment increases only slightly with higher lexical alignment from the chatbot, while user structural alignment increases at a rate similar to that of chatbot structural alignment. Additionally, the alignment measured for the users is overall lower than that measured for the chatbot. However, that does not mean that Konrad aligned more strongly than the users did: because of the information-seeking task, in which users ask questions and Konrad provides answers, it is to be expected that if there were no alignment at all, the answers (by Konrad) would show a higher proportion of repeated words and repeated sentence structures than the questions (user). Konrad's alignment score in the baseline condition is based entirely on random word and structure repetitions. In contrast, the users' alignment score in the baseline condition can be understood as their baseline alignment to a chatbot that only answers in static template replies.

6.1 Alignment & Workload

Our findings are in line with previous literature showing that humans do in fact align to and can be primed by conversational agents. Moreover, this strongly supports the notion that humans will react socially to computers, by showing that they align more strongly when the chatbot is also displaying alignment. Studies have only revealed a one-sided alignment effect in HCI, in which the human interlocutor adapts their linguistic choices to that of the computer, our results show that they will align more strongly when they have reason to believe that the conversational agent is also aligning to them (as it would happen in human conversation).

Linguistic alignment plays an important part in successful communication between humans, especially when it comes to solving tasks: for example, Reitter and Moore [53] found that task success is correlated with syntactic alignment, and Thomas et al. [58] found both lower perceived workload (as measured by the NASA-TLX) and higher user engagement (as measured by the UES) when dialog partners were more strongly aligned to each other in terms of linguistic style. The results of our user study support these findings for HCI, especially in terms of workload: not only are workload as well as perceived effort and frustration lower in the variants with added alignment, but workload also appears to be directly correlated to alignment, mirroring the results reported by Thomas et al. [58] for human interaction. Moreover, our findings also show a difference in terms of user engagement, which was higher in both aligned variants of the chatbot (with no significant difference between the variants). This shows that participants engaged more strongly with Konrad when the chatbot exhibited an alignment effect. However, we could not show a direct correlation between alignment and engagement as Thomas et al. [58] showed. This indicates that the increased engagement for the two alignment variants might not in fact be because of the alignment effect by the chatbot, but instead happens due to the increased variability in chatbot answers, or the more informal language employed by the chatbot.

In addition to their perception of workload and engagement, we also asked participants to rate the chatbot's domain knowledge and their overall opinion of it (including its friendliness, politeness, how natural the language it used was, etc). However, there were no significant differences between the three versions of Konrad. This indicates that despite the aforementioned difference e.g. in user engagement, users did not actively perceive the chatbot's language use to be different. Interestingly, this is in line with previous research that shows that interlocutors are usually not aware of linguistic alignment, or of its effects on communication [12, 35].

We evaluated alignment depending on whether participants used PCs or mobile devices. There have not, to our knowledge, been any previous studies on how alignment is influenced by usage modality. However, chat apps and intelligent assistants are much more commonly used on mobile devices, and the recent increase in interest in conversational agents is arguably tied to the rise of smartphones as many people's main computer. Because of this, if alignment is important in HCI and in particular in natural language conversational agents, we consider it just as important to investigate how alignment on mobile devices differs from classic, desktop-based studies on alignment in HCI. We hypothesized that users would be more likely to align more strongly when on a mobile device, as the interaction would be more similar to a common chatting situation. However, we only had a small number of participants on mobile devices, and user alignment was similar for both kinds of devices.

6.2 Informal Language & Alignment

In analyzing dialogs between the chatbot and the study participants, we also noticed a number of common phenomena that likely influence both the measured alignment, as well as those users' overall interaction with the chatbot.

Firstly, while we expected there to be a lot of informal language in a chat setting, there were also many messages with spelling mistakes (or presumably intentional misspellings especially in regards to punctuation) and non-standard grammar or spelling. These often lead to mistakes by Konrad, either because words were not recognized, or because the dependency grammar it understood was wrong, leading to grammatically incorrect and sometimes nonsensical replies by the chatbot in the transformation condition. These problems would not occur in a chatbot that is based purely on recognizing key terms and that uses static replies - which underlines how important it is for an AI-based chatbot to have strong capabilities of understanding and correcting non-standard language. Especially in a mobile setting, mistakes and non-standard language are understandably quite common. While alignment through transformation produced stronger alignment from the user than both alternative conditions, as well as scoring the lowest workload with the NASA-TLX, it did sometimes produce nonsensical utterances, which one would want to avoid for a real-world commercial application.

Secondly, we noticed differences in how users adapted their choice of language to that of the chatbot. While some quite obviously aligned to the chatbot over the course of the conversation in a similar way to how alignment happens in human dialog, others were clearly aware that they were speaking with a program, and their messages appeared to be informed by their experiences: Some participants, when confronted with mistakes by the chatbot, reverted to noun phrases or one-word command (e.g. 'Batman actors', 'list of movies', 'schedule'), or even attempted to use commands that are common with other chatbot services (e.g. '\restart' or '\stop'). Similarly, a number of participants did not so much align with the chatbot, but instead when they found a specific phrasing that the chatbot understood would keep using this same phrase (e.g. 'Show me the story of [title]'), even when the chatbot used different terms or sentence structures.

7 CONCLUSION & FUTURE WORK

In this paper we presented Konrad, a chatbot that imitates an alignment effect in natural language conversation. In an online study with 75 participants we were able to show that: Firstly, user alignment was correlated with chatbot alignment, showing that users did in fact react to the alignment effect displayed by Konrad. Secondly, user workload was lower and user engagement was higher in the variants of the chatbot with added alignment, and task workload was directly negatively correlated with alignment. The same has been shown in previous work for human conversation [58] - our results show that this connection between alignment and task workload applies to human-computer-communication as well.

Furthermore, we had hypothesized that linguistic alignment would play an even more important role in conversational agents on mobile devices. In future studies we want to further investigate the differences in alignment depending on device and usage modality (including e.g. chatbot vs. voice agent). We tested Konrad in the film domain, because users are familiar with this domain and there is not a huge influence of prior knowledge. However, alignment might differ depending on the domain and user motivation (e. g. solving a task compared to engaging in chit-chat conversation), which we also plan to investigate in future studies. Our implementation based on word similarities has the advantage that this is possible for several (common) domains. Additionally, the length of sentences and conversation might have an influence on the alignment. Therefore, we expect that additional strategies and implementations of alignment will be required for such conversations.

Our results show that the phenomenon of linguistic alignment should be taken into consideration when developing natural language HCI interfaces. Alignment plays an important role in achieving successful communication between humans, and implementing it in conversational agents can clearly have real-world benefits for user interaction.

ACKNOWLEDGMENTS

The research reported in this paper has been supported by the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU Program Horizon 2020.

REFERENCES

[1] Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189.

[2] Janet B Bavelas, Alex Black, Charles R Lemery, and Jennifer Mullett. 1986. " I show how you feel": Motor mimicry as a communicative act. *Journal of personality and social psychology* 50, 2 (1986), 322.

[3] Linda Bell, Joakim Gustafson, and Mattias Heldner. 2003. Prosodic adaptation in human-computer interaction. In *Proceedings of ICPHS*, Vol. 3. Citeseer, 833–836.

[4] Kirsten Bergmann, Holly P. Branigan, and Stefan Kopp. 2015. Exploring the Alignment Space – Lexical and Gestural Alignment with Real and Virtual Humans. *Frontiers in ICT* 2 (2015), 7.

[5] Frank J Bernieri and Robert Rosenthal. 1991. Interpersonal coordination: Behavior matching and interactional synchrony. (1991).

[6] Lars Borin and Jens Edlund. 2018. Language Technology and 3rd Wave HCI: Towards Phatic Communication and Situated Interaction. In *New Directions in Third Wave Human-Computer Interaction: Volume 1-Technologies*. Springer, 251–264.

[7] James J Bradac, Anthony Mulac, and Ann House. 1988. Lexical diversity and magnitude of convergent versus divergent style shifting-: perceptual and evaluative consequences. *Language & Communication* 8, 3-4 (1988), 213–228.

[8] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. *Interactions* 25, 5 (2018), 38–43.

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

[10] Holly P Branigan, Martin J Pickering, and Janet F McLean. 2005. Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 3 (2005), 468.

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

[12] Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of pragmatics* 42, 9 (2010), 2355–2368.

[13] Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, and Clifford Nass. 2003. Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th annual conference of the cognitive science society*. Lawrence Erlbaum Associates, 186–191.

[14] Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, Clifford I Nass, and John Hu. 2004. Beliefs about mental states in lexical and syntactic alignment: Evidence from human-computer dialogs. In *Proceedings of the CUNY Conference on Human Sentence Processing*. University of Maryland College Park, MD.

[15] Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 6 (1996), 1482.

[16] Heloisa Candello and Claudio Pinhanez. 2016. Designing Conversational Interfaces. *Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais-IHC* (2016).

[17] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[18] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems* 92 (2019), 539–548.

[19] Alexandra A. Cleland and Martin J. Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language* 49, 2 (2003), 214 – 230.

[20] Gabriel Doyle and Michael C Frank. 2016. Investigating the sources of linguistic alignment in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 526–536.

[21] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42.

[22] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM* 30, 11 (1987), 964–971.

[23] Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological science* 23, 8 (2012), 931–939.

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

[25] Aimi Shazwani Ghazali, Jaap Ham, Emilia Barakova, and Panos Markopoulos. 2018. The influence of social cues in persuasive social robots on psychological reactance and compliance. *Computers in Human Behavior* 87 (2018), 58–65.

[26] Howard Giles and Peter Powesland. 1997. Accommodation theory. In *Sociolinguistics*. Springer, 232–239.

[27] Ulrich Gnewuch, Stefan Morana, Carl Heckmann, and Alexander Maedche. 2018. Designing conversational agents for energy feedback. In *International Conference on Design Science Research in Information Systems and Technology*. Springer, 18–33.

[28] Eun Go and S Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97 (2019), 304–316.

[29] Stefan Th Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research* 34, 4 (2005), 365–399.

[30] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[31] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139 – 183. https://doi.org/10.1016/S0166-4115(08)62386-9

[32] Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological science* 15, 6 (2004), 409–414.

[33] Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior* 49 (2015), 245–250.

[34] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An End-to-End Conversational Style Matching Agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 111–118.

[35] Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand. 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior* 27, 3 (2003), 145–162.

[36] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. *"I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376175

[37] RG Leiser. 1989. Exploiting convergence to improve natural language understanding. *Interacting with Computers* 1, 3 (1989), 284–298.

[38] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. 2017. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 275–286.

[39] Tong Li, Mingyang Zhang, Hancheng Cao, Yong Li, Sasu Tarkoma, and Pan Hui. 2020. "What Apps Did You Use?": Understanding the Long-Term Evolution of Mobile App Usage. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 66–76. https://doi.org/10.1145/3366423.3380095

[40] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[41] Richard E Maurer and Jeffrey H Tindall. 1983. Effect of postural congruence on client's perception of counselor empathy. *Journal of counseling psychology* 30, 2 (1983), 158.

[42] Charles Metzing and Susan E Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language* 49, 2 (2003), 201–213.

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).

[44] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

[45] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[46] Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. 2008. Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling. In *Ninth Annual Conference of the International Speech Communication Association*.

[47] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28 – 39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[48] Jennifer S Pardo, Rachel Gibbons, Alexandra Suppes, and Robert M Krauss. 2012. Phonetic convergence in college roommates. *Journal of Phonetics* 40, 1 (2012), 190–197.

[49] Jamie Pearson, Jiang Hu, Holly P. Branigan, Martin J. Pickering, and Clifford I. Nass. 2006. Adaptive Language Behavior in HCI: How Expectations and Beliefs about a System Affect Users' Word Choice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1177–1180.

[50] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[51] Martin J Pickering and Victor S Ferreira. 2008. Structural priming: A critical review. *Psychological bulletin* 134, 3 (2008), 427.

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

[53] David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language* 76 (2014), 29 – 46.

[54] Caroline F Rowland, Franklin Chang, Ben Ambridge, Julian M Pine, and Elena VM Lieven. 2012. The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition* 125, 1 (2012), 49–63.

[55] Arabella Sinclair, Adam Lopez, Christopher G Lucas, and Dragan Gasevic. 2018. Does ability affect alignment in second language tutorial dialogue?. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*. 41–50.

[56] Arabella Sinclair, Kate McCurdy, Christopher G Lucas, Adam Lopez, and Dragan Gaševic. 2019. Tutorbot Corpus: Evidence of Human-Agent Verbal Alignment in Second Language Learner Dialogues. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*. ERIC, 414–419.

[57] N. Suzuki and Y. Katagiri. 2007. Prosodic alignment in human–computer interaction. *Connection Science* 19, 2 (2007), 131–141.

[58] Paul Thomas, Mary Czerwinski, Daniel McDuff, Nick Craswell, and Gloria Mark. 2018. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 42–51.

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

[60] Peter Wallis and Emma Norling. 2005. The Trouble with Chatbots: social skills in a social world. *Virtual Social Agents* 29 (2005).