Could you tell a bit more about your experience with the chatbot?

The effect of Conversational Human Voice on User Experience

Eefje de Louw
SNR: 2014295

Master's Thesis
Communication and Information Sciences
Specialization Cognitive Science and Artificial Intelligence (CSAI)

Tilburg School of Humanities and Digital Sciences
Tilburg University, Tilburg

Supervisors:
dr. S. Wubben
MSc L. Bastiaansen

First reader: A.T. Hendrickson

Second reader: dr. M. Atzmüller

January 2019

Abstract

This research examines the effect of Conversational Human Voice on User Experience in chatbots in the context of survey research performed by the municipality of 's-Hertogenbosch (The Netherlands). Literature of suggests that Conversational Human Voice has a positive effect on User Experience, but that never has been tested. Therefore, this study proposes the following first research question: What is the effect of Conversational Human Voice on the User Experience? In the research of Human-Computer Interaction, it is known that humans are likely to attribute human characteristics to the computer when they interact with them and show similarities to Human-Human interaction. It is said that people tend to adapt their language use to that of their conversational agent. Therefore, this study proposes the second research question: To what extent do users alter their language use to the addition of elements of Conversational Human Voice of the chatbot? The results of 551 participants were analysed and the following conclusions are drawn: Conversational Human Voice does not necessarily lead to a higher User Experience score in the context of survey research. For the context of survey research, one can rather stick to the functional communication styles. Compared to the other Conversational Human Voice categories, inviting rhetoric seems most suitable in the context of survey research. Significant differences were found for the "helpfulness" and "clearness" of the chatbots. All in all, chatbots do show potential, but mainly in more bound contexts such as answering frequently asked questions or managing appointment.

Table of Contents

1. Introduction

We contribute to research in the domain of Human-Computer Interaction by performing a (real-life) experiment on the use of Conversational Human Voice in chatbots in the context of survey research performed by the municipality of 's-Hertogenbosch (The Netherlands). Second, we investigate what elements of Conversational Human Voice determine a successful conversation between the citizens and the municipality of 's-Hertogenbosch, where we measure the effect of Conversational Human Voice on User Experience.

Policymakers have always been interested in measuring citizen's satisfaction and experiences. However, classical questionnaires provide little or no room for the spontaneous opinions of respondents. Moreover, it becomes a more costly process to recruit a representative sample of respondents (Ceron & Negri, 2016). This is not only a problem spotted at the municipality of 's-Hertogenbosch, but in research worldwide (National Research Council, 2013).

From earlier research on Human-Computer Interaction, extended interest has arisen in variables, for example Conversational Human Voice (CHV), that determine the effectiveness of the interaction between a human and computer and how chatbots influence the perceptions of the brands in real-life settings (Barcelos, Dantas, & Sénécal, 2018; Liao et al., 2018; Zarouali, Van den Broeck, Walrave, & Poels, 2018). Furthermore, there is still lack of understanding how tone of voice affects the User Experience and previous research provides no consensus among companies what the most appropriate tone of voice entails (Barcelos, Dantas, & Sénécal, 2018). Therefore, this study proposes the following first research question: What is the effect of Conversational Human Voice on the User Experience? (RQ1).

Similarly, in Human-Computer interaction, it is yet unclear how users will converse with chatbots in real life settings. It is said that people tend to adapt their language use to that of their conversational agent. However, participants tend to show more alignment when the agent is perceived more humanly. Therefore, this study proposes the second research question: To what extent do users alter their language use to the addition of elements of Conversational Human Voice of the chatbot? (RQ2).

2. Background

2.1 Conversational agents and dialogue systems

Conversational agents are dialogue systems that imitate human interactions so that they can react accordingly to textual input (Shawar & Atwell, 2007). Research towards chatbots started around 1950 (Shadbolt, Smith, Simperl, Van Kleek, Yang, & Hall, 2013; Coniam, 2014); the main objective in that period was to investigate if it would be possible to convince humans into believing that they are talking to a real person instead of a machine. This started with Alan Turing (1950) who invented the Turing test. This will be further explained in chapter 2.5 'Evaluation of chatbots'.

Since 1980 the interest in chatbots grew rapidly. This had two reasons. First, the amount of data that became available with the increase of mobile internet users, helped drive the adoption of the chatbots and secondly, advances in Artificial Intelligence led to better capabilities of Natural Language Processing, such as maintaining conversations and making chatbots more adaptive to different input styles and tasks (Brandtzaeg & Følstad, 2017).

Nowadays, there is a shift in functionality of interfaces. User Interfaces are moving from mouse-based towards conversation-based and chatbots are seen as a promising alternative to traditional customer service and assistants. For customers, conversations with these bots may feel more natural and efficient than interacting with a mobile app (Brandtzaeg & Følstad, 2017). Moreover, chatbots can increase user engagement, system usability, User Experience and interaction quality (Bergmann, Branigan & Kopp, 2015; Luger & Sellen, 2016; Radziwill & Benton, 2017).

Brandtzaeg and Følstad (2017) researched the motivation behind the usage of chatbots. Productivity and time-saving was named most for why people tend to interact with chatbots. The ease of use, speed, and convenience are named as the main arguments. Other motivations included entertainment and curiosity. However, differences in characteristics of users tend to affect the ratings of the chatbots conversational quality. More specifically, younger users and female users rated the conversations more favourably (Brandtzaeg & Følstad, 2017). Luger & Sellen (2016) named the same motivations, but also found out that the 'hands-free'

component was very convenient. Later in this study, the motivations for using chatbots will provide context to the results that were found in this study.

Conversational agents can be classified as embodied and disembodied conversational agents (Araujo, 2018; Luger & Sellen, 2016; Radziwill & Benton, 2017). Embodied conversational agents (ECAs) have a (virtual) body or face, usually human-like. This enables the bot to use non-verbal cues such as facial expressions, gaze and gestures) in the interactions they have with users. Disembodied agents (or text-based conversational agents), often referred to as chatbots (chatterbot/chatterbox) or natural dialogue systems, communicate with users primarily via a text message. This interface also allows other types of media (including images, cue cards, and videos). However, this chatbot has no dynamic physical representation of the agent (Araujo, 2018). It is, therefore, logical that most research on emphatic bots is regularly performed with embodied agents. However, research of Ciechanowski, Przegalinska, Magnuski, & Gloor (2018) found that their TEXT chatbot was rated more positively compared to their AVATAR chatbot. Their explanation: the uncanny valley effect is higher with embodied agents than with disembodied agents. So until researchers or designers are able to work around this effect by designing a lifelike appearance, disembodied agents might still have the preference and therefore this study focuses on disembodied agents.

2.2 Human-computer interaction (HCI)

With the emergence of the growing interests in Human-Computer interaction, the interest in conversational agents grew similarly. Researchers have come to a consensus that humans are likely to attribute human characteristics to the computer when they interact with them (known as 'Ethopoeia') and show similarities to Human-Human interaction. The effect of Ethopoeia is mainly unconscious and likely due to the social nature of humans (Heyselaar, Hagoort, & Segaert, 2017). Alignment is considered a central aspect of succesful communication and as sign of group cohesion (Hasson & Frith, 2016; Pickering & Garrod, 2006). Alignment is essentially a form of imitation. This commonly occurs when two agents interact in a face-to-face conversation and they mimic each other's gestures or facial expressions. However, alignment can occur at many levels. This is especially notable in conversations where people align their speech rate, their choice of words and ultimately their semantic concepts (Hasson & Frith, 2016).

To date, only a few controlled experiments have been conducted to directly examine the interaction between humans and chatbots (Zarouali, Van den Broeck, Walrave, & Poels, 2018). Research of Hill, Ford and Farreras (2015) found that users who communicated with chatbots were likely to send messages that contained fewer words than messages that were send to a human, but surprisingly, people were inclined to send more than twice the number of messages to chatbots than to humans. This was interpreted as a sign of alignment. Furthermore, they showed that humans who converse with computers often use a less rich vocabulary compared to conversations with other humans and messages contained more profane language compared to the messages that were sent to other people. This effect was probably due to lack of social feedback (Hill, Ford, & Farreras, 2015).

Research of Zarouali, Van den Broeck, Walrave, & Poels (2018) showed that for chatbots that were perceived as human-like, users were more likely to adjust for misunderstandings compared to chatbots that were perceived as machine-like. Research of Liao et al. (2018) found out that despite the question-and-answer set-up of their experiment, users tend to interact spontaneously with their chatbot. Likewise, users were actively involved in giving feedback to the chatbot to help designers eventually compensate for the errors. However, these results differed when the users received the chatbot solely as operating system instead of an anthropomorphic bot. These results are in line with the research of Lee (2010) that states that fostering a personal relationship will be stronger when the chatbot is more human-like. The meta-analysis on this study indeed showed that human-like agents with 'higher realism' result in more positive social interaction, especially when subjective evaluations were employed (Verhagen, Van Nes, Feldberg, & Van Dolen, 2014).

2.3 Tone of voice vs. Conversational Human Voice

Chatbot research in the domain of communication adheres two distinct communication styles. Social oriented and task oriented. Social oriented aims to establish personal relationships with customers and tries to satisfy the customer's emotional needs by personalizing the interaction, whereas task-oriented aims for efficient goal fulfilment with minimized costs, efforts and duration (Dion & Notarantonio, 1992; Luger & Sellen, 2016). The

context of this research is to use chatbots as an alternative to traditional surveys, whereas the communication is likely to be task oriented. However, it is imaginable to start a conversation with some small talk to create a common ground between the chatbot and the participant. The design of the chatbots will be further discussed in chapter 3.3 'Design'.

In the domain of webcare, this is an important concept to grasp because ever since the uprising of Social Media users regularly turn towards social media channels for questions, complaints and freely share their experiences. With these recent advances online service encounters are becoming critical in determining the success of a firm (Verhagen, Van Nes, Feldberg, & Van Dolen, 2014) and to address this issue, many organizations form large customer service teams that respond to the incoming requests. However, this is a time-consuming process and often fails to maintain user's expectations for a quick reply. While chatbots could offer new opportunities in offering users more individual attention and inviting users to keep interacting with the brand, leading to a better user brand experience (Gnewuch, Morana, & Maedche, 2017; Barcelos, Dantas, & Sénécal, 2018). However, the use of chatbots is not without obstacles. Where chatbots are predominantly used for functionality, 40% of user requests to customer service are rather emotional, whereby the user might not even be seeking for a specific solution. Moreover, sharing emotions with public is considered as one of the main motivations for using social media (Xu, Liu, Guo, Sinha, & Akkiraju, 2017; Brandtzaeg & Følstad, 2017).

Similarly, the tone of voice of a firm is also important for the successfulness of online encounters. Tone of voice is commercially used to refer to the language styles or registers that a company uses to express a distinctive personality or set of values that differentiates its brand from those of the competition. The goal is to engage people in the interaction – not solely about products and services they can buy, but also advice about benefits and services that they are invited to take part in or claim (Barcelos, Dantas, & Sénécal, 2018).

Conversational Human Voice, on the other hand, refers to a tone of voice that makes the company or brand feel closer, more real and human. This communication style is characterized as being open to dialog, welcoming conversational communication, providing prompt feedback, communicating with a sense of humour, and admitting mistakes (Kelleher, 2009). Van Hooijdonk & Liebrecht (2018) characterize three different categories of CHV, and this research will continue to use the same terminology:

1. Personalized approach: signatures, personal greetings and personal addressing
2. Informal language use: shortenings, abbreviations, non-verbal cues and interjections
3. Inviting rhetoric: encouraging dialogue, thanking, excusing, and showing sympathy, empathy and humour.

That being said, there is still no consensus among researchers to what the most appropriate tone of voice entails (Barcelos, Dantas, & Sénécal, 2018). This research will experiment with Conversational Human Voice in the context of survey research performed by the municipality of 's-Hertogenbosch and will examine what category of Conversational Human Voice leads to the most successful conversation.

2.4 Architecture of chatbots

When a chatbot obtains user input, it needs to predict the intent of the user (intent recognition) and obtain the information needed to perform the action the user wants to perform (entity extraction). Based on the intent prediction the chatbot chooses what the best action is to undertake. It matters for example whether the user has a follow-up question or the user wants to switch to a different topic. Lastly, the chatbot could look up the information needed by connecting to an API that contains a database with the right information, or it could answer a question with a pre-defined statement.

There are several ways for dealing with these handlings. First of all, there are rule-based chatbots that depend on Artificial Intelligence Markup Language (AIML). AIML is a recursive language that parses a single text input to a pattern that can match to a set of pre-programmed responses (Denecke, Lutz Hochreutener, Pöpel, & May, 2018). If this chatbot cannot find an exact match from input to the predefined patterns (by means of a typo or because the surface form differs from the examples in the trainingset) the chatbot will fail to complete the task.

Secondly, classification-based chatbots are trained on a pre-defined set of questions and answers. It chooses the most probable answer as its output. This chatbot takes the context into account, but cannot produce new output.

A final variant is the more advanced generative chatbot that contains a sequence-to-sequence recurrent neural network algorithm. This means that the chatbot learns conversational rules itself and generates output

autonomously and from scratch. This output is not only dependent from the current input, but also the inputs previous to that. In the next section, the most central elements to chatbots such as text classification, intent prediction and NER will be shortly discussed. Figure 1 shows an example architecture of a chatbot.

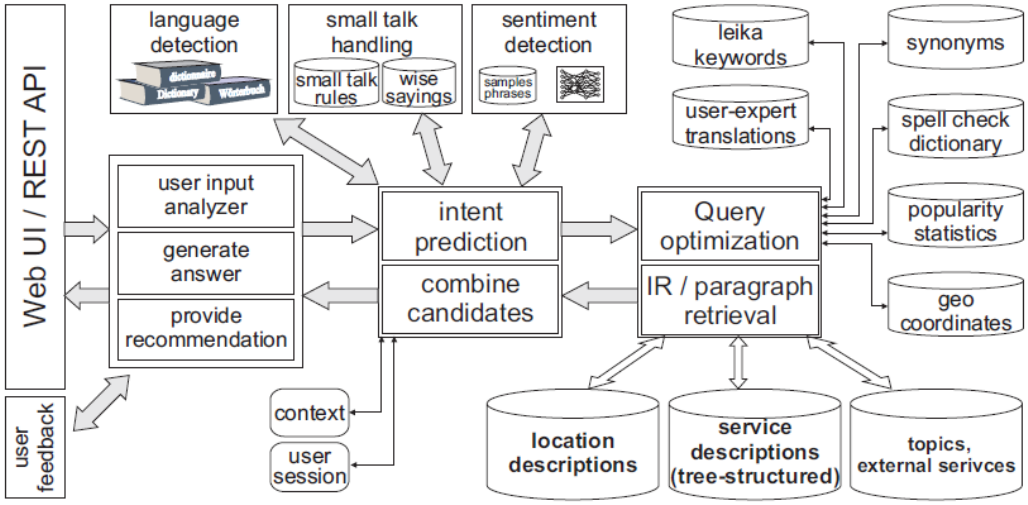


Figure 1. Example Architecture. From: A Next Generation Chatbot-Framework for the Public Administration (Lommatzsch, 2018).

For the intent prediction, the textual input of the user would be divided into classes (intent classification). Intent classification is done via the content of the previous inputs, the context and the current input. These classes can consist of: languages (to detect the input language), locations (such as postal codes, GPS-coordinates or geographical names), frequently asked questions, indicators for follow-up questions (related to services or opening hours or locations) and/or sentiment (Lommatzsch, 2018). Text classifying algorithms that are used for intent classification also need to understand the basics of semantics. Computational semantics often depends on the concept of distributional semantics, which states that similar words occur in similar contexts (also called: co-occurrence). Word similarity is represented in word embeddings via word vectors or a co-occurrence matrix. This matrix consists of columns (terms) and rows (documents) and shows the term frequencies of these combinations. Words are then compared by calculating the Euclidean distance measure.

Sequence labelling is another way of classifying texts. It is, more specifically, a pattern recognition task that involves the labelling of each element of an observed sequence of values. For example, part of speech tagging (POS) assigns grammatical characteristics to each word in an input sentence or document.

For entity extraction, the chatbot depends on entities. Entities represent certain terms in the user's input that provides clarification for the intent of the user. In the case that the chatbot misses a required entity for the intent of the user, the chatbot can ask a follow-up question such as "where would you like the opening hours for?" Named Entity Recognition (NER) is a classifying text algorithm that deals with entity extraction.

NER is a task that seeks to locate words in texts and assign them pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. One strategy to perform NER is the Hidden Markov Model (HMM). This is a statistical model that can predict the next sequence based on the previous sequence, such that it predicts a vector of random variables, given an observed vector. HMM's are typically trained to maximize the joint likelihood of training examples. Another algorithm for NER is Conditional Random Fields (CRF). CRF are used to encode known relationships between observations and construct consistent interpretations (Lafferty, McCallum, & Pereira, 2001). They are similar to HMM in that they can predict sequences of labels for sequences of input samples, but the difference lies in the assumption of CRF that the distinctive elements in fact do affect each other (i.e. CRF takes context into account). In the field of language for example, this assumption might probably work better than HMM.

For chatbots, NER is important for the extraction of entities. Entities represent certain terms in the user's input that provides clarification for the intent of the user. In the case that the chatbot misses a required entity for the intent of the user, the chatbot can ask a follow-up question such as "where would you like the opening hours for?"

2.5 Evaluation of chatbots

For a long time, the evaluation measure of success in chatbots was the Turing test (Turing, 1950). This started as a hypothetical case where a participant talked to a man, woman or a machine and to pass this test, participants should be fooled into believing that they are talking to a human instead of a machine. Similarly, the Loebner Prize (1991) uses this test for their annual competition in the most human-like bot and has thereby driven the natural language user interfaces to become more human-like (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2018; Coniam, 2014).

Since then, several studies have investigated other ways for evaluating chatbots. For example, it was measured how users perceive the personality of a chatbot, responses given by users with an anthropomorphism chatbot, the conversational abilities of a chatbot etc. (Brandtzaeg & Følstad, 2017). Other ways contain: measuring the robustness to manipulation, the accuracy of interpreting commands, the ability to maintain a themed discussion, sensitivity to social concerns or the ability to detect meaning or the right intents. All in all, there are lots of ways to evaluate the abilities of a chatbot and the literature requests for one standard metric.
The research of Radziwill & Benton (2017) reviewed literature since 1990 to 2015 and extracted six quality attributes of evaluating chatbots on: performance, functionality, humanity, affect, ethics & behaviour and accessibility.

The research of Jadeja and Varia (2017) addresses the difficulty in the identification of successful conversational agents and therefore propose their evaluation metric, which consists of four dimensions: the User Perspective, Information Retrieval perspective, Linguistic perspective and Artificial Intelligence perspective. Conversational AI is designed for human-like interaction with end users, but it seems very difficult to achieve this objective practically. Key features of the User Perspective include: usability, level of user satisfaction etc. The Information Retrieval perspective shows the accuracy of information and how quickly the system reacts to the users' query. However, a good information retrieval quality is not only based on the right information, the chatbot also needs to apply the right parameters for personal preferences (for example, item X with the lowest price and the best return policy). The Linguistic Perspective is based on Grice's cooperative maxims (Quality, Quantity, Relation and Manner). Quality adheres to the fact that whatever is said by the speaker must be the truth and can be proved with the help of evidence. Quantity is explained as: the amount of information shared by the speaker must be depending upon the requirement. The speaker must not share too much or too less information. Relation states that the response must always be related to the discussion topic. And lastly, Manner claims that the overall interaction must be direct as well as straightforward and therefore no ambiguity or obscurity is allowed. In this way, a good conversational agent would have a higher degree of support towards Grice's conversational maxims. The Artificial Intelligence Perspective is probably still the most difficult since there is no direct alternative to the Turing test. While the main drawback of this test is that it will not provide guidelines for the overall improvement of the conversational AI.

Research on User Experience, focusing on the user perspective of A.I., can be evaluated in another way. This is done by means of the User Experience Questionnaire (UEQ, Laugwitz, Held, & Schrepp, 2008). It consists of 26 items and takes approximately 3-5 minutes to fill in. In 2017, Schrepp, Hinderks and Thomaschewskie created a shortened version, specially adjusted for the (repeated) measuring of user satisfaction of new applications/ interfaces. This version consists of eight questions based on a 7-point Likert scale with a mean Cronbach's alpha of 0.83, indicating that the different items reliably measure the User Experience. In the current research, User Experience is evaluated based on this short version of the UEQ.

Automatic evaluation

Sentiment analysis can be used as an automatic evaluation metric in unstructured textual data because the algorithm classifies the text in values between -1 (negative) and (+1) positive by comparing each word in the text to a corpus of labelled sentiments. This can be done on the whole document, the paragraph level or at the sentence level. It is possible that the sentiment analysis results with low accuracy because of the following drawbacks: by being fully dependent on a library that classifies words into 'negative' and 'positive', it often fails to detect irony and the numerous nuances that language includes (Ceron & Negri, 2016). Therefore, it can still be best to do the sentiment analysis in complement of another less subjective metric. This study uses sentiment analysis accompanied by the User Experience scores and evaluates the metric by comparing the obtained results to the results of the mean User Experience score.

2.6 E-governance and the application of chatbots

During the past 20 years, electronic Government (eGovernment) has become a political priority (Rijksoverheid, 2018; Steunenberg, 2018). E-governance is primarily characterized as the need for the government to provide, first of all, open government/public administration, which states that citizens have the right to gain access to documents and proceedings of the government with open data. Enabling interested citizens to get directly involved in the jurisdictive process, policymaking (top-down vs. bottom-up) and to encourage transparency and accountability. A second aspect of E-governance is to facilitate online services for citizens (such as appointment making for the retrieval of personal documents or creating notifications) and the third aspect is to decrease the gap between government and citizens by being more open to dialogue. The final task is to introduce New Public Management that states that public service should be more business related and thus improving in efficiency and measurability (Rijksoverheid, 2018; Steunenberg, 2018).

Ceron and Negri (2016) state that in e-governance, the role of Social Media is overlooked. Governments mostly use Social Media as a merely passive way to inform their citizens about new information. However, the chances really lie within the possibility to interact and engage in a conversation with the citizens to enlarge transparency and positively affect trust. They propose using Sentiment Analysis to measure continuously and in 'real-time' citizens reactions of policymaking decisions.

Little research has been done on the implementation of chatbots in governmental contexts. Porreca, Leotta, Mecella, & Catarci (2017) have investigated the possibility of combining a chatbot to open data. For example, citizens could now ask a question to the chatbot and be provided with an answer from the open database as a more user-friendly design. However, this continues to be mainly a prototype, only to evaluate the technical feasibility. They conclude that future work should validate this approach.

Lommatzsch (2018) is the first that provides us with a quite successful framework for a chatbot for question-answering regarding public administration services. Users get specific information related to costs, opening hours, and required documents. The chatbot is online on Berlin's central Service Portal and is evaluated with user feedback. Future works include detecting seasonal trends and Named Entity disambiguation by taking into account the context, to ensure a higher question-answering precision.

The biggest concern of e-governance is the digital divide where the accessibility to internet technologies for the average citizen plays a role (Porreca, Leotta, Mecella, & Catarci, 2017). With this in mind, governments could choose for multi-modal services in which people that are not computational proficient can still conduct their service needs offline. Often this is named as one of the reasons that open government is not yet fully developed. User-centric design could also help solve this problem by first asking potential users what they would like in a service and letting users test the service and provide feedback (Porreca, Leotta, Mecella, & Catarci, 2017).

Present study

Based on the literature, it can be concluded that there is still lack of understanding of how the Conversational Human Voice affects User Experience and how users will converse with chatbots in the context of survey research conducted by the municipality of 's-Hertogenbosch. By recruiting members of the municipality of 's-Hertogenbosch' Digipanel, we obtain a more representative sample than can be done in the recruitment of students on campus in terms of sample size and mean age. Moreover, we achieve a realistic setting since members of the Digipanel are used to filling in surveys for the municipality of 's-Hertogenbosch. It has not yet been measured what happens to User Experience and the language use when a traditional survey is conducted by a chatbot. Therefore, we conduct an experimental study where the following two questions will be researched:

RQ1: What is the effect of Conversational Human Voice on the User Experience?

RQ2: To what extent do users alter their language use to the addition of elements of Conversational Human Voice of the chatbot?

The following three sub questions link to the research questions and will be further investigated in this study: *SQ1*: What is the influence of age? *SQ2*: What is the influence of gender? *SQ3*: Which CHV category is most predictive for a positive UX-score?

Research has shown that the right tone of voice can enhance the user brand experience. Conversational Human Voice is chosen as the communication style rather than Tone of Voice because of the somewhat more corporal status of Tone of Voice. Therefore, the first hypothesis is presented.

H1: The chatbots that use elements of Conversational Human Voice will obtain a higher User Experience score compared to the chatbot that does not use Conversational Human Voice.

Studies examining Human-Computer Interaction have found that when participants experience a chatbot as "human" and therefore trust the chatbot, they write longer messages and show more alignment. The chatbots that use Conversational Human Voice are expected to be experienced as more human. Therefore, the second hypothesis is formed.

H2: The chatbots that use elements of Conversational Human Voice will lead the users to write more words per messages (a), have longer conversation durations (b), express more positive sentiment (c) and use more elements of Conversational Human Voice in their answers (d) compared to the users that interact with the chatbot that does not use Conversational Human Voice.

3. Method

This section describes the setup for the experiment that is conducted in order to answer the research question. An experimental study was executed in which participants evaluated their experience while a chatbot serves as alternative to a conventional survey. A between-subject design was used with the amount of Conversational Human Voice in the chatbot as independent variable. User Experience (with seven questions on a 5-point Likert scale) was used as general dependent variable. Additionally, the success rate, duration of conversation, length of messages, sentiment and alignment of CHV-elements served as dependent variables for the underlying process.

3.1 Participants

The participants were recruited through Digipanel. This is an online panel of active citizens (N = 8,000) of the municipality of 's-Hertogenbosch that are used to fill in surveys for six to eight times a year. For this research, we sent invitations to a total of 1138 participants that participated in a survey of the municipality of 's-Hertogenbosch about environmental stations and had indicated that they would be interested in a follow-up experiment where they would test a chatbot as alternative to a traditional survey about sustainability.

Participants were randomly assigned to either one of the five conditions, which let to blocks of ≈ 227 participants. Invitations including a link with either one of the chatbots were sent in blocks of 80 people every half hour, to maintain a stable flow of new users and guarantee that there was no overload of the server. Moreover, enabling the municipality to answer incoming questions/addressing problems. We analysed the results of total of 551 participants (327 men and 221 women, all citizens of the municipality of 's-Hertogenbosch). All participants were recruited via Digipanel and all participants voluntarily took part without any reimbursement. Only age and gender were recorded as characteristics of the participants.

3.2 Apparatus

Nowadays, it is possible for people to build chatbots without in-depth knowledge of programming, because there exist multiple platforms for building chatbots. By means of a drag-and-drop interface it becomes quite intuitive to put together a conversational flow and intent recognition. For example, the user interface can provide the user with the option to list pre-defined responses the chatbot can produce given a certain input. This is called classification-based, as explained in chapter 2.4 'Architecture of chatbots'. It might also be possible to list multiple sentences that describe the same intent for intent recognition. No machine learning expertise is required. The chatbots in this experiment were built in the Online conversational AI platform Flow.ai and trained on answers of previous versions of surveys on sustainability and input from several forums and social media that discuss the subject of sustainability.

Since the online invitation included a link to a web browser, participants could carry out the experiments on any device of their choosing. The invitations were sent via NetCollector. Appendix 1 shows the exact text of the invitation. Appendix 2 shows some examples of the user interface of the platform.

The User Experience score was based on seven 5-point Likert scale statements, that was based on the User Experience Questionnaire Short (Schrepp, Hinderks, & Thomaschewski, 2017). Table 1 shows the exact statements that were used to measure the User Experience. The Questionnaire's validity is proven by the Cronbach's alpha test (α = 0.88). The reason behind the 5-point Likert scale was simply because the chat window was not large enough to enable 7-point Likert scales.

For the analysis of the various text analysis measures, the packages "nltk" and "sentiment", "statistics", "pandas", "string", "gensim" from Python were used. For the visualizations and statistical analysis in this research, the following R-packages were used: "ggplot2", "tidyverse", "dplyr", "stats", "psych", "qqplotr", "lawstat", "magrittr", "FSA", "Lattice", "rcompanion", "multcompView" and "stringr".

Table 1
Performed User Experience Questionnaire

1	I thought the chatbot was helpful ("ik vond de chatbot behulpzaam werken")
2	I found it easy to work with the chatbot ("ik vond het makkelijk om met de chatbot werken")
3	I could answer the questions quickly with the chatbot ("ik kon de vragen snel beantwoorden met de chatbot")
4	I found the chatbot clear to work with ("ik vond de chatbot overzichtelijk om mee te werken")
5	I liked working with the chatbot ("ik vond het leuk om met de chatbot te werken")
6	I found it interesting to work with the chatbot ("ik vond het interessant om met de chatbot te werken")
7	I found it innovative to work with the chatbot ("ik vond het vernieuwend om met de chatbot te werken")

3.3 Design

Table 2 shows examples of elements of Conversational Human Voice added to the chatbots. Appendix 3 shows the complete overview of the conversational flow of all the chatbots. We tested all chatbots in at least two rounds and adjusted for the feedback accordingly. All CHV-items were replicated from examples from the research of Van Hooijdonk & Liebrecht (2018).

We built five chatbots that each differ in the amount of CHV-elements in the questions about sustainability.

- CH1 used no elements of CHV and provided no feedback to the user's answers. It used the "formal you" ("u/uw") to address the user and "the municipality" to refer to the municipality of 's-Hertogenbosch as organization. This was the most formal chatbot and therefore serves as a baseline. On the basis of this chatbot, all other chatbots were built.
- CH2 used only elements of personalization, and thus was the chatbot provided with a name and asked the chatbot for the name of the user. It used the "informal you" ("je/jij/jouw") to address the user and "I" to refer to itself as spokesman for the municipality.
- CH3 used only elements of informal language use, and thus uses the chatbot smileys and interjections. It used the "informal you" to address the user and "we" ("we") to refer to itself and the municipality as organization. The reason for not using abbreviations and non-verbal cues such as doubled punctuation marks (e.g. ??/!!) and capitalization is because during the test-phase we multiple times received the feedback that such register repelled users to interact with the chatbot, as it is not in line with communication styles one would expect when interacting with the municipality.
- CH4 used only elements of inviting rhetoric, and thus thanks the user, makes excuses and explicitly asks the user to elaborate on its answer. It used the "informal you" to address the user and "we" to refer to itself and the municipality as organization. The reason for not using humour in this chatbot is for the same reasons as of why CH3 did not use abbreviations and doubled punctuation marks.
- CH5 combined all elements in the previous chatbots (CH2-4), and used around 3 times as much CHV as the other chatbots. It used the "informal you" to address the user and "I" to refer to itself as spokesman for the municipality. In the test-phase we also asked the users if the personality was coming of too strong and after deleting smileys, capitalized letters, attempt for jokes and doubled interpunction as "??" and "!!" the test participants were satisfied.

Table 2
Used elements of Conversational Human Voice

CH1	-
CH2	My name is Eefje, what is your name? (“Ik ben Eefje, wat is jouw naam?”) Hi [name], I would like to ask you some questions about sustainability (“Hallo [naam], ik wil je graag wat vragen stellen over duurzaamheid”) Why is that (un)important for you, [name]? (“Waarom is dat (niet) belangrijk voor je, [naam]?”) I think that is important too! (“Ik vind dat ook belangrijk!”) I did not understand you (“Ik begrijp je niet”)
CH3	Oh! That’s interesting to hear! (“Oh! Interessant om te horen!”) Ah like that, we can understand that :) (“Ah zo, we begrijpen het :)”) ... Do you find sustainability important? (“... Vind jij duurzaamheid belangrijk?”) The chatbot did not understand you :((“De chatbot begrijpt je niet :(”)
CH4	Could you tell something more about that? (“Zou je daar iets meer over kunnen vertellen?”) We can understand this is a difficult question to answer (“We begrijpen dat dit een lastige vraag is om te beantwoorden”) Indeed, we find that important too! (“Wij vinden deze inderdaad ook belangrijk!”) Thank you for your answer (“Bedankt voor je antwoord”) The chatbot did not understand you, apologies for that. (“De chatbot begrijpt je niet, excuses hiervoor”)
CH5	My name is Eefje, what is your name? (“Ik ben Eefje, wat is jouw naam?”) Hi [name], nice to meet you! I’d like to ask you some questions about sustainability (“Hallo [naam], leuk je te ontmoeten! Ik zou je graag wat vragen willen stellen over duurzaamheid”) Indeed, I think that’s really important too! (“Ik vind deze inderdaad ook heel belangrijk!”) I understand exactly what you mean. Thanks for your reply :) (“Ik begrijp precies wat je bedoelt. Bedankt voor je antwoord :)”) I did not understand you, sorry :((“Ik begrijp je niet, sorry :(”)

3.4 Procedure

Participants received a brief instruction of the experiment and informed what the participant’s goal was, namely, to experience the interaction with the chatbot and fill in the questionnaire provided by the chatbot. Both the second and fifth chatbots provided a moment to become acquainted with the chatbot (and vice versa), as part of the personalized elements of CHV. The other chatbots started right away with the questions about sustainability, as a way to get familiar with the chatbot and experience the conversation. Participants start with answering the questions about sustainability and fill in the questions about the User Experience afterwards. In total, the participants answered eight questions about sustainability and seven statements that combined the User Experience score.

Lastly, the participants could make last comments about their experience with the chatbot (open-ended question) and were asked to fill in their characteristics (gender and age). Between every question the chatbot took a few seconds, to make the conversation seem more natural. All chatbots except the first one provided prompt feedback according to the user’s input. Chatbot 1 asked every question regardless of the input of the user. The chatbot consistently took turn after a user had sent their answer by pressing enter (or the ‘sent’ button). Participants had a week to take part in the experiment. Appendix 3 shows a complete overview of the conversational flow of the chatbots (in Dutch).

3.5 Analysis

A total of 695 citizens participated in the research (response rate = 61.1%). Only participants who fully completed the questionnaire were included in the analysis (N = 551). Furthermore, 4 participants were excluded for not filling in the User Experience statements accordingly (e.g. one participant gave every statement the maximum rating of 5 and typed in the comment section “it is not possible to erase your answers” and one of them stating: “I did not receive any questions about sustainability”).

For the text analysis the amount of participants differ, since the answers of all open-ended questions were included in the analysis. Afterwards, the text mining analysis was done on either all open-ended questions or the comment section only.

4. Results

This section is divided into two parts: (1) the overall successfulness of the chatbots and (2) deeper understanding of the successfulness. Section one is subdivided into the following subcategories:
overall success rate (1a), duration of conversations (1b), length of messages (1c), alignment (1d) and future intentions (1e). Section two is subdivided into the following subcategories: User Experience (2a), sentiment (2b), effect of gender and age (2c) and additional analyses (1d).

Within this result section, the term 'condition' refers to the five different types of chatbots. The term 'message', refers to the user input (in text) after hitting the 'enter' or 'send' button.

4.1 Overall successfulness of the chatbots

Success rate

To evaluate the overall successfulness of the chatbots, the number of participants that did (not) complete the task was measured. The results of this analysis are presented in table 3. It shows that the success rate drops with the increase of Conversational Human Voice, as condition 5 scores less successful than condition 1 to 4. However, this can also be explained as continuing loops were possible in all conditions except condition 1. This was because condition 1 did not provide any feedback. All in all, condition 1 (introduced as "the baseline") performs better than the adjusted models in the case of completing the survey.

To test if one condition was significantly more successful compared to the overall mean, a Chi-squared analysis was performed. Against the expectations of H1, this did not show any significant difference (p-value: 0.2414). Therefore, H1 is rejected.

Table 3
Success rate per chatbot condition, with variables:

Condition	Participants	Participants that completed the task	Success rate
1 (no CHV)	142	129	90.8%
2 (Personalization only)	154	119	77.3%
3 (Informal language use only)	134	102	76.1%
4 (Inviting rhetoric only)	142	110	77.5%
5 (all CHV-elements)	123	89	72.4%

Duration of conversation

The duration of the conversation was measured in seconds and converted to minutes to investigate whether Conversational Human Voice has an effect on the duration of the chat conversation. The hypothesis stated that the addition of elements of Conversational Human Voice would increase the duration of the conversations that the participants had with the chatbot. To test whether one condition was significantly faster than the other conditions, an analysis of variance was performed. Results[1] show no significant differences (p-value: 0.0999). Therefore, H1 is rejected. Figure 2 shows the boxplots of the duration of the conversations (in minutes) per condition.

[1] While testing the assumptions of the analysis of variance, the assumption of homogeneity was not met (Levene's test, p = 0.002). While testing for the normal distribution, the histogram and density plot showed a positive skewed distribution. The boxplot showed several outliers in every condition. After eliminating the

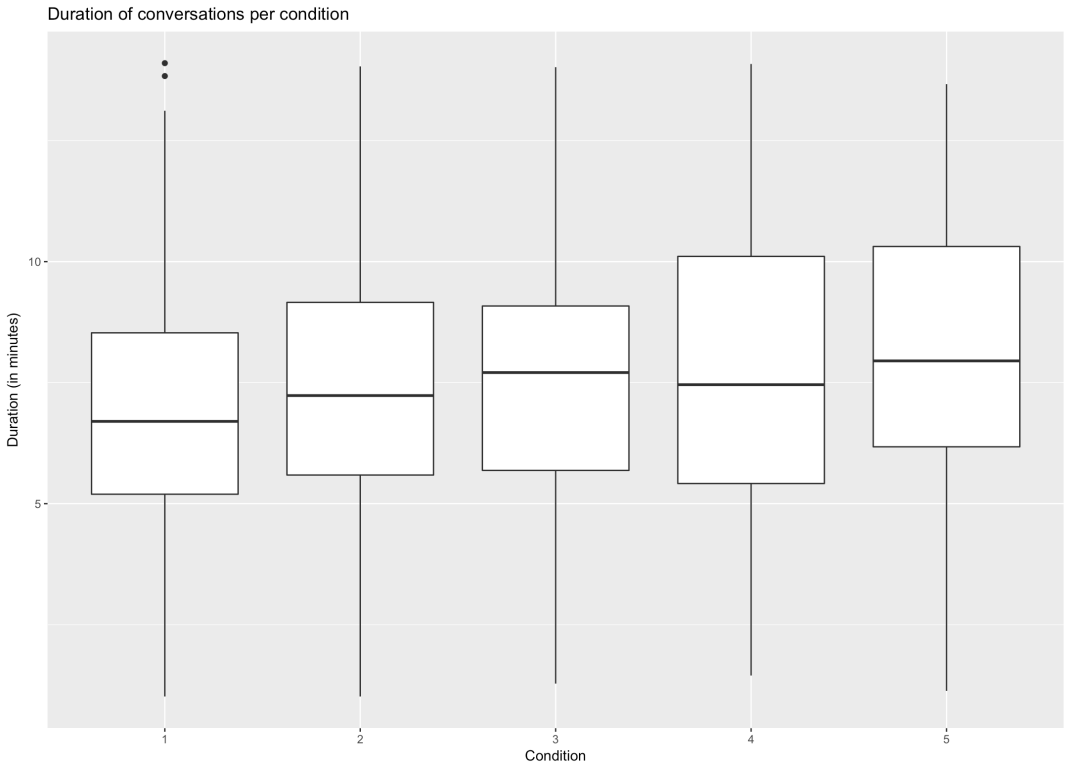


Figure 2. Boxplots of duration (in min) per condition.

Length of messages

To test whether Conversational Human Voice affects the amount of words per produced message by users in every open-ended question[2] was measured. Afterwards, a distinction was made between the analysis of all open-ended questions or the comment section (e.g. "do you have any comments in regard to this chatbot?"). The length of messages was measured in words per message. Answers of one word were not included in the analysis. Table 4 shows the results of the word count analysis.

Table 4

Results of the word count analysis on the open-ended questions

	Comment section only			All open-ended questions		
	N	*M*	*SD*	*N*	*M*	*SD*
Condition 1	129	7.08	8.27	586	8.41	7.25
Condition 2	118	9.45	8.89	517	9.50	6.68
Condition 3	102	8.24	8.45	399	8.91	6.72
Condition 4	110	8.89	8.65	441	9.45	6.91
Condition 5	89	9.44	8.69	342	9.33	7.05

The hypothesis stated that that the addition of elements of Conversational Human Voice would increase the amount of words per message sent by participants. With all open-ended questions, the first and third condition produced the least words per message and these results stay quite consistent in the comment section. Note that in general, participants produced few words per message (M = 9.4), as is visually presented by the positively skewed distributions in figure 3. In line with the results, figure 3 shows that in condition 1 almost every participants used between 5-10 words per message, and in condition 4, the amount of words per message were more spread out to around 5-15 words per message. No significant tests were performed in this analysis.

[2] The open-ended questions include: "do you think sustainability is important?", "why do you find that (un)important?", "do you do more with sustainability?" and "do you have any comments in regard to this chatbot?", since these questions are asked to every participant similarly, in stead of a follow-up question after the participant select one of several options.

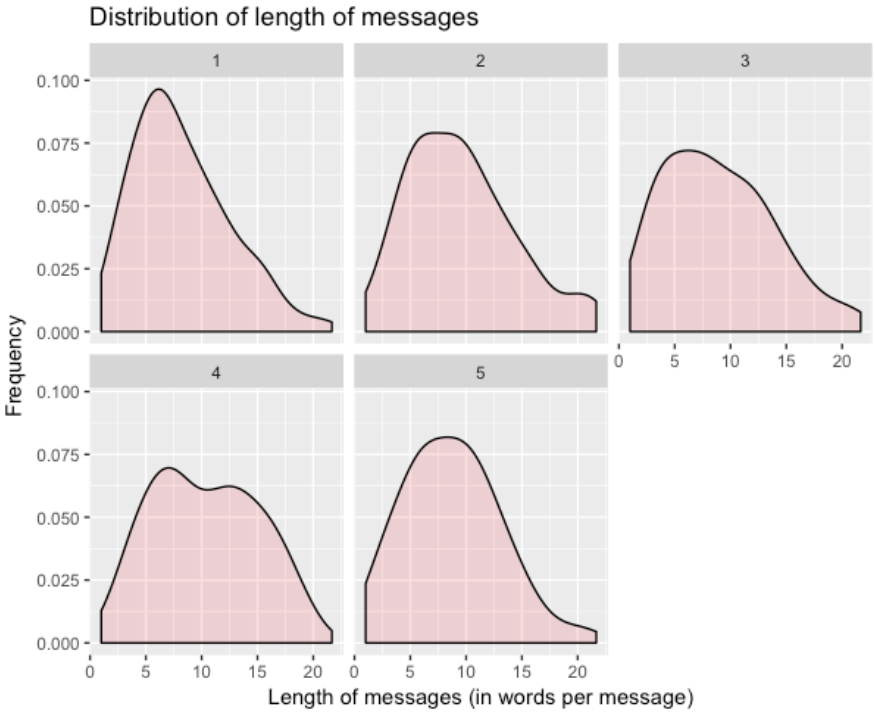


Figure 3. Density distributions of the length of messages (in words per message) per condition.

Alignment

Alignment can be used as evaluation method to measure the perceived humanness of the chatbots. For the analysis of alignment, the text was transformed to non-capitalized words and tokenized. For every open-ended answer the amount of Conversational Human Voice that was produced by users was counted, split per CHV-category and normalized by the total length of the messages. In order to find the elements of Conversational Human Voice produced by the participants, the examples from the research of Van Hooijdonk & Liebrecht (2018) were used in our analysis. That being said, the analysis only included the categories that the chatbot self produced (e.g. no abbreviations, doubled punctuation marks and capitalizations, as is explained in chapter 3.3 'Design'). Therefore, it was no problem that only lowered text was used.

The hypothesis stated that the addition of elements of Conversational Human Voice would increase the perceived humanness of the chatbot en therefore we suspect to find the most alignment in condition five, and similarly, the least alignment in condition one. Furthermore, the theory of alignment would suggest that what words has been produced by the chatbot, would be produced back by the participants, such that we expect that in condition two the most alignment should occur within the category of 'personalisation', in condition three the most alignment is expected within 'informal language use' and condition four should have the most alignment within 'inviting rhetoric'.

Figure 4 presents the results of the analysis and overall, the most alignment was produced by participants in the fifth condition. In all conditions, the most produced category was 'Personalisation', which is likely due to various sentences that include the reference to the participant itself (e.g. "I"). The CHV-category that was the least aligned was 'Informal'. When compared with the other conditions, condition three and five show the most informal language use. Additionally, condition four shows the most inviting language. Oppositely, in comparison with the other conditions, chatbot two does not show the most personal alignment, but condition five and four did. When looking at the comment section, these results stay quite consistent. No test of significance was obtained in this analysis.

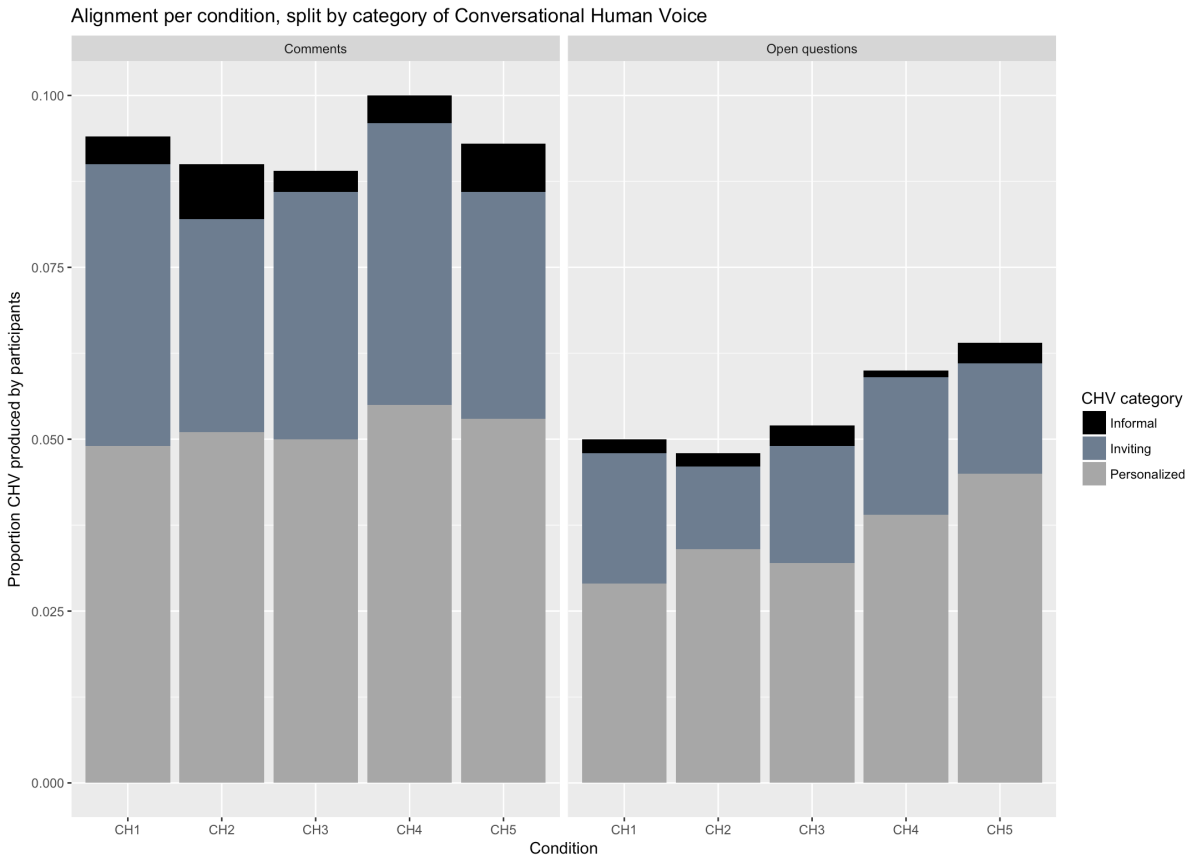


Figure 4. Conversational Human Voice produced by participants on all open-ended questions and comment-section only, per condition, and split by category of Conversational Human Voice

Future intentions

Together with the implicit evaluation metrics described above, it can be helpful to align the obtained results with a more explicit form of evaluation. Therefore this study asked the participants at the end of the experiment the question "would you like to make use of chatbots more often?"

Figure 5 (a) shows the resulting frequencies of the answers to that question and provides this research with extra evidence to what the previous results already seem to suggest: most participants liked interacting with the chatbot and would like to make use of chatbots more often (86%). When distributed over the different conditions, as is visible in figure 5 (b), the second chatbot stands out with 96% of the participants that indicate they would like to use chatbots more often.

4.2 Deeper understanding of the successfulness

Now that we know that participants liked interacting with the chatbots, we would like to know what it exactly is that they liked about it. Therefore we will look at the User Experience and sentiment analysis.

User Experience

The User Experience is to provide insights to what aspect the participants liked and disliked between the different conditions. The hypothesis stated that the addition of elements of Conversational Human Voice should increase the score on the User Experience Questionnaire (UEQ).

Figure 6 presents the results of the analysis. It is shown that there is little difference between the mean User Experience scores of the different conditions. Moreover, with an overall mean score of 3.97, the analysis shows a ceiling effect where all participants rate the chatbots (regardless of CHV) quite high.

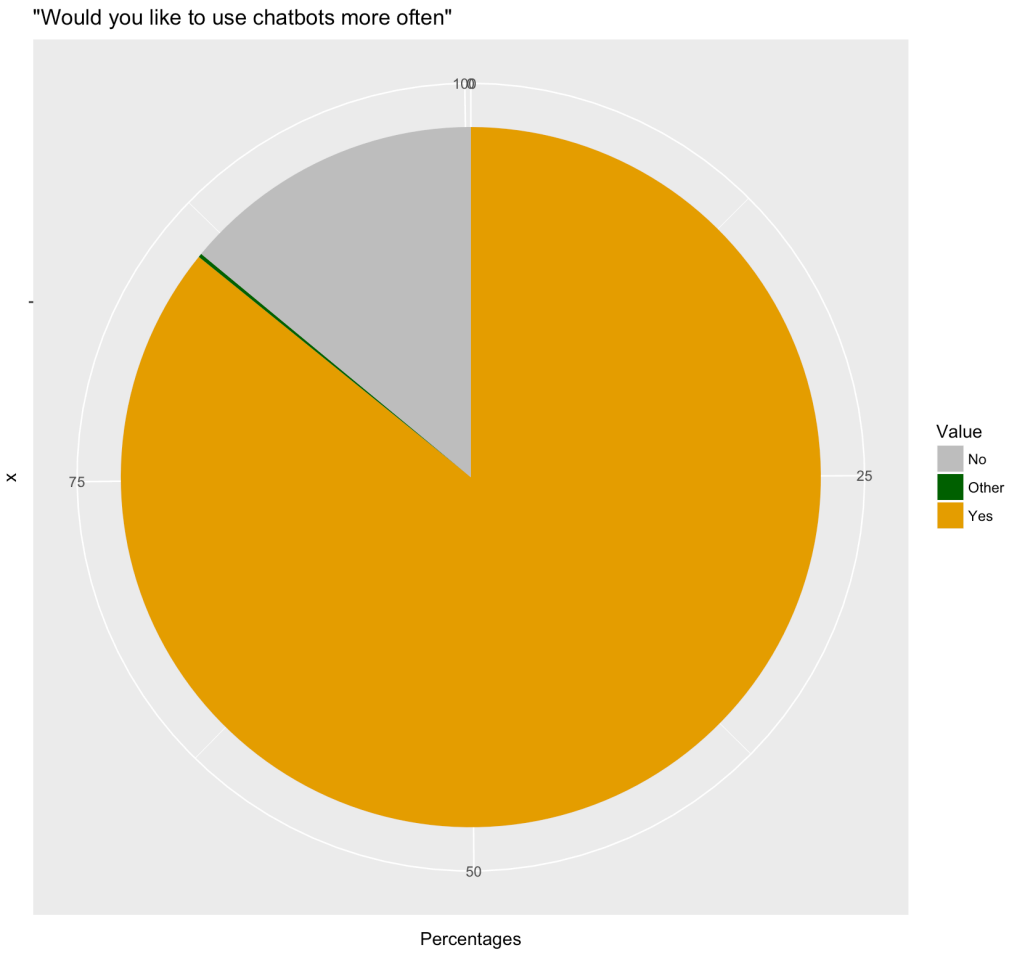


Figure 5. (a) Pie chart of answers in percentages

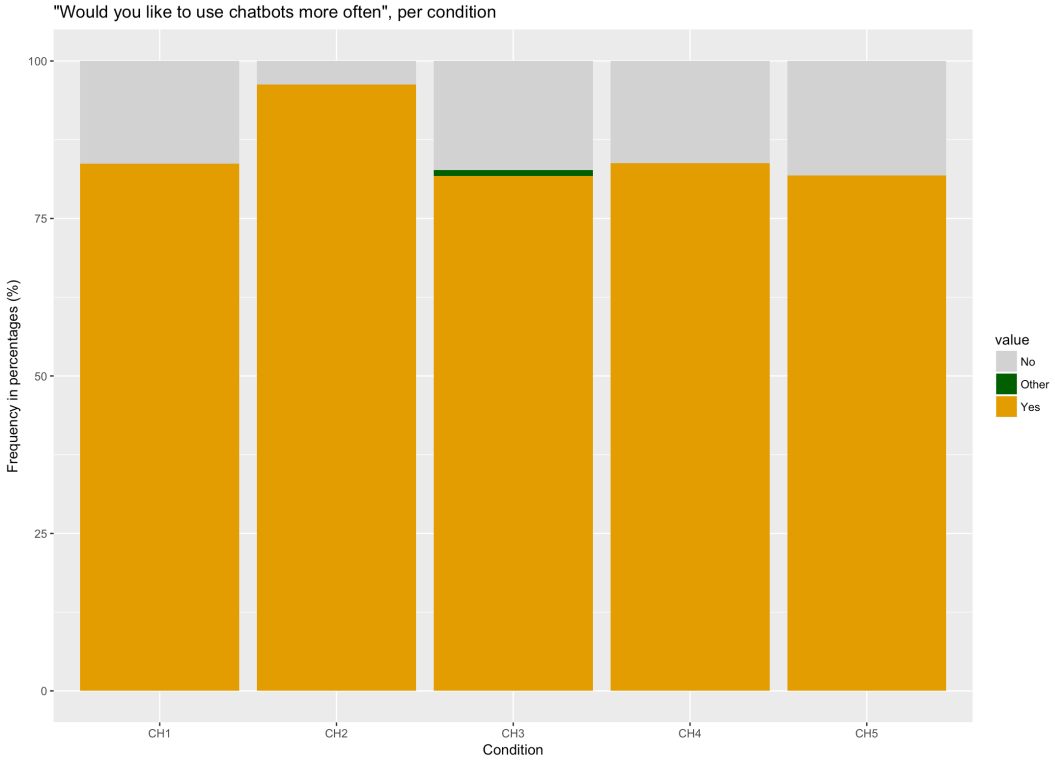


Figure 5. (b) Bar plot of answers in percentages distributed per condition.

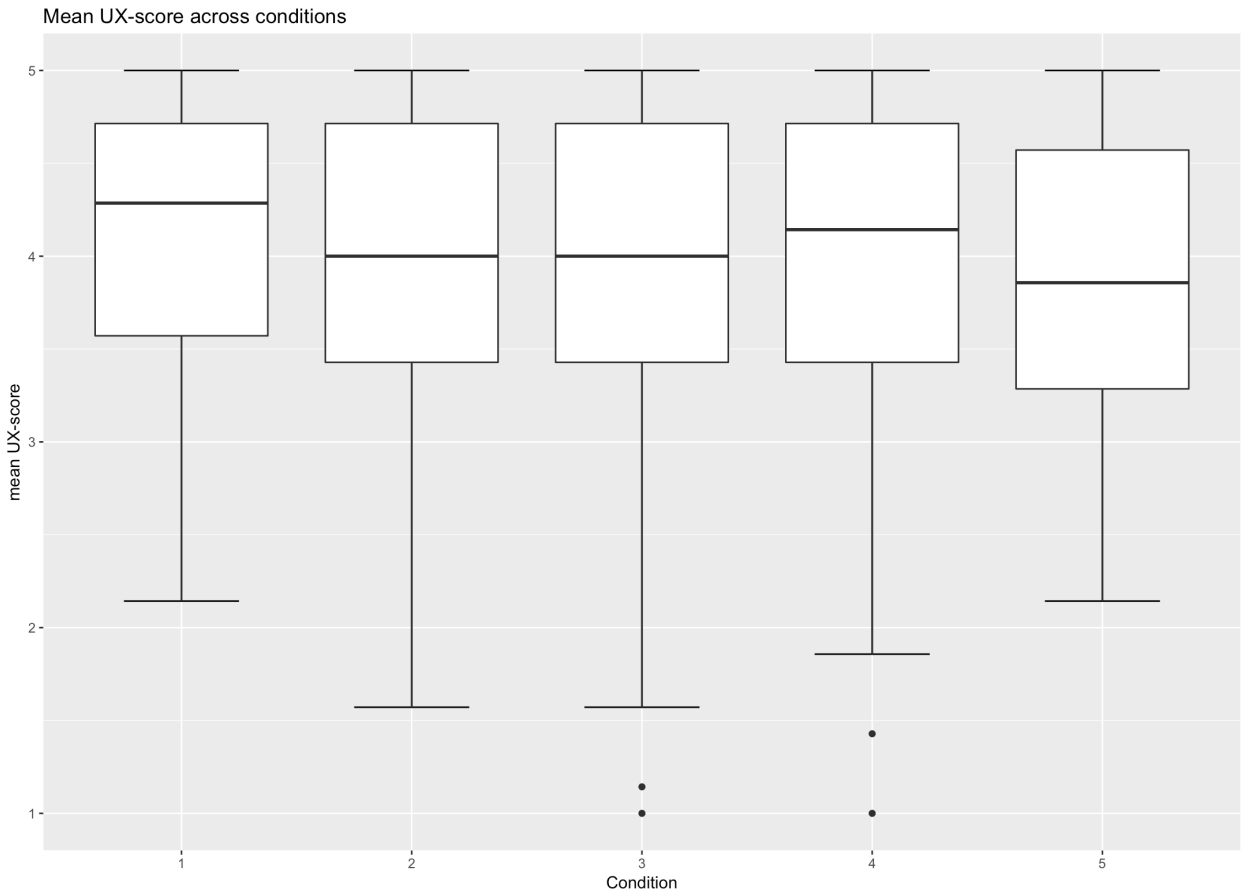


Figure 6. Boxplot of mean UX-score across conditions

To research if Conversational Human Voice has a main effect on the User Experience, an analysis of variance was performed. Results[3] however, show no significant effect of Conversational Human Voice on User Experience on the 0.05 significance level ($p = 0.074$).

The outliers in condition three and four are examined and scored low on user satisfaction due to hitting 'enter' or other symbols too soon and sending the messages before finalizing his/her answer or obtaining the same question multiple times (e.g. ending up in a loop).

Sentiment

To analyse whether Conversational Human Voice affects the sentiment of the users, the sentiment of every answer to the open-ended questions was measured. For the sentiment analysis the text was first tokenized and transformed to lowercase. Only answers that consist of more than one word where analysed through Pattern (De Smedt & Daelemans, 2012). Table 6 shows the result of the sentiment analysis, where the first value provides insight in sentiment and the second value represents the subjectivity of the text. Neutral sentiment was not included in the calculations of the mean and standard deviation.

[3] While testing the assumptions of the analysis of variance, the assumption of homogeneity was accepted (Levene's test, $p = 0.95$), but the distributions of the mean scores on User Experience show a slight negative skewed distribution, due to the ceiling effect. It is therefore that we cannot accept the assumption of normality. This is also presented in appendix 4.

Table 6

Results of the sentiment analysis on the open-ended questions. Mean and Standard Deviation are given for sentiment and subjectivity

	Comment section only			All open-ended questions		
	N	*M* (sentiment; subjectivity)	*SD* (sentiment; subjectivity)	*N*	*M* (sentiment; subjectivity)	*SD* (sentiment; subjectivity)
Condition 1	97	0.11; 0.64	0.40; 0.27	496	0.14; 0.62	0.36; 0.27
Condition 2	109	0.05; 0.72	0.44; 0.25	474	0.11; 0.62	0.35; 0.28
Condition 3	79	0.09; 0.73	0.44; 0.28	351	0.13; 0.60	0.34; 0.29
Condition 4	103	0.11; 0.64	0.38; 0.29	398	0.11; 0.58	0.33; 0.28
Condition 5	94	0.03; 0.66	0.39; 0.27	308	0.13; 0.63	0.36; 0.28

The hypothesis stated that the addition of elements of Conversational Human Voice would lead to more positive sentiment in the messages of the users. Table 6 shows that condition 1 and 4 score in the comment-only section the highest for positive sentiment, such that it seems that chatbot 1 and 4 were perceived most positive. When looking at the section with all open-ended questions: condition 1, 3 and 5 yield the highest score for positive sentiment. However, since scores could run up to -1 (for negative sentiment) and +1 (for positive sentiment), all chatbots are rated quite neutral and differences are very close together. No analysis of significance was performed on this section.

However, it was noteworthy that condition 1 and 4 were also rated highest on the mean User Experience score and these results seem to align pretty well. To test this assumption, a visual inspection of the correlation between User Experience and Sentiment was performed. Neutral sentiment was not included in the analysis. Figure 7 shows a positive correlation effect between sentiment and User Experience.

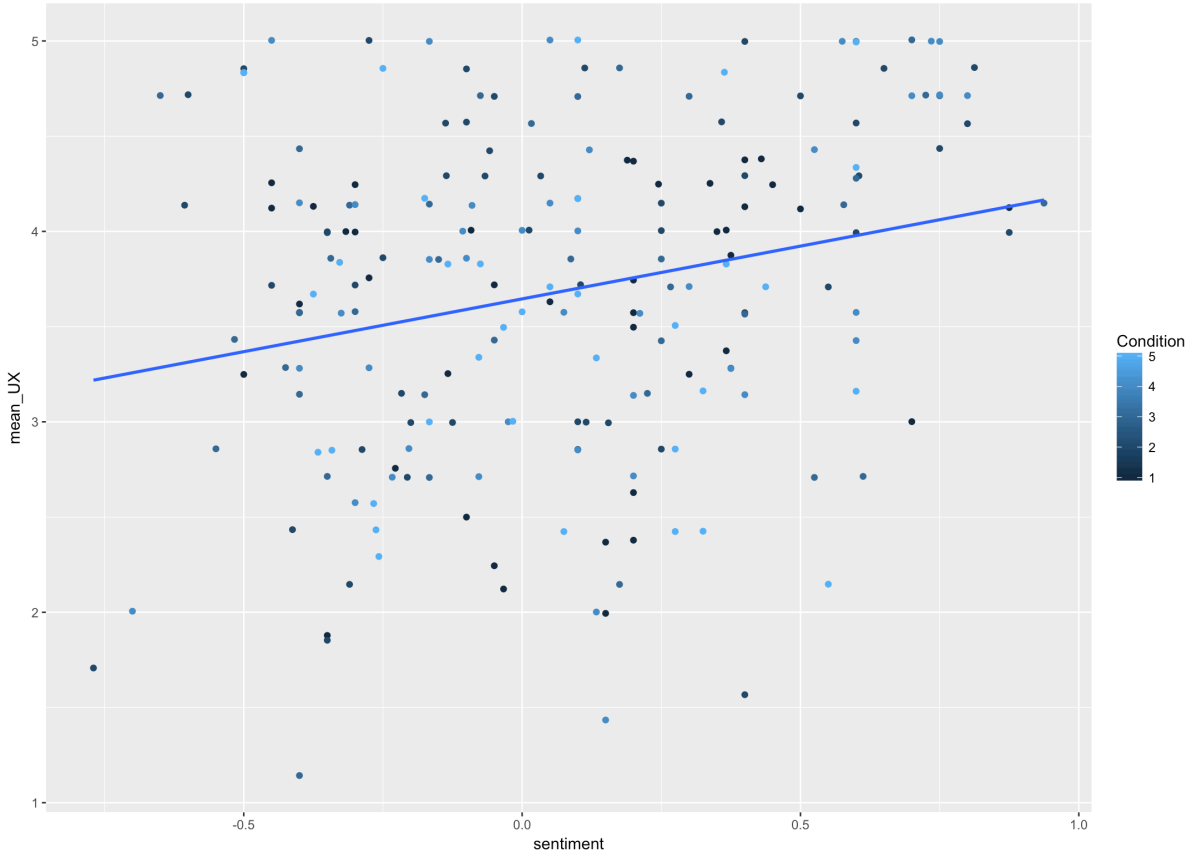


Figure 7. Scatterplot with correlation effect between User Experience and sentiment

Effect of Age and Gender on the mean User Experience

Literature suggests that personal characteristics affect the User Experience. The hypothesis stated that people of younger age would rate the chatbots higher on User Experience. Likewise, women are believed to rate the chatbots higher on User Experience. Figure 8 (a) shows that up until condition five women rate the chatbots consistently higher than men. Figure 8 (b) shows that these results stay consistent for most of the age categories. One exception is the age category of 75 years old or above, where men scored higher than women. As expected, participants between 18-25 years old rated the chatbots as one of the most positive of all age categories (M = 4.10). However, the group size is too small to draw any conclusions (N = 3). Unexpectedly, women and men between 66 and 74 years old and males of 75 years or older score in general the highest on User Experience (M = 4.14). The age categories that in general score the lowest for User Experience are between 56-65 years old (M = 3.85), as can be seen in figure 8 (b). However figure 8 (c) shows that these results differentiate when plotting age across condition. For example, in the personalized condition (condition 2), the 75 years or older group scores second lowest, but the same age category scores highest for the inviting rhetoric condition (condition 4). The ages between 26 and 35 score highest on the first condition and lowest on the fifth, indicating that the effects of Conversational Human Voice on age and gender matter for the User Experience score.

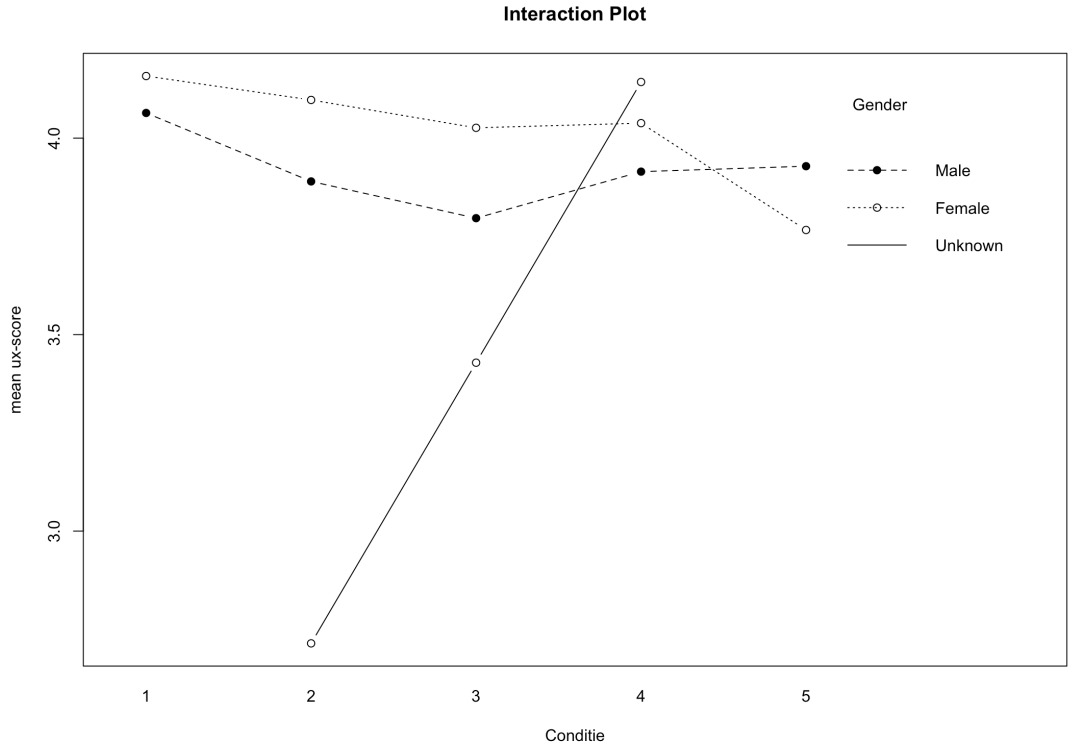


Figure 8. (a) Interaction plot of Condition*Gender

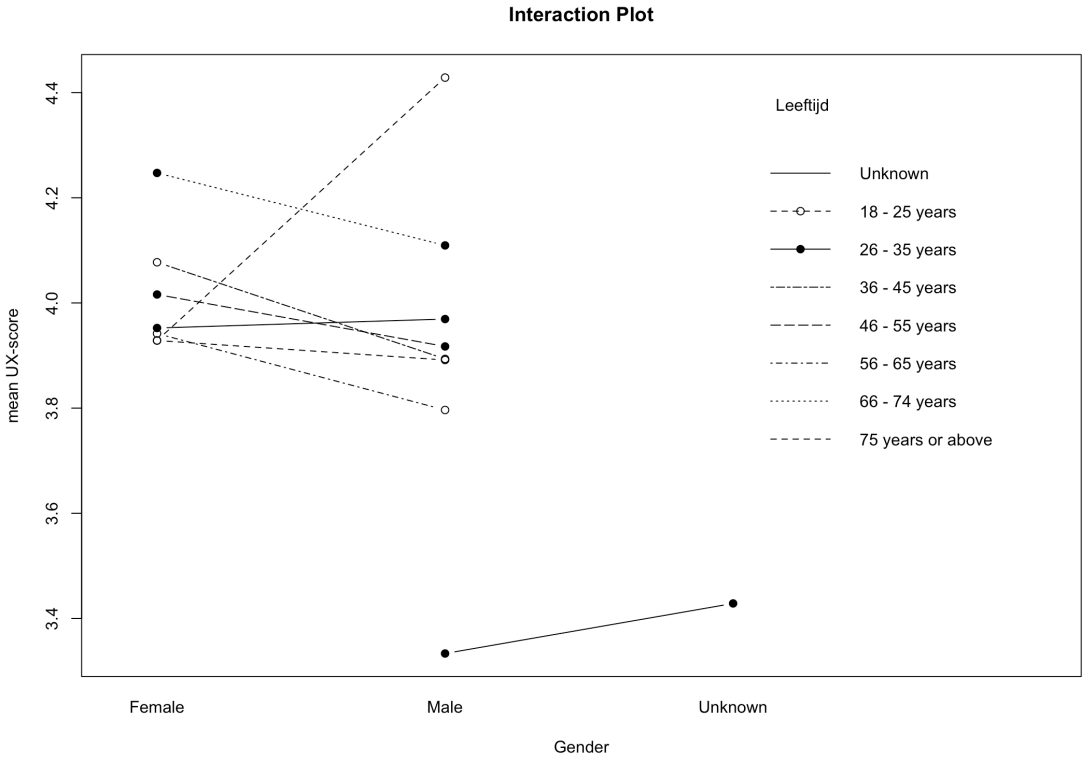


Figure 8. (b) Interaction plot of Gender*Age

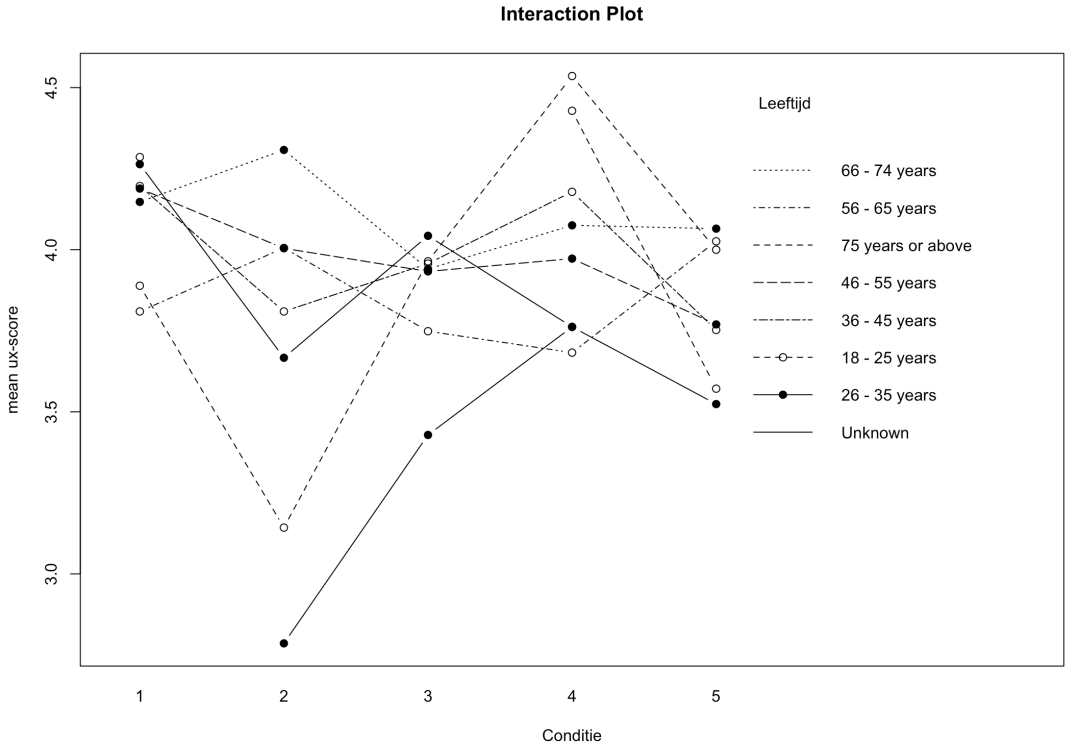


Figure 8. (c) Interaction plot of Condition*Age

To test if the interaction effect between age, gender and condition was significant, the Kruskal-Wallis test was performed rather than an ANOVA[4]. Results on the Kruskal-Wallis test found no significance difference between male and female mean UX-scores or on age at the 0.05 significance level. No significant interaction between age and gender was found (p = 0.2419).

Additional analysis on individual items of the User Experience

While the mean User Experience did not show any significant results, additional analyses of variance were performed for every single item of the User Experience Questionnaire to test if one condition significantly differs from another. The results of this analysis are shown in table 7.

Table 7

Mean scores and results of the analysis of variance across all seven items of the UEQ, examining the effects of Conversational Human Voice on age, gender and their interaction.

	Mean Score	Condition	Age	Gender	Condition*Age*Gender
	M	*p*	*p*	*p*	*p*
1. Quick	4.16	0.46	0.18	0.26	0.20
2. Helpful	3.54	0.02*	0.03*	0.36	0.52
3. Clear	4.06	0.50	0.58	0.04*	0.32
4. Interesting	3.86	0.95	0.53	0.34	0.13
5. Fun	3.88	0.57	0.30	0.17	0.10
6. Easy	4.26	0.61	0.60	0.28	0.50
7. Innovative	3.97	0.90	0.45	0.87	0.09

Notes. * significant effect

Table 7 shows that only two items found significant differences: item 2 “helpfulness” and item 3 “clearness”. The results of the Tukey Post-hoc test with Bonferroni correction for the p-value show that the first condition was rated significantly higher on “helpfulness” (*M* = 3.82) than the fifth condition (*M* = 3.22). Furthermore, the chatbots were for women significantly clearer (*M* = 4.17) than for men (*M* = 4.07), as is shown in figure 9.

For the analysis of the effect of age on item 2 “helpfulness”, the Kruskal-Wallis test was performed rather than an ANOVA[4]. Table 8 shows the results of this test. The results of the Wilcoxon Post-hoc test with Bonferroni correction for the p-value show that woman younger than 55 rated the first chatbot (*M* = 4.18) significantly as more helpful than men of all ages in the same condition (*M* = 3.79). Moreover, women that are younger than 55 years old in condition 5 rated the chatbot significantly less helpful (*M* = 3.00) than both man and women of the same age in condition 1 (*M* = 4.06), as is shown in figure 10.

Table 8

Results of the Kruskal-Wallis significance test examining the effects of Conversational Human Voice, gender and age on item 2 of the UX-score: “helpfulness” and their interaction.

Variables	*p*
Condition	0.01*
Gender	0.27
Age	0.85
Condition*Gender	0.01*
Condition*Age	0.001
Age * Gender	0.99
Condition*Age*Gender	0.01*

Notes. * significant effect

[4] While testing the assumptions of the analysis of variance, the assumption of normality was not met because the sample sizes were too small (e.g. N < 30). Therefore, recoding of the existent dataset was needed to come up with two age categories: younger than 55 years old and older than 55 years old.

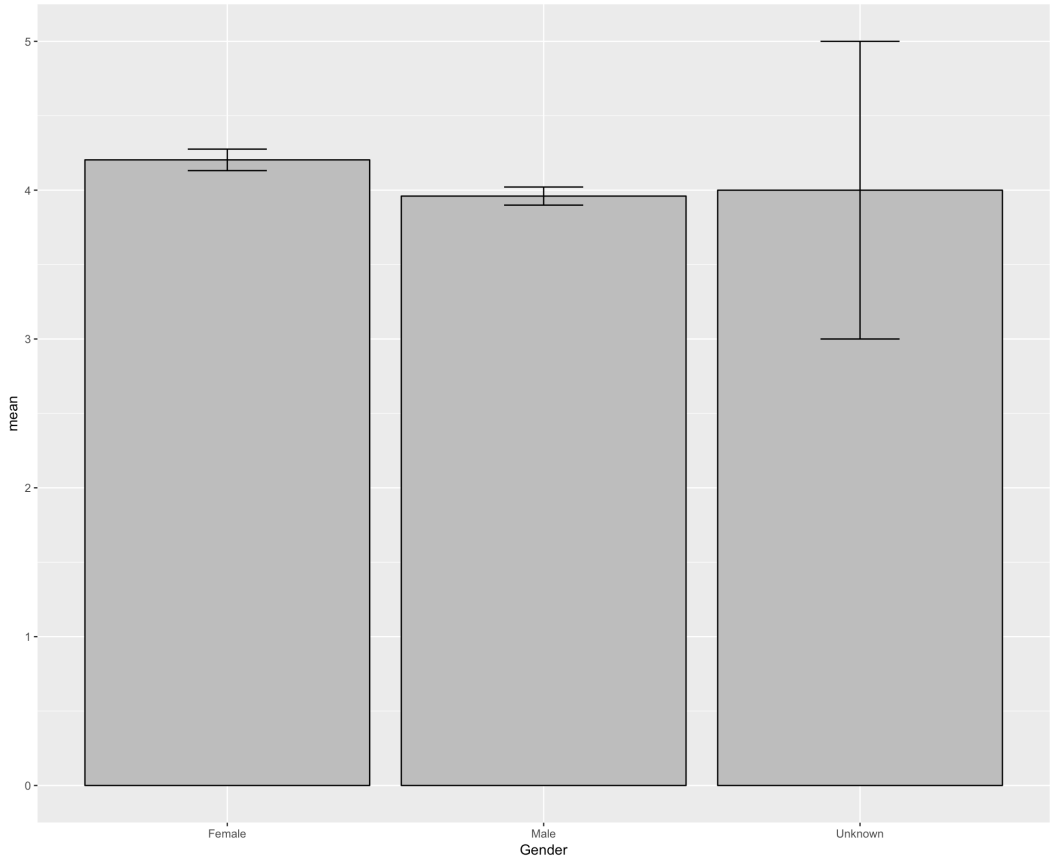


Figure 9. Bar plot of mean scores on item 3: “Clear”, per gender.

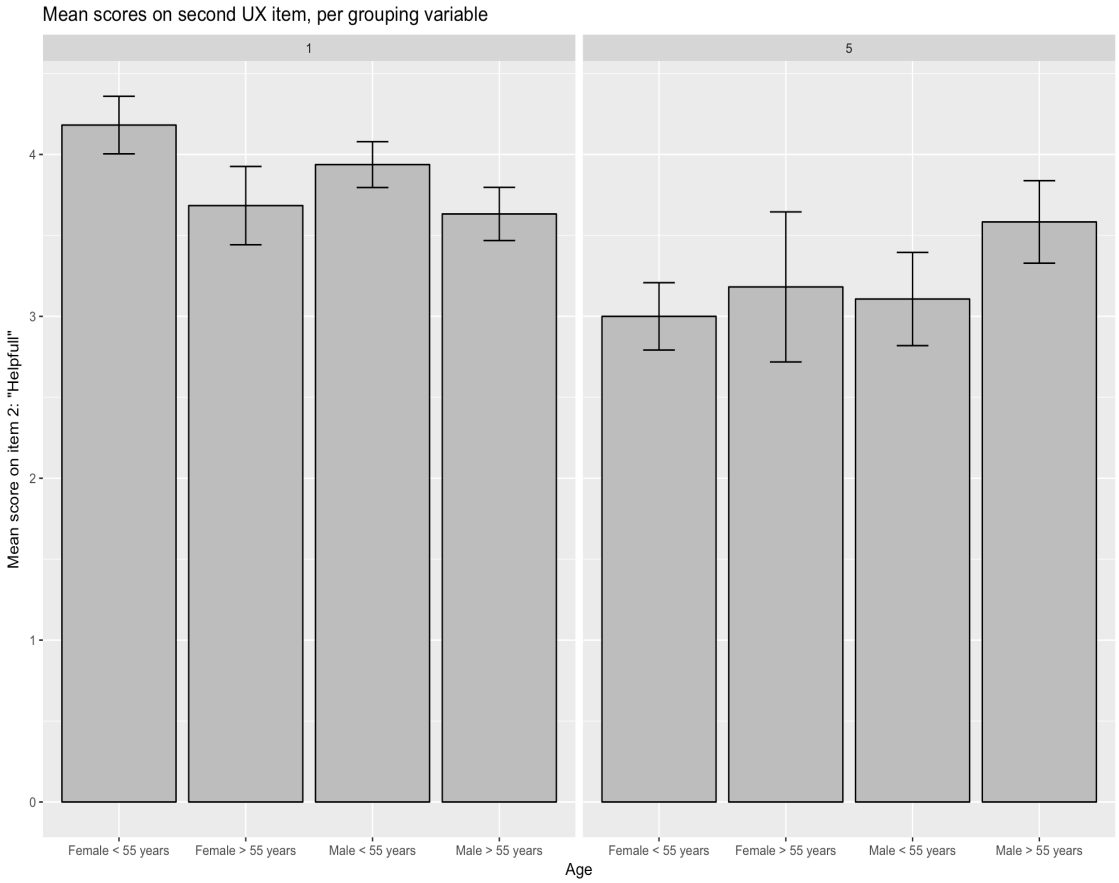


Figure 10. Bar plot of mean scores on item 2: “Helpful”, per gender and age in condition 1 and 5

5. Discussion

The current study was set out to investigate the influence of Conversational Human Voice on User Experience and duration of the chat conversation, length of messages, sentiment and alignment. This was researched by letting participants interact with one of five different chatbots that each differs in use of Conversational Human Voice. The mean score of seven 5-point Likert scale statements concluded the User Experience. At the same time, the underlying process was being monitored by calculations of the overall success rate, duration of conversations, words per message, performing a sentiment analysis and examining the alignment. Results suggest that Conversational Human Voice does not necessarily lead to a higher score on User Experience in the context of survey research. Similarly condition 1, without any added CHV, was the most successful in terms of completing the survey and the condition, together with condition 4, with most positive sentiment. Contrary to our beliefs, condition 5 (that used all elements of CHV) was least successful, took longest to complete and seems to score lowest on User Experience. Condition four (inviting rhetoric) was rated highest on User Experience within the CHV-conditions, and together with condition 1 show the most positive sentiment when the participants were asked for comments regarding the chatbot. Moreover, condition four showed the most words per message, together with condition 2 (personalisation). In the following section, these results will be discussed in the context of past studies leading to suggestions for future research.

Duration of conversations and length of messages

Based on the research of Hill, Ford and Farreras (2015), the hypothesis stated that the addition of elements of Conversational Human Voice would increase the amount of words per message and the duration of the chat conversation. The duration of the chat conversations were longer and the messages contained more words for the conditions that included elements of CHV compared to the condition that did not used CHV. This could be seen as an effect of CHV and thus in support of the hypothesis. However, the odds of ending up in a loop increased with the addition of feedback and might be a more suitable explanation. That being said, condition five took the longest of all chat conversations to complete. This condition included all elements of CHV, thus it can be said that a trend is visible that to seem to suggest that the addition of CHV-elements leads to more extensive messages from the users. Statistical analysis did not show this trend to be statistically significant. Thus, the hypothesis cannot be accepted. No test of significance was performed on the analysis of words per message.

Alignment

Based on the literature (Hasson & Frith, 2016; Heyselaar, Hagoort, & Segaert, 2017; Hill, Ford and Farreras, 2015; Pickering & Garrod, 2006), it was suggested that more Conversational Human Voice would lead to more alignment. Results agree, such that overall the most elements of Conversational Human Voice produced by participants was in the fifth condition, the chatbot that produces the most elements of Conversational Human Voice. These results support the literature stating that when the chatbot is experienced as human, the users show more alignment. However, no significant test was obtained in this analysis, so no definite conclusions can be drawn.

User Experience

Based on the literature (Barcelos, Dantas, & Sénécal, 2018; Gnewuch, Morana, & Maedche, 2017; Kelleher, 2009; Van Hooijdonk & Liebrecht, 2018; Verhagen, Van Nes, Feldberg, & Van Dolen, 2014), the hypothesis suggested that the use of Conversational Human Voice should have a positive effect on the User Experience score. The results obtained in this study showed no significant difference in conditions, indicating that the addition of conversational elements in the context of survey research does not necessarily lead to a higher User Experience score. This could indicate that using chatbots for the context of survey research is seen as a task-oriented communication style (Dion & Notarantonio, 1992; Luger & Sellen, 2016) and therefore too much fuss is not appreciated. Condition four, with inviting rhetoric is rated second highest on User Experience and seems appropriate for the occasion, in opposite to informal language use (condition 2) that for most participants did not feel in line with the communication style of the municipality of 's-Hertogenbosch. In agreement with Barcelos, Dantas, & Sénécal (2018), in the sense that there is not one "right" tone of voice for all companies. We found no supporting evidence that age or gender had significant effect on the User Experience, as was hypothesised by the literature of Brandtzaeg and Følstad (2017).

We did, however, find significant differences for the items “helpfulness” and “clearness” of the User Experience Questionnaire, where the chatbot with no elements of CHV was significantly rated as more helpful than the chatbot that used all elements of CHV. This provides further evidence for the claim made above. Furthermore, there was a significant interaction effect between age, gender and condition, where women younger than 55 rated the first significantly as more helpful than men of all ages in the same condition. Lastly, women found the chatbots significantly more clear than men.

Sentiment

Based on the literature (Verhagen, Van Nes, Feldberg, & Van Dolen, 2014), it was suggested that more Conversational Human Voice would lead to more positive sentiment. However, results show no considerable differences between the conditions. This could be due to the fact that sentiment analysis resulted most often in neutral responds on questions about sustainability. However, when looking at sentiment analysis as a form of evaluating the chatbots, both conditions one and four are rated with most positive sentiment in line with the claims made above. Condition five again scores most negative, arguing that Conversational Human Voice does not automatically lead to a more positive sentiment. Nonetheless, all differences still lie very close together and sentiment analysis has been known to have drawbacks: fully depending on the evaluation of the creater of the library that classifies words into ‘negative’ and ‘positive’, and often failing to detect irony and the numerous nuances that language includes (Ceron & Negri, 2016). That being said, the sentiment scores seemed to align well with the user experience score and is therefore seen as a reliable measurement.

Limitations and future research

Multiple issues may be rising with respect to the current findings. First, no significant effect was found between Conversational Human Voice and User Experience. One solution is to enlarge the sample size and differences between conditions by focussing on only two conditions: with CHV and without CHV. However, in this research there is no direct link between Conversational Human Voice and User Experience, because it was possible for the participants to obtain questions multiple times in conditions 2 to 5 that very likely have affected the User Experience Score. That being said, all scores on User Experience were quite high, due to a ceiling effect and that could be another explanation to why there was no significant difference. A possible explanation for the ceiling effect is that we only recruited participants that indicated that they were interested in testing a chatbot of the municipality of ’s-Hertogenbosch which may have created an bias. To solve the bias, the same experiment should be conducted with a random selection of participants. A solution for the ceiling effect would be to repeat this experiment with participants that are used to new technologies. Future research should have chatbots that do not end up in loops, so a definite conclusion can be drawn. Furthermore, generalization is impossible because this experiment took only into account the citizens of ’s-Hertogenbosch in the context of survey research. Future research should repeat similar experiments and try to find similar results.

Second, besides the test-rounds done in this experiment, participants could still run into a loop since the training sets of several questions about sustainability were too much alike. In condition two some people were addressed with “is” as their name, due to the overload of the sentence “My name is ....” in the trainingset. These flaws fairly decreased the “humanness” of the chatbot and therefore could have interfered with the User Experience score. Despite condition five having the exact same training sets, the name problem did not occur. Because of time constraints, these issues were not solved during the experiment. Future research should entail a larger test panel and more design thinking cycles for a successful chatbot.

Third, the reason for using the members of the Digipanel was to obtain a representative sample and while it was interesting to see the results of adults and elderly, the young adults in this experiment were very much underrepresented. Because we only invited respondents that had already participated in a survey of environmental stations, there could very well be a bias. Future research should focus on a younger group and compare with the results obtained in this study. Another disadvantage of using Digipanel and the link towards a web browser with the chatbot, is that there was no guarantee that the participants ran the experiment in a quiet setting.

Fourth, while we did some implicit analysis, the main objective of this research was to measure the User Experience with a survey of seven 5-point Likert scale items. It has been known that Likert scales are not optimal because participants tend to neglect the maximum scores. Future research should focus on continuous measuring rather than snapshots for more valid results. One drawback of the software used for this experiment

was that participants said they had a maximum amount of words they could enter, which withheld them for answering the question to the fullest. Another disadvantage was that participants could only send only one answer, but could start typing before the question was asked, and when hitting 'enter' they already submitted their answer without the option to rectify. Another complaint was that participants had no way to tell how far in the process they were.

Future research could investigate the option of speech recognition. While typing can still be a cognitive load, it would be interesting to see what results can be obtained. Another idea is to use elements of Conversational Human Voice and in combination with augmented and virtual reality. For both suggestions, this research could serve as a baseline. However, the uncanny valley should always be taken into account.

6. Conclusion

In this research we examined to what extent Conversational Human Voice contributes to a higher User Experience score. This study contributes to understanding how participants will converse with chatbots in real-life settings and experiment with the right communication strategy in the context of a survey for a municipality. Results show no significant differences between the extent of Conversational Human Voice used and the User Experience. However, we did find differences between the helpfulness and CHV and overall differences in gender. The overall conclusion is that the positive effects of Conversational Human Voice do not necessarily translate to every scenario. For the context of survey research, one can rather stick to the functional communication styles. Compared to the other Conversational Human Voice categories, inviting rhetoric seems most suitable in the context of survey research and personalisation seems to predict future intentions the best. Furthermore, informal language use seems the least favourable when citizens interact with the municipality. All in al, interactive surveys do show potential and participants made clear that they would like to make use of chatbots more often.

For the municipality of 's-Hertogenbosch this could mean that chatbots would be more successful in more bound contexts, such as answering FAQ's to reduce the burden of the helpdesk or managing appointments for passports or other personal documents. Another possibility is to keep conversational questionnaires short and simple to reduce the chance for participants to end up in loops. When chatbots can truly manage long-term conversations, the prospect of citizens interacting with the municipality via chatbot can be explored again. That being said, the digital divide should always be taken into account.

Acknowledgements

I would like to thank the municipality of ’s-Hertogenbosch and Flow.ai for making this project possible. Special thanks go out to my supervisors Sander and Lieke who already treated me like one of their colleagues. But I also want to thank Florian who continued to deliver feedback even when he never was one my official supervisors. Lastly, Reina for providing me with the sustainability questionnaire and Ronald, whom I could always drop by to ask questions about statistics.

References

Araujo, T. (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 Behavio*r, *85*, 183-189.

Barcelos, R. H., Dantas, D. C., & Sénécal, S. (2018). Watch Your Tone: How a Brand's Tone of Voice on Social Media Influences Consumer Responses. *Journal of Interactive Marketing*, *41*, 60-80.

Bergmann, K., Branigan, H. P., & Kopp, S. (2015). Exploring the alignment space– lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, *2*, 7.

Brandtzaeg, P. B., & Følstad, A. (2017, November). Why people use chatbots. In *International Conference on Internet Science* (pp. 377-392). Springer, Cham.

Ceron, A., & Negri, F. (2016). The "social side" of public policy: Monitoring online public opinion and its mobilization during the policy cycle. *Policy & Internet*, *8*(2), 131-147.

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2018). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*.

Coniam, D. (2014). The linguistic accuracy of chatbots: usability from an ESL perspective. *Text & Talk*, *34*(5), 545-567.

De Smedt, T., & Daelemans, W. (2012). " Vreselijk mooi!"(terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. In *LREC* (pp. 3568-3572).

Denecke, K., Lutz Hochreutener, S., Pöpel, A., & May, R. (2018, April). Talking to Ana: A Mobile Self-Anamnesis Application with Conversational User Interface. In *Proceedings of the 2018 International Conference on Digital Health* (pp. 85-89). ACM.

Dion, P. A., & Notarantonio, E. M. (1992). Salesperson communication style: The neglected dimension in sales performance. *The Journal of Business Communication* (1973), *29*(1), 63-77.

Gnewuch, U., Morana, S. and Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *Proceedings of the International Conference on Information Systems*. ICIS.

Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693), 20150366.

Heyselaar, E., Hagoort, P., & Segaert, K. (2017). In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior research methods*, *49*(1), 46-60.

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245-250.

Jadeja, M., & Varia, N. (2017). Perspectives for Evaluating Conversational AI. *arXiv preprint arXiv*:1709.04734.

Kelleher, T. (2009). Conversational voice, communicated commitment, and public relations outcomes in interactive online communication. *Journal of communication*, *59*(1), 172-188.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Laugwitz, B., Held, T., & Schrepp, M. (2008, November). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group* (pp. 63-76). Springer, Berlin, Heidelberg.

Lommatzsch, A. (2018). A Next Generation Chatbot-Framework for the Public Administration. In *International Conference on Innovations for Community Services* (pp. 127-141). Springer, Cham.

Liao, V.Q., Hussain, M.M., Chandar, P., Davis, M., Crasso, M., Wang, D., Muller, M., Shami, S.N., & Geyer, W. (2018). All Work and no Play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18). ACM, New York, NY, USA* (Vol. 13).

Luger, E., & Sellen, A. (2016, May). 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* (pp. 5286-5297). ACM.

National Research Council. (2013). *Nonresponse in social science surveys: A research agenda.* National Academies Press.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2-3), 203-228.

Porreca, S., Leotta, F., Mecella, M., & Catarci, T. (2017). Chatbots as a Novel Access Method for Government Open Data. In *SEBD* (p. 122).

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579.

Rijkosverheid (2018). Digitale overheid. Retrieved on November 29th, 2018 from: https://www.digitaleoverheid.nl/

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 12.

Shadbolt, N. R., Smith, D. A., Simperl, E., Van Kleek, M., Yang, Y., & Hall, W. (2013, May). Towards a c lassification framework for social machines. *In Proceedings of the 22nd International Conference on World Wide Web* (pp. 905-912). ACM.

Shawar, B. A., & Atwell, E. (2007). Chatbots: are they really useful?. In *Ldv Forum* (Vol. 22, No. 1, pp. 29-49).

Steunenberg, B. (2018). Adaptieve beleidsontwikkeling: zoeken naar nieuwe vormen van beleidsanalyse voor de digitale overheid, *Beleidsonderzoek Online*. DOI: 10.5553/BO/221335502018000001001

Turing, A. M. (1950). Mind. *Mind*, *59*(236), 433-460.

Van Hooijdonk, C., & Liebrecht, C. (2018). Wat vervelend dat de fiets niet is opgeruimd! Heb je een zaaknummer voor mij?^ EK. *Tijdschrift voor Taalbeheersing*, *40*(1), 45-81.

Verhagen, T., Van Nes, J., Feldberg, F., & Van Dolen, W. (2014). Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, *19*(3), 529-545.

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017, May). A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3506-3510). ACM.

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to a chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking*, *21*(8), 491-497.

Appendix 1. Invitation Experiment

Geacht / best digipanellid,

In het laatste digipanel hebben wij gevraagd of u mee wilt doen aan een nieuwe manier van onderzoek doen. Namelijk met behulp van een chatbot vragen beantwoorden over het thema duurzaamheid. U hebt aangegeven hieraan mee te willen doen. Hartelijk dank daarvoor!

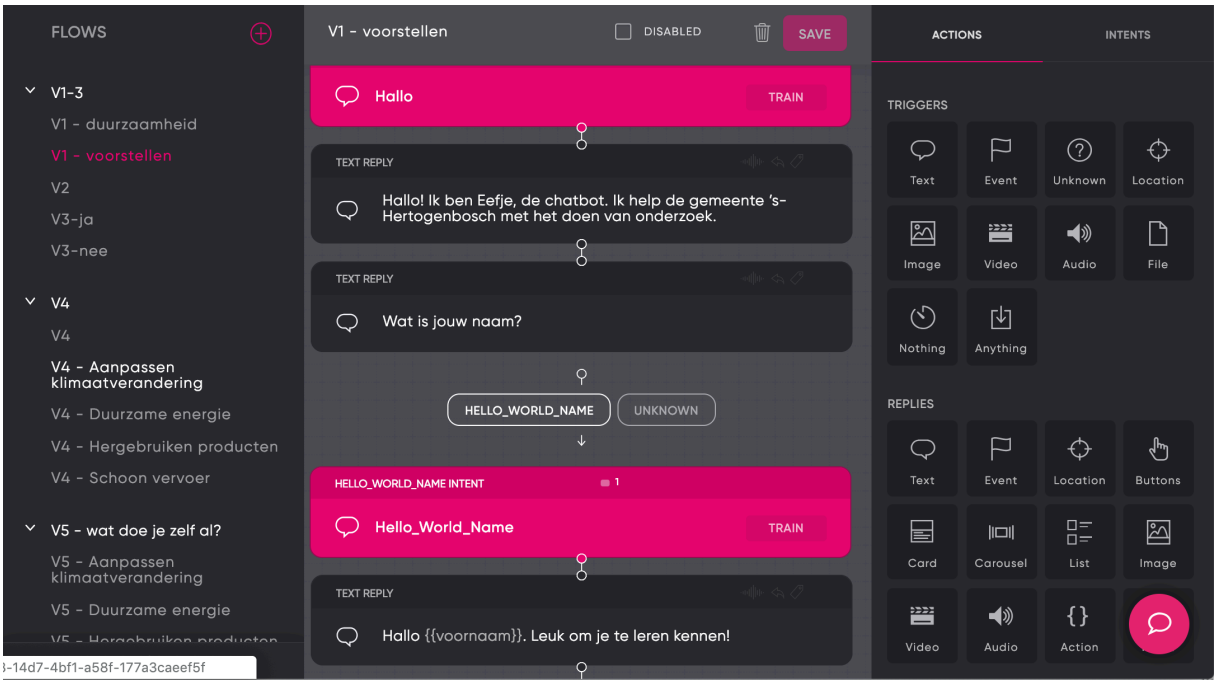
Goed om te weten

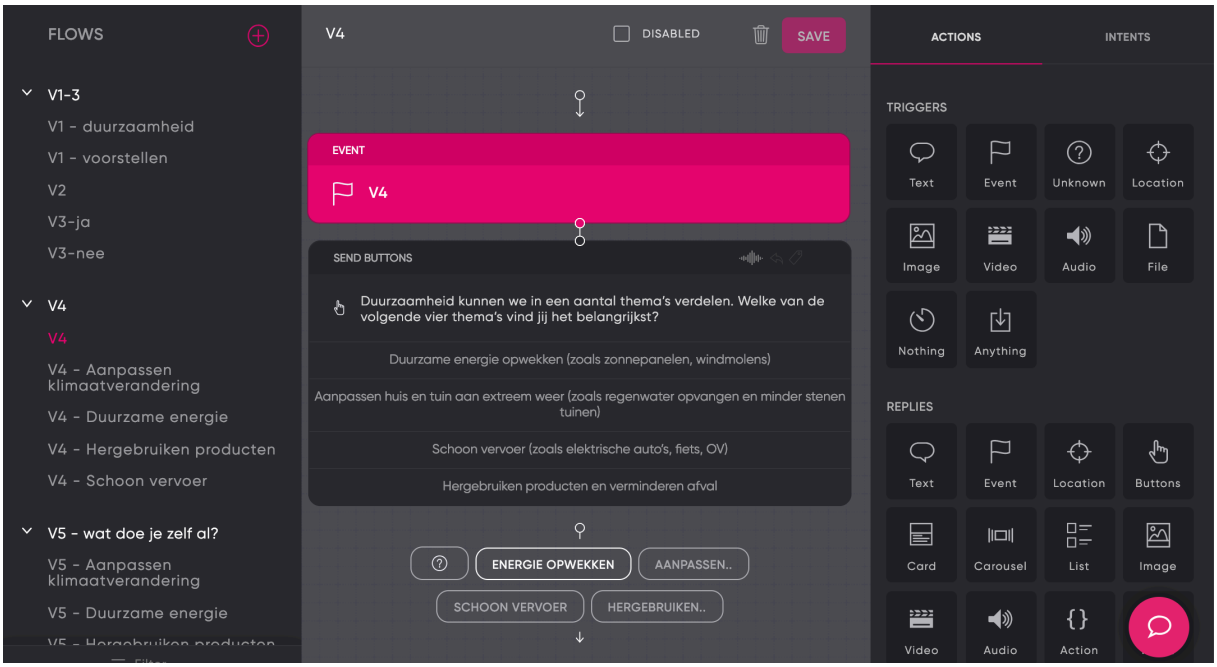
- De vragen worden gesteld via een chatvenster (zoals WhatsApp);
- U start de chatbot door 'hallo' te typen in het chatvenster;
- Nadat de chatbot een vraag heeft gesteld, typt u uw antwoord in het tekstvak;
- U verzendt uw antwoord door op 'enter' te drukken;
- Probeer uw antwoord in één bericht te sturen. Meerdere losse berichten begrijpt de chatbot niet;
- Het kan gebeuren dat de chatbot in herhaling valt. Ga dan door met het gesprek.
- Hebt u een vraag? Stuur dan een e-mail naar digipanel@s-hertogenbosch.nl

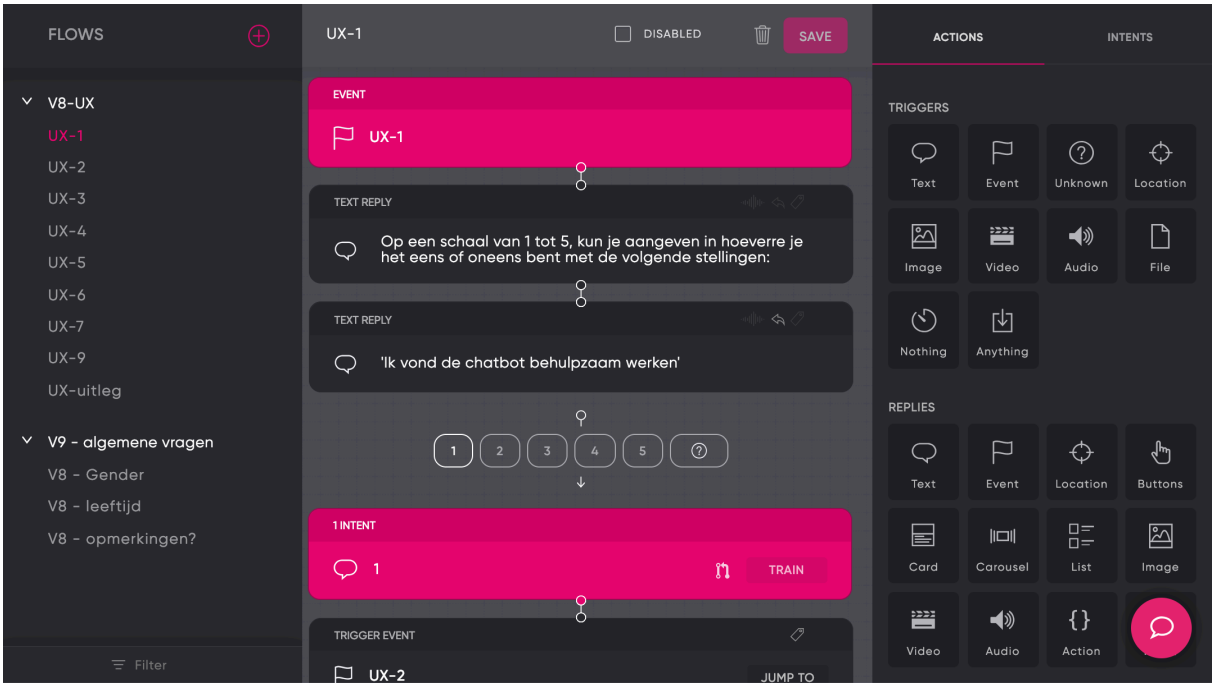
Het onderzoek is ontwikkeld door onze stagiair Eefje, samen met de gemeente, Universiteit van Tilburg en Flow.ai. Uiteraard gaan wij vertrouwelijk met de resultaten om. Het onderzoek duurt ongeveer 5-10 minuten.

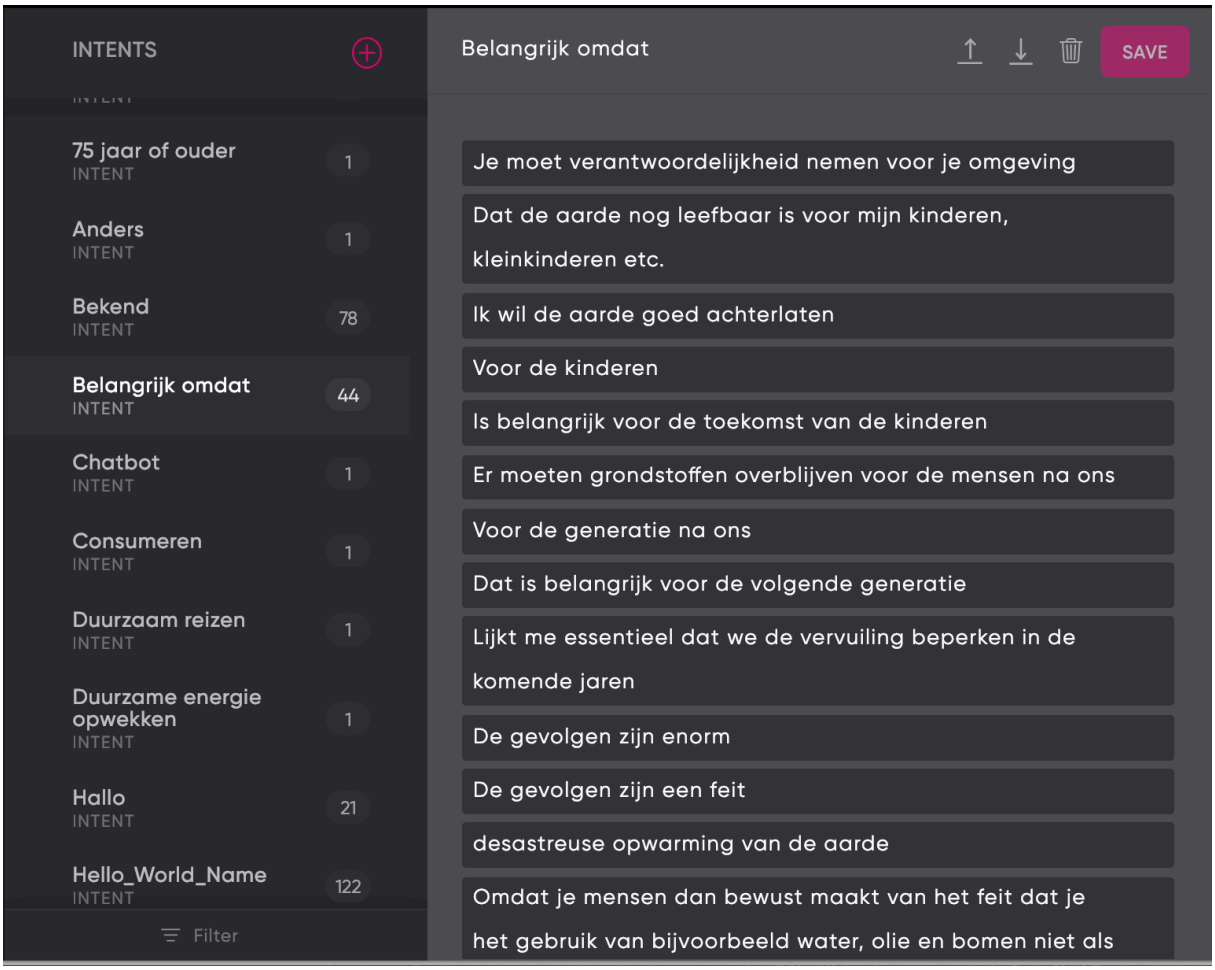
U kunt tot en met **woensdag 28 november** aan het onderzoek meedoen.
Klik op deze link om het onderzoek te starten:
https://widget.flow.ai/try/ODYyMjIzZGMtMWQxZS00NWJhLWI3ODUtN2ZhYTY2ODVlMjcwfGY2NjE3ODMzLWI3N2MtNDhiNi05YjU4LWY0YmI5ZjMyYTgyOA

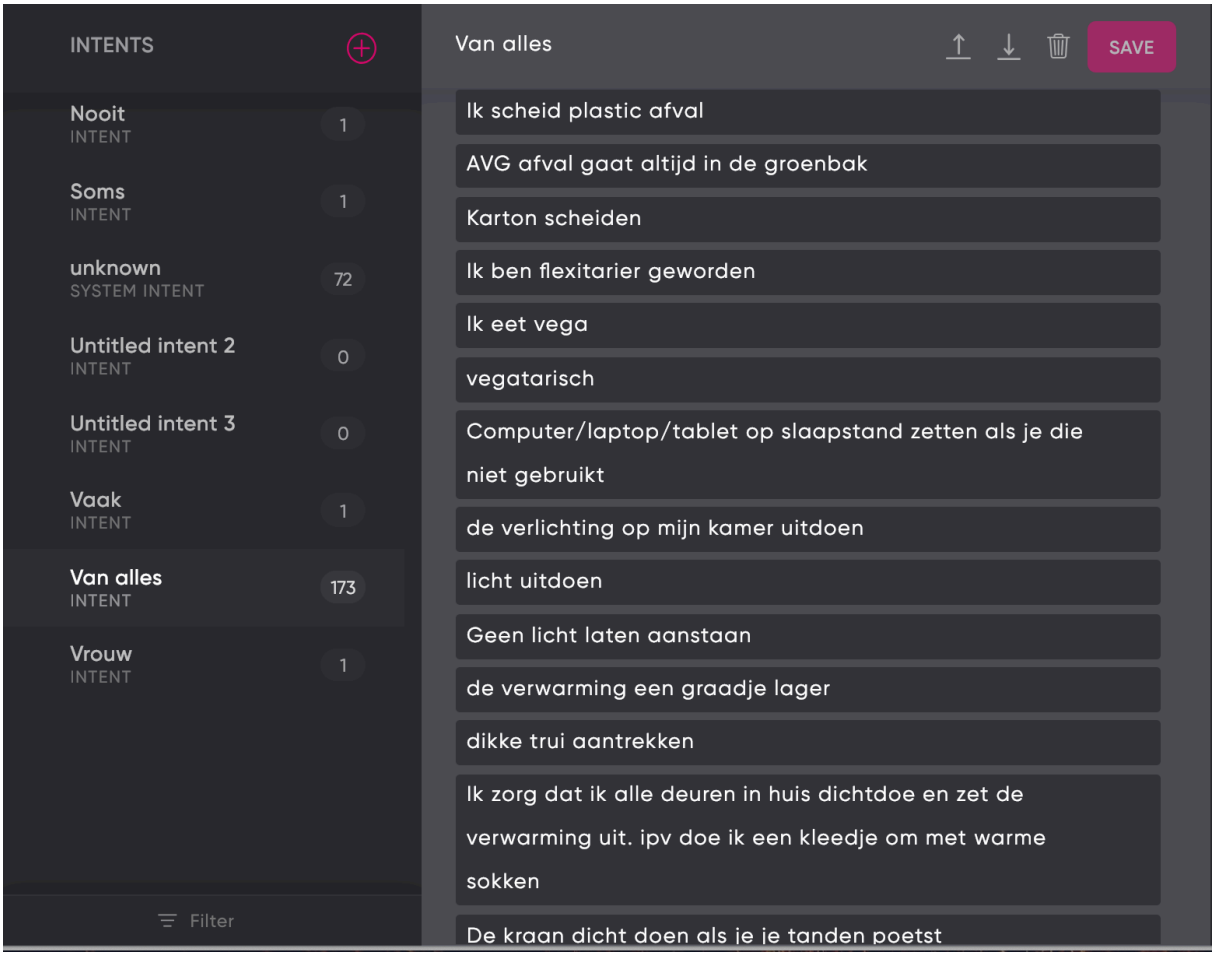
Appendix 2. Examples of the conversational flow and intent recognition in the User Interface of Flow.ai






FLOWS
V8-UX
UX-1
UX-2
UX-3
UX-4
UX-5
UX-6
UX-7
UX-9
UX-uitleg
V9 - algemene vragen
V8 - Gender
V8 - leeftijd
V8 - opmerkingen?
Filter
UX-1
DISABLED
SAVE
EVENT
UX-1
TEXT REPLY
Op een schaal van 1 tot 5, kun je aangeven in hoeverre je het eens of oneens bent met de volgende stellingen:
TEXT REPLY
'Ik vond de chatbot behulpzaam werken'
1
2
3
4
5
1 INTENT
1
TRAIN
TRIGGER EVENT
UX-2
JUMP TO
ACTIONS
INTENTS
TRIGGERS
Text
Event
Unknown
Location
Image
Video
Audio
File
Nothing
Anything
REPLIES
Text
Event
Location
Buttons
Card
Carousel
List
Image
Video
Audio
Action


INTENTS
75 jaar of ouder
INTENT
1
Anders
INTENT
1
Bekend
INTENT
78
Belangrijk omdat
INTENT
44
Chatbot
INTENT
1
Consumeren
INTENT
1
Duurzaam reizen
INTENT
1
Duurzame energie opwekken
INTENT
1
Hallo
INTENT
21
Hello_World_Name
INTENT
122
Filter
Belangrijk omdat
SAVE
Je moet verantwoordelijkheid nemen voor je omgeving
Dat de aarde nog leefbaar is voor mijn kinderen, kleinkinderen etc.
Ik wil de aarde goed achterlaten
Voor de kinderen
Is belangrijk voor de toekomst van de kinderen
Er moeten grondstoffen overblijven voor de mensen na ons
Voor de generatie na ons
Dat is belangrijk voor de volgende generatie
Lijkt me essentieel dat we de vervuiling beperken in de komende jaren
De gevolgen zijn enorm
De gevolgen zijn een feit
desastreuse opwarming van de aarde
Omdat je mensen dan bewust maakt van het feit dat je het gebruik van bijvoorbeeld water, olie en bomen niet als


INTENTS
Nooit
INTENT
1
Soms
INTENT
1
unknown
SYSTEM INTENT
72
Untitled intent 2
INTENT
0
Untitled intent 3
INTENT
0
Vaak
INTENT
1
Van alles
INTENT
173
Vrouw
INTENT
1
Filter
Van alles
SAVE
Ik scheid plastic afval
AVG afval gaat altijd in de groenbak
Karton scheiden
Ik ben flexitarier geworden
Ik eet vega
vegatarisch
Computer/laptop/tablet op slaapstand zetten als je die niet gebruikt
de verlichting op mijn kamer uitdoen
licht uitdoen
Geen licht laten aanstaan
de verwarming een graadje lager
dikke trui aantrekken
Ik zorg dat ik alle deuren in huis dichtdoe en zet de verwarming uit. ipv doe ik een kleedje om met warme sokken
De kraan dicht doen als je je tanden poetst

Appendix 3. Experimental design

De gemeente wil u graag wat vragen stellen over duurzaamheid.

- Personalisatie: Hallo ik ben Eefje; wat is jouw naam?; Hallo, [naam], ik wil je graag wat vragen stellen over duurzaamheid
- Informeel: Hallo! Dit is de chatbot die de gemeente (…)
- Uitnodigend: Hallo! Dit is de chatbot die de gemeente (…)
- Chatbot 5:
  - Hallo! Ik ben Eefje, de chatbot. Ik help de gemeente 's-Hertogenbosch met het doen van onderzoek; wat is jouw naam?; Hallo [naam]. Leuk om je te leren kennen! Ik wil je graag wat vragen stellen over duurzaamheid.

1: In het nieuwste gemeentelijke bestuursakkoord is duurzaamheid een belangrijk onderwerp. Wat verstaat u onder duurzaamheid?

- Personalisatie: (…) wat versta jij onder duurzaamheid?
- Informeel: (…) wat versta jij onder duurzaamheid?
- Uitnodigend: (…) wat versta jij onder duurzaamheid?
- CH5: Wat versta jij onder duurzaamheid?

→ **A**: "Duurzaamheid betekent…"

- Personalisatie: Interessant om te horen. Voor mij betekent duurzaamheid (…)
- Informeel: Oh! Interessant om te horen! Voor de gemeente betekent duurzaamheid (…)
- Uitnodigend: Interessant om te horen. Voor de gemeente betekent duurzaamheid (…)
- CH5: Oh! Interessant om te horen! Voor mij betekent duurzaamheid (…)

→ **(?)**:

- Personalisatie: Voor mij betekent duurzaamheid (…)
- Informeel: Voor de gemeente betekent duurzaamheid (…)
- Uitnodigend: We kunnen ons voorstellen dat dit een lastige vraag is (…) Voor de gemeente betekent duurzaamheid (…)
- CH5: Ik kan me voorstellen dat dit een lastige vraag is om te beantwoorden Voor mij betekent duurzaamheid (…)

2: Vindt u duurzaamheid belangrijk?

- Personalisatie: Vind jij duurzaamheid belangrijk?
- Informeel: …Vind jij duurzaamheid belangrijk?
- Uitnodigend: Vind jij duurzaamheid belangrijk?
- CH5: …Vind jij duurzaamheid belangrijk?

→ **Ja**
→ **Nee**
→ **(?)** Hebt u op 'ja' of 'nee' gedrukt? Dat is wel de bedoeling, anders begrijpt de chatbot u niet. Hier komt de vraag nog een keer:
- Gepersonaliseerd: Heb je wel op 'ja' of 'nee' gedrukt? Dat is wel de bedoeling, anders begrijp ik je niet. Hier komt de vraag nog een keer:
- Informeel: Heb je wel op 'ja' of 'nee' gedrukt? Dat is wel de bedoeling, anders begrijpt de chatbot je niet :(Hier komt de vraag nog een keer:
- Uitnodigend: Heb je wel op 'ja' of 'nee' gedrukt? Dat is wel de bedoeling, anders begrijpt de chatbot je niet, excuses hiervoor. Hier komt de vraag nog een keer:
- CH5: Heb je wel op 'ja' of 'nee' gedrukt? Dat is wel de bedoeling, anders begrijp ik je niet, sorry :(Hier komt de vraag nog een keer:"

3: Waarom vindt u dat belangrijk/ niet belangrijk?

- Personalisatie: Waarom vind jij duurzaamheid belangrijk/ niet belangrijk, [naam]?
- Informeel: Waarom vind jij duurzaamheid belangrijk/ niet belangrijk?
- Uitnodigend: Kun je daar iets meer over vertellen? Waarom vindt je dat (…)?
- CH5: Kun je daar wat meer over vertellen, [naam]? Waarom vind je dat (…)?

→ **A**: "belangrijk omdat…"
- Personalisatie: Ik heb het begrepen.
- Informeel: Ah zo, we hebben het begrepen :)
- Uitnodigend: We begrijpen wat je bedoelt. Bedankt voor je antwoord.
- CH5: Ik begrijp precies wat je bedoelt. Bedankt voor je antwoord :)

→ **A**: "niet belangrijk omdat…"
- Personalisatie: Ik heb het begrepen.
- Informeel: Ah zo, we hebben het begrepen :)
- Uitnodigend: We begrijpen het. Bedankt voor je antwoord.
- CH5: Ah zo, ik begrijp het. Bedankt voor je antwoord :)

→ **(?)**:
- Personalisatie: Ik heb het begrepen.
- Informeel: Ah zo, we hebben het begrepen :)
- Uitnodigend: We begrijpen het. Dank voor je antwoord.
- CH5: Ik begrijp het. Dank voor je antwoord :)

4: Duurzaamheid kunnen we in een aantal thema's verdelen. Welke van de volgende vier thema's vindt u het belangrijkst?
→ personalisatie: Ik vind deze ook heel belangrijk!
→ informeel: Wij vinden deze ook belangrijk!
→ uitnodigend: Wij vinden deze inderdaad ook belangrijk!
→ CH5: Ik vind deze inderdaad ook heel belangrijk!

5: Waarom vindt u [x] het belangrijkst?
- Personalisatie: Waarom vind jij [x] het belangrijkst?
- Informeel: Waarom vind jij [x] het belangrijkst?
- Uitnodigend: Kun je daar wat meer over vertellen? Waarom vind jij [x] het belangrijkst?
- CH5: Kun je daar wat meer over vertellen? Waarom vind jij [x] het belangrijkst?

6: Wat doet u zelf al met [x]?
- Personalisatie: Ik ben heel benieuwd: wat doe jij zelf al met [x]?
- Informeel: Wij zijn heel benieuwd: wat doe jij zelf al met [x]?
- Uitnodigend: Wij zijn heel benieuwd: wat doet jij zelf al met [x]?
- CH5: Ik ben heel benieuwd: wat doe jij zelf al met [x]?

→ **A**: "heel veel"
- Personalisatie: Goed bezig, [naam]!
- Informeel: Goed bezig!
- Uitnodigend: Goed bezig!
- CH5: Goed bezig, [naam]!

→ **A**: "niets"
(in geval van duurzame energie opwekken): "Misschien leuk om te weten: op de website van energiesubsidiewijzer kun je kijken of jij subsidie kunt aanvragen voor het aanleggen van duurzame energie in jouw huis"
(in geval van aanpassen aan extreem weer): "Misschien leuk om te weten: de gemeente had dit jaar een geveltuintjesactie. Er zijn er maar liefst 100 gratis weggegeven!"

(in geval van schoon vervoer): "Misschien leuk om te weten: Arriva biedt sinds kort het stadsbuskaartje aan. Hiermee kun je in het weekend voor drie euro (met twee kinderen) de binnenstad in!"
(in geval van hergebruiken producten): "Het is soms lastig om te weten waar een keurmerk precies voor staat. Hiervoor kun je bijvoorbeeld kijken naar de Keurmerkenwijzer van Milieu Centraal."

→ **(?):** *(hetzelfde als bij "niets")*

7: Wat doe je verder nog meer met duurzaamheid?

- Personalisatie: En wat doe je verder nog meer met duurzaamheid?
- Informeel: En wat doe je verder nog meer met duurzaamheid?
- Uitnodigend: En wat doe je verder nog meer met duurzaamheid?
- CH5: En wat doe je verder nog meer met duurzaamheid?

→ **A**: "heel veel"

- Personalisatie: Jij verdient een duurzame pluim!
- Informeel: Jij verdient een duurzame pluim! ;)
- Uitnodigend: Jij verdient een duurzame pluim!
- CH5: Jij verdient een duurzame pluim! ;)

→ **(?):**

- Personalisatie: Ik heb het begrepen
- Informeel: Ik heb het begrepen :)
- Uitnodigend: Dank voor je antwoord.
- CH5: "Ik heb het begrepen. Dank voor je antwoord."

8: We hadden een hele warme, droge zomer dit jaar. Dit is een voorbeeld van extreem weer. Hoe vaak maakt u zich zorgen over klimaatverandering?

- Personalisatie: Wat hadden we dit jaar een warme, droge zomer. Dat is een voorbeeld van extreem weer. Hoe vaak maak jij je zorgen over klimaatverandering?
- Informeel: Wat hadden we dit jaar een warme, droge zomer hè? Dat is een voorbeeld van extreem weer. Maak jij je vaak zorgen over klimaatverandering?
- Uitnodigend: Wat hadden we dit jaar een warme, droge zomer. Dat is een voorbeeld van extreem weer. Maak jij je vaak zorgen over klimaatverandering?
- CH5: Wat hadden we dit jaar een warme, droge zomer hè? Dat is een voorbeeld van extreem weer. Maak jij je vaak zorgen over klimaatverandering?

a. **Vaak**
b. **Soms**
c. **Nooit**

U bent nu klaar met de vragen over duurzaamheid.

- Personalisatie: [naam], je bent nu klaar met de vragen over duurzaamheid
- Informeel: Je bent nu klaar met de vragen over duurzaamheid :)
- Uitnodigend: Dankjewel! Je bent nu klaar met de vragen over duurzaamheid
- CH5: Dank je wel [naam]! Je bent nu klaar met de vragen over duurzaamheid :)

Er volgt nu nog een aantal stellingen over hoe u de chatbot hebt ervaren.

- Personalisatie: Er volgt nu nog een aantal stellingen over hoe jij de chatbot hebt ervaren.
- Informeel: Er volgt nu nog een aantal stellingen over hoe jij de chatbot hebt ervaren...
- Uitnodigend: Er volgt nu nog een aantal stellingen over hoe jij de chatbot hebt ervaren.
- CH5: Er volgt nu nog een aantal stellingen over hoe jij de chatbot hebt ervaren...

"Op een schaal van 1 tot 5, kunt u/kan je aangeven in hoeverre u/je het eens of oneens bent met de volgende stellingen:"

1) ‘Ik vond de chatbot behulpzaam werken’

2) ‘Ik vond het makkelijk om met de chatbot werken’

3) ‘Ik kon de vragen snel beantwoorden met de chatbot’

4) ‘Ik vond de chatbot overzichtelijk om mee te werken’

5) ‘Ik vond het leuk om met de chatbot te werken’

6) ‘Ik vond het interessant om met de chatbot te werken’

7) ‘Ik vond het vernieuwend om met de chatbot te werken’

Zou je vaker gebruik willen maken van een chatbot?
- ja
- nee

Dank u wel / Dankjewel!

Er volgen tot slot nog twee algemene vragen.

Bent u…?/Ben jij…?
a. **Man**
b. **Vrouw**
c. **Anders**

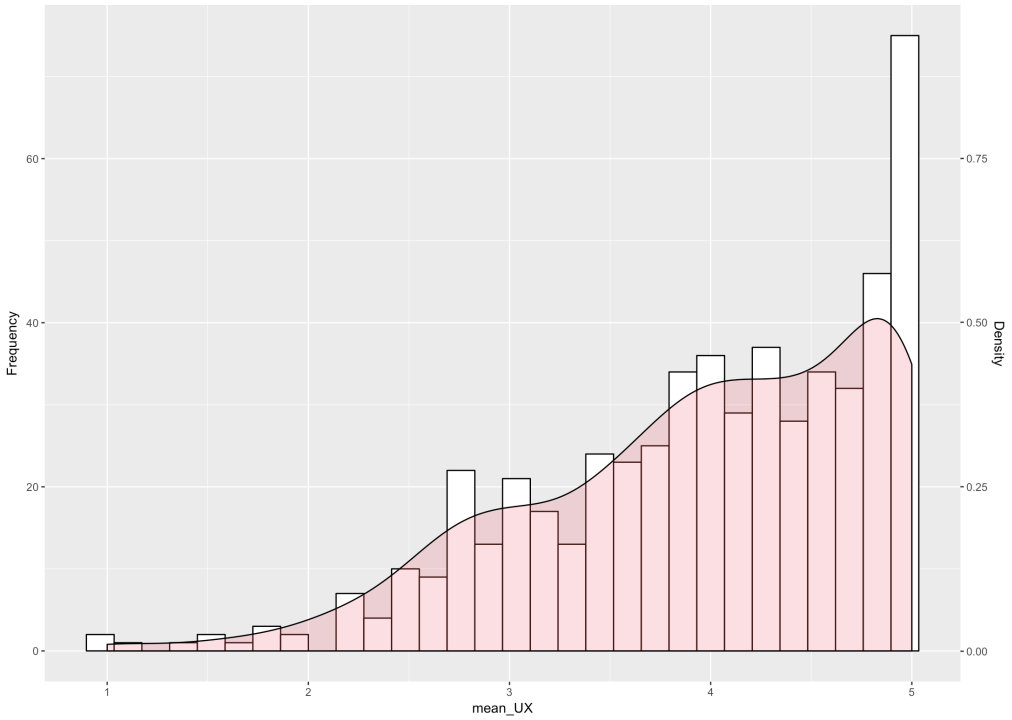
Wat is uw/jouw leeftijd?
a. **18-25 jaar**
b. **26-35 jaar**
c. **36-45 jaar**
d. **46-65 jaar**
e. **66-74 jaar**
f. **75 jaar en ouder**

Heb je verder nog opmerkingen naar aanleiding van deze chatbot? → **afsluiting**

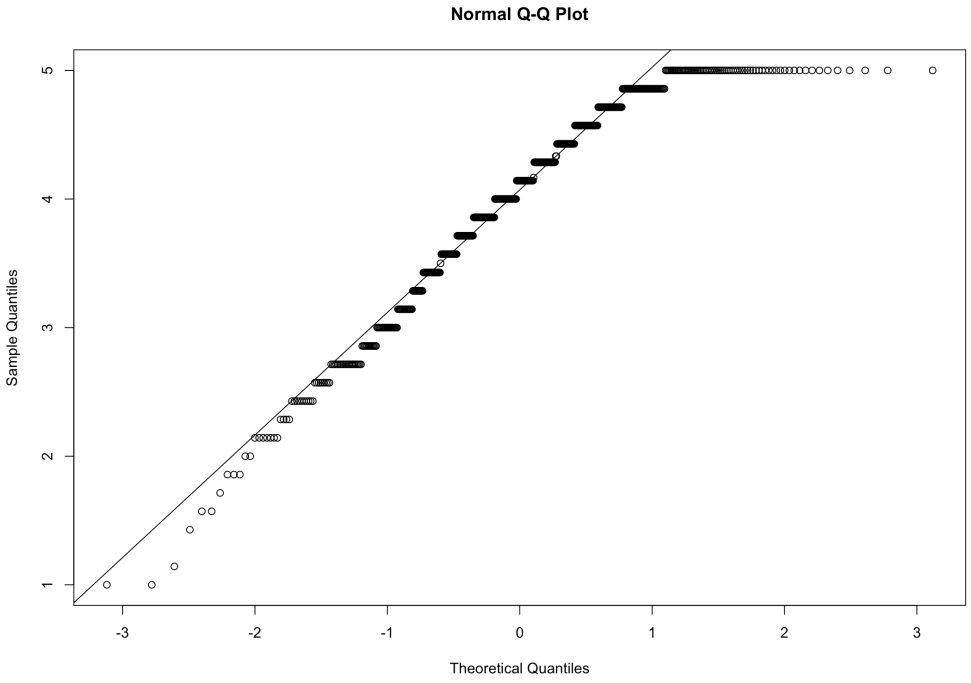
Dankjewel voor het meedoen met dit onderzoek over duurzaamheid! Leuk dat je het gebruik van chatbots bij de gemeente wilde testen.

Met jouw antwoorden gaan we vertrouwelijk om. Je mag dit scherm nu afsluiten.

Appendix 4. Density plot and Q-Qplot of the mean UX-score.

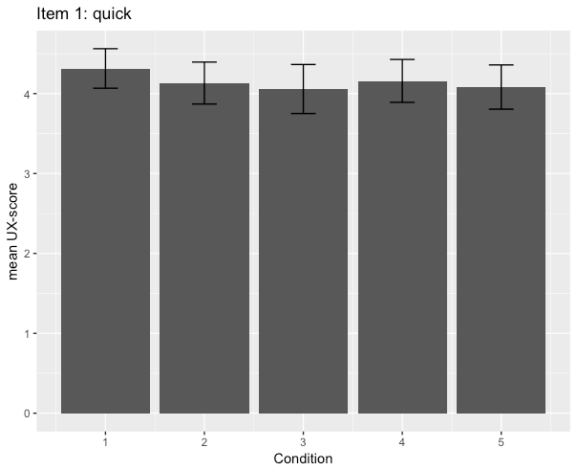


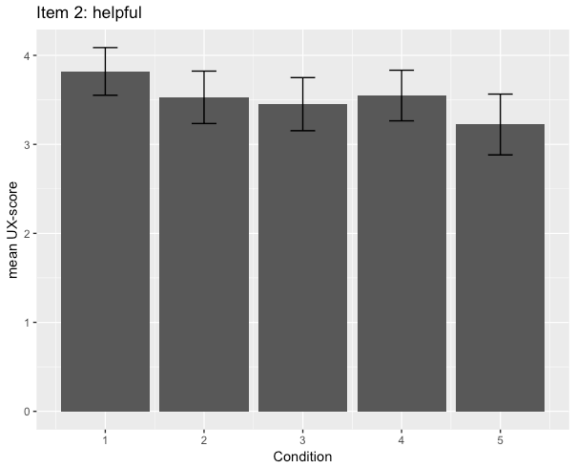
Density distribution of mean UX-score

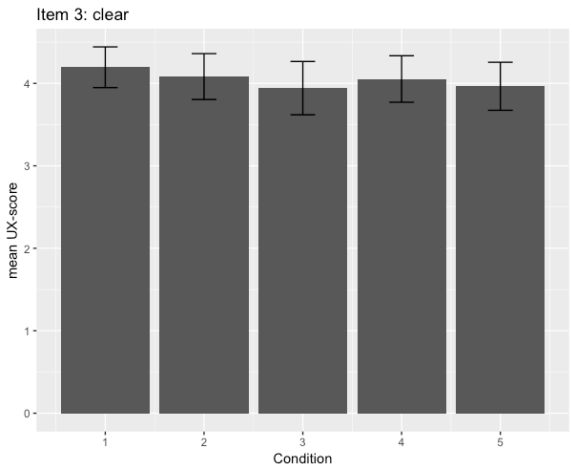


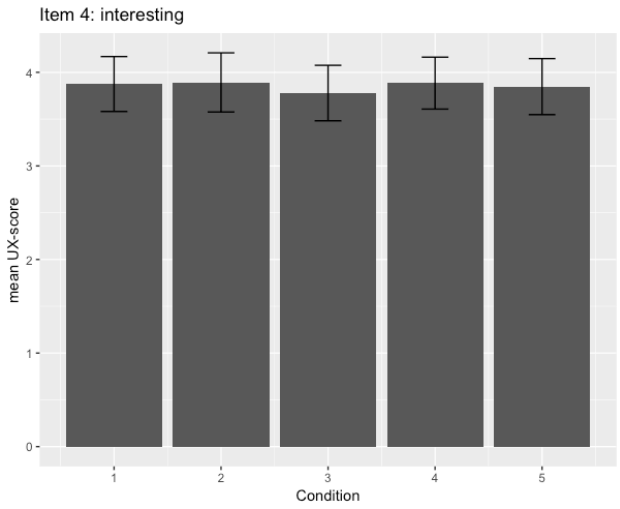
Q-Qplot of the mean UX-score

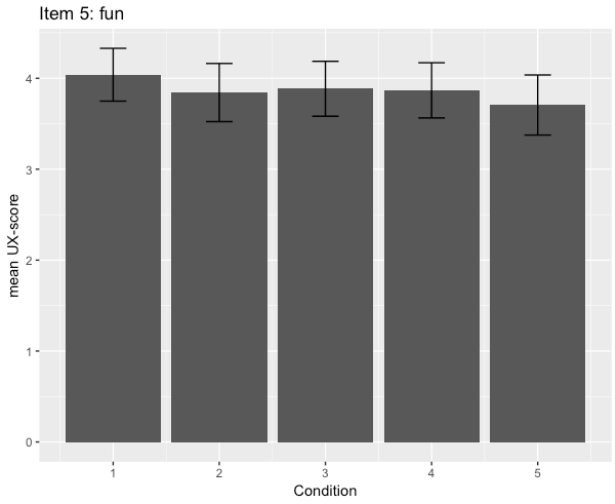
Appendix 5. Bar plots of the 7 items of User Experience

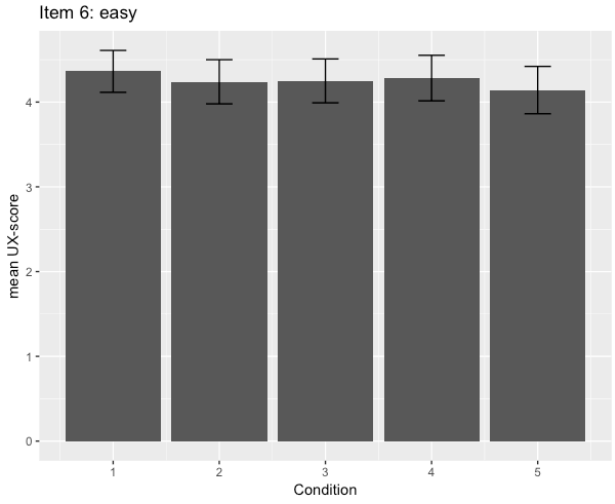


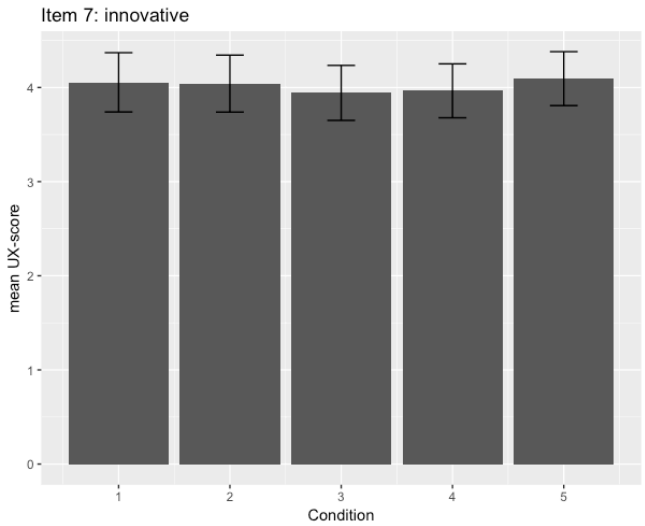





Item 4: interesting
mean UX-score
0
1
2
3
4
1
2
3
4
5
Condition


Item 5: fun
mean UX-score
0
1
2
3
4
1
2
3
4
5
Condition


Item 6: easy
mean UX-score
0
1
2
3
4
1
2
3
4
5
Condition


Item 7: innovative
mean UX-score
0
1
2
3
4
1
2
3
4
5
Condition