# Towards Personality-Aware Chatbots

**Daniel Fernau**[†] **Stefan Hillmann**[†]
**Nils Feldhus**[◇★] **Tim Polzehl**[†◇★] **Sebastian Möller**[†◇]
Technische Universität Berlin[†]
German Research Center for Artificial Intelligence (DFKI)[◇]
{firstname.lastname}@dfki.de

## Abstract

Chatbots are increasingly used to automate operational processes in customer service. However, most chatbots lack adaptation towards their users which may results in an unsatisfactory experience. Since knowing and meeting personal preferences is a key factor for enhancing usability in conversational agents, in this study we analyze an adaptive conversational agent that can automatically adjust according to a user's personality type carefully excerpted from the Myers-Briggs type indicators. An experiment including 300 crowd workers examined how typifications like extroversion/introversion and thinking/feeling can be assessed and designed for a conversational agent in a job recommender domain. Our results validate the proposed design choices, and experiments on a user-matched personality typification, following the so-called law of attraction rule, show a significant positive influence on a range of selected usability criteria such as overall satisfaction, naturalness, promoter score, trust and appropriateness of the conversation.

## 1 Introduction

In today's rapidly emerging technology-driven world, chatbots are becoming a more significant factor in customer interaction. Next to voice-driven assistants, text-based conversational agents—commonly known as chatbots—have attracted significant attention in recent years. Chatbots are designed to interact with humans using natural language and are commonly used on messaging platforms and websites (Dale, 2016; Gnewuch et al., 2018). With recent advancements in the field of artificial intelligence (AI), organizations are starting to realize the potential of chatbots to automate their customer service operations and hence reduce costs (Adam et al., 2020). Furthermore, it was predicted that 80% of organizations would have deployed a chatbot by 2020 (Sandbank et al., 2017). However, the quality of today's systems does not seem to meet customer expectations (Gnewuch et al., 2018). A key obstacle preventing most chatbots from being successful is that the interaction lacks humanness and naturalness (Schuetzler et al., 2014; Gnewuch et al., 2018). Several studies have investigated social cues and their positive effect on users' perceived social presence, trust, enjoyment, and usage intentions (Zumstein and Hundertmark, 2017; Ahmad et al., 2020). However, it has also been shown that social cues may have a negative effect that ends up irritating the user (Louwerse et al., 2005).

Studies about the nature and quality of human-machine interactions have identified personality as an essential factor for this issue (Chaves and Gerosa, 2021). Personality is a stable pattern that provides a measure for a person's behavior (und Gregory J Feist, 2002). Traditionally, personality is assessed by questionnaires; current approaches, however, make it possible to use human-generated data from social media or online forums (Boyd and Pennebaker, 2017). A person's language can provide information about the user's personality (Pennebaker and King, 1999; Boyd and Pennebaker, 2017; John et al., 1988).

To address these challenges and leverage modern technologies, the development of a personality type-indicator adaptive chatbot that automatically adapts to a user's presumed personality type is proposed in this work. The studies analyzes the impact of the so-called "law of attraction," according to which users reported higher communication interaction, human-likeliness, preference, and friendliness when interacting with a chatbot that has equal personality traits (Ahmad et al., 2020; Park et al., 2012). However, the studies introduced did not produce statistically significant results except for Ahmad et al.'s (2020) work (Ahmad et al., 2020). Their study did not require full interaction with an

★ Corresponding authors



*Proceedings of the SIGdial 2022 Conference*, pages 135–145
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 

applied chatbot, but rather examined the perception of different personalities in a chatbot by showing their participants screenshots of the interactions. In our empirical quantitative user study, we therefore evaluate how adapted personality types are perceived by chatbot users for the domain of a job recommender chatbot and whether or not personality type-based adaptation can lead to higher overall satisfaction, usability, trust, and appropriateness.

Furthermore, there exist very few works about design criteria for how to realize personality in terms of chatbot design. This paper seeks to contribute to this area by giving design implementation details.

## 2 Related Work

### 2.1 Personality and MBTI Typification

Looking in the psychologically motivated literature of personality assessment and analysis the predominantly used model is the so called five factor model (FFM) (McCrae and Costa, 1987; McCrae and John, 1992). However, and despite overt scientific criticism, e.g. (Pittenger, 1993; Boyle, 1995), when looking into concurrent practical application outside the scientific community the application of Myers-Briggs Type Indicator (MBTI) as a pre-employment assessment in career and job seeking processes, all originating to (McCaulley and Martin, 1995), has gained substantial popularity. In this work, we therefore adopt and extend the principles of MBTI typification into a job recommender chatbot interaction while taking good care of MBTI validity and type indicator selection for our experiments. MBTI is a personality theory classifying people into the combination of four types resulting in one of 16 distinct classifications (McCrae and Costa, 1987), rather than continuous dimensions native to FFM. This distinction leads to a difference in the meaning of each combination. The MBTI consists of four dichotomies: Extroversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P) (Myers-Briggs et al., 1998). (McCrae and Costa Jr, 1989) examined the degree of empirical convergence between the Big 5 and the MBTI. Their results show that each MBTI type is correlated to at least one Big 5 trait. The largest study in Furnham (1996) shows large correlations[1] for I/E with Extroversion, and P/J with Conscientiousness and medium correlation between N/S with Openness and T/F with Agreeableness.

| **Introversion** | **Extroversion** |
|---|---|
| problem talk | pleasure talk |
| single topic | many topics |
| few semantic errors | many semantic errors |
| *few self-references* | *many self-references* |
| *formal* | *informal* |
| *many tentative words* | *few tentative words* |
| *many nouns*, *adjectives* | *many verbs*, *adverbs* |
| *prepositions* | *pronouns* |
| *many words per sentence* | *few words per sentence* |
| *many articles* | few articles |
| many negations | few negations |
| *few positive words* | *many positive emojis* |
| *less emojis* | *few negative emotions* |
| *many negative emotions* | *affiliative humor* |
| (*bad emojis*) | |

(cues in *italic* were used in our study)

Table 1: Overview of linguistic cues for I/E as by (Ruane et al., 2020; Mairesse et al., 2007; Pennebaker and King, 1999; Mehl et al., 2006; Scherer, 1979; Furnham, 1990; Gill and Oberlander, 2002)

### 2.2 Link between Personality and Language

According to John et al. (1988), a modern approach to infer personality is inferring it from language based on the lexical hypothesis (John et al., 1988). Over the years, subsequent research has refined this theory. As a system, the lexical hypothesis is considered to be a general approach with implications for cross-cultural diversity, cognitive theories, and other areas of psychology (Digman, 1990). The hypothesis states that each person has different opinions and preferences which are expressed in a person's language (John et al., 1988). Thus, in language analysis based on personality vocabulary, one should use a clearly defined list of the most important characteristics (John et al., 1988). Which characteristics to utilize to design a chatbot's personality is explained in the following.

Prior work has mapped linguistic cues for each of the personality traits (Boyd and Pennebaker, 2017; Pennebaker and King, 1999; Mairesse et al., 2007; Ruane et al., 2020). As already indicated, the application of these cues in the literature is predominantly derived from the FFM, as research lacks linguistic cues based on the MBTI (Furnham,

[1] According to Cohen (1988), a correlation > 0.1 is considered as low, > 0.3 as medium, and > 0.5 as large (Cohen, 1988)

| Thinking | Feeling |
|---|---|
| *swearing* | *longer words* |
| *anger* | *shorter sentences* |
| negations | *positive emotions* |
| *references to facts* | *cheerful* |
| *less mentions to emotions* | *many self-references* |

(cues in *italic* were used in our study)

Table 2: Overview of linguistic cues for T/F as by (Ruane et al., 2020; Pennebaker and King, 1999)

1996).

Selecting carefully our experimentation scope, this study focuses on two of the four dichotomies, namely I/E and T/F, for a essential reasons. Both dichotomies show respective correlations to extroversion and agreeableness offering well established linguistic cues (Ruane et al., 2020; Mairesse et al., 2007; Pennebaker and King, 1999; Mehl et al., 2006; Scherer, 1979; Furnham, 1990; Gill and Oberlander, 2002) drawn from the FFM. The I/E dichotomy has the strongest correlation to the FFM's extroversion scale. Among all four MBTI dichotomies, however, with the correlation to the FFM being the lowest between T/F and agreeableness, there is no significant difference to the other scales when compared with McCrae and Costa's study (1989) (McCrae and Costa Jr, 1989). Table 1 and 2 show the overview of linguistic cues for extroversion and agreeableness as adapted to I/E and T/F for the presented study.

Obtaining MBTI types is typically done by questionnaires, e.g. Form M (93 items). Due to availability and transparency reasons, this study excerpts from the open-source Open Extended Jungian Type Scales (OEJTS) questionnaire (Jorgenson, 2015) provided from openpsychometrics[2].

### 2.3 The Law of Attraction

The law of attraction is the central theory to adapt a chatbot in order to achieve greater usability. According to this theory, people seek out those similar to them and prefer to interact with people with similar traits. As explained by (Infante et al., 1997), the perceived similarity is the degree to which we believe someone's characteristics are similar to our own. These characteristics can include several factors such as demographics, political views, and personality. Many studies in psychology and communication have confirmed this rule (Blankenship et al., 1984; Nass and Lee, 2001). Originating from the observations of Human-Human Interaction (HHI), this concept is frequently applied to Human-Computer Interaction (HCI) as well.

Transferred to HCI, the law of attraction states that a user prefers to interact with a computer that has matched personality types rather than mismatched ones. When matched, information from the computer has also been rated as better and more trustworthy (Zumstein and Hundertmark, 2017). Specifically for the Big 5 theory, a study found that for a sub-dimension of the trait extroversion, dominant people prefer to interact with a dominant counterpart, and vice versa for the submissive trait (Moon and Nass, 1996). Several other studies in the field of HCI also confirmed the law of attraction (Ahmad et al., 2020; Smestad, 2018; Lee and Nass, 2005). However, some studies do not support the law of attraction in the area of HCI (Isbister and Nass, 2000; Liew and Tan, 2016), suggesting that the applicability may also depend on a concrete scenario or application. A supporting argument comes from the field of Human-Robot Interaction (HRI), e.g. the analysis of task dependency in (Tay et al., 2014).

## 3 Chatbot Design

Our personality-adaptive chatbot prototype is based on the Microsoft Azure Bot Framework and is built in the browser, allowing it to be embedded into various channels. Depending on the input personality, the respective conversation tree is activated for the task of job recommendation divided into two sub-dialogs. The first sub-dialog generally greets the user, while the second one asks job-related questions to give a personality-based recommendation.

To design the chatbot's personality type, the previously introduced linguistic cues were used. Table 3 and 4 show the applied cues including their degree for the four differently designed characters of the chatbot. For the analysis of the chatbot responses, both the Python library *spaCy*[3] and the service *Count Wordsworth*[4] were utilized. In contrast to other studies, this table enhances the transparency of the degree of linguistic cues applied, whereas related work oftentimes does not include a description of exact design choices.

[2]The Open Extended Jungian Type Scales (OEJTS) can be accessed under: https://openpsychometrics.org/tests/OEJTS developed by Jorgenson (Jorgenson, 2015)

[3]https://spacy.io

[4]https://countwordsworth.com

| | ET | EF | IT | IF |
|---|---|---|---|---|
| **Manipulating E and I** | | | | |
| Percentage of I/we | 3.81% | 3.83% | 2.41% | 2.67% |
| I, me | 12 | 11 | 11 | 10 |
| first person | 44 | 41 | 32 | 30 |
| verbs | 54 | 44 | 64 | 53 |
| verbs by WR | 17.14% | 15.33% | 14.04% | 14.17% |
| adverbs | 19 | 20 | 24 | 17 |
| adverbs by WR | 6.03% | 6.97% | 5.26% | 4.55% |
| pronouns | 51 | 48 | 66 | 48 |
| pronouns by WR | 16.19% | 16.72% | 14.47% | 12.83% |
| affiliative humor | 1 | 1 | 0 | 0 |
| informal words | 18 | 13 | 1 | 2 |
| Total words | 315 | 287 | 456 | 374 |
| articles | 6 | 4 | 36 | 29 |
| articles by IR | 1.90% | 1.39% | 7.89% | 7.75% |
| nouns | 41 | 23 | 80 | 69 |
| nouns by IR | 13.02% | 8.01% | 17.54% | 18.45% |
| adjectives | 13 | 18 | 39 | 30 |
| adjectives by IR | 4.13% | 6.27% | 8.55% | 8.02% |
| prepositions | 35 | 31 | 83 | 62 |
| prepositions by IR | 11.11% | 10.80% | 18.20% | 16.58% |
| tentative words* | 2 | 1 | 8 | 9 |
| third person (formality) | 5 | 5 | 11 | 7 |
| **Manipulating T and F** | | | | |
| words per sentence | 8.75 | 8.46 | 12.32 | 11.32 |
| emojis emotion negative | 2 | 0 | 2 | 0 |
| words related to swearing/anger | 6 | 0 | 3 | 0 |
| aggressive humor | 0 | 0 | 1 | 0 |
| references to facts | 2 | 0 | 2 | 0 |
| average length of words | 3.98 | 4.11 | 4.54 | 4.70 |
| words related to emotion | 8 | 13 | 5 | 11 |
| emojis emotion positive | 7 | 25 | 0 | 1 |
| emojis neutral | 13 | 26 | 0 | 0 |
| neutral humor | 0 | 0 | 0 | 1 |

WR: word ration, IR: interaction ratio, *e.g. *would/could*

Table 3: Linguistic cues applied for personality expression

Overall, due to the short nature of the interactions, the metrics concerning word counts, sentence length, and word length were hard to manipulate when designing the messages, as there were too many dependencies on other metrics such as references to facts.

## 4 User Study

The experiment consists of five steps. (1) Users fill out a short 12-item personality self-report according to *OEJTS*. (2) Participants interact with our chatbot, with random assignment of matched or mismatched personality type. (3) Users assess first interaction by nine usability items. (5) Participants again interact with our chatbot, this time seeing the alternative personality type as in step 2. (6) User again assess nine usability items plus questions on preference of one version over the other.

| **Report**: | ET | EF | IT | IF |
|---|---|---|---|---|
| **Metrics of linguistic cues where T higher than F** | | | | |
| words per sentence | 8.75 | 8.46 | 12.32 | 11.32 |
| emojis emotion negative | 2 | 0 | 2 | 0 |
| words related to swearing/anger | 6 | 0 | 3 | 0 |
| aggressive humor | 0 | 0 | 1 | 0 |
| references to facts | 2 | 0 | 2 | 0 |
| **Metrics of linguistic cues where F higher than T** | | | | |
| average length of words | 3.98 | 4.11 | 4.54 | 4.70 |
| words related to emotion | 8 | 13 | 5 | 11 |
| emojis emotion positive | 7 | 25 | 0 | 1 |
| emojis neutral | 13 | 26 | 0 | 0 |
| neutral humor | 0 | 0 | 0 | 1 |

Table 4: Overview of the metrics of linguistic cues to design personality for T/F.

The first part of the study is a survey is a 12-item personality self-report based on the OEJTS. For this study, each of the nine highest scoring items on the I/E and the T/F scales are used in this experiment. Additionally, each dichotomy has further been divided into six items of the E/I types and six items of the T/F types. The selected items were assessed by using a five-point Likert scale in between"Strongly agree," "Agree," "Neither agree nor disagree," "Disagree," and "Strongly disagree."

Depending on the users personality type, two chatbots were automatically selected to be tested in step 2 and step 5, of which one is designed to be perceived the same personality type as the user (matched), whereas the other one represents the opposite option settings (mismatched).For example, if a user is classified as EF (extroverted-feeling), they interacted with both an EF and an IT designed chatbot, in random order. The extroverted chatbot was named Carla and the introverted one was named Sophia to achieve the effect that users are more likely to share personal information if the chatbot appears to be female (Toader et al., 2020).

The topic of the interactions in step 2 and 5 is to chat about personal and job-related preferences to recommend a suitable job. The job recommendations given by the chatbot in the end of the conversation are hand crafted and based on the personality of the user. Note that we do not analyze the performance of any recommendation accuracy, nor the users' acceptance towards it. In this work, we focus on the impact of personality on the usability of the interaction explicitly. In more detail, the

conversation starts with some general questions regarding the name, origin, and personal preferences. Afterwards, the chatbot commences asking about job-related preferences. Three questions are asked that are based on additional items of the OEJTS. For example, one item of the OEJTS to measure extroversion assesses whether the user "works best in groups" or "works best alone."

Further, the chatbot is designed to be between the edges of an intra- and an interpersonal chatbot within a closed domain, offering limited functionality (Nimavat and Champaneria, 2017). Hence, the chatbot only allows the user to answer the questions instead of providing functionality that answers custom questions of the user. This limitation was explicitly clarified at the beginning of the survey to avoid false expectations. Moreover, the users have also been instructed of another limitation of the current state of chatbot prototype implementation, namely that writing multiple messages is not supported. This means that all information has to be put into a single message.

The usability questionnaire applied consists of nine items that are asked after each chatbot interaction: two items that compare both chatbots with each other and five general items about the participants. First, the nine items that are asked directly after each conversation with the chatbot are introduced. These items are split into four items derived from ITU telecommunication standardization sector (ITU-T) Recommendation P.851 (Rec, 2003), while the other five items are custom-designed. Adapted to the personality domain, four items were selected that are related to the following factors: acceptability, naturalness, and promoter score. For these items (among others), it was demonstrated that acceptability and naturalness are well generalized (Möller et al., 2007). The personality factor from ITU-T was not suitable for the experiments at hand due to the strong focus on personality type differentiation of this study. Hence, five custom items were designed to measure whether the design choices applied could be perceived by the participants when interacting with the different chatbots. These nine usability items were assessed using the same five-point Likert scale from above. In addition, two items were designed to directly compare Carla (extroverted) and Sophia (introverted) head-to-head. The first item assesses which chatbot is being perceived as more adapted toward the users' preferences, while the second asks for the general preference when comparing both directly. For both items, users had the option to choose Carla, Sophia, both, or none. Eventually, five profiling questions were asked at the end of the survey regarding gender, age range, experience with chatbots, native language, and their current profession. All items are shown in **Appendix A**, also including the items used for comparison and general profile data.

### 4.1 Participants

300 participants were recruited using the the Crowdee (Naderi et al., 2014) crowdsourcing platform [5] across the U.S., Great Britain, and Australia. Participants were paid equally by minimum floor wage based on the estimated work duration of the task at hand.

From the general profile items we see, that 90% of the participants were English native speakers.52% of the participants were women and 46% men, while a minority was diverse (1%) or did not like to share their gender (1%1). All participants were older than 18 years, and the distribution among age classes was as follows: 18–25 (20%), 26–35 (36%), 36–45 (24%), 46–55 (15%), and <55 (5%). Regarding their experience with chatbots, a minority of 13% had never been in touch with a chatbot before. Moreover, 5% use a chatbot on a daily basis, while 20% interact with one at least monthly and 62% occasionally. In total, out of 300 crowd workers who participated, 266 valid responses can be considered. 32 participants did not complete the interactions or the questionnaires, or interactions could not successfully be logged. Furthermore, 2 participants were excluded from the study as outliers due to their scores being three times higher than the interquartile range.

From a preliminary analysis of the qualitative feedback we feel confident that the participants could solve the task as expected and generally enjoyed the study. The overall tone in qualitative feedback was positive, e.g. "Carla was the best one, [...] It was cool but scary.", "Sophia was great. Sounded like a real person was on the other end.", or "It was pretty fun speaking with the first one [extroverted], she was way more accurate with her job recommendations than Sophia.".

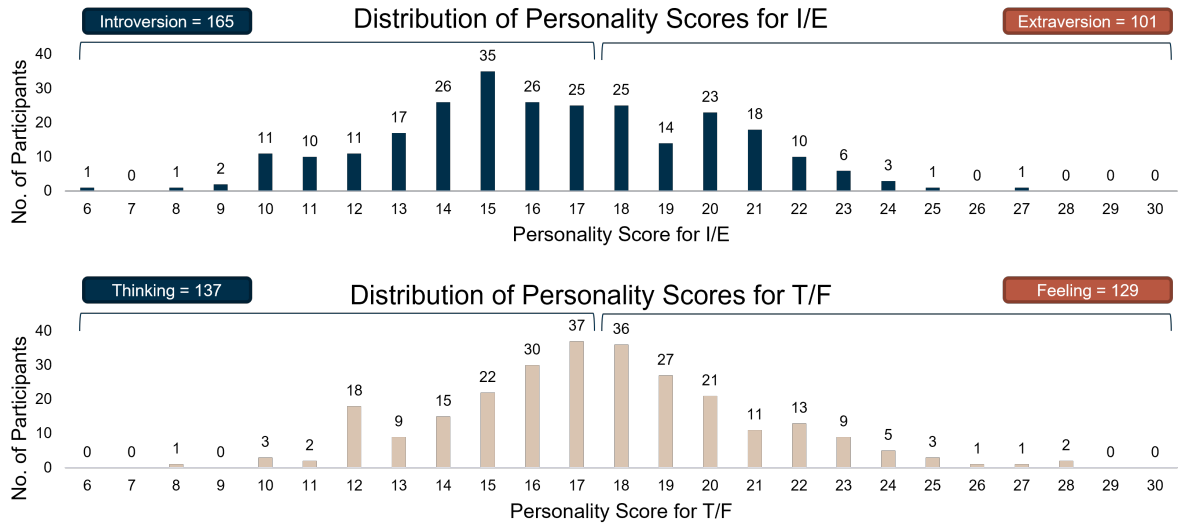


Figure 1: Distribution of personality type scores and counts, including classification boundaries; top for E/I, bottom for T/F dichotomies.

## 5 Results

### 5.1 Personality Type Distribution

Figure 1 shows the distributions of personality scores measured with the *OEJTS*. Both bar charts show the number of participants by personality score between 6 (low = **I**ntroversion or **T**hinking on the left) and 30 (high = **E**xtraversion or **F**eeling on the right), and the equal space binning threshold of 18 to differentiate the values into binary classes. The upper chart regarding I/E shows that the ratio between I and E classified participants is 62:38. More balanced is the distribution of T/F with a ratio of approximately 51:49 in the lower bar chart. All types are represented by at least 47 participants, with ET being the minority with 18% (47 participants), followed by EF with 20% (54). Among the introverted participants, IF represents 28% (75) and a majority of 34% are classified as IT (90). As no class is equal to or greater than twice the size of another, there are no imbalances in the overall distribution.

### 5.2 Results for the Law Of Attraction

In order to analyze the effect of the law of attraction, a one-sided $t$ test was used to examine the statistically significant difference between the matched and mismatched scores of Q1-9 (see Appendix **??**). The test for significance was done at the level of $\alpha = 0.05$ for the following $t$-tests. It was not necessary to apply the Bonferroni correction, as we analyze the means of different items (i.e., data) between two groups. The one-sided test was applied, as we have expected higher usability ratings for all items (Q1-9) in the matched-condition due to the law of attraction. Additionally, a Chi-square test was used to examine whether the matched bots were preferred and whether an adaption of the matched bot could be perceived when both are directly compared.

| | matched | | mismatched | |
|---|---|---|---|---|
| **Usability Item** | **mean** | **SD** | **mean** | **SD** |
| Q1 Overall Satisfaction* | **3.94** | 1.06 | 3.58 | 1.19 |
| Q2 Naturalness* | **3.68** | 1.07 | 3.45 | 1.12 |
| Q3 Promoter Score* | **3.56** | 1.12 | 3.20 | 1.19 |
| Q4 Dialogue Length | 3.50 | 0.10 | **3.53** | 1.01 |
| Q7 Trustworthiness* | **3.52** | 0.95 | 3.38 | 0.94 |
| Q9 Appropriateness* | **3.74** | 1.11 | 3.24 | 1.24 |

Table 5: Descriptive statistics ($N$=266) for Q1-4, Q7, and Q9 comparing matched with mismatched personality. * denotes a statistically significant difference of means ($p < 0.05$).

Shown in Table 5, there is a significant difference between the overall satisfaction (Q1) of the matched personality is significantly higher compared to the mismatched personality, t(265) = 4.016, $p$ = < .001, $d$ = .246. Moreover, the perceived naturalness (Q2) of the matched chatbots is significantly higher compared to the mismatched personality ones, t(265) = 2.782, $p$ = .003, $d$ = .171. Similarly, the matched personality type chatbot is more likely to be recommended to a friend (Q3)

[5] www.crowdee.com

compared to the mismatched personality, t(265) = 3.894, $p$ = < .001, $d$ = -.239. Furthermore, there is a significantly higher trustworthiness (Q7) in the matched personality than the mismatched one, t(265) = 2,015, $p$ = .022, $d$ = .124. Finally, also the matched personality scores significantly higher in appropriateness for the task at hand than the mismatched personality, t(265) = 4.572, $p$ = <.001, $d$ = .280.

These results support our assumption that a matched personality has a positive influence on the perceived usability of our job recommender chatbot. However, it seems that there is only a small effect of the matched personality adaption.

Despite explicit manipulation, results also show that no significant difference was perceived by the participants with respect to the dialogue length, t(265) = -0.373, $p$ = .355.

### 5.3 Validation of Design Choices

Table 6 shows the results of our analysis on the impact of the design choices.

The one-sided $t$ test found that the formality (Q5) of the introverted bot is significantly higher compared to the extroverted bot, t(265) = 24.571, $p$ = < .001, $d$ = 1.507. This strongly supports the assumption that the introverted bot is perceived as more formal than the extroverted, which corresponds to the design choices.

Moreover, the perceived trustworthiness of the introverted bot is significantly higher compared to the extroverted bot, t(265) = 6.840, $p$ = <.001, $d$ = .419, while there is also a significantly higher appropriateness of the introverted bot compared to the extroverted bot, t(265) = 9.190, $p$ = <.001, $d$ = -.563.

Message length (Q8) and Emotionality (Q6) were not perceived significantly differently, although messages from the introverted bot are perceived as longer compared to the extroverted bot, t(265) = -2.778, $p$ = < .003, $d$ = -.170. Finally, the bot design of Feeling (Q6) was also not perceived as significantly more emotional than the bot designed as Thinking, t(265) = -0.356, $p$ = .361.

Finally, a direct comparison of both bots was examined with a Chi-square test to assess which chatbot was perceived as most adapted to the user. The results show no significant difference between the I/E personality type and a perceived adaption in the chatbot's behavior, $\chi^2(3) = 2.523$, $p$ = .471.

| | **Introverted** | | **Extroverted** | |
|---|---|---|---|---|
| **Usability Item** | **mean** | **SD** | **mean** | **SD** |
| Q5_Formality* | **3.94** | 0.95 | 1.97 | 1.15 |
| Q7_Trustworthiness* | **3.68** | 0.83 | 3.22 | 0.10 |
| Q8_Message_Length | 3.36 | 1.10 | **3.55** | 0.96 |
| Q9_Appropriateness* | **3.94** | 0.88 | 3.04 | 1.31 |
| | **Feeling** | | **Thinking** | |
| Q6_Emotional | 2.73 | 1.01 | 3.56 | 1.06 |

Table 6: Descriptive statistics ($N$ = 266) for Q5-9 MOS comparing the introverted and extroverted bot. * denotes a statistically significant difference of means (p < 0.05).

## 6 Discussion

In general, our results and expectations are in line with the law of attraction within a text-based conversational agent (Park et al., 2012) domain such that overall satisfaction, trustworthiness and appropriateness are significantly higher for the matched personality-based chatbot.

Also, the difference between the combination of ET and IF is much smaller compared to a scenario in which the user interacts with the bots EF and IT. For the first scenario, the messages of the bots only differ by 59 words; however, the second scenario offers 87 words in message length through the overall course of the dialogue.

A set of preliminary results may shed some light on the unexpected results. When looking at a one-sided $t$ test within the sub-sample of EF and IT classified participants, the effect of perceived message length is also greater compared to the whole sample (t(144) = 3,863, $p$ < .001, $d$ = -.321). However, it is natural that the ET and IF types are more similar to each other compared to the EF and IT. Surprisingly, the differently designed emotionality of the messages did not yield significant results in terms of distinction. A possible explanation for this could be that the perception of emotionality is biased by the use of emojis, which are perceived as an emotional variable. The difference in the usage of emoticons between IF and ET is in favor of the ET type. Hence, the ET type could be perceived as more emotional given the higher number of emojis, which is also related to feeling. Therefore, similar to the design aspect of message length, the other combinations of IT and EF should show clearer results as EF is designed to be feeling and uses numerous emojis. A paired $t$ test also supports this assumption where the EF type is per-

ceived as significantly more emotional than the IT, t(144) = 1,967, $p$ = .026, $d$ = .163. Hence, there might be an interference with the usage of emojis and the relationship towards feeling that was not designed clearly enough for those participants that were interacting with ET Carla and IF Sophia. Another possible reason for the lack of perceived emotionality, in general, could be that this study designed the T/F dichotomy under the assumption that there is a correlation with Big 5's agreeableness. Due to the lack of research modelling thinking and feeling linguistically, the linguistic cues of agreeableness were used to design T/F. The two traits correlate with each other (0.47) according to a study by McCrae and Costa (1989) (McCrae and Costa Jr, 1989). Nevertheless, they are not equal, which might result in an information loss or false interpretation other than what was intended. Further, the separation of the extroverted and introverted bot is also dependent on whether they were rated as the matched or the mismatched interaction, respectively. Our study shows, the law of attraction has an impact on the perception of the two chatbots. However, a subliminal study showed that there are no major differences when analyzing the scores within the samples of only matched interactions, the samples of only mismatched interaction, and the whole sample.

When investigating the results, regardless of the matched or mismatched personality, the introverted and formal-designed chatbots (introverted Sophia) were rated higher than the more informal ones (extroverted Carla). This also fits into the domain of job recommendation which is usually associated with professionalism where formality is required. The more formal bot also scores better on appropriateness and overall satisfaction.

For the evaluation, 266 people have interacted with it in a realistic scenario, and have rated the interaction by means of MOS. Similar studies either did not provide a direct interaction with the chatbot (Ahmad et al., 2020) (users only rated screenshots) or could only show tendencies with small sample sizes (Smestad, 2018; Ruane et al., 2020). Hence, to the best of our knowledge, this is the first study to show a statistically significant positive effect, though small, of automatically adapted matched personality of a chatbot (N = 266) toward usability, trust, and appropriateness for the task of job recommendation.

In addition, linguistic cues that correlate with certain personality traits were introduced (Pennebaker and King, 1999; Mairesse et al., 2007; Ruane et al., 2020) and the results presented in this paper further contribute to this body of research. They indicate that personality differences embodied in language were significantly perceived in two out of three design choices. These findings further validate that matched personality results in significantly higher usability scores (in all but one of the items used in our study) of a chatbot. Apart from that, trustworthiness and appropriateness (for the task of job recommendation) were also shown to be significantly better when matching the personality type compared to mismatching it. Our results are in line with previous research (Moon and Nass, 1996; Ahmad et al., 2020; Smestad, 2018; Lee and Nass, 2005; Zumstein and Hundertmark, 2017), while at the same time quantitatively demonstrating the effect of the law of attraction for a high number of participants (Park et al., 2012). In contrast to other studies, our study enhances the transparency of the degree of linguistic cues applied by precisely stating the numbers of linguistic cues; related work on chatbots with personality only described their exact measures briefly.

### 6.1 Future Research

In future work, we aim to examine whether a chatbot that automatically classifies the user's personality could become more accurate over time with a growing body of textual language to result in a personalized user experience. Additionally, it would be interesting to apply natural language generation (NLG) for the matched response generation of the chatbot to achieve even higher usability scores and higher overall flexibility. A similar approach to automatically create utterances that express a certain personality was developed with PERSONAGE (Mairesse and Walker, 2010).

A potential practical future experiment could be the steady recalculation of the user's personality for the saved conversation logs. This would allow a personality classification model to iteratively verify the user's personality traits with increasing text size. Based on the assumption that larger text samples will improve the accuracy of the predicted personality, the usability of the system could also be improved over time while it is in usage. However, storing the users' texts in business contexts to calculate their personality raises ethical as well as legal questions which have to be studied too.

Eventually, a more dedicated work comparing the selected dichotomies from MBTI along their impact on usability to scales and constructs derived from the FFM would be desirable in order to contribute to further personality theory validation.

## References


Martin Adam, Michael Wessel, Alexander Benlian, et al. 2020. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 9(2):204.

Rangina Ahmad, Dominik Siemon, and Susanne Robra-Bissantz. 2020. Extrabot vs introbot: The influence of linguistic cues on communication satisfaction. In *AMCIS*, pages 0–10. Researchgate.

Virginia Blankenship, Steven M Hnat, Thomas G Hess, and Donald R Brown. 1984. Reciprocal interaction and similarity of personality attributes. *Journal of Social and Personal Relationships*, 1(4):415–432.

Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: a new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68.

Gregory J Boyle. 1995. Myers-briggs type indicator (mbti): some psychometric limitations. *Australian Psychologist*, 30(1):71–74.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences. hoboken.

Robert Dale. 2016. The return of the chatbots. *Natural Language Engineering*, 22(5):811–817.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Adrian Furnham. 1990. Faking personality questionnaires: Fabricating different profiles for different purposes. *Current psychology*, 9(1):46–55.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and individual differences*, 21(2):303–307.

Alastair J Gill and Jon Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24, page 6. eScholarship.

Ulrich Gnewuch, Stefan Morana, Marc TP Adam, and Alexander Maedche. 2018. "the chatbot is typing…"–the role of typing indicators in human-chatbot interaction. In *Proceedings of the 17th Annual Pre-ICIS Workshop on HCI Research in MIS*, pages 0–5. Researchgate.

Dominic A Infante, Andrew S Rancer, and Deanna F Womack. 1997. Building communication theory. *Waveland PressInc.*

Katherine Isbister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International journal of human-computer studies*, 53(2):251–267.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

Eric Jorgenson. 2015. Development of the Open Jungian Type Scales.

Kwan-Min Lee and Clifford Nass. 2005. Social-psychological origins of feelings of presence: Creating social presence with machine-generated voices. *Media Psychology*, 7(1):31–45.

Tze Wei Liew and Su-Mae Tan. 2016. Virtual agents with personality: Adaptation of learner-agent personality in a virtual learning environment. In *2016 Eleventh International Conference on Digital Information Management (ICDIM)*, pages 157–162. IEEE, The International Review of Research in Open and Distributed Learning.

Max M Louwerse, Arthur C Graesser, Shulan Lu, and Heather H Mitchell. 2005. Social cues in animated conversational agents. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(6):693–704.

François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Mary H. McCaulley and Charles R. Martin. 1995. Career assessment and the myers-briggs type indicator. *Journal of Career Assessment*, 3(2):219–239.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.

Sebastian Möller, Paula Smeele, Heleen Boland, and Jan Krebber. 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech & Language*, 21(1):26–53.

Youngme Moon and Clifford I Nass. 1996. Adaptive agents and personality change: complementarity versus similarity as forms of adaptation. In *Conference companion on Human factors in computing systems*, pages 287–288. Association for Computing Machinery.

Isabel Myers-Briggs, Mary H McCaulley, Naomi L Quenk, and Allen L Hammer. 1998. *MBTI manual: a guide to the development and use of the Myers-Briggs Type Indicator*, volume 3. CPP.

Babak Naderi, Tim Polzehl, André Beyer, Tibor Pilz, and Sebastian Möller. 2014. Crowdee: mobile crowdsourcing micro-task platform for celebrating the diversity of languages. In *Fifteenth Annual Conference of the International Speech Communication Association*, pages 1496–1497. International Speech Communication Association.

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

Ketakee Nimavat and Tushar Champaneria. 2017. Chatbots: An overview. types, architecture, tools and future possibilities. *International Journal for Scientific Research & Development*, 5(7):1019–1024.

Eunil Park, Dallae Jin, and Angel P del Pobil. 2012. The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 9(2):35.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

David J. Pittenger. 1993. The utility of the myers-briggs type indicator. *Review of Educational Research*, 63:467 – 488.

ITUT Rec. 2003. P. 851, 2003. subjective quality evaluation of telephone services based on spoken dialogue systems. *International Telecommunication Union, Geneva*.

Elayne Ruane, Sinead Farrell, and Anthony Ventresque. 2020. User perception of text-based chatbot personality. In *International Workshop on Chatbot Research and Design*, pages 32–47. Springer, Cham.

Tommy Sandbank, Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, John Richards, and David Piorkowski. 2017. Detecting egregious conversations between customers and virtual agents. *arXiv preprint arXiv:1711.05780*, pages 1–9.

Klaus Rainer Scherer. 1979. *Personality markers in speech*. Cambridge University Press.

Ryan M Schuetzler, Mark Grimes, Justin Scott Giboney, and Joesph Buckman. 2014. Facilitating natural conversational agent interactions: lessons from a deception experiment. *Information Systems and Quantitative Analysis Faculty Proceedings and Presentations*, pages 0–16.

Tuva Lunde Smestad. 2018. Personality matters! improving the user experience of chatbot interfaces-personality provides a stable pattern to guide the design and behaviour of conversational agents. Master's thesis, NTNU.

Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38:75–84.

Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T Rădulescu. 2020. The effect of social presence and chatbot errors on trust. *Sustainability*, 12(1):256.

Jess und Gregory J Feist. 2002. *Theories of Personality*. 7th edition. McGraw-Hill Humanities/Social Sciences/Languages.

Darius Zumstein and Sophie Hundertmark. 2017. Chatbots–an interactive technology for personalized communication, transactions and services. *IADIS International Journal on WWW/Internet*, 15(1):96–109.

**Appendix A: Item Setup for Overall Study**

| No. | Type | Item |
|---|---|---|
| PQ1 | IE | I consider myself to be energetic rather than relaxed. |
| PQ2 | IE | I would describe myself as a talker rather than a listener. |
| PQ3 | IE | I oftentimes like to stay home rather than going out to town. |
| PQ4 | IE | Speaking in public is more likely to frighten me than to entertain me. |
| PQ5 | IE | I describe myself as a calm person rather than being impulsive. |
| PQ6 | IE | I would describe myself as an open person instead of being guarded. |
| PQ7 | TF | I am more a skeptical person than a believer. |
| PQ8 | TF | I rather strive to have an mechanical mind than striving to let my thoughts run free. |
| PQ9 | TF | I am easily hurt and not emotionally thick-skinned. |
| PQ10 | TF | I prefer to follow my heart rather than my head. |
| PQ11 | TF | I rather value emotions instead of feeling uncomfortable with (expressing) them. |
| PQ12 | TF | I rather use reason over instinct. |
| Q1 | ITU-T | Overall, I was satisfied with the chatbot. |
| Q2 | ITU-T | The chatbot reacted naturally. |
| Q3 | ITU-T | I would advise my friends to also use the chatbot. |
| Q4 | ITU-T | The overall dialogue course was too long. |
| Q5 | Custom | The chatbot was formal. |
| Q6 | Custom | The chatbot was emotional. |
| Q7 | Custom | The chatbot was trustworthy. |
| Q8 | Custom | The messages were too long. |
| Q9 | Custom | The chatbot was appropriate according to my expectations. |
| C1 | Comparison | Do you believe the interaction was adapted to you personally? |
| C2 | Comparison | Which chatbot do you like more? |
| G1 | General | How often do you use chatbots? |
| G2 | General | Please tell us about your age range. |
| G3 | General | Is English your native language? |
| G4 | General | Please tell us about your gender. |
| G5 | General | What is your current profession? |

Table 7: Overview of all items used throughout the study.