Knowledge-Based Systems 163 (2019) 794–799

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

The effect of context-dependent information and sentence constructions on perceived humanness of an agent in a Turing test

Roy de Kleijn [a,*], Marjolijn Wijnen [a], Fenna Poletiek [a,b]

[a] Leiden Institute for Brain and Cognition, Leiden University, The Netherlands
[b] Max Planck Institute of Psycholinguistics, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 17 May 2018
Received in revised form 8 August 2018
Accepted 4 October 2018
Available online 16 October 2018

Keywords:
Turing test
Humanness
Linear grammar
Center-embedded grammar
Context dependence
Chatbots
Human–computer dialog

ABSTRACT

In a Turing test, a judge decides whether their conversation partner is either a machine or human. What cues does the judge use to determine this? In particular, are presumably unique features of human language actually perceived as humanlike? Participants rated the humanness of a set of sentences that were manipulated for grammatical construction: linear right-branching or hierarchical center-embedded and their plausibility with regard to world knowledge.

We found that center-embedded sentences are perceived as less humanlike than right-branching sentences and more plausible sentences are regarded as more humanlike. However, the effect of plausibility of the sentence on perceived humanness is smaller for center-embedded sentences than for right-branching sentences.

Participants also rated a conversation with either correct or incorrect use of the context by the agent. No effect of context use was found. Also, participants rated a full transcript of either a real human or a real chatbot, and we found that chatbots were reliably perceived as less humanlike than real humans, in line with our expectation. We did, however, find individual differences between chatbots and humans.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Ever since Alan Turing proposed his *imitation game* [1], now colloquially known as the *Turing test,* it has been a popular topic in the debate on artificial intelligence. Turing proposed that the imitation game could identify intelligent machines through a conversation between a human judge and both a machine and a human confederate, whose identities are unknown to the judge. After the conversation, the judge is to decide which of the two agents is human and which one is a machine. If there is no detectable difference between a human and a machine, it could be argued that this machine can think. Turing [1] predicted that by the year 2000, an appropriately programmed machine would be able to fool at least 30% of human judges into thinking it was human after a 5-minute conversation.

Turing tests are conducted regularly, with one of the biggest and most frequent editions being the annual *Loebner Prize contest*. Typically, the Loebner Prize contest consists of two rounds. First, a selection round is held in which chatbots are judged on their humanness by human judges. This round consists of a set of the same 20 questions asked to each chatbot. The judges rate chatbot responses on three characteristics: relevance (is the response relevant to the question being posed?), correctness (is the response correct, either factually, or in the character of the agent?), and plausibility and clarity of expression (is the response grammatically correct and correct in the context or the character of the agent?). The four highest scoring chatbots will advance to the actual contest, consisting of four 25-minute conversations between a judge and both the chatbot and a human confederate, similar to the original imitation game proposed by Turing [2].

Of course, language is only one of the domains in which AIs can display seemingly intelligent behavior. Modern AIs approach or even surpass human performance on face recognition [3], chess [4], and even Go [5], and can assist humans in a wide range of activities such as driving [6], surgery [7], music composition [8], and, amusingly, evaluating Turing test performance [9]. Artificial agents can even *appear* intelligent without performing anything that we would consider "thinking". Recent studies have shown that humans attribute intelligence to artificial agents based on physical appearance [10], goal-directed behavior [11], and their own cognitive states and traits [12]. Early chatbots such as ELIZA [13] used clever tricks to fool humans into thinking they were actually intelligent, and it could be argued that most modern chatbots are not that different. How exactly one can fool a human judge is investigated in the current study, in which we will return to the paradigm of the original Turing test: a conversation between a human judge and an agent that is either human or artificial.

* Correspondence to: Wassenaarseweg 52, 2333 AK Leiden, The Netherlands.
E-mail address: kleijnrde@fsw.leidenuniv.nl (R. de Kleijn).

https://doi.org/10.1016/j.knosys.2018.10.006
0950-7051/© 2018 Elsevier B.V. All rights reserved.

1.1. Cues of judgment

While ample studies (e.g. [14]) have focused on the chatbot and its abilities to pretend to be human, the role of the judge often remains underappreciated. As such, much remains unknown about the factors that influence the perception of the judges of their partners' humanness. The purpose of the present study is to investigate the cues judges use to determine the identity of conversation partners in a Turing test context. Knowledge about these factors cannot only help us understand the characteristics of mutual understanding in human conversations, but might also contribute to the development and improvement of chatbots so that they can appear more humanlike in conversations.

Lortie and Guitton [15] analyzed transcripts from past conversations in the Loebner Prize contest in which humans were incorrectly judged as machines by at least one of the judges. They found a relationship between the number of questions asked by the judge and the perceived humanness of the agent, and concluded that reciprocity of the exchanges between humans and agents is an important factor in judging humanness. Another interesting finding was reported by Warwick and Shah [16], who analyzed eight transcripts of humans judged as a machine taken from the practical Turing tests held on the 100th anniversary of Turing's birth. They found that judges can be fooled into judging a human as a machine based on a lack of shared knowledge between judge and agent. The judge may lack specific knowledge that the agent does know or the judge assumes specific knowledge to be common which the agent does not possess. Furthermore, they found that judges specifically base their decision upon only one sentence or question, commonly the first or last one, regardless of the rest of the conversation [16]. Judges might also start to converse differently with the agent after having made a decision about the agent's identity. Indeed, Hill et al. [17] found that humans converse differently online when they are aware of conversing with a chatbot instead of with another human.

Warwick and Shah [14] investigated which cues judges use as indicators of humanness in the responses of the agent. They evaluated 13 Turing test transcripts in which hidden human foils were judged as machines, and extracted some guidelines for humanness: the agent should not answer out-of-the-box, not be boring, not dominate the conversation, let the judge be in control, not show too much knowledge and not be humorous because the judge may not understand it.

Saygin and Cicekli [18] analyzed conversations from the Loebner Prize contest based on the cooperative principle of Grice [19]. This principle consists of four sub-principles which characterize meaningful conversations: be relevant (relevance), do not make your contribution less or more informative than is required (quantity), try to make your contribution one that is factually true (quality) and be perspicuous (manner) [19]. They found that the subprinciple of *relevance* should never be violated in order to come across as a human while violating the subprinciple of *manner* is perceived as displaying human emotions. The latter two subprinciples together are a rather good predictor of performance. Furthermore, they found that providing more information than required gives the impression of being a machine while providing less information than required might not always create a machinelike impression. Strikingly, the validity of information has no predictive power. Lortie and Guitton [15] analyzed transcripts of the free conversation part of the Loebner Prize contest and found that fewer long words, fewer articles and fewer words per message were taken as indicators of being an artificial agent.

The indicators studied in previous research mainly regard the *content* of the agent's utterances. The *manner* in which the speech is structured might also be taken by listeners as a cue for humanness. In particular, according to a seminal linguistic theory, recursive complex grammatical structures are the defining feature of human language (as compared to other, e.g. animal language; [20]). The use of recursive hierarchical constructions versus linear constructions in sentences is interesting, because it has been suggested that primates can learn linear humanlike grammar constructions (right-branching) of the type ABABAB, while this is not the case for hierarchical constructions of the type AAABBB involving long-distance dependencies [21].

Moreover, it has been suggested that the recursive property characterizing human language also applies at the level of information exchange in human conversations [22]. Characteristics of linguistic communication that make it unique for humans have been derived theoretically [23] but only incidentally tested empirically, by comparing animal language processing with human language. Here, the "humanness" of presumably defining characteristics of human language is tested, for the first time, using the Turing test procedure.

If we observe that recursive complexity is crucially perceived to be associated with humanness, this would be evidence for its contribution to making language uniquely human. The same can be predicted for conversations: conversations having a recursive structure should be perceived more often as human than non-recursive (linear) conversations, if the recursive feature is indeed defining for human information exchange.

1.2. Contextual and grammatical cues

Recursive complex grammars and conversations involve structures in which a basic pattern can be applied repetitively within that pattern, creating multiple levels of information within one unit of information (i.e. sentences within a sentence, or conversations within a conversation). Typically, these structures require non-linear processing – moving attention backward and forward in the information stream – for decoding. For example, complex grammatical embedded constructions as in *[The boy the girl kisses, laughs]* require the listener to get back all the way to *[the boy]* when arrived at *[laughs]* to bind the subject to the action. In contrast, the linear construction *[The girl kisses the boy who laughs]* can be processed linearly, on the fly.

Similarly, in conversations, a piece of conversation can be embedded within a piece of conversation requiring a listener to retrieve and interpret previous contextual information to make sense of the new one. Consider a conversation between two teachers:

"John is quite uncomfortable in contact with the girls in my class".

"– Oh, but yesterday I saw a girl kissing a boy of your class, and the boy the girl kissed laughed"

"– That boy was not John, I suppose".

To follow this conversation, linear processing without a memory buffer does not suffice. Human listeners selectively retrieve past referent information to make sense of the current one. The appreciation by judges of presumably specific human structural features of linguistic communication used by the agents in Turing test conversations has not been investigated before.

Regarding the structure of conversations, we distinguish two levels at which contextual information might play a role: a global level (the conversation) and a local level (the paragraph). For both levels, if a question refers to a previously asked question and accompanying response, and the agent gives a plausible response, i.e. an answer that is factually correct in the context that is presented, we hypothesize that this has a positive influence on the perceived humanness by the judge [24]. In contrast, giving a non-plausible response by neglecting relevant context information given in the preceding paragraph is predicted to have a negative influence on the perceived humanness by the judge.

The influence of grammatical complexity on perceived humanness is investigated by comparing agents using either linear constructions (right-branching), such as: *[the girl kisses the boy that*

laughs], or complex center-embedded constructions: *[the boy the girl kisses laughs]* [25,26]. To avoid a confound of the plausibility of the information (complex sentences seeming less plausible), we manipulated plausibility independently of sentence structure. In the current study, we investigate two predictions. First, in line with the theory that hierarchical structures are unique for human language, we might expect that hierarchical constructions favor perceived humanness. On the other hand, though, linear constructions are more frequent in natural languages than hierarchical ones [27]. The second, opposed, prediction is therefore that the high frequency of linear constructions in language might make them more humanlike. If an agent uses the simpler construction, like humans most frequently do, this might be perceived as an indication of humanness by the judge. Thus, complex constructions might increase perceived humanness because they are a typical – even defining – feature of human language. But linear constructions might also be perceived as more human, just because they are more frequently used by human speakers.

Sentences can also differ in plausibility with regard to world knowledge. Words referring to objects and subjects have typical roles in line with their actual roles in the world. We would, for example, not say "the bread ate Mary", because the thematic roles of the words get assigned based on their position in the sentence, and the grammatical rules determining role assignment in the English language make this sentence very unlikely. However, the role-reversed sentence "Mary ate the bread" is very easy to interpret for any English language user. If language users have shared knowledge on what are plausible events in the world, we might predict that speech violating this world knowledge is perceived as not-human. This relates to the sub-principle of *quality* as proposed by Grice [19], which states that a contribution to a conversation needs to be a factually true one in order to be perceived as humanlike. In the current study, we manipulated the plausibility of sentences by swapping the grammatical positions of objects and subjects. In complex sentences with high syntactic–semantic congruency, i.e. "matched" sentences such as "the banana the girl ate was fresh", the grammatical (syntactic) structure matches the semantic roles (object vs. subject) assigned to words. In sentences with low syntactic–semantic congruency, such as "the girl the banana ate was fresh", the grammatical structure does not match with the semantic meaning of the sentence. We hypothesized that the latter would be perceived as being less humanlike because humans might be well aware that mismatched – or implausible – constructions are very hard to process, while chatbots do not have this awareness. Hence, our prediction was that mismatched sentences are perceived as less humanlike than matched sentences.

In conclusion, the subject of interest in this study is whether the usage of context-dependent information and of complex recursive grammatical constructions – theoretically assumed to be defining features of human language – about plausible and implausible topics, are used as cues for judges to determine if the agent they are conversing with in a Turing test is human.

2. Method

2.1. Participants

A total of 53 participants (mostly undergraduate students) were recruited (9 males, 44 females) in exchange for either course credit or a payment of €6.50. There were no drop-outs.

2.2. Materials

To test our hypotheses, we divided the study in three experimental tasks for which we used the following materials.

	Conditions	
Task 1	Judge the humanness of sentences varying in grammatical constructions and plausibility	
	$N = 53$	
Task 2	Judge the humanness of a conversation	
	Context	No context
	$N = 27$	$N = 26$
Task 3	Judge the humanness of a conversation	
	Chatbot (1, 2, 3 or 4)	Human (1, 2, 3 or 4)
	$N = 26$	$N = 27$

Fig. 1. Overview of the procedure.

In the first experimental task, participants were presented with a set of generated sentences that were manipulated to contain either right-branching (linear) or center-embedded grammatical constructions. These sentences were further manipulated to either have a syntactic–semantic congruency (match) or a syntactic–semantic incongruency (mismatch), leading to four different types of sentences: (1) center-embedded, matched, (2) center-embedded, mismatched, (3) right-branching, matched, and (4) right-branching, mismatched (see Appendix A for the full set of sentences).

For the second task, two conversations consisting of questions and responses were created. One of these conversations contained the responses of an agent using the context-dependent information of the conversation correctly, and the other conversation contained the responses of an agent using this information incorrectly or not at all. These conversations were based on the questions and responses of real chatbots in the selection rounds of earlier Loebner Prize contests. The responses given by the agents during this contest were manipulated in order to have the two conversations be exactly the same, except the agents' use of context in their responses (see Appendix B for the full conversations).

For the third task, full transcripts of the Loebner Prize 2015 selection round were used. The responses to the questions of this round given by the four highest scoring chatbots as judged by the Loebner Prize judges were included: Mitsuku, Lisa, Izar and Rose. In addition, four actual humans answered the same set of questions and these answers were used as the "human answers" in this experimental task (see Appendix C for the full transcripts), leading to eight different sets of answers.

2.3. Procedure

The three tasks were presented to the participant in a random order on a computer screen. We numbered them for the clarity of this report. See Fig. 1 for a short summary of the procedure.

For the first experimental task, the participant was presented with the full set of independent sentences of Appendix A in a random order. For each sentence, the participant judged the humanness of this sentence on a scale from 1 (certainly a machine) to 7 (certainly a human). The participant was able to do this by clicking on the corresponding number and subsequently on the OK button on the screen using their mouse.

For the second task, participants were presented with a manipulated, full conversation. Participants were randomly divided into two conditions. In the first condition, participants were only shown the conversation in which the agent used the context-dependent information correctly ($N = 27$), while the participants in the second condition only judged the conversation with no correct use of the context ($N = 26$), see Appendix B. The participant rated the response of the agent to each question on humanness in the same way as in the first experimental task.

In the final task, participants were randomly divided into two conditions. Participants in the first condition were randomly presented with one of the four Loebner Prize 2015 transcripts (see Appendix C), referred to as the chatbot condition ($N = 26$). In the human condition ($N = 27$), the participants were randomly presented with the responses from one of the humans to the same questions (see Appendix C). After each response, the participant rated the humanness of the response in the same way as in the first task. After having rated all the individual responses to the questions, participants were asked to make a final judgment on the humanness of the agent, based on the previously seen questions and responses. Participants used the same scale as the previous tasks, ranging from "certainly a machine" to "certainly a human".

3. Results

3.1. Task 1: Grammatical structures

A two-way ANOVA with grammatical structure of the sentence and syntactic–semantic congruency as independent variables and the perceived humanness of these sentences as the dependent variable showed a significant main effect of grammatical structure on the perceived humanness of the sentences, $F(1, 52) = 231.2$, $p < .001$, $\eta^2{}_G = .58$. Sentences with a center-embedded structure are perceived as significantly less human-like compared to right-branching sentences. Also, a significant main effect of the plausibility of the sentence (semantic–syntactic congruency) on the perceived humanness of the sentences was found, $F(1, 52) = 87.6$, $p < .001$, $\eta^2{}_G = .12$. The implausible (mismatched) sentences were perceived as being significantly more machinelike than plausible (matched) sentences. The significant interaction effect shows that the effect of the plausibility of the sentence on perceived humanness was smaller for center-embedded sentences than for right-branching sentences, $F\ (1, 52) = 40.8$, $p < .001$, $\eta^2{}_G = .045$. Main and interaction effects are shown in Fig. 2.

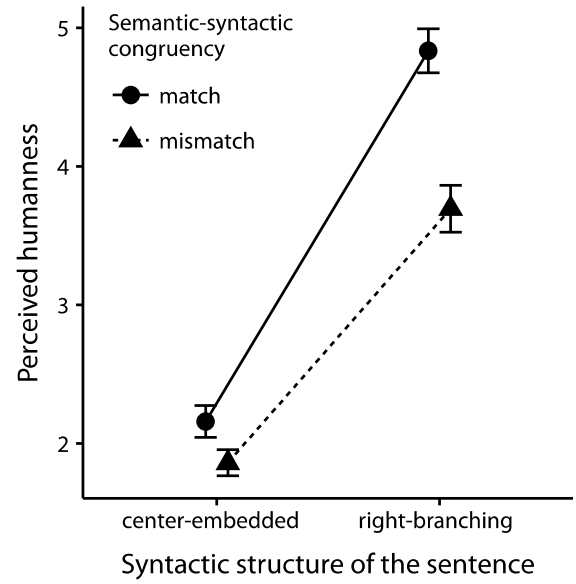

Fig. 2. The effect of grammatical structure and plausibility of sentences on the perceived humanness of the sentences in Task 1. Error bars indicate within-subject 95% CI.

3.2. Task 2: Use of context

Five of the responses given to the questions in the conversations used for the second task were the same for both the context and the no context condition (see Appendix B). The data of these stimuli were excluded for the following analysis, because we were only interested in the differences between using context correctly or not.

In a one-way ANOVA with context use as the independent variable and perceived humanness as the dependent variable, we did not find a significant effect of the use of context on perceived humanness, $F(1, 51) = 3.34$, $p = .074$, $\eta^2{}_G = .061$. Whether the agent uses the context in the conversation correctly or not does not influence the perceived humanness of the agent significantly.

To find out if particular questions and corresponding responses have an influence on the perceived humanness of the agent for the participant, we calculated the differences between the perceived humanness on each question-response combination and the humanness on the follow-up question-response as a measure of how much any given question changes the judge's existing opinion. We then compared these difference scores on the pairs of questions between the context and the no-context condition using independent t-tests. When corrected for multiple comparisons, only the difference score between the questions 11 and 12 (suggesting that the answer to the question "The car couldn't fit into the parking space because it was too small. What was too small?" had the largest effect on ratings of humanness, see Appendix B) compared between the context and no-context condition was significant, $t(51) = 3.12$, $p = .003$, while all other difference scores had $ps > .021$.

3.3. Task 3: Transcripts of chatbots and humans

In the third task, participants were asked to give a final decision on the humanness of the agent. This final decision is a summary of the separate decisions on humanness on each question-response combination made by the participants themselves. The data of the final decision was not included in the following analyses, because we regarded the overall decision of the judge as a variable separate from the decisions on humanness on each question-response combination.

A one-way ANOVA with the category of the agent (chatbot or human) as the independent variable and perceived humanness as the dependent variable showed a significant effect of the category of the agent on the perceived humanness, $F(1, 51) = 27.77$, $p < .001$, $\eta^2{}_G = .35$. The perceived humanness of responses given by chatbots was significantly lower than the perceived humanness of responses given by humans.

Within these categories, we compared the different chatbots with each other in a one-way ANOVA with the identity of the chatbots as independent variable and the perceived humanness as dependent variable. This showed a significant effect of the identity of the chatbots on perceived humanness, $F(3, 22) = 6.22$, $p = .003$, $\eta^2{}_G = .46$. In other words, there were differences between chatbots in human-likeness. We also compared the different humans with each other in a one-way ANOVA with the identity of the humans as independent variable and the perceived humanness as dependent variable. This showed no significant effect of the identity of humans on perceived humanness, $F(3, 23) = .070$, $p = .976$, $\eta^2{}_G = .009$. This indicates that the different humans are perceived as being equally humanlike (see Fig. 3).

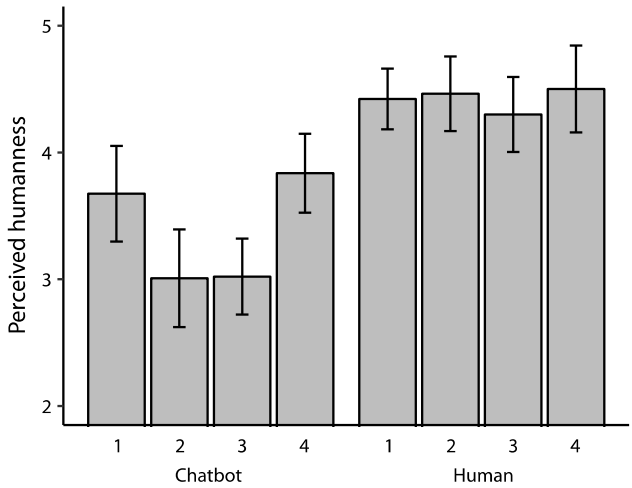

Fig. 3. Perceived humanness for each chatbot and human on all question-response combinations in Task 2. Error bars indicate 95% CI.

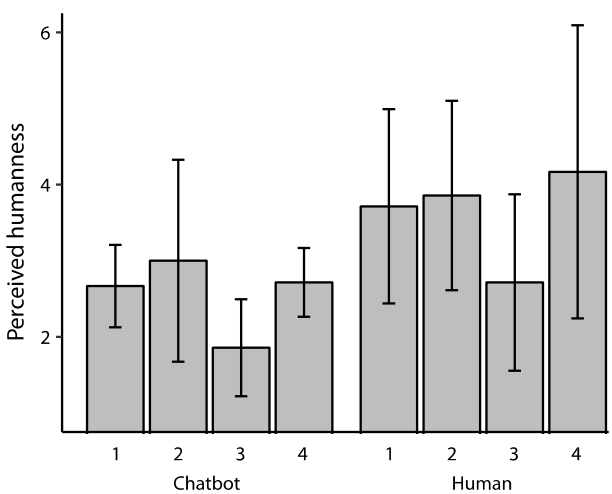

Fig. 4. The final decision on humanness for each chatbot and human given by participants. Error bars indicate 95% CI.

Next, we compared only the final decisions of the participants on the humanness of the chatbots and humans. A one-way ANOVA with the category of the agent (chatbot or human) as independent variable and the final decision as dependent variable showed a significant effect of agent category on the final decision made by the participant, $F(1, 51) = 10.01$, $p = .003$, $\eta^2{}_G = .16$. This means that, overall, human agents were reliably recognized as humans, and chatbots were reliably recognized as chatbots.

A one-way ANOVA with the identity of the chatbots as independent variable and final decision as the dependent variable showed no significant effect of the identity of the chatbot on the final decision made by the participant, $F(3, 22) = 2.61$, $p = .08$, $\eta^2{}_G = .26$. The chatbots did not significantly differ from each other with regard to the final decision made by the participants. Within the category of humans, a one-way ANOVA with the identity of the human as independent variable and the final decision as the dependent variable also showed no significant effect, $F(3, 23) = 1.26$, $p = .31$, $\eta^2{}_G = .14$. The humans did not significantly differ from each other with regard to the final decision made by participants (see Fig. 4).

4. Discussion

4.1. Grammatical structure

The results of the first task, in which participants rated the humanness of sentences, showed that center-embedded hierarchical sentences are perceived as being less humanlike than sentences with simple linear right-branching constructions. Hence, the ratings of the used Turing test conversations do not provide evidence for the uniqueness of hierarchical recursive complexity in human languages. Perceived humanness appears to increase when agents do what humans mostly do: using simple grammatical constructions, rather than when they use typical human features in their speech. This seems to be in line with Warwick and Shah's [14] work, who advised any human trying to convince the judge that they are indeed a human to not write sentences out-of-the-box, but just write what would be expected from you based on what the average human would do.

However, Saygin and Cicekli [18] found that being perspicuous does not enhance the perceived humanness of the agent, presumably because it decreases emotional expression. Showing emotion, they further argue, is only expected of humans and not of machines. While simple grammatical constructions might be more perspicuous and therefore show less emotion, we found they were perceived as more humanlike than complex, center-embedded grammatical constructions, possibly because these constructions are more often used by humans in real life.

Our results show that implausible (mismatched) sentences are perceived as less humanlike than plausible (matched) sentences. This means that to be perceived as human, it is important to make sentences that are in line with general, shared background knowledge. It could be that humans are expected to only make sentences that are indeed semantically plausible, while machines make sentences that are semantically implausible simply because they do not have the proper background knowledge to determine semantic plausibility. This does not imply that speech should always be true necessarily (*quality* subprinciple) to be perceived as humanlike [18]. Plausible but factually false statements could also be perceived as being humanlike. Our data suggest that the effect of plausibility is larger for simple than for complex recursive constructions, which might be caused by complex constructions obscuring the plausibility assessment, while plausibility is more readily clear in simple sentences.

4.2. Use of context

We hypothesized that correct retrieval and processing of previous context information during the conversation would enhance perceived humanness of the agent compared to using the context incorrectly or not at all. This prediction was not supported by our data. Possibly, the manipulation of context was too weak; the correct use of context and presumed incorrect use of context might have differed insufficiently to have an effect on perceived humanness. Another possibility is that, in contrast to our assumption, humans typically do not remember or correctly retrieve every bit of information that has been communicated in the conversation. In fact, machines could be expected to have a bigger memory capacity than humans,[1] and judges might therefore actually expect them to remember everything that has been said in the conversation for use later on.

Thus, paradoxically, the selective use of distant context information might seem very machinelike to judges just because

[1] Although Bartol et al. [28] note that this is not yet the case: the human brain is estimated to be able to hold as much as 10^{15} bytes, whereas the largest available consumer computer storage can hold 10^{12} bytes as of writing.

it is correct. We did not find this inverse effect on the overall results of Task 2, but the suggestion might be supported by the fact that to question 13 ("Where do I live?", see Appendix B), the *correct* response in the context of the conversation ("Exeter") influenced the perceived humanness of the agent negatively, while the *incorrect* response ("New York", which is a more plausible answer, because New York is a more commonly known city than Exeter) influenced perceived humanness slightly positively. The difference between these influences on perceived humanness was significant. This suggests that retrieving previous information and binding to current information correctly is not necessarily a cue for humanness, but rather an indicator of being an artificial agent. In conclusion, we found no evidence that the hierarchical non-linear recursive structure of language – both at the syntactic and at the conversational level – are perceived as typically associated with humanness. This result, that to our knowledge provides the first data about the perceived human character of recursive structures in actual language use, contrasts with theoretical assumptions about the status of these structures in the human language system.

In the third task, we compared the responses of real chatbots and real humans with each other. In this task, two questions (6 and 7, see Appendix C) required the use of local context in order to respond to them correctly. The *incorrect* use of the local context was rated as less humanlike than the *correct* use of the context in this task. These results seem to suggest that using the local context correctly and therefore giving a factually correct answer in the context provided, does indeed improve the perceived humanness of the agent. This was contrary to the results we found in Task 2, where local and global context uses were combined, and no effect of context was found. In other words, failing to use *distant* previous contextual information does not interfere with humanlikeness, but *recent* context is assumed by judges to be available and used in conversations with humans on a regularly basis, and failing to use it is seen as an indication of being an artificial agent.

4.3. Chatbots compared to humans

The responses given by chatbots were perceived as less humanlike than the responses given by humans to the same questions, illustrating the relatively poor performance of chatbots in modern Turing tests.

Comparing individual chatbots, we found that not all chatbots were rated equally humanlike. Surprisingly, in the Loebner Prize contest, chatbot 4 (Rose) was rated as least humanlike, while in the current study this chatbot was rated most humanlike out of all four of them. A possible explanation for this could be that the judges in the Loebner Prize contest are supposed to be experts in judging the humanness of agents, while our participants were all naive undergraduate students. Moreover, in the Loebner Prize contest only four judges decide on the humanness of the agents, while in the current study 26 participants distributed over four chatbots ($N_1 = 6, N_2 = 6, N_3 = 7, N_4 = 7$) judged the humanness of these chatbots. No difference was found between human agents on perceived humanness, indicating that the actual humans were perceived as equally humanlike.

In conclusion, we found that grammatical, recursive structures, a feature that has been claimed to be defining for human language, were in fact not rated as more humanlike than linear structures. Similarly, conversations with recursive structure requiring the use of previous contextual knowledge were not considered more humanlike than conversations without cross-references to contextual information. In fact, we found a tendency in judges to rate simple linear sentences and conversations as more humanlike. Secondly, plausible sentences are judged more humanlike than implausible ones, but sentences need not be factually true to be rated as humanlike. More variance is found in perceived humanlikeness between chatbots than between humans, indicating that some chatbots are better in behaving as humans than others. These findings could be used to improve the quality of chatbots in human–computer interaction software.

Declaration of interest

Declarations of interest: none.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2018.10.006.

References

[1] A.M. Turing, Computing machinery and intelligence, Mind 59 (1950) 433–460.
[2] M.M. al Rifaie, Loebner prize. Retrieved on November 7, 2015 from http://www.aisb.org.uk/events/loebner-prize, 2015.
[3] P.J. Phillips, A cross-benchmark assessment of a deep convolutional neural network for face recognition, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2017.
[4] M. Campbell, A. Hoane, F. Hsu, Deep blue, Artificial Intelligence 134 (2002) 57–83.
[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, et al., Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359.
[6] L. Abdi, A. Meddeb, In-vehicle augmented reality system to provide driving safety information, J. Visualization 21 (2017) 163–184.
[7] J. Shuhaiber, Augmented reality in surgery, Arch. Surg. 139 (2004) 170.
[8] C. Kerdvibulvech, An innovative real-time mobile augmented reality application in arts, in: L. De Paolis, P. Bourdot, A. Mongelli (Eds.), Augmented Reality, Virtual Reality, and Computer Graphics. AVR, in: Lecture Notes in Computer Science, vol. 10325, Springer, Cham, 2017.
[9] R. Lowe, M. Noseworthy, I.V. Serban, N. Angelard-Gontier, Y. Bengio, J. Pineau, Towards an automatic Turing test: Learning to evaluate dialogue responses, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1116-1126.
[10] M. Mori, The Buddha in the Robot, Charles E. Tuttle Co., 1982.
[11] L.M. Oberman, J.P. McCleery, V.S. Ramachandran, J.A. Pineda, EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots, Neurocomputing 70 (2007) 2194–2203.
[12] N. Epley, A. Waytz, J.T. Cacioppo, On seeing human: A three-factor theory of anthropomorphism, Psychol. Rev. 114 (2007) 864–886.
[13] J. Weizenbaum, ELIZA—A computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1966) 36–45.
[14] K. Warwick, H. Shah, Human misidentification in Turing tests, J. Exp. Theor. Artif. Intell. 27 (2015) 123–135.
[15] C.L. Lortie, M.J. Guitton, Judgment of the humanness of an interlocutor is in the eye of the beholder, PLoS One 6 (2011) e25085.
[16] K. Warwick, H. Shah, Assumption of knowledge and the Chinese Room in Turing test interrogation, AI Commun. 27 (2014) 275–283.
[17] J. Hill, W.R. Ford, I.G. Farreras, Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations, Comput. Hum. Behav. 49 (2015) 245–250.
[18] A.P. Saygin, I. Cicekli, Pragmatics in human–computer conversations, J. Pragmat. 34 (2002) 227–258.
[19] P.H. Grice, Logic and conversation, in: P. Cole, J. Morgan (Eds.), Syntax and Semantics 3: Speech Acts, Academic Press, New York, 1975, pp. 41–58.
[20] M.D. Hauser, N. Chomsky, W.T. Fitch, The faculty of language: what is it, who has it, and how did it evolve? Science 298 (2002) 1569–1579.
[21] W.T. Fitch, M.D. Hauser, Computational constraints on syntactic processing in a nonhuman primate, Science 303 (2004) 377–380.
[22] S.C. Levinson, Recursion in pragmatics, Language 89 (2013) 149–162.
[23] N. Chomsky, On certain formal properties of grammars, Inf. Control 2 (1959) 137–167.
[24] B. Wilcox, S. Wilcox, Making it real: Loebner-winning chatbot design, ARBOR Ciencia 189–764 (2013) a086.
[25] R. Hudson, The difficulty of (so-called) self-embedded structures, UCL Working Papers in Linguistics, Vol. 8, 1996, pp. 283–314.
[26] E. Gibson, Linguistic complexity: Locality of syntactic dependencies, Cognition 68 (1) (1998) 1–76.
[27] F. Karlsson, Constraints on multiple center-embedding of clauses, J. Linguist. 43 (2007) 365–392.
[28] T.M. Bartol, C. Bromer, J. Kinney, M.A. Chirillo, J.N. Bourne, K.M. Harris, T.J. Sejnowski, Nanoconnectomic upper bound on the variability of synaptic plasticity, eLife 4 (2015) e10778.