The Gricean Maxims of Quantity and of Relation in the Turing Test

1st Baptiste Jacquet*
CHArt (P-A-R-I-S)
Université Paris VIII & EPHE
Paris, France
baptiste.jacquet28@gmail.com

2nd Jean Baratgin
CHArt (P-A-R-I-S)
Université Paris VIII & EPHE
Paris, France
Institut Jean Nicod (IJN)
Paris, France
jean.baratgin@univ-paris8.fr

3rd Frank Jamet
CHArt (P-A-R-I-S)
Université Paris VIII & EPHE
Paris, France
Université de Cergy-Pontoise (UCP)
Cergy-Pontoise, France
frank.jamet@u-cergy.fr

***Abstract*—Previous research in the field of cognitive science has demonstrated the relevance of measuring reaction times to describe the cognitive cost of processing information, yet it has seldom been studied in the context of conversations and, to our knowledge, never in free flowing, interactive conversations. This study presents a way of analyzing entire online conversations in a protocol inspired by the Turing Test to investigate the relation between violations of Grice's Cooperation Principle and the response times of the participants. We hypothesized that response times are directly correlated to the cognitive cost required to generate implicatures from a statement. Our results show that violations of the maxim of Relation significantly increased the response time, especially for female participants. This confirms that measuring response times during a conversation can be a simple and relevant way of inferring the cognitive cost of processing an utterance.**

***Index Terms*—Conversational Expectations, Pragmatics, Relevance, Turing Test, Natural Language, Cooperation**

I. Introduction

Linguistic structures have been extensively studied in the context of *Natural Language Understanding* and *Natural Language Generation* [1], [2], and such tools are becoming increasingly useful as interacting with virtual agents and processes becomes more and more common.

However, evaluating the quality of the generated responses of automated assistants is quite complicated, in particular without a specific task to complete, which is the case for *chatbots*: agents for which the main goal is conversing itself. Because of the dynamic aspect of conversations, finding a gold standard to evaluate the naturalness of a conversation partner remains difficult. Many different evaluation methods exist, yet few explore the pragmatics of conversations [3, for a review].

It is arguable that the analysis of the pragmatics in a textual conversation is complex. Indeed, it consists in the construction of inferences based on clues that are not specifically present within the words used, but in all of what surrounds them. In oral conversations, it can for example involve the tone of the voice or the gestures [4], which help determine the intentions of the speaker [5] or their confidence in what they are saying as well as their level of understanding of it [6].

Pragmatic clues are not absent in written conversations though, and ignoring them has been shown to produce a feeling of machine-like behavior [7]. Indeed, participants rely much more on the clues contained in the structure of the conversation itself than they do in oral conversations, for it is all they have access to.

We suggest a new experimental paradigm based on the Turing Test to evaluate the humaneness of a conversational partner through their use and misuse of the pragmatic clues in an online conversation.

In the field of pragmatics, [8] was the first to describe a set of rules that he thought were shaping the dynamics of a meaningful conversation, which he called the Cooperation Principle. This principle is itself divided into sub-principles, or maxims. It offers a framework useful in experimentation, for it is possible to tell when a maxim is used and when it is ignored.

The Relevance theory [9] further develops this concept and expands it to much broader applications, in particular by providing a cognitive explanation of the behavior of humans in a conversation that the Cooperation Principle alone did not have, by defining the relevance of an utterance as the interaction of its cognitive cost and of its contextual effect on the conversational partner.

Our hypothesis is that the relevance of an utterance A in a conversation can be measured through the delay between the instant the utterance it is responding to was sent and the instant A itself is sent.

We expect that violations of the gricean maxims will lead to a higher difficulty in retrieving the meaning of an utterance, implying a higher cognitive cost, and will delay the response of the participants accordingly. In particular, we expect violations that highly reduce the relevance of an utterance, such as violations of the maxim of Relation[1], to increase the response time with a greater magnitude than violations of the first maxim of Quantity[2].

*Corresponding Author.

[1] "Be relevant." [8, p. 46]

[2] "Make your contribution as informative as is required (for the current purpose of the exchange)." [8, p.45]

978-1-5386-5024-0/18/$31.00 2018 IEEE

II. Theoretical Background

A. Pragmatics of conversation

Reference [8] explores conversations to define the basic conditions that apply to them, regardless of the context or the topic. He argues that for a conversation to happen, the interlocutors must be cooperating with each other. They expect the other interlocutor to respect a set of principles or rules in order to generate conversational implicatures (elements of meaning inferred from an utterance). Grice calls this the Cooperation Principle, which he defines as follows:

> "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." [8, p. 45]

This principle is divided into four different maxims, which represent the expectations of one interlocutor regarding the productions of the other, and what they believe the other expects from theirs. The first maxim refers to the truthfulness of the information provided in the contribution and is called the Maxim of Quality, defined in the following terms:

a) Maxim of Quality:: "Try to make your contribution one that is true." and its specific maxims: "Do not say what you believe to be false." and "Do not say that for which you lack evidence."

The maxim of quantity refers to the appropriate amount of information that should be provided within an utterance. It is defined as follows:

b) Maxim of Quantity:: 1: "Make your contribution as informative as is required (for the current purposes of the exchange)." and 2: "Do not make your contribution more informative than is required."

Grice argues that while giving too much information might not seem to be a violation of the Cooperation Principle, it can potentially confuse and mislead the interlocutor. The Maxim of Relation gives a solution to this potential debate, as being over informative becomes a waste of time for both interlocutors that brings nothing to the progression of the conversation. It is simply defined by Grice as follows:

c) Maxim of Relation: : Be relevant.

Finally, Grice introduces the Maxim of Manner defined by four more specific sub-maxims:

d) Maxim of Manner: : 1: "Avoid obscurity of expression.", 2: "Avoid ambiguity.", 3: "Be brief (Avoid unnecessary prolixity)." and 4: "Be orderly."

He explains that the speaker within a conversation will do one of the following actions during each utterance: (1) they will observe the maxims; (2) they will opt-out of the maxims by using cues to inform their conversational partner of this fact; (3) they will flout the maxims, with the knowledge of their conversational partner; (4) they will violate the maxims.

It is because we expect cooperation that we will assume that any utterance contains a meaningful information. Grice does indeed consider cases where his maxims seem to not be respected. Here is an example: In a conversation between three interlocutors (*A*, *B* and *C*), *A* might be tempted to become obscure in their utterances if they do not wish to give a specific piece of information to *C*, while not making it too obscure to let *B* understand what *A* is talking about. In this case one might be tempted to say that *A* violated the Maxim of Manner, and transgressed the Cooperation Principle, yet in this context, the obscurity of the statement had a purpose within the conversation, as it carried the additional information that *A* does not wish *C* to know something that he expects B to know. While it can be argued that in doing so *A* is violating the maxim of manner with *C* (even if temporarily), they do not violate it with *B*. Yet, it is likely that *C*, while trying to infer the meaning of the obscure utterance will be able to understand that *A* attempted to bypass cooperating with them, but the piece of information *A* wanted to avoid giving, might remain out of reach. If this was the intention of *A*, they did not transgress the cooperation principle, as they just flouted the maxim.

1) Quantity: Reference [10] uses an experimental protocol similar to the one used by [11] which consists in presenting to two participants items to move, and spots where they can be moved to. In this protocol, one participant needed to describe where the items should go, while the other had to move them accordingly. It has shown that participants had a tendency to over-describe the items to move in 30% of cases, and in doing so violated the second maxim of Quantity. Yet, participants actually moving the items were directly affected by the violations of the maxim in the instruction given to them. An under-description resulting in a confusion visible in the tracking of their ocular fixations (Participants' gaze remain longer on the wrong items). A similar confusion was observable in the cases of over-description. Similar results had already been published [12].

In the context of artificial intelligence in conversations, the over-description is often considered to be a rather mechanical and artificial behavior, while (and this can seem surprising) the lack of information, which can cause ambiguity, is considered to be more human-like [7].

2) Relation and Relevance Theory: The maxim of Relation is, according to Grice, "exceedingly difficult" to detail, especially in the cases of topic switching during the conversation [8, p. 46]. Initially trying to investigate this maxim further, [9] eventually developed their own theory to unify the different maxims, by theorizing that the underlying dynamic of the Cooperation Principle is the balance between the relevance of an utterance (defined by its effect on the mental representations of the target) and the cognitive cost associated with the process of producing it or inferring its meaning, for the processing capacity of the brain is assumed to be limited. Since every utterance will need to be analyzed, the most relevant utterances within a conversation are the ones that require the least cognitive effort. Such utterances are considered to have reached Optimal Relevance. This theory is able to predict the effects of the maxim of Quantity and of the maxim of Manner, and while it often differs from the predictions given by the maxim of Quality, its predictions in these differing cases are often more appropriate [13]. The relevance theory is also able to give good predictions in many cognitive tasks that were previously believed to involve a bias in human cognition by revealing that humans are indeed coherent in their heuristics, but use contextual information as well as information stored in their memory to reason on specific tasks. This can be observed in many reasoning tasks studied in the field of cognitive psychology, such as connectors logic [14]–[16, for examples], on the bias in probability judgment [17], [18, for examples] and in decision making [19], [20, for examples].

Violations of the maxim of relation have an important effect on the conversation partners. Since they will still assume that violating it gives an information on the intention of the speaker. People will indeed assume that they are not comfortable with the topic and want to change subjects. Yet when they believe the speaker is a machine, they simply assume that it did not understand them and will qualify it as not being humanlike [7].

B. Turing Test

Studies on such aspects of language are often only qualitative because of the difficulty of producing operational protocols with statistically analyzable measures [21], [22, for examples]. Yet with the advances of Artificial Intelligence, new methods can be developed to study the natural language, as artificial entities are now common enough and efficient enough to be believably used in the context of psychological experiments, or at the very least, it is not improbable for a participant in a study to believe they are indeed used.

The Turing Test [23] describes an experiment where the participants would act as a judge, trying to find which one of two interlocutors is a machine, and which one is the human. Because it is centered around participants making a comparison between two interlocutors through interactive conversations with text messages, it is an interesting tool to investigate what conversational features are expected of another human by the participants. We argue that this protocol can be used to study human conversations without the need to involve an Artificial Intelligence, by just letting the participants believe they will be conversing with one.

It is important to consider that the Turing Test is criticized in the field of Artificial Intelligence for not being relevant enough to its alleged initial goal of detecting intelligence. We will not discuss further such critics, as our goal is not to test the intelligence of a conversational partner, but rather to test the humanness of its behavior.

We will not be using any Artificial Intelligence in our experiment to avoid potential interactions between the violations of the Gricean Maxims, and other factors like vocabulary and grammar, on the delay between utterances. Yet, presenting one interlocutor as being an Artificial Intelligence will remain a part of our protocol.

In order to do so, we will focus on two Gricean Maxims that are often violated by virtual conversational agents: the first maxim of Quantity and the maxim of Relation.

C. Reaction Times in Cognition

Reaction Times (RT) are commonly used in experimental psychology and in cognitive sciences in general. Yet, to our knowledge, they have never been investigated in the context of online conversations between two interlocutors.

We do not claim that RT measures are a perfect representation of what is actually happening inside of our brain while processing information, and thus the actual numerical values of RT should be carefully interpreted. Yet their recording is much more easily done than using imaging techniques or electroencephalography and depending on the context can suffice to show the influence of different factors on information processing [24]–[26, for examples].

The RT are also not constant with age in most information processing tasks, and the age factor should be considered for any of such measures, especially below 15 years old, if different age groups are present [27].

To avoid potential age biases, our experiment will focus on the response times between an interlocutor and the participant's utterances as an indicator of the cognitive cost of processing the interlocutor's utterance, for participants above 18 years old.

III. Methods and Procedures

A. Participants

40 native English-speakers used to textual conversations through messenger softwares participated in the experiment.

The majority lived in the United States of America (23), 2 lived in the United Kingdom, 4 in Germany, 2 in Canada, 2 in France. The other participants came from countries including Austria, Italy, Greece, Turkey, Czech Republic.

The age of the participants varied between 18 and 40 ($M = 24, SD = 5.0$). 20 participants were males, while 20 were females. Participants were recruited by contacts in the respective countries which had to find one or more participants, ideally of different gender.

Participants had very varied career fields (when they had a job) or fields of study (when they were students), including Computer Sciences, Engineering, Biology, Tourism, Music, Accounting, Design, Security, Management, Economy, Linguistics and others. 4 chose not to answer this question.

B. Variables

1) Main Factor - Maxim Violations: The main factor was the type of the Gricean maxim violated. The conversation order (first or second), and the gender of the participants were also considered to be potential factors to control.

2) Main variable - Response Times: The main recorded variable was the delay (in seconds) between the participants sending their utterances and the utterances they were responding to.

In order to avoid taking into account the influence of the length of the messages in the response time, we calculated a model of our data with a multiple linear regression with an interaction between the length of the previous utterance and the length of the current utterance on the observed delays within utterances. We calculated this model on the conversations with the human (no violations) to constitute the reference for the expected delay without violations.

A theoretical delay was then calculated based on this model for each of the participants' utterances (D).

$$D = (a \times C_p) + (b \times C_c) + (c \times C_c C_p) + d \tag{1}$$

...where C_p is the previous utterance's length (in characters), C_c the current length of the participant's utterance (in characters). a, b, c and d are the coefficients of the model.

The difference between observed delay and theoretical delay was then used in the statistical analysis.

$$\Delta d = d - D \tag{2}$$

... where d is the observed delay.

3) Secondary variable - Identification Ratio: The binary answer to the Turing Test was also recorded and compared to the expected answer to define a successful or failed identifications of the machine actor.

4) Other recorded variables: Other control variables were recorded, among which the participants' age, gender, the duration of each of the two conversations, the self-graded knowledge about artificial intelligence and self-graded knowledge about computer sciences, and the self-graded confidence in their guess in the Turing Test.

C. Procedure

The experiment was carried out online. Participants only needed a computer with an internet connection, and did the experiment at home. Mobile devices were not allowed to avoid additional differences in typing speed depending on machines. The website they connected to was hosted on a private server located in France and had been designed specifically for this experiment [28, for the source code of the website]. Participants information was recorded at the beginning of the experiment through a questionnaire before agreeing to the online consent form. The information was only sent to the server if consent was given.

Once the questionnaire filled, participants were sent to a *ChatBox* where the experimenter (displayed as *Moderator*) explained to the participants the rules they would have to respect during the two conversations and their goal during the experiment [28, for the standardized interactions of the moderator during the experiment].

The participants played the role of the judge in the Turing Test and had to converse with an interlocutor in two different conversations of up to 15 minutes. Participants could interrupt a conversation whenever they had made their decision.

The participants' interlocutors were presented as being a human and an artificial intelligence, both trying to portray the same fictional character (called Andrew) in the most convincing way they were able to. Using the fictional character allowed to avoid questions targeted toward the interlocutors themselves, and also to make sure that not only the machine would have to claim to be someone it was not. For this reason, they were both presented as actors.

The participants were deceived, since no artificial intelligence was in fact present, and both actors were the same experimenter (A male student in experimental psychology). The difference between the experimenter's two roles was that in one case (human actor) he could portray Andrew with no restrictions (except for the pre-defined traits and story of the character). In his other role (machine actor), the experimenter was constrained in the way he could answer, but not in the actual content (except again for the pre-defined traits and story of the character). The type of constraints was expected to trigger different violations of the gricean maxims during the conversation, and thus to produce targeted changes to the feeling of humanness of a participant, its behavior expected to be closer to that of an artificial intelligence. The actor recorded whether or not he had introduced a voluntary violation for each sentence.

Each Participant had a conversation with each of the two interlocutors, in a random order. Smileys, double-posts and hypertext links were not allowed. The maximum length of messages, both for the participant and for the actor, was set to 255 characters.

D. Conditions

a) Quantity: In this condition, the experimenter (when in the role of the artificial intelligence) was instructed to respond briefly to the participant's utterances, providing too little information than would be required to naturally answer the participant when possible.

We expected this condition to cause violations of the first maxim of Quantity[3]. The expected results for this condition were that it would produce a small increase in the response time of the participants.

b) Relation: In this condition, the experimenter (when in the role of the artificial intelligence) was instructed not to use any contextual information from the previous posts of the participant. Whenever not enough information was provided in the participant's last utterance to answer in a relevant way, the experimenter answered with a generic utterance or by switching topics.

We expected this condition to cause violations of the maxim of Relation[4]. The expected results for this condition were that it would produce a large increase in the response time of the participants.

IV. Results

Among all 985 participants' utterances recorded, 467 were from the conversations with the human actor and 518 from the conversations with the machine actor. Among the utterances from the conversations with the machine actor, 189 followed utterances with violations of the maxim of Quantity, 76 followed utterances with violations of the maxim of Relation and 253 followed utterances without violations.

[3] "Make your contribution as informative as is required (for the current purposes of the exchange)."

[4] "Be relevant"

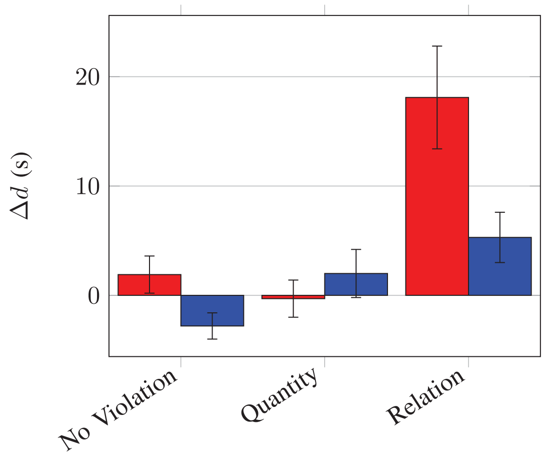

Fig. 1. Differences between observed delay and predicted delay (Δd) in seconds until the participant's utterance is sent, depending on the previous utterance's violation type and the participant's gender (females in red and males in blue).

Utterances from the conversations with the human actor were used to generate the linear model, Utterances from the conversations with the machine actor were used to compare our conditions in the ANOVA (Type III). Contrasts were analyzed using t-tests of comparison of means and the p-value was corrected with a Holm-Bonferroni correction [28, for the complete analysis procedure].

The final answers of the participants to the Turing Tests were analyzed with Chi-Squared tests.

Self-grading on artificial intelligence knowledge and computer sciences knowledge, as well as the confidence scores were analyzed with the Kruskal-Wallis test by Ranks.

A. Linear Model

The regression calculated on the utterances from the conversations with the human actor (no violations) gave the following equation ($R^2 = .4$):

$$D = (0.20 \times C_p) + (0.36 \times C_c) - (0.0005 \times C_pC_c) + 7.4 \quad (3)$$

B. Response Times

A significant interaction was found between the type of violations and the gender ($F_{2,512} = 4.6, p < .01$).

A main effect of the type of violation was found on the delay ($F_{2,512} = 14.9, p < .001$). A main effect of the gender was also detected by the analysis ($F_{1,512} = 2.2, p =< .05$).

1) Interaction with the Gender: The graphical representation of this interaction is shown in Fig. 1.

Δd had a tendency to be higher following a violation of the maxim of Relation for females than for males ($t(57.5) = -2.42, p_{Holm} = .06$).

Δd had a tendency to be higher following a no violation for females than for males ($t(229.9) = 2.26, p_{Holm} = .06$).

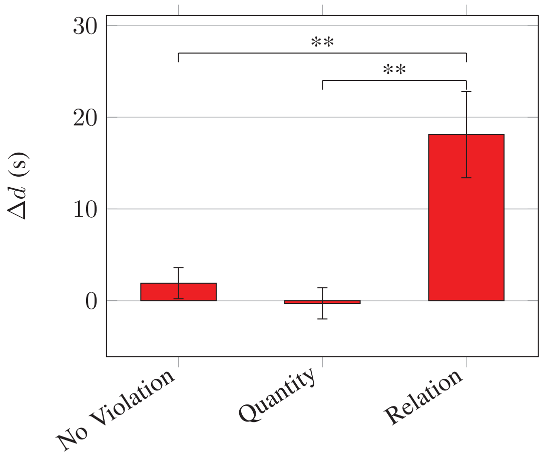

Fig. 2. Differences between observed delay and predicted delay (Δd) in seconds until the participant's utterance is sent, depending on the previous utterance's violation type (Females Only).

$^{**}p < .01$

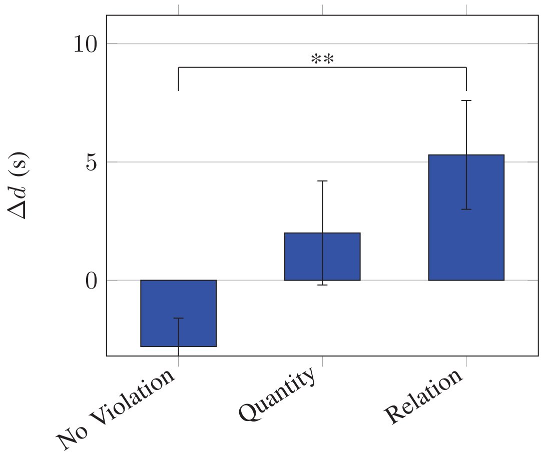

Fig. 3. Differences between observed delay and predicted delay (Δd) in seconds until the participant's utterance is sent, depending on the previous utterance's violation type (Males Only).

$^{**}p < .01$

2) Utterances from female participants: For females (Fig. 2), Δd was significantly higher following utterances with a violation of the maxim of Relation than following utterances with a violation of the maxim of Quantity ($t(50.5) = -3.66, p_{Holm} < .01$) and than following utterances without violations ($t(50.9) = -3.21, p_{Holm} < .01$).

Δd was significantly higher following utterances with a violation of the maxim of Relation than following utterances with a violation of the maxim of Manner ($t(65) = -3.14, p_{Holm} < .01$).

3) Utterances from male participants: For males (Fig. 3), Δd was significantly higher following utterances with a violation of the maxim of Relation than following utterances without violations ($t(53.1) = -3.12, p_{Holm} < .01$).

C. Turing Test

No significant difference on the number of successful identification of the machine actor was found between the conditions with the Pearson Chi-Square Test ($\chi^2(1) = 0.43, p = .51$), yet in the condition of violations of the maxim of Relation, the success rate (70% of correct identifications) shows a tendency to differ from a random chance ($\chi^2(1) = 3.2, p = .07$) while no difference could be found in the condition of violations of the maxim of Quantity (55%, $\chi^2(1) = 0.2, p = .65$).

V. Discussion

A. Quantity

Our results show that violations of the maxim of quantity did not have a significant effect on the response time of our participants. This can be explained by the fact that we only selected the first maxim of Quantity: "Make your contribution as informative as is required." indeed, this specific submaxim has been shown to produce a feeling of humanness when violated [7]. This confirms that violating the maxim of Quantity to provide less information than what is required is not something humans are bothered by in the context of a conversation. Indeed, violating it gave the participants the impression that the actor was not very talkative, bored or upset, but not inhuman. This may seem to be at odds with the experiment of [10], for they have shown that violating the maxim of quantity to provide less information than required confused the partners. Yet the context of the experiments was not comparable. In their experiment, the participants had a task to accomplish, making them rely heavily on the information they received from their partner. In our experiment this was not the case, as the participants only had to keep the conversation going rather than actually rely on the information given.

This can be further explained by the design of the experiments. In [10], the speaker could only say one sentence, which constituted the only information available to the listener. There was no possibility for the listener to ask more details or further explanations. In our experiment, the conversations were free-flowing, and the participant could ask the actor to elaborate at any point, even though he would also give less information than what was required in this elaboration.

Our results for the response times are also coherent with the results of the Turing Test, where no effect could be found on the perceived humanness of the actor in this condition.

B. Relation

A significant increase of the response times in the utterances of male participants was also observable ($M = 5.3$s, $SD = 13.7$s), although it was much smaller than with female participants ($M = 18.1$s, $SD = 30.3$s). The difference between the two genders was not significant though, although a tendency was noticeable.

This increase in the response times validates our hypothesis that violations of the maxim of relation have a strong cognitive effect on the participants of our study, which could be caused by the participant's attempt to retrieve the meaning of the contextualized answer of the actor, and by the detection of an unexpected behavior. This would be coherent with the results of the Turing Test in that condition, which seem to indicate a tendency to diverge from a simple random chance to correctly identify the machine actor, indicating that in this condition, the machine actor did seem less human, which also confirms the findings of [7] regarding this maxim.

C. Linear Model

It is worth noting that the linear model used to predict the normal response time represented the data from the conversations with the human actor rather well, even if it was not perfect ($R^2 = .4$). Perfection was definitely not the goal in this matter, as many other factors are likely to contribute to the variability of the response times, including the processing of the semantic content, moments of memory recollection and other thoughts that are not easily predictable. Thus, our model seems to have been quite efficient at reducing the bias induced by the length of utterances.

D. Limitations

We've shown that response times are a relevant measure of the cognitive cost, yet improvements to the protocol could certainly be made. We do not claim that reading, thinking and typing happen in a perfect sequence in time either. It is highly probable that the participant is already preparing an answer while he or she reads an utterance, and that he or she continues to think about it while he or she types it, maybe even modifying it halfway through. Yet, on average, and as our results confirm, the response times can be relevant enough, provided enough data can be analyzed.

E. Future Works

As this study is the first of its kind on this particular topic we chose not to involve too many conditions, yet many other conversational implicatures could most certainly be tested, including those involving the remaining maxims of Grice. Indeed, violations of the second maxim of Quantity "Do not make your contribution more informative than is required" have a strong negative effect on the humanness of the conversational partner [7]. Other maxims, like the fourth maxim of Manner "Be orderly" would likely have an effect on the response times as well as the order of clauses in an utterance can trigger specific implicatures like cause and effect or temporal continuity [22], and an unexpected ordering of the clauses would certainly make the machine actor feel quite odd to humans.

Another potential area of expansion would be to investigate the differences between different genders of the character and of the actor portraying it. Indeed in our study we've only involved a male actor, portraying a male character: Andrew. Since our results indicate that a gender bias was present, testing the behavior of males and females in same-gender and mixed-gender conversations would provide more data to interpret our findings.

F. Conclusion

This experiment shows the relevance of studying conversations within the framework of the Turing Test by measuring the response times of participants. Indeed, our results show that using response times is sensible enough to catch smaller effects that the binary answer to the Turing Test cannot, and in a much more quantifiable way.

Furthermore, unlike other evaluation methods, using a protocol based on the Turing Test allows participants to take an active role in the conversation.

This could constitute an additional method of evaluating conversational agents within online conversations, including comparing artificial agents between each other. Besides, compared to other evaluation methods, the response times can be recorded directly while the conversation is going on.

Declaration of Interests

The authors of this study declare not having any conflicts of interests relative to the results of this research.

Privacy and Consent

The data collected has been made anonymous to preserve the privacy of the participants. All participants gave their consent and were informed that they could retract it at any moment of the experiment.

Acknowledgment

We would like to thank the services of International Relations of the University of Paris VIII and the University of Western, Ontario, for providing us with international contacts in order to recruit participants for this study in addition to our own personal contacts.

References

[1] X. Yang, Y.-N. Chen, D. Hakkani-Tür, P. Crook, X. Li, J. Gao, and L. Deng, "End-to-end joint learning of natural language understanding and dialogue manager," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 5690–5694. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7953246

[2] D. Braun, A. Hernandez-Mendez, F. Matthes, and M. Langen, "Evaluating natural language understanding services for conversational question answering systems," in *Proceedings of the SIGDIAL 2017 Conference*, 2017, pp. 174–185.

[3] P. Paroubek, S. Chaudiron, and L. Hirschman, "Principles of evaluation in natural language processing," *Traitement Automatique des Langues*, vol. 48, no. 1, pp. 7–31, 2007. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00502700

[4] O. Masson, J. Baratgin, and F. Jamet, "Nao robot, transmitter of social cues: What impacts? the example with endowment effect," in *Advances in Artificial Intelligence: From Theory to Practice. IEA/AIE 2017. Lecture Notes in Computer Science*, S. Benferhat, K. Tabia, and M. Ali, Eds., vol. 10350. Springer, Cham, 2017, pp. 559–568. [Online]. Available: https://doi.org/10.1007/978-3-319-60042-0_62

[5] N. Hellbernd and D. Sammler, "Prosody conveys speakers intentions: Acoustic cues for speech act perception," *Journal of Memory and Language*, vol. 88, pp. 70–86, 2016. [Online]. Available: https://doi.org/10.1016/j.jml.2016.01.001

[6] K. J. Wells, "Noticing students conversations and gestures during group problem-solving in mathematics," in *Teacher Noticing: Bridging and Broadening Perspectives, Contexts, and Frameworks*. Springer, 2017, pp. 183–204.

[7] A. P. Saygin and I. Cicekli, "Pragmatics in human-computer conversations," *Journal of Pragmatics*, vol. 34, pp. 227–258, 2002. [Online]. Available: https://doi.org/10.1016/S0378-2166(02)80001-7

[8] H. P. Grice, *Logic and Conversation*. New York: Academic Press, 1975, pp. 41–58.

[9] D. Sperber and D. Wilson, *Relevance: Communication and Cognition, 2nd Edition*. Oxford: Blackwell, 1995.

[10] P. E. Engelhardt, K. G. Bailey, and F. Ferreira, "Do speakers and listeners observe the gricean maxim of quantity?" *Journal of Memory and Language*, vol. 54, pp. 554–573, 2006. [Online]. Available: https://doi.org/10.1016/j.jml.2005.12.009

[11] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, pp. 1632–1634, 1995. [Online]. Available: http://www.jstor.org/stable/2888637

[12] M. J. Spivey, M. K. Tanenhaus, K. M. Eberhard, and J. C. Sedivy, "Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution," *Cognitive Psychology*, vol. 45, pp. 447–481, 2002. [Online]. Available: https://doi.org/10.1016/S0010-0285(02)00503-0

[13] D. Wilson, "Is there a maxim of truthfulness?" *UCL Working Papers in Linguistics*, vol. 7, pp. 197–212, 1995.

[14] I. A. Noveck, "When children are more logical than adults: experimental investigations of scalar implicature," *Cognition*, vol. 78, no. 2, pp. 165–188, 2001. [Online]. Available: https://doi.org/10.1016/S0010-0277(00)00114-1

[15] D. Sperber, F. Cara, and V. Girotto, "Relevance theory explains the selection task," *Cognition*, vol. 57, pp. 31–95, 1995. [Online]. Available: https://doi.org/10.1016/0010-0277(95)00666-M

[16] G. Politzer, "Laws of language use and formal logic," *Journal of Psycholinguistic Research*, vol. 15, no. 1, pp. 47–92, 1986. [Online]. Available: https://doi.org/10.1007/BF01067391

[17] J. Baratgin and I. A. Noveck, "Not only base rates are neglected in the engineer-lawyer problem: An investigation of reasoners underutilization of complementarity," *Memory & cognition*, vol. 28, no. 1, pp. 79–91, 2000. [Online]. Available: https://doi.org/10.3758/BF03211578

[18] J. Baratgin and G. Politzer, "The psychology of dynamic probability judgment: Order effect, normative theories and experimental methodology," *Mind & Society*, vol. 5, pp. 53–66, 2007. [Online]. Available: https://doi.org/10.1007/s11299-006-0025-z

[19] M. Bagassi and L. Macchi, "Pragmatic approach to decision making under uncertainty: The case of the disjunction effect," *Thinking & Reasoning*, vol. 12, no. 3, pp. 329–350, 2006. [Online]. Available: https://doi.org/10.1080/13546780500375663

[20] H. Bless, T. Betsch, and A. Franzen, "Framing the framing effect: the impact of context cues on solutions to the 'asian disease' problem," *European Journal of Social Psychology*, vol. 28, no. 2, pp. 287–291, 1998. [Online]. Available: https://doi.org/10.1002/(SICI)1099-0992(199803/04)28:2¡287::AID-EJSP861¿3.0.CO;2-U

[21] S. Herring, *Relevance in computer-mediated conversation*, 01 2013, pp. 245–268.

[22] D. Blackmore and R. Carston, "The pragmatics of sentential coordination with 'and'," *Lingua*, vol. 115, pp. 569–589, 2005. [Online]. Available: https://doi.org/10.1016/j.lingua.2003.09.016

[23] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.

[24] S. M. Bowyer, L. Hsieh, J. E. Moran, R. A. Young, A. Manoharan, C. cheng Jason Liao, K. Malladi, Y.-J. Yu, Y.-R. Chiang, and N. Tepley, "Conversation effects on neural mechanisms underlying reaction time to visual events while viewing a driving scene using meg," *Brain Research*, vol. 1251, pp. 151–161, 2009. [Online]. Available: https://doi.org/10.1016/j.brainres.2008.10.001

[25] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, 1996. [Online]. Available: https://doi.org/10.1038/381520a0

[26] P. M. Fitts, "Cognitive aspects of information processing: Iii. set for speed versus accuracy." *Journal of experimental psychology*, vol. 71, no. 6, pp. 849–857, 1966. [Online]. Available: https://doi.org/10.1037/h0023232

[27] S. Hale, "A global developmental trend in cognitive processing speed," *Child development*, vol. 61, no. 3, pp. 653–663, 1990. [Online]. Available: https://doi.org/10.1111/j.1467-8624.1990.tb02809.x

[28] B. Jacquet, J. Baratgin, and F. Jamet. (2018, May) The gricean maxims of quantity and of relation in the turing test - data repository. [Online]. Available: https://doi.org/10.17605/OSF.IO/ETSFW