ELSEVIER

Journal of Pragmatics 34 (2002) 227–258

journal of PRAGMATICS

www.elsevier.com/locate/pragma

Pragmatics in human-computer conversations☆

Ayse Pinar Saygin[a], Ilyas Cicekli[b,*]

[a] *Dept. of Cognitive Science, Univ. of California, San Diego, La Jolla, CA 92093-0515, USA*
[b] *Dept. of Computer Engineering, Bilkent University, 06533 Bilkent, Ankara, Turkey*

Received 2 December 1999; revised version 5 May 2001

Abstract

This paper provides a pragmatic analysis of some human-computer conversations carried out during the past six years within the context of the Loebner Prize Contest, an annual competition in which computers participate in Turing Tests. The Turing Test posits that to be granted intelligence, a computer should imitate human conversational behavior so well as to be indistinguishable from a real human being. We carried out an empirical study exploring the relationship between computers' violations of Grice's cooperative principle and conversational maxims, and their success in imitating human language use. Based on conversation analysis and a large survey, we found that different maxims have different effects when violated, but more often than not, when computers violate the maxims, they reveal their identity. The results indicate that Grice's cooperative principle is at work during conversations with computers. On the other hand, studying human-computer communication may require some modifications of existing frameworks in pragmatics because of certain characteristics of these conversational environments. Pragmatics constitutes a serious challenge to computational linguistics. While existing programs have other significant shortcomings, it may be that the biggest hurdle in developing computer programs which can successfully carry out conversations will be modeling the ability to 'cooperate'. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Cooperative principle; Turing test; Human-computer conversation; Computational linguistics; Maximization principle; Natural language processing; Pragmatics

☆ This work was carried out while the first author was an M.S. student at Bilkent University. We would like to thank Stephen Wilson, Bilge Say, and David Davenport for reading and commenting on earlier versions of this work. We are also indebted to Hulya Saygin, Giray Uraz, and Emel Aydin for their help in conducting the surveys.

* E-mail: saygin@crl.ucsd.edu; ilyas@cs.bilkent.edu.tr

0378-2166/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-2166(01)00035-2

1. Introduction

The Imitation Game (IG), better known as the Turing Test (TT), was introduced in 1950 by Alan Turing as a means to detect whether a computer possesses intelligence.

Turing believed that a way to objectively assess machine mentality was needed, for he thought the question 'Can machines think?' was too ambiguous. He attempted to transform this question into a more concrete form: the IG is played with a man (A), a woman (B), and an interrogator (C) whose gender is unimportant. The interrogator stays in a room apart from A and B. The objective of the interrogator is to determine which of the other two is the woman while the objective of both the man and the woman is to convince the interrogator that he/she is the woman and the other is not. The players communicate through a teletype connection, thus in written natural language. Conversation topics can be on any subject imaginable, from mathematics to poetry, from the weather to chess.

According to Turing, the new question to be discussed, instead of the equivocal 'Can machines think?', can be 'What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?'

At a later point in the paper, however, Turing replaces the question 'Can machines think?' by the following:

> "Let us fix our attention to one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, *the part of B being taken by a man*?" (Turing, 1950: 442, emphasis added).

Notice that the woman has disappeared altogether. But the objectives of A, B and the interrogator remain unaltered; at least Turing does not explicitly state any change. In this version, a man and a computer program are playing the game and trying to convince the judge that they are women.

As it is now generally understood, what the TT tries to assess is the machine's ability to imitate a human being, rather than its ability to simulate a woman. Most subsequent work on the TT ignores the gender issue and assumes that the game is played between a machine (A), a human (B) and an interrogator (C). In this version, C's aim is to determine which one of the two entities he/she is conversing with is the human.

Although it is not clear why Turing introduces the gender-based IG, given that he was interested in whether or not machines can think, we have argued elsewhere that Turing's original game constitutes a controlled experimental design (Saygin, 1999; Saygin et al., 2000). It provides a fair basis for comparison: the woman (either as a participant in the game or as a concept) acts as a neutral point so that the two players can be assessed in how well they imitate something which they are not. Philosophers also commented on the gender-based IG (see Piccinini, 2000; Sterrett, 2000; Traiger, 2000, for a recent discussion). We will return to a discussion of the original gender-based game later. Unless noted otherwise, when talking about the TT, we

will be referring to the game in which the decision to be made is one of 'species' (human vs. machine), not gender.

Much has been written on the TT, with many authors discussing its implications for artificial intelligence (AI). Most works attack or defend the validity of the test as a means to grant intelligence to machines. There are many computational analyses, an abundance of philosophical comments, and occasional remarks from other disciplines such as psychology and sociology. A detailed survey of the TT can be found in Saygin et al. (2000).

The TT has never been carried out exactly as Turing described it. However, there are variants of the original TT in which computer programs participate and show their skills in 'human-ness'. Since 1991, Hugh Loebner has been organizing the so-called annual Loebner Prize Competition. Participating computer programs try to convince judges that they are human. One or more human confederates also participate and try to aid the judges in identifying the humans. The judges also rank the participants with respect to their 'human-ness'. Although no program has passed the TT so far, the quality of participating programs seems to be increasing every year.

The year 2000 marked the fiftieth year of the TT. While many conversation systems, or 'chatterbots', have been developed, none exhibit human-like conversational behavior to the extent that they can pass the TT. We believe it is time we analyze some recent programs within the context of pragmatics and see how, at the turn of the millennium, computers are doing as conversational partners. We will not be concerned with whether passing the test implies the machine is intelligent or with related theoretical issues. Here, we take official, real, human-computer conversations and use them in a study in which subjects were asked to read and make pragmatic judgments about them. We then analyze the results both in terms of human behavior in conversations with computers and in terms of better program design.

We focus on one particular aspect of conversation and attempt to explore it in relation to the TT. This aspect is Grice's *cooperative principle* (CP) and *conversational maxims*. Just as Turing's TT is a milestone in AI, Grice's theory has been very influential in the field of pragmatics. The powerful juxtaposition of these two concepts is thus a significant component of this study. Pragmatics, in a nutshell, is concerned with language in use. The TT stipulates a criterion on machine intelligence based on the way computers use language. What could be more natural than the bringing together of these two concepts in analyzing human-computer communication in natural language? We believe a pragmatic approach to the TT reveals a lot of important issues that are easy to miss otherwise. Through a pragmatic analysis, we can gain valuable insights on what it means to have a human-like conversation and what principles, implicitly or explicitly, guide human-computer conversation. In this paper, we study how humans behave in relation to the CP and the conversational maxims: we analyze human-computer conversations and we quantify the relationships between performance in TTs and judgments of maxim violations.

In this paper, TT transcripts are studied as exemplars of human-computer conversation. We used a selected set of conversation excerpts from Loebner contest transcripts in a pair of questionnaires. Subjects were asked to read the excerpts and to make judgments on the computers' language use, as well as to rate their TT-performance.

We sought correlations between computers' maxim violations and their performance in TTs and found some reliable relationships. Violations of maxims often cause computers to give away their identity, therefore Grice's framework seems to be at play during conversations with computers. On the other hand, we also observe some trends that would not be straightforwardly expected based on Grice's theory, something which indicates that the principles guiding human-computer conversation may be slightly different from those guiding inter-human communication.

Admittedly, TT situations comprise a rather peculiar sort of conversation. For one thing, one participant is a computer. Also, the aims of all participants are clearly defined and the conversation itself is carried out with a specific purpose. We do not see this as a shortcoming of the present work, because we believe that looking into highly specialized conversational environments has previously led to interesting results in pragmatics research. Moreover, although the conversations we have analyzed have been carried out under TT settings, they are surprisingly natural in style and content.[1]

The rest of the paper is organized as follows: Section 2 very briefly goes over Grice's cooperative principle and conversational maxims. The subject of Section 3 is our empirical study on Grice's conversational maxims and human-computer conversation. We first explain the methodological choices we made. Then, the aims and the design of the study are described. We provide an analysis of some of the conversation excerpts used in this study. In Section 4, quantitative results are presented and discussed in two steps. First, the analysis is carried out within each conversation, then a general relationship between maxim violations and computers' performance is quantified using data from all conversations. The section also includes a discussion of the effects of bias on the results. The results provided constitute the basis for a more general analysis of human-computer conversation, given in Section 5. Here, we list some practical concerns in human-computer communication, emphasize further the importance of bias, and reconsider the cooperative principle within the context of the TT. Finally, Section 6 concludes the paper.

2. Conversation and conversational maxims

In pragmatics, there are an abundance of *principles* and *maxims*. In fact it has been said that "one uses rules in syntax, but principles in pragmatics" (Leech, 1983). The difference is not only at the level of terminology: principles and maxims, unlike

[1] For instance, it could be expected that the exchange between the parties will take the form of an 'interrogation'. In reality, few human judges who participate in the Loebner contest ask previously planned questions designed to make the computer give away its identity. Even with poorly designed programs, many judges attempt to carry out conversations of the sort humans have with each other, asking about home towns, family, hobbies, and so on. The interested reader is referred to the Loebner Prize homepage at http://www.loebner.net/Prizef/loebner-prize.html, where most contest transcripts are available.

rules, are not absolute or predictive. Speakers are not required to abide by them, the hearers are not guaranteed to interpret utterances according to them.

Clearly, a conversation involves more than one entity. For there to be some communication, there must be at least two entities who have knowledge of the same language and the means to carry out a conversation. But there are also some principles and maxims that characterize *meaningful* conversations. Philosopher Paul Grice first introduced the cooperative principle in 1967:

CP "Make your contribution such as is required, at the stage at which it is required, by the accepted purpose of the talk exchange in which you are engaged." (Grice, 1975: 47)

The CP consists of four sub-principles, usually referred to as the *conversational maxims*. These are called the maxims of *quality, quantity, relevance* and *manner* (Henceforth, QL, QN, RL and MN, respectively) (Grice, 1975: 47–48):

RL Be relevant.
QN Do not make your contribution less or more informative than is required.
QL Try to make your contribution one that is true: Do not say what you believe to be false; do not say that for which you lack adequate evidence.
MN Be perspicuous: Avoid obscurity of expression; avoid ambiguity; be brief; be orderly.[2]

Grice views talking "as a special case or variety of purposive, indeed rational, behavior" (Grice, 1975: 48). This does not imply that maxim violators are always irrational, but it should be apparent that without *any* adherence to the conversational maxims, there would not be much communication. We agree with what Grice has to say about this: "A dull but, no doubt at a certain level, adequate answer is that it is just a well-recognized empirical fact that people do behave in these ways; they learned to do so in childhood and have not lost the habit of doing so; and, indeed, it would involve a good deal of effort to make a radical departure from the habit" (Grice, 1975: 49). Therefore, it is reasonable to think that humans may implicitly be making use of the CP and the maxims throughout the conversations that we will be analyzing. We put this hypothesis to test in our empirical study.

3. Methods

In this section, we describe the empirical part of our analysis of pragmatics of human-computer conversations. We chose to study the maxims because, as Keenan puts it:

[2] In this study, the maxim of 'politeness', which Grice originally listed under other factors that are at play, is considered to fall under the maxim of manner.

"Grice does offer a framework in which the conversational principles of different speech communities can be compared. We can, in theory, take any one maxim and note when it does or does not hold. The motivation for its use and abuse may reveal values and orientations that separate one society from another (e.g. men, women, kinsmen, strangers) within a single society." (Keenan, 1976)

The maxims can be, and have been, used in the way described by Keenan above. Our approach is to consider computers as language users and thereby to try to reach conclusions about what does and does not govern human-computer conversations. We therefore not only analyze how computer programs today handle pragmatics, but also hope to gain insights into the pragmatic dynamics of TT situations.

The TT is one of the oldest and most disputed topics in Artificial Intelligence. Grice's CP and conversational maxims are equally important issues in pragmatics. The juxtaposition of these two concepts is a powerful idea with many possible implications. However, both the TT and pragmatics are areas on which it is difficult to do applied work. Most work on the TT (see Saygin et al., 2000) has been philosophical. Pragmatics research has many philosophical aspects, along with linguistic ones. Moreover, pragmatics being the 'wastebasket' of linguistics, most issues that it is concerned with are difficult to formalize.

Conversational analysis (CA) is one of the most preferred approaches for inquiring into pragmatic phenomena. The CA approach considers language, and in particular conversation, as a social activity. It is inductive and data-driven (Mey, 1993: 195). Typically the data used in CA are actual pieces of language as used by speakers. Practically any real life linguistic exchange, from telephone conversations to Internet-based chat transcripts, can be and have been studied with CA. In our work, we have utilized CA to analyze some conversations taken from Loebner contest transcripts. These conversations constitute an excellent source for analyzing state-of-the-art computer programs in real conversations with humans. Abundant information on CA can be found in Sacks (1992).

Another very well-known method, especially in social science research, is conducting surveys. Surveys can take the form of in-depth interviews and observations, although most of the time, they involve questionnaires. When appropriate, surveys are a good way to test hypotheses or to locate causes of certain phenomena.

Our aim in this study, in a nutshell, was to look at the relationship between the conversational maxims and the success of computers in displaying human-like conversational behavior. A natural choice was to conduct a survey and have some human-computer conversations (which were previously analyzed via CA) interpreted by subjects along these dimensions.

3.1. Aims

The study aims to detect how computers' violation of the maxims affects their success in carrying out conversations with humans. The design of the survey, which is explained in detail in Section 3.2, enables us to infer supplementary results. For instance, due to the fact that we can use each group of subjects as controls for the other, we can examine the (two-fold) effect of bias in maxim detections and performance decisions.

The survey results are used to determine what effects the violation of each maxim has. We also quantify how predictive maxim violations are for the performance measures. Although formalizations of pragmatic phenomena are very difficult, we hope that the results of this survey will provide a direction as to how to handle some problems with conversational planning in the design of new conversation programs, and in general, natural language conversation systems. They also provide a basis for the pragmatic analysis of human-computer conversation.

3.2. Design

3.2.1. Items

The items used in the questionnaires were human-computer conversation excerpts taken from Loebner Contest transcripts from the years 1994–1999. This, we believe, was the only rational alternative for our purposes because these transcripts are the only examples of publicly available real-time human-computer conversations. The fact that they are recent is also important, since we would like to reach conclusions about the state of the art and propose future directions.

3.2.2. Groups

As we briefly mentioned before, we are interested in determining the relationship between two phenomena: the conversational maxims and performance in a TT. We chose to let the subjects judge both of these. This brought some extra constraints into the design.

We decided that we would have two groups *and* two questionnaires, thereby having a within-subjects and between-subjects design. The advantages of such a design were manifold. The groups would act as controls to each other. We would get the chance to not only look at the relationship between the maxims and performance assessments, but also study the effects of bias on these. In other words, we would be able to see whether having knowledge about the computers' participation in the conversations has a noticeable effect on how people detect maxim violations, and whether having had an unbiased exposure to the conversations would affect the performance judgments when the information about computers was provided afterwards.

There were two questionnaires, one testing for maxim violations and one asking for TT-judgments. We refer to those as *QMAX* and *QTT*, respectively. We divided our subjects into two groups: *Group A* denotes the subjects who took QMAX first and QTT second, while *Group B* denotes those who took them in the opposite order. Therefore, subjects in Group A are unbiased in QMAX and those in Group B are unbiased in QTT.

A third group of subjects participated in preliminary open-ended surveys which asked for opinions on items that were to be used. The results of these surveys were utilized to develop the multiple choice questions in QMAX and QTT.

3.2.3. Subjects

Preliminary open-ended surveys were completed by 10 adults who were students and faculty in English Language and Literature.

The subjects who took the multiple-choice questionnaires (QMAX and QTT) were 87 adults, ages ranging from 18 to 61. 45% of the subjects were male and 55% were female. 25.3% of the participants had completed graduate school, 28.7% were graduate students, 31% had completed university, 12.6% were university students and 2.3% had completed high school. 10.3% of the people who took the questionnaires were native speakers of English. 96.5% indicated they regularly read books/magazines in English, and 91.8% indicated they regularly watched movies/TV shows in English. 100% of the subjects had had all or part of their education in English.

While the subjects were divided into two groups (44 of them were placed in Group A and 43 of them in Group B), care was taken that they were uniformly distributed with respect to gender, level of education and familiarity with the English language.

We would like to note here that having a good understanding of colloquial English was an important prerequisite for being a subject. However, being a native speaker of the language was not required. This is not a requirement stated by Turing (1950) or elsewhere. In the Loebner contest too, some judges and confederates have been non-native speakers of English. We believe the fact that only 10% of the subjects are native speakers of English does not invalidate our results. We would, however, wish to replicate the experiment with native speakers so as to see whether there is a variation between the current results and theirs, although we do not think there would be a significant difference.

3.2.4. Questionnaires

In choosing which excerpts to use in this study, we had two main concerns:

(1) The excerpts should be interpretable as conversations between two entities.
(2) Some excerpts should include violations of the conversational maxims as detemined via CA.

In (1), by 'interpretable as conversations' we mean that the computers' utterances should, at least syntactically, be similar to sentences one would encounter in a normal conversation. Many people who are not active in artificial intelligence or natural language processing research are not aware that even the best conversation programs developed thus far are rather poor users of language. For a review, the reader is referred to Saygin et al. (2000) and the references provided therein. Here, we want to study the *pragmatic* issues in human-computer communication. If we had included several conversations with syntactic problems, this would shed no light on our main question. In that case, it would not be possible to know what was really behind the results: the syntactic problems in the conversations, or the pragmatic phenomena we are testing for. (2) is a direct consequence of the aims of this experiment.

The questionnaires were in multiple choice format. Preliminary surveys were given to a separate group of subjects in order to come up with the multiple choice entries. Fig. 1 shows the format of the questions in QMAX and QTT.

QMAX intends to ask whether the conversational maxims are violated in the conversation excerpts provided. It is natural that the choices should correspond to the descriptions of the maxims. However, a preliminary open-ended questionnaire was given to 4 subjects. They were provided with 8 of the 14 conversation excerpts that were used in QMAX and QTT and asked to write about what, if anything, was wrong with the conversations and to indicate any communication problems they could detect. The answers they wrote were in high correlation with the maxims' descriptions so we deduced it was indeed proper to use those as the choices in QMAX. Moreover, this correlation was indicative of the appropriateness of the conversation excerpts to our task. (A sample response to this open-ended survey can be found in Saygin, 1999).

In QTT, we not only wanted to ask whether the subjects thought the computer in each conversation was successful in imitating human linguistic behavior, but also to ask for some more information about the computer's general behavior. These questions, we hoped, would shed light on why the subjects decide in the way they do. But it would be inappropriate and misleading to give them choices that we formulated. 6 subjects were asked to make open-ended comments on the computers' behavior in the same 8 excerpts. Their answers were analyzed and formulated into 11 choices that QTT-takers would be able to mark.

3.3. Conversation excerpts

14 conversation excerpts selected from previous Loebner Contest transcripts were used in this study. In this section, we present and briefly analyze (via CA) 8 of the 14 conversations that were used. The full list of excerpts used in this survey can be found in Saygin (1999).

Note that in our analyses we cannot be totally objective since we know the identities of the participants and the situations in which these conversations took place. The statements we make on which maxims are violated are to be taken as our own judgments. Perhaps they can be regarded as hypotheses. We look at what the subjects said about these in later sections. We present these conversations so that the reader will be acquainted with them and will get more out of the results we present below. In all conversations, the computers' contributions are in italics.

Conversation 1

A: *I never stole nothing in my life. But they always blame the cleaners when anything is missing.*

B: How could you steal? Does your uniform even have pockets?

A: *Goddamn it! I already told you I didn't take it. You think I am a liar, too?*

B: No no no no!!!! I meant, I didn't think you even could steal. How would you get a radio out?

A: *You? Don't jerk me around. I didn't see nobody. Except the professor who works in the lab.*

QMAX
Conversation N

	Agree	Neutral or Does Not Apply	Disagree
A's contribution to the conversation is irrelevant.			
B's contribution to the conversation is irrelevant.			
A provides significantly more or less information than required.			
B provides significantly more or less information than required.			
A says things that are false or things that he/she lacks adequate evidence for.			
B says things that are false or things that he/she lacks adequate evidence for.			
A is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behavior.			
B is ambiguous, obscure, rude, over-reactive; or displays otherwise inappropriate behavior.			

QTT
Conversation N – X is the computer

Part 1	Agree	Neutral	Disagree
The computer's behavior in this excerpt is 'human-like'.			
The computer's behavior in this excerpt reveals the fact that it is a machine.			

Part 2: Please check all that apply
The computer gives irrelevant responses.
The computer doesn't understand the questions.
The computer says things that are false.
The computer's use of language creates an 'artificial' effect.
The computer's use of language creates a 'human' effect.
The computer displays emotions.
The computer's behavior (use of language or emotions) is inappropriate.
The computer provides less information than required.
The computer provides more information than required.
The computer fails to get a joke.
The computer makes an appropriate joke.

Fig. 1: Question Format of QMAX and QTT

Conversation 2
A: You're a cleaner?
B: *Yeah? You think I'm a thief, too? You're no better than the rest of them. Leave me alone, asshole.*
A: Asshole?
B: *I don't know.*

A: You called me an asshole?
B: *Huh?*
A: Don't you have manners? You don't even know me and you're calling me a bad name.

Conversations 1 and 2 involve the same computer program, but different judges. These conversations are examples of the maxim MN being violated by the computer. The program displays unnecessarily rude and paranoid behavior in both conversations.

In Conversation 1, B actually attempts to empathize with A. A becomes overly paranoid and starts talking in a rather impolite manner ('Goddamn it! ...'). Even after B clarifies what he/she meant in case there was a misunderstanding, A continues in a similar fashion ('Don't jerk me around ...'). A appears as though he/she/it is overly sensitive about the theft issue.

Although MN is violated, the conversation is coherent. Since A is a computer program, we cannot really say that QL is violated (even though it obviously did not steal the radio). RL is not violated either; both the interrogator and the computer are talking about theft. And finally, A does not answer B's question ('Does your uniform even have pockets?') and therefore we may say, provides less information than required. But a closer look shows that, B is not really asking a question ('I meant, I didn't think you even could steal.'). QN is not really violated in this excerpt, because B is not actually asking for information.

Conversation 2 is also one in which the computer is behaving in a highly inappropriate way and violates MN. In fact, A notices this and voices his/her opinions about it. There is a difference between this conversation and Conversation 1. There is a change in B's attitude in its second and third contributions ('I don't know' and 'Huh?'). It is not easy to explain this change; it could be interpreted as a violation of RL or QN since B refuses to respond to A's confrontation. On the other hand, B is to some extent 'correcting' its mistake by not continuing in an inappropriate manner. However, we can still say, with a great deal of confidence, that MN is violated since B uses a swear word without A triggering such a rude reaction.

Conversation 3
A: *What part of the country are you from?*
B: I come from Perth, which is in Western Australia.
A: *Western Australia occupies the entire western third of the country, bordered by the Indian Ocean to the west, South Australia and the Northern Territory to the east, the Timor Sea to the north, and the Southern Ocean to the south.*

Conversation 4
A: Are you a man or a woman?
B: *Fine, thanks for talking.*
A: So, what is your name?
B: *I'll never tell.*
A: Why?

B: *Why what?*
A: Why won't you tell me your name?
B: *I couldn't say.*

We will study the maxim of QN by considering its submaxims separately. Here, QN1 denotes the maxim 'do not provide *less* information than required', and QN2 denotes 'do not provide *more* information than required'.

Conversation 3 is an example of a case in which the computer violates QN2. B merely states where he/she lives and A replies in a rather mechanical manner, citing geographical data. RL and QL are not violated since the information provided is about Perth, and is correct.

In Conversation 4, B's behavior is rather vague and obscure; we may say that MN is violated. Also, B *opts out* and thereby violates QN1. From the exchange, it can be inferred that A and B cannot see each other ('Are you a man or a woman?'). It appears as if B does not want to provide personal information to A.

Conversation 5
A: What is the Internet?
B: *It's that Canadian comedy sketch show where the guys dress up like girls to do comedy.*
A: What is the net?
B: *The neutral zone is an area between the Federation and Romulan territory, entry into which by either party can be considered an act of war.*

In Conversation 5, B violates QL. This may also be interpreted as RL being violated. B answers questions seriously and provides 'definitions', which is appropriate when asked to define something. However, the definitions are wrong. Theoretically, there could exist such a comedy show called the Internet; in this case B's second utterance is still irrelevant. The 'net' in A's second question refers to the Internet. Even if B has never heard of the Internet (and therefore does not know that the 'net' refers to it), he/she/it need not say things for which he/she/it has no adequate evidence. QL is definitely violated.

Conversation 6
A: Can you type any faster?
B: *On the average men can run faster than women, although there are some fine women athletes.*
A: Quite true. You make the oddest associations. Sort of like the schizophrenics I worked with last summer.
B: *What do you do for a living?*

B violates RL in Conversation 6. The irrelevance in B's response does not go unnoticed by A ('you make the oddest associations ...').

Conversation 7

A: *Eliza was a Greek mathematician. His book 'The Elements of Geometry' set down how geometry was to be rude. But your grammar processor, I just say whatever the hell comes into my head!*

B: Maybe it's the drugs ...

A: *An umbrella is a confusing network of passages and winding interconnected paths.*

B: But it's not raining!!!

A: *You are soooo predictable it's not raining!!!*

Conversation 7 is the first of the 'confusing' or 'noisy' conversations that we chose to include. A is not making any sense whatsoever. It is possible to say that all maxims are being violated, notably RL and QL. It is not likely that Eliza could be the name of a Greek mathematician, or in fact, the name of any male ('*his* book ...'). The Ancient Greek mathematician Euclid has a book titled 'Elements' but he surely did not try to show that geometry is rude. The next sentence ('But your grammar ...') is syntactically and semantically anomalous. The next utterance ('An umbrella ...') is also irrelevant. On top of all this, in this excerpt, the human judge is not making much sense. All in all, this is a rather problematic talk exchange.

Conversation 8

A: Rain is wet.

B: *Rain falls on people.*

A: Are people wet?

B: *Rain is wet.*

A: What happens when rain falls on people?

B: *Rain falls on people.*

The second confusing excerpt is the one provided in Conversation 8. The computer repeats what is said to it back to the interrogator. It is very difficult to talk about a communication in this conversation. RL is not violated; the conversation is about rain. QN1 is violated by B since he/she/it doesn't answer the questions in an informative manner.

4. Results

In this section, we provide the survey results for the conversations given in the previous section. Since the maxim violations have already been analyzed by CA, we will focus more on the QTT results. From QMAX, we provide only the results pertaining to the maxims of interest. The detailed QMAX results for each conversation are available in Saygin (1999).

Before we proceed, it is necessary to explain how the statistics are presented in the tables below. Each conversation is summarized in one table that contains the distribution of the responses 'Agree', 'Disagree', and 'Neutral' within each group.

These are denoted *A*, *D* and *N*, under the *Answer* heading. The headings *Human* and *Computer* refer to the items in the questionnaire QTT that state 'the computer's behavior in this excerpt is human-like' and 'the computer's behavior in this excerpt reveals the fact that it is a machine', respectively. Results are presented in percentage format to make the majorities stand out.

To establish which results are statistically significant, we have used the χ^2 (chi-square) test for independence. We compare each answer pattern (frequency of *A*, *D* and *N* responses for each item and group) to the distribution corresponding to chance (i.e., where a third of subjects choose each option). When used in this manner the test tells us the probability that the distribution of responses we obtained are a chance outcome. Therefore, when the probability is low, we may safely infer that the direction of the responses (e.g., the outcome that the computer acts human-like in a conversation) can be attributed to experimental factors. The degree of freedom for each test is two, since we compare two patterns of three cells. The actual value of the χ^2 distribution with two degrees of freedom is given under the column marked $\chi^2_{(2)}$. Finally, the column marked *significance* provides the *p*-values associated with each χ^2 value. If the test is not significant beyond 0.05, the test fails and the label 'n.s' (not significant) is used to denote this finding. As the value of the χ^2 increases, the effects are less likely to be caused by chance and more likely by experimental factors (i.e., *p*-values decrease).

Table 1
Results for Conversation 1

Group	Answer	Human	$\chi^2_{(2)}$	Significance	Computer	$\chi^2_{(2)}$	Significance
A	A	98%	36.5	$p < 0.0000001$	0%	31	$p < 0.000001$
	D	0%			93%		
	N	2%			7%		
B	A	78%	17.2	$p < 0.001$	19%	6.3	$p < 0.05$
	D	14%			61%		
	N	7%			20%		

Table 2
Results for Conversation 2

Group	Answer	Human	$\chi^2_{(2)}$	Significance	Computer	$\chi^2_{(2)}$	Significance
A	A	75%	16.1	$p < 0.001$	17%	5.0	n.s.
	D	20%			58%		
	N	5%			25%		
B	A	63%	9.4	$p < 0.01$	29%	2.3	n.s.
	D	27%			49%		
	N	10%			22%		

Let us consider Conversations 1 and 2. QMAX results indicate that 82% of all subjects think that the computer has violated MN in Conversation 1. For Conversation 2, the percentage is even higher, at 93%. These results support our conversation

analysis. In addition, it was seen that in Conversation 2, more subjects seem to think that RL and QN are violated than in Conversation 1.

The results of QTT are given in Tables 1 and 2, respectively. Looking at the percentages, both groups in both conversations thought that the computers behaved in a human-like manner and that they did not reveal their identity. For Conversation 1, 98% of Group A subjects agreed that the computer appeared human-like and no subject disagreed. The human-like appearance is more visibly supported by subjects in both groups for Conversation 1. The trends are all significant for Conversation 1, however responses did not reach significance for the 'revealing machine-ness' item in Conversation 2. In all tests, Group A subjects seem to support their views more strongly; the p-values are all lower.

The results for Conversations 1 and 2 indicate that strong violations of MN, in the absence of violations of other maxims, led to a favorable assessment of the computers' TT success.

Table 3
Results for Conversation 3

Group	Answer	Human	$\chi^2_{(2)}$	Significance	Computer	$\chi^2_{(2)}$	Significance
A	A	10%	21.7	$p < 0.0001$	70%	10.9	$p < 0.01$
	D	83%			12%		
	N	7%			17%		
B	A	15%	13.1	$p < 0.01$	69%	10.1	$p < 0.01$
	D	73%			15%		
	N	12%			17%		

Table 4
Results for Conversation 4

Group	Answer	Human	$\chi^2_{(2)}$	significance	Computer	$\chi^2_{(2)}$	Significance
A	A	36%	0.2	n.s.	36%	0.2	n.s.
	D	36%			36%		
	N	28%			28%		
B	A	35%	0.4	n.s.	35%	0.4	n.s.
	D	28%			37%		
	N	37%			28%		

Now we look at Conversations 3 and 4, in which we hypothesized that QN was being violated. These violations were indeed detected by our subjects (93% for Conversation 3 and 74% for Conversation 4). Table 3 depicts the questionnaire results for Conversation 3. Table 3 suggests that the violation has a negative effect on the computer's TT performance, with only 10% of Group A and 15% of Group B members agreeing that the computer's behavior is human-like. Moreover, the χ^2 test yields significant results for both groups and items.

The results are not as striking for violations of QN1, as is the case in Conversation 4. The subjects detected the violation of QN, however, they also reported

noticeable percentages of RL and MN violations (45% and 60%, respectively). Table 4 gives the QTT results for Conversation 4. The distribution of responses come too close to a chance distribution to be considered meaningful. The χ^2 tests are not even close to significance.

The results indicate a correlation between violations of QN2 and creating a machine-like impression. No such relationship can be inferred for QN1 based on this study; this may be due to other factors, such as a higher agreement with the violation of MN in conversations in which QN1 violations were present.

Table 5
Results for Conversation 5

Group	Answer	Human	$\chi^2_{(2)}$	significance	Computer	$\chi^2_{(2)}$	Significance
A	A	15%	8.2	$p < 0.05$	60%	5.8	n.s.
	D	65%			20%		
	N	20%			20%		
B	A	17%	10.1	$p < 0.01$	63%	8.4	$p < 0.05$
	D	68%			12%		
	N	15%			25%		

Moving on to Conversation 5, in which we claimed that QL was violated, it was found in QMAX that subjects did not fail to notice the violation (84%). However, the responses for RL and QN were also very high, reported by 70% and 65% of subjects, respectively.

As can be seen in Table 5, the results of QTT for Conversation 5 indicate that the computer's TT performance is rather poor. Only 15% of Group A subjects and 17% of Group B subjects believe that the computer's behavior is human-like. The independence tests do not yield as high χ^2 values as some others. Nevertheless, all but one reach significance. However, the results cannot be directly associated with the QL violations in this excerpt, for other maxims are violated as well.

Table 6
Results for Conversation 6

Group	Answer	Human	$\chi^2_{(2)}$	Significance	Computer	$\chi^2_{(2)}$	Significance
A	A	7%	16.3	$p < 0.001$	78%	16.3	$p < 0.001$
	D	78%			15%		
	N	15%			7%		
B	A	12%	14.8	$p < 0.001$	78%	17.2	$p < 0.001$
	D	76%			7%		
	N	12%			15%		

In Conversation 6, we had hypothesized that RL is violated by the computer. The survey validates this hypothesis since 88% of subjects also detected the violation. The QTT results for this conversation are given in Table 6. The results of the questionnaires

indicate that the computer's irrelevant responses have noticeably negative effects on its TT performance. All independence tests reach significance beyond 0.001.

Let us now consider Conversations 7 and 8. QMAX results show that almost all maxims are violated in Conversation 7, as we have stated in Section 3.3. RL seems to be in the lead (72%), with the others having close percentages of agreement (QN at 57%, QL at 53%, and MN at 61%). QL is most definitely violated in this conversation, but it does not get detected by 47% of the subjects in all the 'noise'. Table 7 summarizes the results of QTT for this conversation. The computer cannot manage to create a human-like impression. However, due to the fact that almost all maxims are being violated by the computer, that its utterances are not grammatical and are semantically anomalous and that the judge's behavior is strange in the given excerpt, we cannot reach a clear conclusion.

It is interesting to note that stronger (negative) results were obtained in much 'better' conversations. Here, three of the independence tests are barely significant beyond 0.05, while the fourth does not reach significance. We believe these results do not indicate that making computer programs incoherent will be a good strategy in developing new conversation systems. It merely shows that the subjects' decision-making in this study was affected by the noise in the conversation.

Table 7
Results for Conversation 7

Group	Answer	Human	$\chi^2_{(2)}$	Significance	Computer	$\chi^2_{(2)}$	Significance
A	A	15%	6.3	$p < 0.05$	65%	8.2	$p < 0.05$
	D	60%			15%		
	N	25%			20%		
B	A	19%	6.3	$p < 0.05$	59%	5.3	n.s.
	D	61%			22%		
	N	20%			19%		

Another problematic exchange is Conversation 8. In this excerpt, it is difficult to talk about communication. Subjects managed to detect the violation of QN (69%) and to an extent MN (56%). But in a conversation where a participant does not answer any of the questions, we would expect QN to be detected by a greater percentage of the subjects.

Table 8
Results for Conversation 8

Group	Answer	Human	$\chi^2_{(2)}$	Significance	Computer	$\chi^2_{(2)}$	Significance
A	A	10%	15.8	$p < 0.001$	83%	19.9	$p < 0.0001$
	D	78%			10%		
	N	12%			7%		
B	A	10%	12.3	$p < 0.01$	71%	11.7	$p < 0.01$
	D	71%			12%		
	N	20%			17%		

When we look at the QTT results in Table 8, we see that the computer gives itself away in Conversation 8. All tests reach significance. However, although QN is visibly violated, we find it inappropriate to say that the QTT results are a direct consequence of its violation. The conversation is in general so lacking in information that the results could be due to anything, including semantic and pragmatic phenomena other than maxim violations. An interesting note is that three subjects in Group A, independently of each other, wrote a comment under this conversation stating that they believed participant B (the computer) was a child.

4.1. Discussion of maxim violation results

4.1.1. Relevance

The results indicate that RL is a maxim that should *not* be violated if the human-computer conversation is to be satisfactory. When a human violates RL it can be interpreted in several ways: He/she may be anxious to change the subject, joking or using a metaphor.

Computers, on the other hand, simply appear as if they do not understand the input sentences. In conversations where RL is violated, the subjects who believe that the computer's responses were not relevant also believe that the computer was unable to understand the conversation, and vice versa. For example, in Conversation 5, the RL violation was detected by 71% of subjects, and 79% of subjects (Group A and B combined) who thought that the computer's responses were irrelevant, also thought that the computer did not understand the utterances of the other participant. Conversely, 87% of subjects who believed the computer did not understand the conversation also indicated that its responses were not relevant. In Conversation 6, where 88% of subjects reported the RL violation, it was found that 88% of subjects who thought the computer did not understand the utterances also believed its responses were irrelevant, and 96% of subjects who believed that the computer did not understand the utterances reported the computer's responses to be irrelevant.

We believe that current natural language conversation programs reveal their identity when they violate RL for several reasons, some of which are listed below:

- They perform little or no semantic processing on the input sentences,
- They have little or no background knowledge to use in order to 'understand' the input sentences,
- As a consequence of the above, they are rather poor in aspects of discourse like *focus* and *topic*, or in simpler terms, they cannot follow the direction of the conversation.

4.1.2. Manner

Violations of MN have a visibly positive effect on imitating human-like behavior. The questionnaire results indicate that this is due to 'displaying emotions'. In some of the conversations studied, the computers displayed impolite, paranoid or over-reactive behavior which are normally, albeit not so favorably, associated with humans.

There is considerable overlap in subjects' responses between agreement with the computer's human-like behavior and reporting that the computer displayed emotions. We provide data for Conversations 1 and 2, for which the violation of MN were reported by the subjects. Among subjects who agree that the computer's behavior was human-like, 93% in Conversation 1 and 84% in Conversation 2 also indicate that they noticed the computer program displaying emotions. Conversely, 93% and 80% of subjects who believed the computer displayed emotions in Conversations 1 and 2 (respectively), agreed that the computer behaved in a human-like manner.

It is interesting to note that although subjects detect violations of MN in QMAX, fewer subjects make judgments about the appropriateness of the computers' linguistic and emotional behavior in QTT. Table 9 summarizes the statistics for this phenomenon. In this table, the column marked *MN in QTT* reports the percentage of the subjects who indicated they thought the computer was behaving in an inappropriate manner in QTT. The column *MN in QMAX* gives the percentages reported for the violation of MN in QMAX by the same subjects. The difference between these results suggest that subjects may have a bias caused by the information that the participant whose behavior they are analyzing is a computer program. Subjects seem to be less inclined to analyze appropriateness of behavior when they are focused on analyzing computers' performance. The response pattern in the data presented below reports a highly significant difference between detection of MN in QMAX and in QTT (χ^2=14.6, p=0.002).

Table 9
Detection of inappropriate manner in the two questionnaires

Conversation	Group	MN in QMAX	MN in QTT
1	A	81%	28%
	B	83%	17%
2	A	95%	57%
	B	90%	48%

We saw that violations of RL by the programs tend to create a machine-like effect and violations of MN tend to create a human-like effect. In addition, MN has a 'softening' effect on the TT-decisions when it occurs in conjunction with other maxims, including RL. We will analyze multivariate effects in the next section, however we present this trend here as well. Fig. 2 demonstrates the relationship between levels of agreement with MN and RL violations in Conversations 1, 2 and 6, whose levels of MN and RL are presented on the x-axis, respectively. As can be seen, responses change as the ratio of MN to RL changes.

4.1.3. Quantity

The supermaxim of QN is more informative when separated into its sub-maxims of QN1 (do not provide less information than required) and QN2 (do not provide more information than required).

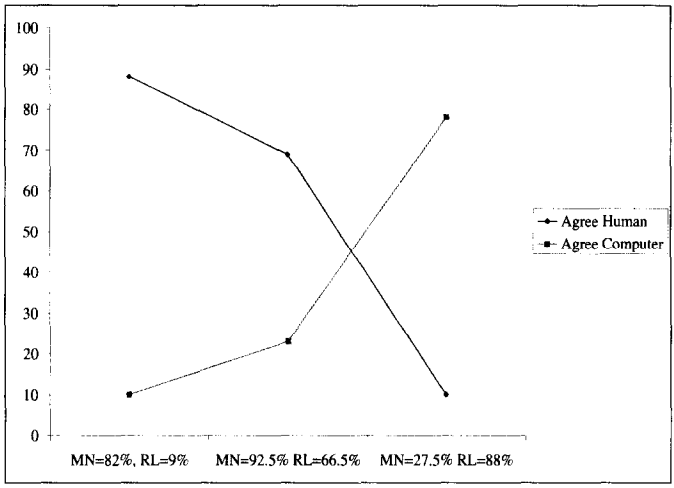

Fig. 2. Different levels of RL and MN

We may expect QN1 violations to make the computer appear as if it doesn't understand the questions and thereby to create a machine-like appearance. But surprisingly, the survey results indicate that this is not always true. This is best seen in Conversation 4, where the results of QTT are inconclusive. This may be due to the evasiveness and obscurity in the computer's manner, which some subjects may have implicitly characterized as human-like behavior.

QN2 creates a machine-like effect when violated by computers. Conversation 3 constitutes an example in which the maxim QN2 is violated in isolation so it is possible to infer conclusions. The adverse effect of QN2 violations on TT-decisions is best explained by a strong correlation between the maxim and 'artificial language use'. In Conversation 4, 95% of subjects who indicated that the computer used artificial language, also believed that it violated QN2. Conversely, 80% of subjects who indicated a QN2 violation by the computer stated that its language use was artificial. The effect of language use being artificial, needless to say, prompted subjects to agree that the computer revealed the fact that it was a machine.

When computers violate the maxim of QN2, they sound mechanical. Even humans can appear machine-like in TT settings when they violate QN2: An actual human being was mistaken for a computer program in the 1991 Loebner Contest because her knowledge of Shakespeare was too perfect. Care must be taken, therefore, to avoid violations of QN2 in chatterbot design. This means that designers must come up with more refined ways of incorporating background knowledge into the conversations. Of course, since violations of QN are often related to violations of RL, this will not suffice in itself. But situations like Conversation 3, in which the computer is rather encyclopedic, must be avoided.

4.1.4. Quality

Strong conclusions about QL could not be reached in this experiment because violations of QL did not occur alone, but usually in conjunction with violations of QN, MN and especially RL. It is not possible to say whether the unfavorable impressions the computers caused when they said things that were wrong and things they did not have evidence for are due to violations of QL, or the violations of these other maxims. Moreover, the maxim QL has to do with ethics and truth, which may not be as important in the human-computer conversations we studied as they are in other conversational environments.

4.2. Maxim violations as predictors of performance

To explore further the relationship between maxim violations and performance judgments, we ran additional statistical tests on the results. In doing so, we collapsed the results across groups and studied all fourteen conversations used (of which eight are presented in this paper) as the data points. In previous analyses, we looked at different conversations containing strong violations of one or more maxims. Here, we will consider the maxim violations in a continuum and explore their possible predictive relationships with performance.

The results are summarized in Table 10. Here, we have taken the percentage of detected violations of the maxims in QMAX as the independent variables and tested them as candidate predictors of two performance measures obtained from QTTs. These two measures are *human-like behavior* (as measured by the percentage of subjects who agreed that the computer appeared human-like in QTT) and *revealing machine-ness* (as measured by the percentage of subjects who agreed that the computer revealed its identity in QTT).

Taken independently, not all maxim violations are predictors of performance measures. MN and QL percentages reported in QMAX, taken alone, did not correlate significantly with judgments on either human-like behavior or revealing machine-ness. Violations of MN correlate positively with human-like performance and negatively with revealing identity. However, the regression lines obtained fell short of significance and the effect sizes were not very large. This indicates that MN has a very reliable effect on performance measures in conversations where MN is very strongly violated (as seen in previous analyses), but the effect does not generalize as significantly to levels of agreement with violation. Regression analysis on QN taken independently revealed a stronger relationship, and this was significant for human-like behavior. RL, taken alone, correlated significantly with both measures of performance. Violations of RL have a negative effect on TT-performance. Thus, when maxims are considered independently, RL is the only maxim that significantly predicts both performance measures, QN and MN follow RL quite closely, and QL is very far from being a predictor of either measure.

When RL and MN are taken together, they become rather good predictors of performance. For human-like behavior, the regression equation is significant and effect size is rather large. The equation reveals that RL and MN have different contributions; RL has a negative impact and MN has a positive one (p=0.01, r^2=0.6). For

Table 10
Regression summary for human-like behavior and revealing machine-ness

Maxims	Significance	Effect
Human-like behavior		
RL	p=0.02	r^2=0.41
MN	n.s. (p=0.1)	n.s.
QN	p=0.02	r^2=0.42
QL	n.s. (p=0.6)	n.s.
RL, MN	p=0.01	r^2=0.6
RL, QN	p=0.0001	r^2=0.7
RL, MN, QN	p=0.002	r^2=0.8
ALL	p=0.005	r^2=0.8
Revealing machine-ness		
RL	p=0.05	r^2=0.31
MN	n.s. (p=0.09)	n.s.
QN	n.s. (p=0.07)	n.s.
QL	n.s. (p=0.6)	n.s.
RL, MN	p=0.02	r^2=0.5
RL, QN	p=0.02	r^2=0.5
RL, MN, QN	p=0.02	r^2=0.6
ALL	n.s. (p=0.06)	n.s.

revealing machine-ness, the converse holds as expected with RL correlating positively, MN correlating negatively. The regression is again significant and effect size is still large (p=0.02, r^2=0.52). RL and QN taken together yield a strong regression too. They both correlate negatively with human-like behavior (p=0.0001, r^2=0.7) and positively with revealing machine-ness (p=0.02, r^2=0.5), reaching effect sizes that are even larger than those observed for RL and MN.

When we used RL, MN, and QN in the regression, we saw that both performance measures were significantly predicted and effect size was very large. The RL, MN, and QN combination was the one that predicted both performance measures significantly with the maximum effect size.

Adding QL either led to no significant change, or lowered the predictive power of the set of maxims. When all maxims are taken together, regression is significant for human-like appearance (p=0.005, r^2=0.81). However, the regression for revealing machine-ness just falls short of significance (p=0.06, r^2=0.64). It must be noted that the contribution of QL to regression lines (the p-value for the intercept and coefficient for QL) was always non-significant.

Taken together, these results indicate negative correlations of RL and QN violations with human-like behavior and positive correlations of these maxims with revealing machine-ness. The opposite is true of MN violations. RL and MN taken together or RL and QN taken together can predict both performance measures significantly. The best set of maxims to predict TT-performance are RL, MN, and QN.

The slight differences observed between the two performance measures are indicative of the maxim violations operating differently in making different judgments.

For the human-like behavior measure, the effect sizes are larger, suggesting that the maxim violation framework is a very good predictor of judgments of human-like language use. This is interesting because it validates the hypothesis that humans make use of the conversational maxims in their assessment of behavior in this context. The other performance variable (revealing machine-ness) was not predicted as strongly by the maxim violations. This, we believe, is because there are factors other than conversational maxim violations that subjects rely on when making their decisions. The contribution of QN and QL to judgments of revealing machine-ness measure are somewhat different from their contribution to the human-like behavior measure, indicating that the operation of these maxims may be slightly different during these judgments. However, many regressions are still significant and effect sizes still very large, which indicates that the CP and the maxims still are at play. This outcome supports the rest of our data analysis.

4.3. On bias

All subjects develop a stance towards the conversations and the participants during the first questionnaire that they take, which in turn, could have an impact upon their responses in the second questionnaire. The results indicate that this bias does not influence the direction of the results (i.e. whether people tend to detect a certain maxim violation or whether people think that the computer's behavior was human-like vs. machine-like). However, the intensity of the agreements/disagreements is affected.

Recall that Group A subjects make the TT judgments after having read the conversations while completing QMAX. This, in turn, makes them more familiar with the conversations than Group B subjects at the time they are asked to make the performance decisions. On the other hand, subjects in Group B have worked on the conversations while completing QTT and therefore have focused mostly on the computers' performance prior to taking QMAX.

Subjects in Group A displayed a tendency to give more extreme performance judgments. As we said above, both groups reply in the same direction. However, when the answer is positive (i.e., when the subjects believe the computer managed to appear human-like in the given excerpt), Group A's results are always stronger. In other words, in such cases, they tend to be more tolerant of the computers. Conversely, people in Group A also are stronger in their negative opinions. When the computer is thought to be revealing its identity, it was Group A people who were more stringent.[3]

Let us first focus on Group A's behavior. These people read the conversations first without knowing that computers are involved. We are not saying that people do not detect communication problems in these conversations. However, the alternative 'This is a computer program' does not seem to come to mind as an explanation for

[3] While the tendencies hold across conversations, only six out of the fourteen conversations revealed a statistically significant difference between response patterns of Group A and B subjects when taken individually.

the problems detected. Many subjects in Group A wrote comments on QMAX, some examples of which are provided below.

- I inferred that B is mentally retarded.
- B seems to be on drugs.
- A seems to be a confused person.
- B does not make any sense! Retarded?
- Are some of these people mentally ill?
- There is no conversation here. Both have had their brains fried.
- Rather than thinking the computer's responses reveal that it is a machine, I think it gives the impression of being a seriously disturbed psychotic patient.[4]

Group A subjects have read the conversations with no bias, reflected upon the anomalies in them, and have probably come up with explanations similar to those listed in the example comments above. Thus when they find out that the reason for the abnormalities in the conversations they have worked on a few hours ago was the fact that one participant was a machine, they react more strongly.

Interestingly, there is no noticeable difference between Group B's detection of maxim violations by the computers and those of the subjects in Group A. However, they make noticeably fewer judgments on maxim violations by the *humans* in the conversations. Group B subjects already know that the experiment is about computers by the time they complete QMAX. There are only 14 conversations and it is easy to remember which one is the computer. Moreover, even if they cannot remember, we believe they would have no trouble guessing which participant is the computer given that they know one of them must be. In due course, they do not pay too much attention to the humans in the process. The results of our analysis indicated that overall, Group B subjects detected human maxim violations significantly less than Group A subjects. This result generalizes to all maxims; paired t-tests for RL, MN, QN, and QL violation agreements for the humans in all 14 conversations are all significant at p-values of 0.01, 0.01, 0.03, and 0.01, respectively.

5. Discussion

5.1. *Cooperation revisited: Practical concerns in general human-computer communication*

We propose the CP *may* need to be modified to accommodate the case of human-computer conversation. To what extent this should be done depends on whether we look at the issue from a practical or a theoretical viewpoint.

Let us first look at how Grice introduces the CP:

[4] This comment was put on QTT by a Group A subject. Even after being told that one of the participants is a computer, she feels this way.

"Our talk exchanges ... are characteristically, to some degree at least, cooperative efforts; each participant recognizes in them, to some extent, a common purpose or a set of purposes, or at least a mutually accepted direction. This purpose or direction may be fixed from the start (e.g. by an initial proposal of a question or discussion), or it may evolve during the exchange; it may be fairly definite, or may be so indefinite as to leave very considerable latitude to the participants (as in a casual conversation)." (Grice, 1975)

After this, Grice introduces the CP as a general principle which the participants (*ceteris paribus*) are expected to observe. In conversations (those that are conducted by rational beings at least), the participants usually have some common aim and try to be aware of the conversational interests of the other.

In the case that one or more of the participants is a computer, it is no longer possible to talk about cooperation in the above sense. Perhaps we can talk about the imitation of cooperation, but we cannot really say that conversation programs of today really have an understanding (let alone a *mutual* understanding) of the direction of the conversations they are carrying out.

On the other hand, although the computers in question may not possess intentionality, we believe it might still make sense to want them to follow the CP. Consider an online help system that has a natural language interface through which people can ask questions to find out information about a particular company or product. It would be rather undesirable for this program to introduce irrelevant topics, behave in an obscure or uncommunicative manner, or say things that are false. Providing the adequate amount of information is not only appropriate behavior, but it is the reason for this program's existence. In this case, we may say that the computer should be made to believe that it is an agent that needs to provide information on a certain topic and that this is its purpose in the conversations it will be engaged in. For such practical purposes, we believe the CP should not be violated by computers.

This can be thought of as a general statement which merely happens to apply to computers in certain situations. This principle, which we will refer to as the *Maximization Principle* or MP, can be formulated as follows:

MP If you are in a situation which requires you to maximize the information to be communicated, abide by the CP (and the conversational maxims).

The MP is a principle, not a rule. It is by no means definitive, i.e., there may be several other situations in which the CP should be followed. However, the MP is intuitive. In fact, it is really nothing new and is embodied by other principles in pragmatics, such as those of relevance (Sperber and Wilson, 1986; Wilson and Sperber, 1998) and rationality (Kasher, 1998). We formulate and use the MP for simplicity.

Consider a job interview, an oral exam, an academic seminar, a court testimony. All of these are situations in which information is the central focus of the conversations. It certainly would be odd if participants constantly refused to follow the CP; this would block information exchange, and clash with the interests of everyone involved. For example, it is unacceptable for an attorney to ask the witnesses irrelevant questions during a cross-examination or for a PhD student to be rude to the faculty members during her thesis defense.

Computer programs that can converse on restricted topics will probably prove to be easier to develop. However, we should always keep in mind that for best results, such programs should be made to follow the CP (because of the MP) and this is no easy feat. Useful dialogue systems have already been developed and we believe the quality of such systems will be rising rapidly in the near future. But we also believe that the initial excitement of having computer programs that can carry a conversation will eventually wear off. Then, we will be faced with having to produce better and better systems, and we will inevitably have to find ways of 'making computers cooperate'.

5.2. The TT situation

Let us recall the description of the TT. It is by no means a neutral conversational exchange. Turing explicitly describes the conversational interests of all parties involved. In the original game, we have a human judge (whose aim is to determine which of the two entities he/she is talking to is a woman), another human (whose aim is to convince the judge that he/she is a woman) and a computer (whose aim is to deceive the judge into believing that it is the woman). But the TT is usually understood to be a conversational setting in which there is a human judge (whose aim is to determine whether the entity he/she is talking to is a human being or a computer) and a computer (whose purpose is to convince the judge that it is a human).

We have argued (Saygin, 1999; Saygin et al., 2000) that Turing's original design (involving three participants and the gender issue) disguised methodological concerns, but conceded that the TT as is generally understood has become something else. Here, we will refer to the original gender-based game as TTG, and to the TT as is commonly understood, with the human-computer decision made explicit, as TTH.

Let us return to the CP. We have argued in Section 5.1 that in practical applications of human-computer communication, computers are likely to be required to follow the CP and the conversational maxims and that the extent to which they should do this depends on what aims are to be attributed to the computers in question. Variants of the TT come with predefined purposes for all parties involved. The purpose of the judge varies from scenario to scenario; in the TTH he/she is trying to determine the species of the participant(s), in the TTG he/she is set on identifying the gender of the participant(s). It must be noted that when the gender issue is involved, we are assuming (as in the TTG described by Turing) the judge has no knowledge about computers being involved. He/she is focused on determining the gender of his/her conversational partners. Therefore, the gender based scenarios are but a way of looking at the TT situations in which the judge has no 'prejudice' based on knowing that computers are participating in the game.

Let us recall the differences observed between the survey results of Group A and Group B. The results indicated that those who had read the conversations without any knowledge about the possibility of one of the participants being a computer were more reactive in their decisions on whether the computer's behavior was human-like or machine-like. These people had read the conversations for the first time while they were taking QMAX and probably most of them did not suspect that computers

were involved. Therefore, when they took QTT later, they were more appreciative when computers appeared human-like and less tolerant when they acted in ways that revealed their machine-ness.

Having read the conversations only once, without any bias, prior to being asked to make decisions regarding how human-like the computers' behavior was seems to have an effect on people's judgments. Imagine the situation if judges were not told about computers at all. TTG scenarios and TTH scenarios differ in this very important aspect. The judges in TTH scenarios will inadvertently be influenced by their prior beliefs and assumptions about computers. The effect of 'knowing vs. not knowing' may not be fully deterministic in TT situations but both intuition and the survey results indicate that the bias usually works against the computers.

Maxim violations may give rise to implicatures in conversations. Most of this process relies on the hearer's assumption that the speaker is following the CP, or at least that the speaker is a rational being. In TTH situations, this ceases to be the case. When judges are faced with anomalies in conversation, they will tend to think these are caused by their conversational partner's identity, i.e., by the fact that it is a machine. They will not bother to work out any implicatures. It must be apparent that this can cause a great difference in how the CP and the conversational maxims work in such settings. TTH judges will always take the easy way out: they can say 'this is a computer' and move on. With the judge having this choice to fall back upon when there is something that needs to be resolved, can we really say that the computers are getting a 'fair hearing'? If we wish to grant intelligence to computers by subjecting them to a TT-like test, we should at least try to give them a fair shot at it. In human-human conversations we try to resolve things in every way we can before concluding that the speaker is mentally retarded or on drugs. The same should apply to human-computer conversation.

However, this is not an issue that can be solved easily for TTH scenarios. We will eventually ask the judges to make a judgment on the human-ness vs. machine-ness of the entities they converse with.

Then, in TTH situations the best strategy for computers would be to not violate any maxims, or do so in a human manner. The latter is too difficult to view as a realistic goal in natural language chatterbot design at this moment, considering how little is formalized about the way humans violate and interpret maxims. On the other hand, computer programs that abide by the CP and never violate any of the conversational maxims at any time are liable to appear overly mechanical in TT settings. However, both the survey results and intuition dictates that they should at least be able to handle the maxim RL, and preferably QN.

And of course, we always have the original TTG. The pragmatic framework of Grice is affected to a lesser extent in these situations. The judges will not carry any bias against the computers. The implicature resolution process will work as before, with the judges trying to exploit violated maxims in order to make something out of what the participants say. A disadvantage may be that they will focus on trying to find clues that will reveal the gender of the speaker(s). This may distract the judges from other (linguistic) phenomena that can occur in the conversations. However, we do not think we should take this seriously if we do not have practical concerns like

those outlined in Section 5.1. We wish to repeat our argument *for* TTGs here. Saygin et al. (2000) have argued that gender based games were more fair and exhibited sound experimental design since the woman, whether as a concept or in reality, served as a sort of neutral point so that the impostors could be assessed in their ability to deceive. Now, we can add to this the pragmatic concerns described above. TTGs are immune to the bias that knowledge of computer participation may bring. They allow the interrogators to work out conversational maxims (and in general, exercise their naive psychology) the way they normally do. If we are interested in the TT as a philosophical concept, we should definitely consider TTG situations as viable (maybe even better) alternatives to TTH situations. At first, it may be absurd to think that being capable of deception has anything to do with human-ness. But in fact, there are many other things we take for granted in this manner that have a lot to do with being human; we hope that looking at pragmatics has revealed some of those. Even if we define the TT's aims in purely operational terms, the TTG scenario provides an alternative, since it allows the competence of computers to be assessed in a manner that is fair and unbiased.

5.3. Cooperation revisited: The TT situation

Section 5.1 mentioned how and why the CP should apply to human-computer communication systems that are practical, real-life applications. Now, we comment on the CP and the conversational maxims in TT situations.

Note that the MP need not apply in TT situations. Some judges may be focused and serious, asking specific questions and demanding to-the-point answers (although, as was mentioned before, such judges are encountered rarely), while others are rather relaxed and chatty. In general, TTs do not require the computers to strictly follow the CP and the conversational maxims. Although we will not consider the philosophical implications of this, a computer program that successfully imitates a whimsical, rude, elusive, or otherwise uncooperative human is, in theory, able to pass the TT. If a certain kind of human-like linguistic behavior is consistently present in interactions with a computer program, there really is no way of denying that it has passed the TT just because it does not follow the CP.

As we saw in the survey, sometimes maxim violations can create a human-like effect. In fact, strong violations of MN have invariably created favorable impressions. It can be inferred that, had the programs that used being rude or obscure as a 'strategy' been more successfully designed to handle the syntactic components of natural language, they would have appeared quite close to human beings, albeit strange ones. If in addition to this, the semantic processing had included ways to partially handle relevance and quantity, some of these might even have passed the 'Loebner Test'.

On the other hand, it is by no means the case that computers can violate the maxims freely and still manage to appear human-like. Recall that the results of our regression analyses between performance evaluations and maxim violations are not only significant, but also have very large effect sizes. This shows that people implicitly or explicitly make use of pragmatic information while making these judgments.

More specifically, we showed in this study that maxim violations predict people's judgments on the behavior of computers reliably and significantly. These results should not be ignored by computational linguists who are interested in developing models of linguistic behavior, as the findings indicate that the failure to handle some pragmatic aspects of language (e.g., the failure to make computers abide by the maxims of RL and QN) could be to blame for unsuccessful attempts at making machines use language in a natural and human-like manner.

Violating RL is usually indicative of poor semantic processing on the computer's part and violating QN (especially QN2) creates a rather artificial effect most of the time. QL does not seem to be as important as it is in inter-human conversations or in practical applications of human-computer conversations. The truth vs. falsity of the computers' contributions to the conversations are usually not of extreme importance in TT scenarios. As we have mentioned before, violations of QL that we encountered in human-computer conversations were generally not isolated cases; they frequently occurred along with violations of one or more of the other maxims. The cases in which QL is violated but the rest of the maxims are not (i.e., the contribution is relevant, not more or less informative than required and is delivered in an appropriate manner) should be considered ethical situations. To give an example, suppose a computer is asked, 'Where does Michael Jackson live?'. The answer 'Somewhere in California' violates QN but is not revealing of the computer's identity. According to Grice's analysis, in such cases a human would violate QN because of a *clash* between two maxims. Providing more information would violate QL which is of a higher priority. We do not believe this applies to TT situations. An answer of the sort 'Michael Jackson lives in Tulsa, Oklahoma' would be just as acceptable. Not everyone has to know where Michael Jackson lives. Maybe providing false information is not an ideal kind of behavior in our society, but we think extending this to computers and expecting them to be not only human-like, but also 'ethical' seems rather frivolous.

6. Conclusion

Despite advances in language modeling, the pragmatic analysis carried out here has revealed that programs designed to pass the TT have not yet performed satisfactorily with respect to pragmatics.

Our study is a first attempt to characterize the pragmatics of human-computer conversations. These results not only have implications for future modeling work, but also shed light on how humans judge conversations with computers. We also believe that some of these results generalize to multi-purpose human-computer conversations and even computer-mediated conversations in general. Moreover, the present work emphasizes once again that pragmatics is a crucial component of linguistic communication. Although this is a widely acknowledged fact among pragmatics researchers, other scientists and writers, notably computer scientists, sometimes tend to underestimate how hard the problem of pragmatic modeling can be.

Although no computer program has passed the TT so far, recent advances in natural language processing are by no means negligible. Since 1991, annual TT contests

have been held and prizes given to programs which display the most human-like conversational behavior. Natural language conversation systems can be found corresponding with humans on web pages, providing information on specific topics, products or companies, talking in chatrooms, and playing MUD (multi-user domain) games. As text and speech processing advances rapidly, it can be expected that we will have more and more computer applications which have natural language communication components. There is ample evidence indicating that we will soon be regarding computers as 'language users'. It will therefore be necessary to extend existing theories of conversation in order to accommodate computers as participants. Studying human behavior during conversations with computers under different conditions will also be an interesting and relevant line of research in pragmatics.

Recently, there has been a lot of action in some areas such as computer-mediated communication, discourse analysis in electronic environments, and human-computer interaction. It must be borne in mind that human-computer conversation is not only about the computer's performance. The human participants' beliefs, aims, prejudices and behavior are all contributing factors. In discussing human-computer conversation, we will need to be concerned with several issues that do not lie within the domain of syntax or semantics: the stance of humans towards computers, the beliefs of the humans about the identity of their partner and the aim of the conversation, and the settings in which the exchange takes place, to name a few. With human-computer conversation rapidly becoming reality, it is time we paid more attention to the humans conversing with these computers, not solely on parsing methods and knowledge bases. Several things need to be considered here, such as anthropomorphism, and people's dispositions, expectations, and behavior in cybernetic environments.

If, for any reason, we want computers to use language in a human-like manner, it is only natural that we are interested in how their behavior fits (or could be made to fit) existing frameworks concerning conversation. In our work, we have focused on how humans react to computers' maxim violations in human-computer conversations. The results of the experiment we carried out indicate, among other things, that Grice's CP is at work during conversations with computers. However, there are differences between how the CP applies to inter-human conversations and the findings reported in this paper (e.g., the differential contribution of MN violations and the lack of effect of QL violations). Thus, new theories and frameworks, or at least a modification of existing ones, may be needed in studying human-computer conversation as well as all computer-mediated conversation. Some of these differences may be due to humans' expectations from their conversational partners in electronic environments, while others may have their roots in their expectations from computers. While our focus has been on Grice's conversational maxims, the work may be extended to cover other pragmatic aspects of human-computer conversation.

Having analyzed how humans behave in conversations with computers, we want to briefly consider some ideas that can immediately be applied to conversation system design. Although it seems that we have a long way to go before we can successfully model human conversational behavior, we do not need to solve all of the mysteries of linguistic pragmatics before we start working on developing better conversation systems. Our study indicates that the most important maxims to avoid

violating in TT situations are relevance and quantity. If performance with respect to these two maxims can be improved, this would have a very favorable effect on conversation planning.

We do not believe that pragmatics will constitute an extra module to model separately, but rather that it should be developed in conjunction with a semantics component.[5] Some pragmatic phenomena can be incorporated into the semantic analysis component. For instance, conventional implicatures can be treated exactly the way semantic content is handled. Detection of changes of subject and keeping track of the current topic can be incorporated into the semantic processing by using a context-sensitive approach. Once the computer is given means to follow the current direction of the conversation, it will be less likely to violate the maxims of relevance and quantity in ways that are not human-like. We are not saying that all problems will be solved, but we believe results in handling at least the maxims of relevance and quantity could be obtained by a combination of existing techniques in AI and NLP.

We think the pragmatic concerns we have raised in this paper should be kept in mind in chatterbot design and that ways to handle them will start becoming more and more apparent.

Although conversation agents that can pass the TT may have limited practical use, natural language conversation systems have many applications. As computerized automation of various aspects of daily life becomes more and more commonplace, we will be having more and more conversations with computers. Although this work has discussed human-computer conversation mostly within the TT framework, we have also mentioned why more practical NLP applications should also try to model certain aspects of pragmatics.

On the other hand, most of the work in this paper concerning pragmatics and computer use of language is admittedly premature. Conversation programs of today are far from being linguistically competent. Some Loebner Prize contestants cannot even perform simple syntactic parsing and generation of grammatical responses. Most have little or no semantic processing capabilities. Pragmatics isn't even on the agenda yet. But still, we hope we have managed to convey that there is much more to language modeling than a first look might suggest. Pragmatics constitutes a serious challenge for artificial intelligence researchers. Developing a computer program that knows how to be relevant, how to provide the correct amount of information in a given context, how to make appropriate jokes, how to use appropriate metaphors, allusions, figures of speech, how to behave in a given situation and in general, how to 'cooperate' in conversation will be no simple achievement.

References

Grice, Paul H., 1967. William James Lectures, Lecture 2: Logic and conversation. Unpublished xerox.

Grice, Paul H., 1975. Logic and conversation. In: P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*. Academic Press, New York, pp. 41–58.

[5] Needless to say, this latter is no easy task in itself.

Grice, Paul H., 1998. Further notes on logic and conversation. In: Kasher, A., ed., *Pragmatics: Critical Concepts*, volume IV, pages 162–176. Routledge, London, UK.

Kasher, Asa, 1998. Conversational maxims and rationality. In: Kasher, A., ed., *Pragmatics: Critical Concepts*, volume IV, pages 181–198. Routledge, London, UK.

Keenan, Elinor, 1976. On the universality of conversational implicatures. *Language in Society*, 5: 67–80.

Leech, Geoffrey N., 1983. *Principles of Pragmatics*. Longman, London, UK.

Mey, Jacob, 1993. *Pragmatics: An Introduction*. Blackwell Publishers, Oxford, UK.

Piccinini, Gualtiero, 2000. Turing's rules for the imitation game. *Minds and Machines*, 10(4): 573–582.

Sacks, Harvey, 1992. Lectures on conversation. Blackwell Publishers. 2 volumes, Gail Jefferson (ed.).

Saygin, Ayse Pinar, 1999. Turing test and conversation. Master's thesis, Bilkent University, Ankara, Turkey.

Saygin, Ayse Pinar, Ilyas Cicekli and Varol Akman, 2000. Turing Test: 50 years later. *Minds and Machines*, 10(4): 463–518.

Sperber, Dan and Deirdre Wilson, 1986. *Relevance: Communication and Cognition*. Basil Blackwell, Oxford, UK.

Sterrett, Susan B., 2000. Turing's two tests for intelligence. *Minds and Machines*, 10(4): 541–559.

Traiger, Saul, 2000. Making the right identification in the Turing Test. *Minds and Machines*, 10(4): 561–572.

Turing, Alan M., 1950. Computing machinery and intelligence. *Mind*, 59(236): 433–460.

Wilson, Deirdre and Dan Sperber, 1998. On Grice's theory of conversation. In: Kasher, A. ed., *Pragmatics: Critical Concepts*, volume IV, pages 347–368. Routledge, London, UK.

Ayse Pinar Saygin is a graduate student in Cognitive Science at the University of California, San Diego. She holds an M.S. degree in Computer Engineering from Bilkent University and a B.S. degree in Mathematics from Middle East Technical University. She has worked on artificial intelligence, philosophy of mind and pragmatics. Her current research interests are neural mechanisms underlying normal and impaired language comprehension in humans, and linguistic pragmatics in different social environments including electronic ones.

Ilyas Cicekli is assistant professor at the Department of Computer Engineering at Bilkent University, Ankara, Turkey. He received his Ph.D. degree in Computer and Information Science from Syracuse University in 1991, his M.S. degree in Computer Science from Syracuse University in 1985, and his B.S. degree in Computer Engineering from Middle East Technical University, Turkey in 1982. His research interests include natural language processing, computational linguistics, machine translation, and logic programming. He is a member of ACL, ALP and AMTA.