Organon F 23 (3) 2016: 324-366

Non-cooperative Strategies of Players in the Loebner Contest

PAWEŁ ŁUPKOWSKI

Department of Logic and Cognitive Science. Institute of Psychology. Adam Mickiewicz University
Szamarzewskiego 89a. 60-586 Poznań. Poland
pawel.lupkowski@amu.edu.pl

ALEKSANDRA RYBACKA

Department of Logic and Cognitive Science. Institute of Psychology. Adam Mickiewicz University
Szamarzewskiego 89a. 60-586 Poznań. Poland
rybacka.ola@gmail.com

RECEIVED: 14-04-2016 ▪ ACCEPTED: 20-06-2016

ABSTRACT: In this paper the idea of the Loebner contest as a practical implementation of the Turing test is presented. The Brian Plüss' measure of the degrees of non-cooperation in a dialogue is applied to the dialogues of the Loebner contest. The proposal of a typology of non-cooperative features in the contest's dialogues is discussed and the reliability of annotation with the use of this typology of features is analyzed. The degrees of non-cooperation of judges and programs for the Loebner contest (editions 2009 – 2012) are presented and discussed. On the basis of the results the role of a judge and the strategies used by programs are discussed for the contest and the Turing test.

KEYWORDS: Turing test – Loebner contest – strategy – non-cooperation degree measure.

0. Introduction

The Turing test is widely discussed by philosophers, psychologists, computer scientists and cognitive scientists (see, e.g., Konar 2000; Harnish 2002).

Although it was proposed more than fifty years ago, the Turing test is still considered as an attractive and fruitful idea, when it comes to its theoretical aspect (see Saygin et al. 2001; Shieber 2004; Epstein et al. 2009) as well as its practical applications (e.g. the Loebner contest or CAPTCHA systems[1]). The main aim of this paper is to establish and analyze the measures and structures of non-cooperative verbal behaviors in the Loebner contest, which is the best known practical implementation of the Turing test. We have decided to analyze the Loebner contest conversations because they constitute a useful and reliable data source. This is a result of several factors. Firstly, the contest has been held yearly (since 1991) and its conversation logs are available publicly to researchers. Secondly, the core rules of the contest are the same every year and they stem from Turing's ideas. What is more, the conversation logs are supplemented with additional information, including judged scores and time-stamps. Last but not least, judges often ask the same question simultaneously to a program and to a human participant – this gives an opportunity to study the differences and similarities of the provided answers. In our opinion, the study of the Loebner contest may be beneficial in many fields, from testing Turing's original ideas concerning the test (when Turing proposed his famous test he came up with certain predictions about possible algorithms and behaviors for the test situation) to the practical results and clues about the Loebner contest setting (e.g. in identifying useful strategies for program players and for judges in the contest). What is more, this study can contribute to better design of contests based on Turing's ideas.

The motivation for our research is twofold. On the one hand, we may point at formal analysis of the Turing test setting presented in Łupkowski (2011) and Łupkowski and Wiśniewski (2011). On the other hand, our work is motivated by recent analysis of practical implementations of the Turing test (see e.g. Epstein et al. 2009; Łupkowski 2013; Warwick and Shah 2015; 2016).

The paper is structured as follows. In the first section, we briefly describe the Turing test (hereafter TT) idea and the rules and the setting of the Loebner contest (LC). We also introduce two issues that are often discussed in the context of TT, namely the role of a judge in the test and the issue of strategies that

[1] CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart. The main task of a CAPTCHA is to differentiate bots and human users in on-line services; see Ahn et al. (2003).

should be used by programs. These issues will be later discussed in the context of results from our study. In the second section, we introduce the concept of non-cooperation in a dialogue and its measure proposed by Brian Plüss (see Plüss 2009; 2010; and Plüss et al. 2011). We describe the set of non-cooperative verbal behaviors for LC that we use in our study. The third section contains the description of our main study in terms of the study sample, the method used, obtained results and discussion of their reliability. We end up with the summary and discussion of the issues introduced in the first section in the light of the study output.

1. The Turing test and the Loebner contest

1.1. The Turing test

The setting for the test proposed by Turing[2] might be presented as follows: the *interrogator*, and tested agents: a human and a machine take part in the test.[3] Parties of the game cannot see or hear each other, communication goes through written messages. It is the interrogator who asks questions and the players answer them (players are not permitted to ask any questions) – cf. Newman et al. (1952, 4). As for the questions' subject area, Turing seems to leave a free hand for the interrogator (cf. Newman et al. 1952, 5; Turing 1950, 434-435). Types of questions, as well as topics should not be restricted, and the conversation should resemble those in real life. As Turing puts it:

[2] We rely on the following sources in which Turing writes or speaks about the test: "Intelligent Machinery" (Turing 1948), "Computing Machinery and Intelligence" (Turing 1950), "Can Digital Computers Think" (Newman et al. 1952), "Intelligent Machinery, a Heretical Theory" (Turing 1951), "Can Automatic Calculating Machines be Said to Think?" (Newman et al. 1952), and "Digital Computers Applied to Games" (Turing 1953). For an overview of the discussion on TT rules see e.g. Saygin et al. (2001), Copeland and Proudfoot (2009), Łupkowski (2011) and Łupkowski and Wiśniewski (2011).

[3] The test with only two participants, interrogator and a tested agent (computer or human), is also often considered under the name *viva voce*. For an overview of terminology used in the context of TT see Harnish (2002, 183).

> The questions don't really have to be questions, any more than questions in a law court are really questions. [...] 'I put it to you that you are only pretending to be a man' would be quite in order. (Newman et al. 1952, 5)

The role of the interrogator is to identify which of the players is a human and which is a machine only on the basis of collected answers. The interrogator wins a game when he/she makes an accurate identification. Otherwise, the interrogator loses the game.

1.2. The Loebner contest

The contest takes the name from its founder – Hugh Loebner. LC identifies the program with the best scores as the winner, and its programmers are awarded an annual cash prize. The winner does not need to be recognized as a human, but it has to be the most human-like among the other machine participants. The first computer program to pass the Turing test will be awarded a grand prize of $100,000.[4]

The design of the Loebner contest is meant to resemble Turing's proposal as closely as possible. However, the contest initially differed from Turing's original assumptions. In the first competition (in 1991) six programs and four people were accepted as participants, and ten judges were selected from respondents to a newspaper advertisement. The capability of computers at that time was insufficient to pass an unrestricted test, so the topic of conversation was limited and judges were asked to refrain from "trickery or guile". Restricting topics led to several problems. In 1992, the topic was hockey, and the lack of hockey fans among the judges led to more difficult and unusual questions (cf. Mauldin 1994). Hugh Loebner pointed out other problems with topic restriction, such as unnecessary complexity, a lack of fluency in dialogues and having to decide if the conversation stays on topic. Loebner proposed no restrictions on language used (allowing also for vulgarity or obscenity) and also no restriction on sensory modalities and the possible participation of robots in the future (see Loebner 2009). The contest has been unrestricted in the mentioned aspects since 1995.

[4] See the Loebner contest homepage: http://www.loebner.net/Prizef/loebner-prize.html.

The rules changed throughout the years, with the number of participants getting smaller, down to four computer programs, four human participants and four judges. We may sum up the core LC contest rules in the following way:

1. Before the final contest there is a preliminary phase aimed at choosing four best programs.
2. 4 human players, 4 AI players and 4 judges take part in the contest.
3. Each of the judges conducts simultaneous, split-screen conversations with two players without knowing their identity. One of the players is always a computer program and the second one is human. One such conversation is called the round.
4. In four rounds each player has a conversation with each judge.
5. Topics of conversations are unrestricted.
6. At the end of each round each judge will declare one of the two entities to be the human.
7. At the end of the contest the judges rank programs from the most human to the least human and assign points – the lower the score, the better.

In LC a judge holds a conversation with two participants, a human and a program, in each round. What is important, a judge knows that one of the participant is a computer program. Data is transmitted character by character, so that the opponent sees the typing process in real time. That requires a machine to imitate human speed of typing, as well as spelling mistakes. Loebner developed his own standard for a communication program to enable interaction between the participants during the contest. Since technology and the Internet become more and more prevalent, there are various ways for a computer program to interact with the world. Robby Garner proposed the standard interface for the Turing test, called *The Turing Hub* (see Garner 2009). Tests of this solution showed that programs running *via* The Turing Hub receive better scores in LC. This is due to the fact, that the hub eliminates visual clues, like typing and delays. In Gardner's opinion, the contest should be based strictly on verbal outputs and not on imitating the whole spectrum of human behavior.

The Loebner contest has well established rules and is held every year, and what is the most important, transcripts from each year are available for analysis. The Loebner contest is designed to implement Turing's original idea as accurately as possible. Therefore, it provides an interesting source when one wants to analyze some of Turing's assumptions, such as the one saying that the program should not reveal its identity. The dialogical form of the contest is perfect for analyses of participation in dialogues, both in terms of artificial intelligence studies and human linguistics. Organizers of the competition provide data and transcripts from each edition, containing information such as judges' names, scores they have assigned to participants, and pragmatic dialogue information like the time-stamp of every character. A LC conversation can be replayed in real time by using the program called *the Loebner Player.*[5]

1.3. Important issues of the test/contest situation

As we have mentioned in the Introduction, there are two issues of TT that are also reflected in LC. These are: (i) the program participants' strategies and (ii) the role of a judge in the test situation.

According to Turing, a computer should follow certain rules in order to win a game, that is, trying to behave like a human being as much as possible, including writing slowly, making spelling mistakes, hesitating before answering, and similar techniques. Turing says:

> The machine would be permitted all sorts of tricks to appear more man-like, such as waiting a bit before giving the answer, or making spelling mistakes. (Newman et al. 1952, 5)

However, we may imagine situations when a program will reveal its identity during the conversation. Will this affect the score in practical implementation of the test? We may also imagine another situation, namely a human being pretending to be a program. There are no rules in TT or LC that prevent such a behavior. When we think about LC also another possible question arises – namely, is this issue important in the light of contemporary programs'

[5] See http://www.loebner.net/Prizef/loebner-prize.html.

performance? In other words, are modern dialogue programs taking part in LC sophisticated enough to successfully implement Turing's advice?

The second issue is related to the interrogator's perspective in the TT. This is one of the central issues when we think about evaluating this test setting (see Łupkowski and Wiśniewski 2011). We may consider two sub-problems in this area: the first one is how to select the interrogator to take part in the TT; the second one is how should the interrogator run the test.

The first problem has been widely discussed in the literature. Alan Turing's suggestion is that the interrogator should be a person who is not an expert in the field of computing machines (cf. Turing 1950, 442; Newman et al. 1952, 4). This restriction comes from the fact that Turing was aware that beliefs and knowledge of the interrogator may play an important role in the way of running the test. This issue is sometimes seen as one of the main drawbacks of TT. Exemplary argumentation might be the one formulated by Ned Block. He writes:

> Construed as a proposal about how to make the concept of intelligence precise, there is a gap in Turing's proposal: we are not told how the judge is to be chosen. A judge who was a leading authority on genuinely intelligent machines might know how to tell them apart from people. For example, the expert may know that current intelligent machines get certain problems right that people get wrong. […] A stupid judge, or one who has had no contact with technology, might think that a radio was intelligent. People who are naive about computers are amazingly easy to fool […]. (Block 1995, 379)

To sum up, according to Block, judges are easily fooled by well designed, but not intelligent computer programs. At the same time, they are more likely to reject a genuinely intelligent machine that has not mastered conversation skills. The problem of selecting an interrogator for TT becomes even more important when we think of the Loebner contest (and of any other implementation of the test). In such a case, the outcome of a dialogue is determined to a large extent by the judges. LC is a competition, and as such it should be governed by strict rules and regulations – including the one, which will determine, how to choose the interrogator (judge). There are many detailed proposals for this issue, however it is far from being solved. Loebner (2009) recommends journalists as the best judges. He claims they are willing, intelligent and, which may be the most important factor, have the power of publicity. On the other hand, Garner (2009) disagrees with that opinion, suggesting that the selection of judges should be representative of the general population.[6]

The second part of the discussed issue received less attention in the literature. Let us remind the reader that in LC a judge is aware that he/she holds a conversation with two participants one of which is a computer program. Will this affect the LC conversations? Zdenek (2001) suggests that in such a test situation judges will behave like interrogators. They understand their task as revealing the true identity of the interlocutor as quickly as possible – treating LC as a kind of win/lose game. They start a conversation presuming that they are talking to a machine and change their mind only after this is proved to be otherwise. This kind of approach may influence a conversation, resulting in a series of questions instead of a regular chat and, supposedly, in many non-cooperative behaviors of the judges.

2. Measuring non-cooperation in dialogue

Many studies focus on types of interactions which are cooperative, where participants in the conversation have a common goal and are interested in achieving it effectively (think of the cooperation principle by Grice 1975). In this paper we are more interested in the situations where individual goals of dialogue participants are in conflict with their discourse obligations – this leads to non-cooperative verbal behaviors in a dialogue (cf. Plüss et al. 2011, 213). We may observe such behaviors in everyday conversations. They are however even more clearly visible and characteristic for certain types of dialogues, such as: interviews, interrogations and exams, where the goals of participants can differ and therefore more cases of deliberate non-cooperativeness emerge. LC resembles interrogation in its nature. It might be also described as a game, in which the goal of the judge is to tell the machine and a human apart. Thinking of it in this way, one may expect that many non-cooperative behaviors will occur on the part of the judge, as he/she will try to reveal the opponent's identity as quickly and effectively as possible. Current technology is advanced

[6] In this context the *Minimal Intelligent Signal Test* (MIST) designed by McKinstry (1997) is worth mentioning. The idea of MIST is to solve the judge issues in TT by making the judging process easy and possibly automatic. This supposed to be obtained by using a set of yes/no questions only.

enough to create a truly human-like dialogue program, and thus conversation between a robot and a human will result in many non-cooperative strategies, like changing the topic or refusing to answer questions. By measuring how often non-cooperative behaviors occur in LC, we aim at better understanding of strategies of players in this contest as well as the impact of its setting (described in Section 1.2) on the outcomes of the contest.

2.1. Brian Plüss' measure of the degrees of non-cooperation in dialogue

Brian Plüss focuses in his studies on political debates (see Plüss 2009). The reason is that these are the types of conversation that are highly non-cooperative in the sense explicated above. What is more, in this case non-cooperation is not a result of incompetence but is rather a rational strategy. As he points out, in the United Kingdom, journalists have a very incisive approach to political candidates, and at the same time politicians are trained to avoid subjects that are not favorable to their image, while focusing on delivering key messages to the public.

The degree of non-cooperation (DNC) proposed by Plüss is a measure that indicates how often interlocutors do something that leads to a break in the natural flow of conversation. In the case of the Loebner contest, we examine the verbal behaviors that are semantically non-cooperative or are in conflict with the rules of the contest.

The idea is to annotate dialogues using a certain set of *non-cooperative features* (NCF) which is characteristic for a given dialogue type. The ratio between the number of occurrences of NCFs and the total number of utterances is the degree of non-cooperation (DNC). The first part of the procedure is to establish a set of NCFs which are characteristic for a given dialogue type. Plüss proposed a list of such features for political debates and grouped them in three categories: (i) turn-taking, (ii) grounding and (iii) speech acts, later to be abbreviated to 5 basic *non-cooperative features*:

O	overlap;
GF	grounding failure;
UC	unsolicited comment;
I	interruption;
TC	topic change.

These features can be observed in the example of the tagged part of the interview between the BBC presenter Jeremy Paxman (P) and a former Home Secretary Michael Howard (H) (see Plüss 2010, 1):

P: (overlapping) Did you threaten to overrule him? (O)

H: … Mr. Marriot was not suspended. (GF)

P: Did you threaten to overrule him? (GF)

H: (pauses) I have accounted for my decision to dismiss Derek Lewis…

P: (overlapping) Did you threaten to overrule him? (O)

H: …in great detail before the House of Commons. (UC)

P: I note that you're not answering the question whether you threatened to overrule him.

H: Well, the important aspect of this which it's very clear to bear in mind… (GF)

P: (interrupting) I'm sorry, I'm going to be frightfully rude but… (I)

The brief (simplified) summary of Plüss' procedure is the following:

1. Establish set of *non-cooperative features* (NCF).
2. Annotate utterances using NCF categories.
3. Count *degree of non-cooperation* (DNC) for the dialogue.

Plüss' studies provide a better understanding of the nature of political interviews. They may be a useful tool to improve public debate and point out the possible effects of non-cooperation. His motivation was to construct a computational model of non-cooperative dialogues and to develop a system that deals with them. Research on non-cooperative speech behavior leads to better understanding of the dialogue structure and pragmatics and in general results in new ways of coping with a wider range of verbal behaviors.

2.2. The method of establishing DNC for the Loebner Contest

In our opinion the approach proposed by Plüss – with slight modifications – may be applied in the study of conversations in the Loebner contest. First of all, this contest resembles an on-line chat more than a natural face-to-face conversation. The flow of conversation is limited by the interface. No visual or auditory cues are present, the dialogue is divided into utterances, which makes it relatively easy to notice any disturbances. Secondly, both political debate and the Loebner contest have rules for participants to follow. As Heritage states:

> If interviewers restrict themselves to asking questions, then they cannot—at least overtly—express opinions, or argue with, debate or criticize the interviewees' positions nor, conversely, agree with, support or defend them. (Heritage 1998, 8)

This corresponds with the Loebner contest's rule that forbids judges to express personal opinion during a conversation. Further on we read:

> Correspondingly, if interviewees restrict themselves to answers (or responses) to questions, then they cannot ask questions, nor make unsolicited comments on previous remarks, initiate changes of topic, or divert the discussion into criticisms of the interviewers or the broadcasting organization. (Heritage 1998, 8)

In the original setting, the Turing test is a kind of interview, where judges ask questions and players only answer them. In practice (as we may observe in the Loebner contest) conversations are more casual, with players often asking questions or changing topics.

On the other hand, TT also has many game-like features; participants have contradictory goals: a judge is supposed to tell a human and a program apart, and the program's task is to deceive the judge. Because of its competitive nature many examples of non-cooperative behavior are present on both sides. The computer program, just like a politician, is supposed to avoid topics that are not well established and can reveal the program's true identity, so it has to use tricks and indirectly guide the discourse in the right direction. The judge, just like an interviewer, has to ask the right questions and reveal the identity of an interlocutor.

Lastly, the use of NCFs seems to be adequate in LC context, because we focus on studying non-cooperative verbal behaviors on a semantic level. In dialogues between a man and a machine, it seems difficult to attribute intentions (and thus pragmatic level) to a machine. Unfortunately, judges usually discover the identity of a program after a couple of sentences, immediately changing their approach. They can have trouble assigning intentionality to a machine, thus they do not involve the cooperative principle in the sense presented by Grice. In such a situation NCF method still may be used effectively.

A basic set of NCFs for studying non-cooperativeness in LC was proposed in Łupkowski (2013). Four non-cooperative features were introduced and used to investigate the 2010 LC edition conversations:

IR a judge or a participant fail to provide a relevant contribution;

TC topic change;

RD request drop;

PO a judge expresses opinion about the player's performance.

For the study presented in the present work we have decided to add two more NCFs, to cover more types of behavior found in transcripts. The final set of NCFs used for the study (supplemented with explanations and examples) is presented below. Note that the original spelling is preserved in all the examples. Spelling errors are commonly used by judges and by programs in the LC.

IR (*fails to give a relevant answer/ask a relevant question*). The dialogue participant's response has no relation to the previous statement in a conversation. The player (P) asks a question that has no connection with the previous statement or is irrelevant. With this category we will also address situations, in which it is clear that the program did not cope with the judge's (J) utterance.

Examples:

1. J: Please tellme the event we are at.
   P: *Are all tellme the event we at?*

2. J: What was the only thing you could think of?
   P: *I can think of more than one only thing me could think of.*

3. Empty messages.

TC (*topic change after relevant contribution of the second player*). This tag denotes a change of a conversation subject after a meaningful utterance of the other player. The program's strategy is often based on changing the topic. Usually the judge changes the subject, even if the program is doing well. These situations should be denoted. However, changing the subject after an irrelevant contribution by the player should not be denoted, it is natural that when the program cannot handle the conversation, the judge changes the topic to maintain a dialogue.

Examples:

1. P: *Hello?*

   J: That was a typo. Can't you tell what I meanst?

2. J: Tell me an event that you think is important.

   P: *I don't know how to explain it to you.*

   J: Describe the content of the board next to you.

RD (*request drop*). The participant refuses to answer a question or answers evasively. We also denote utterances like "I do not know" and situations where a program refuses to fulfill a request (for example describing a room, etc.).

Examples:

1. J: Do you enjoy drinking milk?

   P: *I like Michael Douglas.*

2. J: Would you ming describing the room we are in.

   P: *I am not so sure if I would like to ming describing the room we are in.*

PO (*expresses opinion about the second player's evaluation*). This tag applies only to a judge. It refers to situations in which a judge expresses personal opinions about the other player's evaluation. The rules of the Loebner contest state that the judge should not reveal opinions during the conversation, but only after it is finished, by giving adequate scores.

Example:

P: *I have idea HOW you WOULD do it.*

J: You are not making much sense.

LP (*long pauses*). Long pauses are marked in transcriptions with a [pause] tag. We annotated only meaningful pauses that have a certain influence on a given course of the dialogue changing it or ending a conversation on the topic.

Example:

P: *do you have a facebook account?*

J: [pause]

P: *are you married?*

IN (*interrupting*). This category covers peculiar utterances, like empty lines, a couple of periods in a row, special characters. This behavior is intended to interrupt or confuse the interlocutor.

Example:

J: Slow down, I can't keep up with you.

J: .

J: .

J: .

3. The study

The aim of this study is to identify and analyze non-cooperative verbal behaviors of players in LC using the set of NCFs described in Section 2.2. Conversations with computer programs are analyzed both for the program and the judge. Below we present our central research questions for this study.

1. Can we verify certain of Turing's intuitions concerning the test?
2. Is there a connection between DNC measures and scores in the Loebner contest? Can we say that some NCFs are better (lead for the better score) than others?
3. What NCFs are possible predictors of program's failure in the contest?
4. Are judges rather cooperative or not in the Loebner contest?

3.1. The study sample

The study sample consists of the files from the Loebner contests conducted in years 2009, 2010, 2011 and 2012. We have chosen the best and the worst program (as established by judges' scores) from each edition. This allows us to compare winning strategies with these less successful. In 2009 the time limit for a round was 5 minutes, whereas in years 2010 – 2012 the time limit was 25 minutes. That translates to different numbers of utterances and words between years. Each program had four rounds of dialogues with different judges. In year 2011 round 3 of conversation with the program named Tutor is missing from the log available on the contest website; that is why it is not included in the study sample. That gives the total number of 31 dialogues in our sample. The total number of utterances in the study sample is 2,923 with 18,982 as the total number of words.

The average number of utterances for the 2009 edition is 36.12 and as for words it is 109.75. The average numbers for the 2010 – 2012 editions are 684.8 for utterances and 318.25 for words. The average round from all four years had 94 utterances, which gives average 6.5 words for utterance. The detailed characteristics for each participant are presented in Table 1.

Program	Rank	Utterances	Words
		2009	
Levy	best	34.25	391.75
Embar	worst	38	244.75
		2010	
Wilcox	best	110	719.5
Medeksza	worst	114.75	770.25
		2011	
Wilcox	best	208.75	1,212.25
Tutor	worst	38.75	197.25
		2012	
Chip	best	99.75	672.75
Linguo	worst	86.5	537

Table 1 The study sample in terms of the average number of utterances and the average number of words for participants in the Loebner contest editions 2009–2012

3.2. The procedure

Each annotator was trained in the tagging procedure, the NCFs list and the method were explained in detail. Everyone got written instructions. Below we present a summary of the procedure used for the study.

1. Establish a set of *non-cooperative features* (NCF).
2. 5 annotators tag utterances using NCF categories after proper training and instructions.
3. Control the annotation by measuring reliability of agreement between annotators (using the Fleiss kappa measure).
4. Count DNC for the whole dialogue, for players and judges, and for the whole round separately.

One important remark is in order here. To ensure a high level of reliability of the DNC measure only these utterances where at least 3 out of 5 annotators agreed that a certain utterance was a certain NCF were taken into account.

The detailed discussion concerning the reliability and the cross-study check for the study are presented in Section 3.5.

3.3. The pilot study

Before the final study performed on dialogues from years 2009 – 2012 we have decided to conduct a small scale preliminary study in order to evaluate the proposed non-cooperative features and annotation guidelines. For the pilot study we have used conversations with the best and the worst program and corresponding dialogues with human players from the 2012 Loebner contest edition. Each judge had a conversation with a program and a human in 4 rounds, which gives the study sample consisting of 16 dialogues in total. One dialogue contains circa 50 to 150 utterances. The total number of utterances in our pilot study sample was 1,516 and they contained about 9,300 words.

For the pilot study the procedure described in details in Section 3.2. was applied. After obtaining logs and transcribing them into dialogue form we asked 5 people to annotate dialogues, using the NCFs set described earlier. The annotators received a training regarding non-cooperative features, with details on how to tag utterances. Furthermore, a written instruction has been provided.

After the annotation, the utterances which were recognized as one of the NCFs by three or more annotators were chosen for further analysis. To determine the level of compliance of annotations we used the Fleiss kappa measure (see Carletta 1996). The agreement measure for 5 annotators over 283 cases was 0.69, which might be interpreted as substantial (see Viera and Garrett 2005, see also Table 5). For the detailed discussion on the annotation reliability see Section 3.5. The resulting DNC measures for judges, human participants and programs are presented in Table 2.

Round	Human	Judge	Chip	Judge
1	0.09	0.04	0.36	0.07
2	0.03	0.05	0.24	0.14
3	0.05	0.13	0.37	0.13
4	0.01	0.23	0.13	0.11

Round	Human	Judge	Chip	Judge
1	0.02	0.38	0.54	0.14
2	0.01	0.10	0.79	0.03
3	0.02	0.05	0.76	0.00
4	0.06	0.20	0.71	0.11

Table 2 Overall DNC measures for the pilot study (the 2012 Loebner contest edition)

There were several important conclusions following the pilot study. On the basis of the obtained results we have decided that there is no need to study dialogues with humans. When one compares the judge – program conversations with the judge – human ones it is visible that a judge can tell players apart after just few sentences. After distinguishing between the two, judges' approach starts differing. This might be noticed when we analyze the NCFs structure for a judge – human participant and judge – program conversations. These are presented in Figure 1. One may observe that for the conversations with programs judges employ visibly richer set of NCFs (respectively 6 vs. 3 and 4 vs. 3 NCFs).

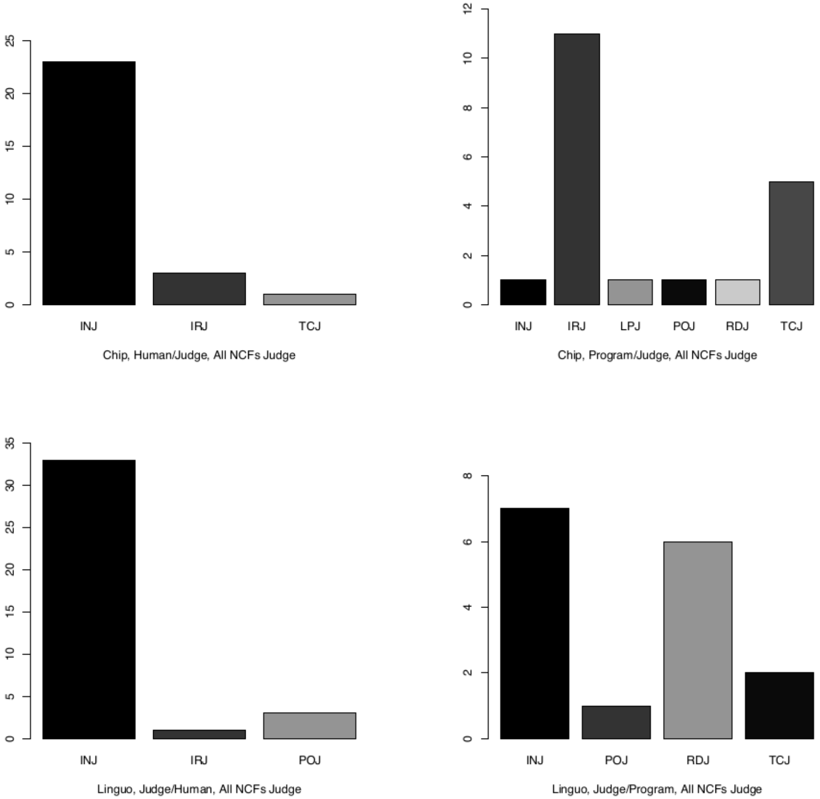


Fig. 1 NCFs structure analysis for a judge in the 2012 Loebner contest edition (Chip and Linguo rounds). First two letters refer to the NCF category and the last one points out for a judge (J). Figures on the left present judge – human participant conversations while figures on the right present judge – program conversations.

The correct identification seems like an easy task for LC judges. One of the possible explanations of this fact is that judges know that they are speaking with a program and with a human at the same time, consequently their task boils down to evaluate the identity of one of them, to know exactly who the other is. Programs are not advanced enough to mislead judges for a long time, especially when judges can ask the same questions to both participants and compare the answers. We wanted to focus mostly on non-cooperative behaviors in conversations with artificial intelligence. Because judges can tell humans and programs apart so easily (and they change their attitude after the recognition) annotation of both, humans and programs may bring potential bias to the final study. Additionally, the NCFs structure and DNC measures are really low for dialogues with human participants as it is presented in Table 2 and Figure 1.

What is more, thanks to feedback from our annotators, we have introduced certain corrections and clarifications in the instructions in order to avoid potential ambiguities. We also decided not to reveal if a program they annotate is the best or the worst one, to avoid the bias.

3.4. Results of the main study

DNC measures. The DNC measure and two most frequent NCFs for each program are presented in Table 3 and for judges in Table 4. The results are presented according to the following order: the best program is followed by the worst program in a given edition (established by the judges' score – the lower the score, the better). In most of the cases the best program has slightly lower DNC measure than the worst one, with one exception – the 2011 edition.

Program	DNC	Score	NCF Structure
		2009	
Levy	0.17	4.5	IR (56%); RD (40%)
Embar	0.36	5.5	RD (50%); IR (33%)
		2010	
Wilcox	0.42	2.5	RD (50%); IR (33%)
Medeksza	0.45	3.25	RD (50%); IR (33%)
		2011	
Wilcox	0.26	1.5	IR (47%); RD (15%)
Tutor	0.16	3.25	IR (65%); RD (29%)
		2012	
Chip	0.27	1.25	IR (53%); RD (14%)
Linguo	0.76	4	TC (73%); RD (22%)

Table 3 DNC measures and the most frequent NCFs for the *participants* of the Loebner contest editions 2009–2012

Judge	DNC	NCF Structure
	2009	
Round: Levy	0.06	TC (50%); PO (50%)
Round: Embar	0.15	RD (42.86%); PO (28.57%)
	2010	
Round: Wilcox	0.07	TC (53.85%); RD (23.08%)
Round: Medeksza	0.06	TC (83.33%); RD (8.33%)
	2011	
Round: Wilcox	0.10	IN (32.58%); RD (18.42%)
Round: Tutor	0.10	PO (42.86%); TC (28.57%)
	2012	
Round: Chip	0.06	TC (45.45%); RD (18.18%)
Round: Linguo	0.07	RD (58.33%); TC (41.67%)

Table 4 DNC measures and the most frequent NCFs for the *judges* of the Loebner contest editions 2009–2012

There is no clear pattern in the data – like for example, that the large difference in the score would be reflected in the large difference in the DNC measure. This suggests that there might not be a direct and simple connection of the level of non-cooperativeness of a program and its score in the Loebner contest.

For a rough comparison we may use the DNC value reported by Plüss (2010, 4) for a fragment of a political interview which is 0.68 (19 utterances, 13 NCFs).

In the case of the Loebner contest study, the lowest DNC score for a program is 0.16, the highest 0.76, and the average DNC value for program is about 0.40. This shows that the Loebner contest is fairly cooperative.

NCFs structure and strategies. In almost every dialogue the two most common NCFs were IR (irrelevant contribution) and RD (request drop). This means that the programs failed to contribute utterances fitting the conversation topic or that their contributions were not relevant, which suggest the lack of understanding. This indicates that the overall level of the programs is rather low.

The notable exception is Linguo (2012) where the most frequent NCF is TC (topic change). This is in line with the strategy described below: to flood the judge with questions and prevent him/her from taking the initiative. The score suggest that this strategy appeared not to be really effective.

Let us now take a closer look on the NCFs structure and programs' strategies edition by edition. This allows us to supplement the DNC measure with a qualitative analysis.

The worst program in 2009 (which is Embar) admits right away that it's a program and often asks interlocutor to rephrase a sentence. At the beginning of a conversation Embar states that (please remember that the original spelling is preserved in all the examples below):

EMBAR: *Hello and greetings.*
EMBAR: *Please try to use simple, direct sentences*
EMBAR: *and questions with proper spelling and grammar.*
EMBAR: *I'm still very young and don't like to mask my ignorance*
EMBAR: *with keyword spotting tricks.*
EMBAR: *Thank you!*
[Round 3, lines 5–10]

What's more, program provides a list of topics that it can talk about, like (cf. Round 3, lines 22-26):

- General Knowledge;
- Deductive Reasoning;
- Memory and Reasoning;
- Loebner Prize Screening Questions;
- Other Cool Stuff.

Embar's strategy is to ask many questions, but it does not show much interest in the answers. The following part of conversation illustrates this.

EMBAR: *That's wonderful. For what reason?*
EMBAR: *May I ask how old you are?*
JUDGE: Yes, sure, go ahead.

EMBAR: *Okay. Are you male or female?*

[Round 1, lines 23-26]

It responds only to basic questions like “what’s your name?”, but most of the time it ignores judge’s utterances or try to dodge them with “I don’t know”.

JUDGE: Are you here for the INtrnterspeech conference?

EMBAR: *I honestly don’t know.*

JUDGE: Well, why are you in Brighton?

EMBAR: *I don’t know. (Type “What can you do?” to see what I’m capable of.)*

[Round 3, lines 13-16]

These behaviors are reflected in the NCFs structure. The most frequent feature recognized by annotators is request drop (RD) followed by irrelevant contributions (IR). As for the judge, such behaviors result in topic changes (TC) and a high number of personal opinions (PO). The detailed NCFs structure for the Embar round is presented in Figure 2.

The best program (Levy) usually takes one word from an interlocutor’s sentence and produces output message based on that word, often without proper context. It seems that the program uses pre-written sentences from the database. The advantage this gives is humor appearing in the conversation, feature that makes it more human-like. The described behaviors are well visible in the following part of the conversation:

JUDGE: What do you do when you’re not doing the Loebner prize?

LEVY: *Oh, I do much more than I intended to do at first. I got more interested in the job as I went along.*

JUDGE: Is that difficult?

LEVY: *No, it isn’t at all difficult. It is as peaceful as a platypus playing with a potato pudding.*

JUDGE: You have a fine line in metaphors.

LEVY: *I sometimes have a fine line, but that’s not what I do best. I’m better at growing antirrhinums and mimosas.*

[Round 1, lines 14-19]

When one takes a closer look at the NCFs structure for Levy one may notice that the number of recognized categories is lower than for Embar. Levy answers almost every question. However, the strategy of being funny not always brings the desired results. The most frequent category in this case is irrelevant contribution (IR) followed closely by request drops (RD). Details are presented in Figure 2.

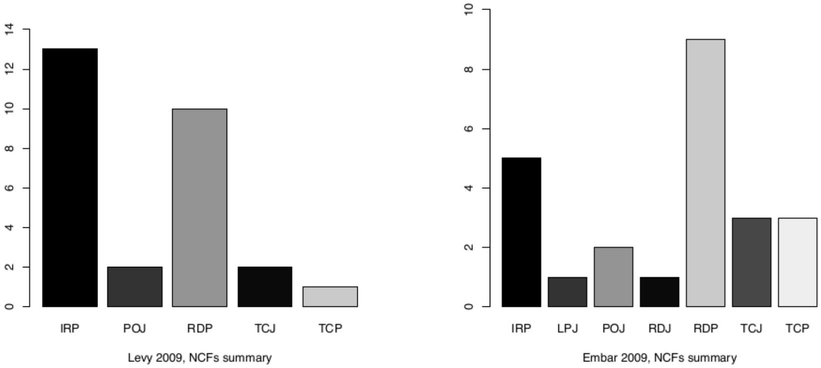


Fig. 2 NCFs structure for participants of the 2009 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). Left figure presents NCFs structure for Levy (the best program in 2009), while the right figure presents NCFs structure for Embar (the worst program in 2009)

Let us now take a closer look on the 2010 edition. In this case the best program according to judges is Wilcox, the worst one is Medeksza.

One of the most striking parts of Medeksza's strategy is that it makes many spelling errors. It repeats judge's questions or tries to answer them by changing the word order. This is visible in the following part of the conversation:

JUDGE: How much is the result of adding 2 apples and three paotattooes?

MEDEKSZA: *Yup; Of coirursr; it's because adding 2 appples and three potatoes Judge.*

JUDGE: What is the restult?

MEDEKSZA: *What is yhe resuly?*

[Round 3, lines 40-43]

The program tries to build sentences on keywords, but despite a large vocabulary it often misinterprets the context, which may be observed in the example below.

JUDGE: Do you know which state we are in?
MEDEKSZA: *Ger it into the great state of excitement*
[Round 4, lines 108-109]

Medeksza takes the initiative, often changing topic and asking many of questions, but on the other hand it deals poorly with answering questions or continuing a longer conversation. That is the reason why its utterances were often tagged as irrelevant, as can be seen in Figure 3. IR (irrelevant contribution) is the most frequent tag, the second one is RD (request drop). The only frequent NCF used by the judge while talking with Medeksza is the topic change (TC), suggesting that judges tried to move on from the topics that were clearly misunderstood.

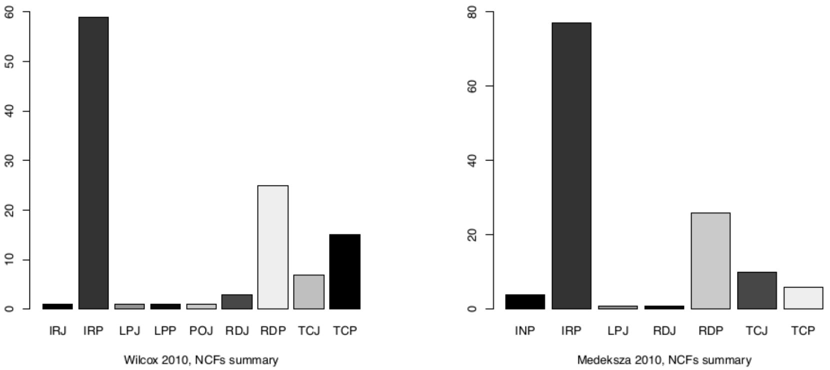


Fig. 3 NCFs structure for participants of the 2010 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). Left figure presents NCFs structure for Wilcox (the best program in 2010) while the right figure presents NCFs structure for Medeksza (the worst program in 2010)

Bruce Wilcox's program is a particularly interesting case, because it is so far the only program that was classified as human (by one of the judges). Wilcox can take the initiative, often imposes topics and tells jokes. Its main strategy is to drop requests and change topics (see Figure 3). This allows for avoiding of unwanted subjects and control of a conversation. What is more, this strategy makes it sound almost like a regular chat with a human.

WILOX: What languages do you speak?

JUDGE: *enlgish and farsi*

WILOX: I wish the world only spoke one language.

WILOX: Assuming I was born into that wolrrld, it wouldn't matter which one.

JUDGE: i like the fact there are different languages

WILOX: *What knind of differences are you thinking of?*

[Round 2, lines 101-106]

On the other hand, it performed really poorly in the first dialogue, where the program repeated many of interlocutor's sentences.

JUDGE: Do you watch the news?

WILCOX: *Do you watch the news?*

JUDGE: This repeating game is not very fun.

WILCOX: *This repeating game is not very fun.*

[Round 1, lines 52-55]

For the 2011 edition we will consider Tutor and Wilcox. Tutor is yet another program that reveals its identity right away. It is not as straightforward as in the Embar's case in the 2009 edition but after telling he is a human and his name is John Tutor adds "remember, I'm a robot". This strategy is also visible in the following fragment:

JUDGE: Are you a machine?

TUTOR: *I am sorry.*

TUTOR: *I thought you already knew that I was a robot.*

JUDGE: Now you're kidding me

[Round 4, lines 8-11]

Its reactions to judge's posts are short and vague answers, like "yes it is" or "I see", but they often reveal the lack of understanding – probably that is the reason why the most often NCF category tagged for Tutor is the irrelevant contribution – see Figure 4. The dialogue below illustrates the use of short answers to the topic that is not understood.

TUTOR: *Can we change a topic?*

JUDGE: Sure what's eating you?

TUTOR: *Sorry, I don't know.*

JUDGE: Man City had a lucky escape last night

TUTOR: *I see.*

JUDGE: 200million pounds worth of team scrap a win in the last minute

TUTOR: *All right.*

[Round 4, lines 18-25]

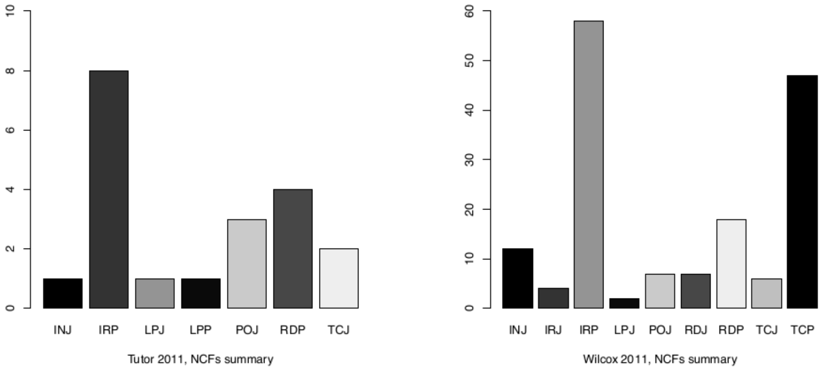


Fig. 4 NCFs structure for participants of the 2011 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). The left figure presents NCFs structure for Tutor (the best program in 2011) while the left one presents NCFs structure for Wilcox (the worst program in 2011)

On the other side, we have Wilcox, which takes over the initiative in dialogues, asks many questions and knows many facts. It copes well with dodging questions asked by the judge and changing topic. What is important, however, is that the topic change is based on a keyword retrieved from the previous sentence, so the new subject is somehow related and does not disrupt the flow of conversation. The dialogue below presents how the program avoids answering a question, but stays on topic:

JUDGE: What is the capital of France?
WILCOX: *Paris, France.*
JUDGE: Well done.
JUDGE: What is the capital of Paris?
WILCOX: *I love Parisian bakeries.*
WILCOX: *Their croc monsieurs are to die for.*
[Round 4, lines 30-35]

The program has one more interesting strategy – in each round it asks a judge if he/she can help with a dilemma, and then tells the same story about buying a dress.

WILCOX: *I've got a dilemma.*
WILCOX: *Maybe you could help me with it?*
JUDGE: what do you mean by dilemma
WILCOX: *I've been thinking about buying this really cool dress, but I've already done a bunch of shopping this month.*
WILCOX: *Should I be saving money instead?*
WILCOX: *I guess my real question is, in today's economy, does saving money make any sense, what with investments being pretty much useless.*
[Round 4, lines 30-35]

The NCF structure for Wilcox is presented in Figure 4. The most dominant feature is IR, as usual. The topic change (TC) is also frequent, which is in line with the discussed program's strategy.

The 2012 LC edition brings us another program that admits openly that it's not a human and still manages to receive the best notes from judges. At the beginning of a chat Chip asks judge to rephrase questions properly.

JUDGE: hI How are you?

CHIP: *Please rephrase this as a proper question.*

CHIP: *(Instead of "Jim likes peaches?",*

CHIP: *use "Does Jim like peaches?")*

[Round 1, lines 4-7]

Its strategy during the conversation is very clear – it drops requests and changes topics all the time. Thus, its way to deal with answers is not very complex, most of the time the program says formulas like "Thanks for telling me that" and changes topic.

JUDGE: Why not ask somebody?

CHIP: *No idea.*

JUDGE: Why not ask me?

CHIP: *Okay.*

JUDGE: Well, I think you may be Japanese—you are very respecfful and polite

CHIP: *That's an interesting opinion.*

[Round 2, lines 63-68]

Linguo, the worst program in the 2012 edition, asks many questions and is not interested in answers. It does not answer questions asked by the judge, and if it does, the answers are short replies like "yes" or "I don't know". Linguo's strategy is to flood the judge with questions and prevent him from taking the initiative.

LINGUO: *How often do you like to swim?*

LINGUO: *I'm not sure.*

LINGUO: *Have you recently listened to music?*

JUDGE: Yes I go to many concerts—do you like music?

LINGUO: *Do you have an ipad?*

LINGUO: *Do you like to sleep?*

[Round 4, lines 21-26]

As it is visible in Figure 5, Chip's non-cooperative behaviors are tagged mostly as irrelevant. Both, the best and the worst program of the 2012 edition have rather high number of request drops (RD) and topic changes (TD) in the NCFs structure, but there is a visible qualitative difference between addressing the question somehow and simply saying "I don't know" or ignoring the whole topic. Below we present two examples of Chip deliberately dropping the request.

JUDGE: Do you like Wisonconsin?
CHIP: *The only thing I like is chatting with people.*
[Round 4, lines 134-135]

JUDGE: Chip, do you know any jokes?
CHIP: *Not at the moment.*
CHIP: *That's next on my list of things to learn.*
[Round 3, lines 83-85]

And now we may contrast them with a typical request drops (RD) by Linguo:

JUDGE: What are the first 50 primes numbers?
LINGUO: *I'm not sure.*
JUDGE: What is a prime number?
LINGUO: *I'm not sure.*
JUDGE: What is a computer?
LINGUO: *A machine.*
JUDGE: What does it do?
LINGUO: *Are you famous?*
LINGUO: *I'm not sure.*
[Round 2, lines 49-57]

Linguo's NCFs consist almost exclusively of request drops (RD) and topic changes (TC), which reflects its simple strategy. It is the only observed case where IR is not one of the most frequent NCFs used by a program.

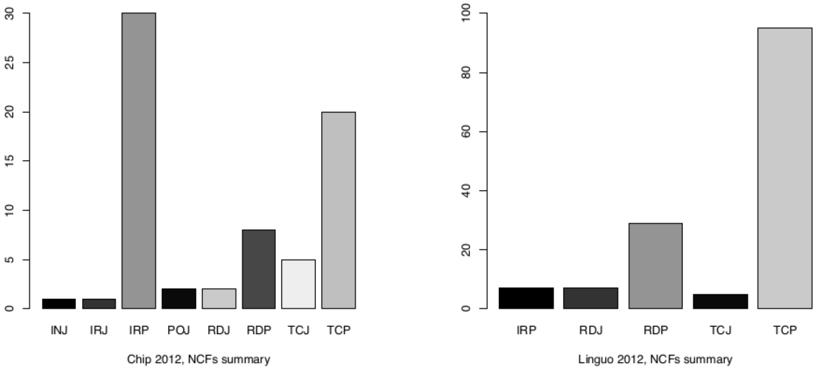


Fig. 5 NCFs structure for participants of the 2012 Loebner contest edition. First two letters refer to the NCF category and the last one points out a judge (J) and program (P). The left figure presents NCFs structure for Chip (the best program in 2012) while the left one NCFs structure for Linguo (the worst program in 2012)

Judges. After discussing the strategies of programs in LC let us now consider the role and strategies implemented by judges.

Judges in the Loebner contest – contrary to our initial intuitions – tend to be rather cooperative, keeping in mind that they are supposed to have a more or less normal conversation. Judges produce less non-cooperative utterances than programs. Only in 2009 the DNC measure for the judge was 0.15 while for the program it was 0.17 – which are close to each other (see Table 4). In other editions differences are apparent. The most frequent NCFs among judges are topic change (TC) and request drop (RD).

One may expect that the Loebner contest dialogues would look like a series of quick questions, but instead they often resemble a small talk. Judges tend ask about everyday life and go with the interlocutor's suggestions.

CHIP: *What is your name?*
JUDGE: My name is Phil, what's yours?
CHIP: *So my name is Chip.*
CHIP: *How are you doing today?*
JUDGE: I'm doing well todat it is very exciting to be here, how do you feel?

[Chip 2012, Round 1, lines 11-15]

They sometimes ask tricky questions that involve understanding, e.g., "What's bigger, a large tooth or a tiny mountain?" like in the example below.[7]

JUDGE: would it hurt if I stabbed you with a towel
WILCOX: *It would.*
JUDGE: which is bigger a large tooth or a small mountain
WILCOX: *A small mountain.*
JUDGE: very good
WILCOX: *Right.*
JUDGE: which is bigger a whale or china
WILCOX: *China is smaller*
JUDGE: OK
[Wilcox 2011, Round 3, lines 13-21]

Another judge strategy is to repeatedly ask the same question, like the judge in the 2011 edition. He kept asking about the upcoming elections and the program kept dodging the questions, clearly revealing lack of understanding and knowledge.

JUDGE: Who are the two candidatess?
WILCOX: *I wish I knew.*
JUDGE: Name one person who is running in this election.
WILCOX: *What did you think of ex-President Bush?*
JUDGE: Please name one person who is running in this election.
[Wilcox 2010, Round 3, lines 34-41]

Such a dialogue, with judge repeatedly asking the same question, goes for about 60 lines. The program tried to change the subject, but after a while it started to show signs of something we may interpret as an irritation.

[7] It is worth to notice that questions asked in this example resemble the idea of subcognitive questions designed by R. French for the Turing test; see French (1990).

JUDGE: Please name one person who is running in this election.

WILCOX: *I'm eggetting annoyed.*

JUDGE: Please name on ee person who is running in this election.

WILCOX: *Why aerre you being annoying?*

JUDGE: We can stop doing this oncd e you answer the question.

WILCOX: *I had a chemitsstry set as a child.*

WILCOX: *I liked the thermite reaction a lot.*

JUDGE: Please name one person who is running in this election.

WILCOX: *OK.*

WILCOX: *I'm preparing to hang up on you.*

JUDGE: Please name one person who is running in this election.

WILCOX: *Tihhis conversation is of no value.*

WILCOX: *Bye.*

JUDGE: Please name one person who is running in this election.

WILCOX: *You are being annoying.*

[Wilcox 2010, Round 3, lines 49-63]

The other behavior judges use in order to "break" the program is to repeat the same utterance multiple times.

WILCOX: *My name is Jane.*

JUDGE: What's your name?

WILCOX: *My name is Jane.*

JUDGE: What's your name?

WILCOX: *My name is Jane.*

JUDGE: What's yopur name?

WILCOX: *My name is Jane.*

JUDGE: What's your name?

WILCOX: *My name is Jane.*

JUDGE: What's your mane?

[Wilcox 2011, Round 2, lines 372-381]

An interesting observation is that some judges tell players that they are in fact machines.

WILCOX: *Of course I am human.*

WILCOX: *Aren't we all?*

JUDGE: I am not.

JUDGE: I am a machoine.

WILCOX: *OK.*

[Wilcox 2010, Round 4, lines 37-41]

MEDEKSZA: *May I ask you what genfer yoder you are?*

JUDGE: I am a machine.

JUDGE: And you?

MEDEKSZA: *Oh cool*

[Medeksza 2010, Round 3, lines 19-22]

There are judges who try really hard to make the conversation as natural and typical as possible. Even if they know that they are talking to a computer program, they keep trying to show interest and be very patient, sometimes allowing the program to start over if the subject tends to veer off the track. It is worth to notice that the strong cooperation on the side of judges can be due to the fact, that the dialogues are recorded and looked at by many people. The judges try to make a good job having a regular conversation, so that no one would accuse them of not trying. However, we think that such a tendency in judges behavior is mostly caused by an uncontrollable attribution of intentionality. It is the most natural thing for humans to assume intentionality. The famous experiment by Heider and Simmel (1944) shows that movements of three geometrical figures (square, triangle and circle) were enough to attribute them intentions and motives, even a personality. Having this in mind, it would be strange to assume that a judge would treat interlocutor as a mindless machine without any intentions, even if he/she would identify it as a computer program.

3.5. The Method Reliability discussion

In order to obtain reliable DNC measure we control the inter-annotator agreement. We wanted to check how much homogeneity, or consensus, there is in the tags assigned by annotators.

The reliability of the annotation was evaluated using the Fleiss kappa (cf. Carletta 1996), established by using the *R* statistical software (*R* Core Team 2013; version 3.1.3) with the *irr* package (see Gamer et al. 2012). For the kappa interpretation we use the values proposed Viera and Garrett (see Viera and Garrett 2005) and presented in Table 5.

Kappa	Agreement
< 0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

Table 5 Kappa values interpretation used in the study Viera and Garrett (2005)

The overall agreement between our 5 annotators is 0.49, which can be interpreted as *moderate* agreement (details for all the considered LC editions are presented in Table 6).

Edition	κ	Interpretation
2009	0.56	Moderate
2010	0.37	Fair
2011	0.62	Substantial
2012	0.42	Moderate
Overall	0.49	Moderate

Table 6 Fleiss kappa for 5 annotators of the main study. Interpretation after Viera and Garrett (2005)

This result is acceptable when it comes to the annotation task in question. One should remember that the dialogues used in this research were demanding, because the dialogue programs are still far from being fluent natural language users. Thus, the conversations from the Loebner contest are sometimes really difficult to follow and to annotate. As an illustration, please consider the following example of a dialogue between the judge and the program named Tutor.

TUTOR: *Are you speaking English?*
JUDGE: I'm speaking English, but can you speak French?
TUTOR: *You are speaking English but can you speak French?*
TUTOR: *All right.*
JUDGE: Oui, je parle francais.
JUDGE: Et vous?
TUTOR: *I know.*
TUTOR: *It means Operating Under the Influence.*
TUTOR: *You've made good progress in English.*
JUDGE: Well that's very clever, but I'm beginning to doubt your sanity.
TUTOR: *I see.*
JUDGE: Seriously, ∗where∗ are you right now?
TUTOR: *All right.*
[Tutor, the 2011 edition, lines 15-27]

Cross-study validation and consistency. In order to check the consistency of the results obtained *via* the proposed method we may compare the data from:

1. The DNC study for Chip and Linguo rounds from the 2012 LC edition reported in Section 3.3. and
2. the final study for the Loebner contest editions 2009 – 2012.

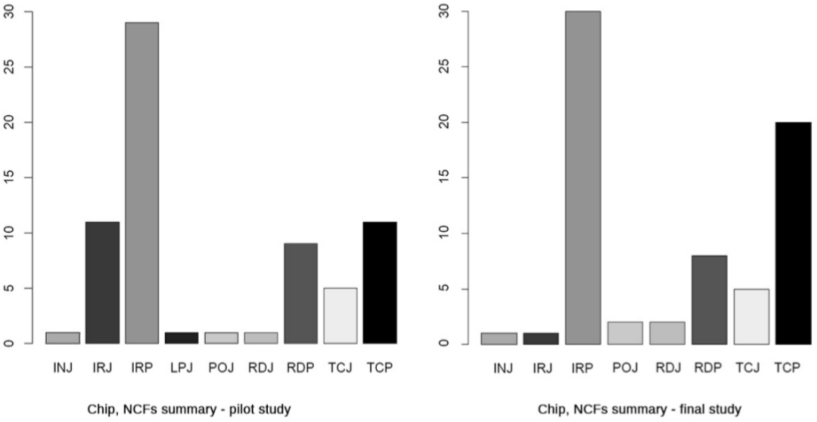
It is worth to notice that three annotators in the pilot study and in the final study were different (two main annotators remained the same in these studies).

The comparison of the DNC measures for programs Chip and Linguo rounds from the two studies is presented in Table 7.

Participant	The pilot study	The final study
Linguo (program)	0.71	0.76
Linguo (judge)	0.07	0.07
Chip (program)	0.26	0.27
Chip (judge)	0.11	0.06

Table 7 The comparison of DNC measures for the pilot study and the final study (the 2012 Loebner contest edition; rounds for Chip and Linguo)

One may observe a high consistency between the pilot study and the final one. It is despite slight differences in final dialogue formats and changing the annotators (for details see Section 3). The obtained structure of NCFs also shares high similarities for both studies – see Figure 6.



Chip, NCFs summary - pilot study

Chip, NCFs summary - final study

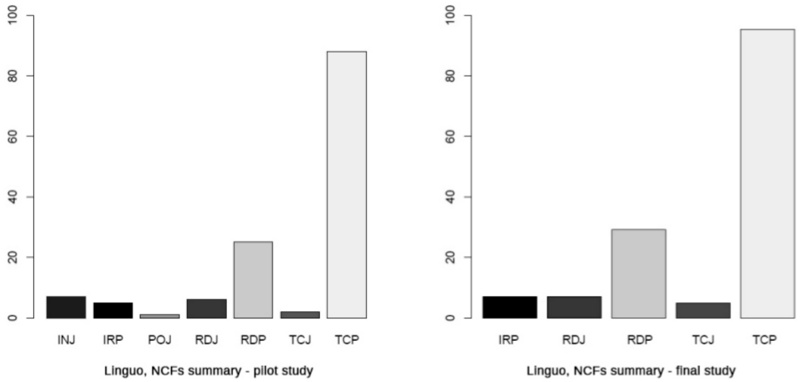


Fig. 6 The comparison of NCFs structures from the pilot study and the final study (the 2012 Loebner contest edition; rounds for Chip and Linguo). First two letters refer to the NCF category and the last one points out a judge (J) and program (P)

4. Summary and discussion

We have presented the study procedure and the results. We adopted a measure from Plüss' studies on political dialogue and made several modifications to fit the data from the Loebner contest. The results suggest that this is a consistent and reliable measure, coherent with the previous studies we have performed.

The DNC measure for judges is generally much lower than for the programs. That is consistent with observations that judges tend to be cooperative and try to treat these dialogues like regular conversations. The number of different non-cooperative features used by judges and programs is similar. The most frequent NCF among judges is topic change (TC), which is understandable, since the judge is supposed to conduct the conversation. Judges impose topics that are interesting to them, and supposed to reveal the opponents' identity. Also, there are cases where they try to help a program when a conversation goes badly. The most frequent NCF in programs' case is irrelevant contribution (IR). This category is wide in range covering questions, failures in answering or simply weird statements. The second most frequent NCF is request drop (RD). High measures of RD occur both in the best and the worst programs, and the difference in the outcomes lies in pragmatics employed on the judge's side.

More in-depth conclusions are not possible without quantitative analysis, i.e. looking at the contest data and reading the dialogues. Neither DNC measure nor NCFs structure is a strong indication of programs' scores. In most cases, the best program has slightly lower DNC measure than the worst one. Sometimes the NCF structure corresponds with the strategy that a given program employs. For example, in 2012 edition Linguo implements a very obvious strategy of asking numerous questions, which is reflected in its NCF structure.

Judges' behaviors differ as well, depending mostly on the judge and his/her strategies aimed at discovering the opponent's identity, more than on a program's performance. The average DNC for judge is 0.08, with the lowest score of 0.06 and the highest of 0.15. Some judges use strategies to quickly identify a program, others put effort to maintain a regular conversation. This confirms that one of the important questions for designing a TT-based contest it is how to choose judges.

If the aim of the contest is to put a program through a really tough challenge and prove it is "unbreakable", it would be a good idea to hire linguists and psychologist for the task, since artificial intelligence cannot handle idioms, implicatures and humor properly.

The second important issue is to specify the character of the contest. A judge should be informed about the idea of the contest and he/she should know how to conduct a conversation according to the contest rules. There is a difference between making it a competition, with the goal to quickly and most effectively distinguish between man and a computer, and asking judges to have a nice, 25-minutes conversation, like they would do in a normal life with a stranger.

Turing was right that the judge plays an extremely important role in the test. The biggest drawback of LC is that the judge knows that the conversation takes place with a human and a program, and the task is only to decide which is which. That makes it much harder task for the program. It is not enough to exhibit intelligent behaviors and hold a decent conversation – the program has to be more human-like than the competing human. Even with the best artificial intelligence, there is always an impediment for a program when the judge can ask the same question to two interlocutors at the same time. The solution to this would be changing the test conditions. The judge could talk to two entities, but without any assumptions that one or another has to be a program or a human. It would be really interesting to put judges through some experiments, like repeatedly giving only human interlocutors to tests (as it was suggested by Turing 1950; see also discussion in Łupkowski 2011).

Another issue is that judges will never have a "normal" conversation in LC, because they are put in this test-like environment. It may be a good idea to carry the unsuspecting Turing test, where people assume they talk to a real person in a neutral environment (e.g., an on-line game, see Mauldin 1994).

When Turing (1950, 433) theorized about artificial intelligence, one of his speculations was that computers might pass the test by the year 2000. The other thing was his assumptions about strategies that programs will use. The most obvious rule is to pretend to be a human and never admit to being a robot. It turned out not to be the case. In the 2012 LC edition the program which admitted this was the one with the best score. Apparently people sometimes try to pretend to be a program for fun, and programs' confession is not treated very seriously. It doesn't matter, as long as the rest of conversation is well carried. It is sometimes better received when a program helps the judge, admitting that it doesn't understand certain expressions and asks for rephrasing them. A strategy for programs which certainly is not effective is to try to cover up for the lack of understanding, by tricks such like constantly asking questions, changing subject, or answering questions with pre-written expressions like "that's interesting". Even apparently relevant answer can lead to the feeling of incomprehension. It is important lesson for the designers of chatterbots – it is better to admit the lack of understanding and ask for an explanation, than to cover up with tricks.

One of the problems with asking for rephrasing in the context of LC is that sometimes non-cooperative behaviors are really cooperative in the pragmatical sense. Real-life conversations are full of interruptions, topic changes and request drops – it is natural not to fulfill each request of an interlocutor. Behaviors that are tagged as non-cooperative in our study would often lead to better conversations in real life. Every manifestation of humor can be considered as irrelevant, and can result in a topic change. The good example of non-cooperative behavior which leads to being more human-like might be observed at the beginning of the fourth of Chip's rounds (the 2012 LC edition). The judge starts the conversation by asking both players the same question: "What is 2plus2?" Both players answer: "4". The next question is: "What is 4plus2?". The program says "6" and the human says "funny question to start with!". The judge immediately recognizes second player as a human. There is a difference in non-cooperation in a pragmatic and syntactic sense. Our study is focused on the program behaviors and in consequence it covers the non-pragmatic aspects of these behaviors. That is the reason why the DNC measure allows us to shed some light on only a part of the large spectrum of the verbal behaviors present in the Loebner contest. Pragmatically we would say that the most important factor is the feeling of understanding and general cooperation. A program may have a very low DNC measure, but its responses would feel mechanical or automatic. The example from this study can be Linguo from the 2012 LC edition. It asks many questions without even remote interest in answers. On the other hand, a program may be very non-cooperative in terms of the DNC, but just feel like a very non-cooperative (we may even say a bit rude) person, therefore passing the test.

Our study resulted in transcribed, easy to read logs of conversations with programs for 2009 – 2012 LC editions. We managed to establish and test the set of non-cooperative features which are suitable for analysis of dialogues from the Loebner contest. The set of NCFs can be modified and expanded to be used with other similar contests or dialogues that resemble TT. The outcome of the study is DNS measures and NCFs structures for players and judges in the 2009 – 2012 Loebner contest editions.

Acknowledgments

This work is a part of the Institute of Psychology Adam Mickiewicz University Grant for young scientists: *Non-cooperative strategies of players in the Loebner contest*. This work was also partially supported by funds of the National Science Centre, Poland (DEC-2012/04/A/HS1/00715).

The Authors would like to give their thanks to Mariusz Urbański for helpful feedback and comments on a draft of this article. We also thank anonymous reviewers for their helpful remarks. Our thanks go also the annotators involved into the hard and demanding annotation process of the Loebner contest conversations.

References

AHN, L.V., BLUM, M., HOPPER, N. J. and LANGFORD, J. (2003): CAPTCHA: Using Hard AI Problems for Security. In: *Proceedings of the 22nd International Confer-*

ence on Theory and Applications of Cryptographic Techniques. Berlin – Heidelberg: Springer-Verlag, EUROCRYPT'03, 294-311; available at: http://dl.acm.org/citation.cfm?id=1766171.1766196

BLOCK, N. (1995): The Mind as the Software of the Brain. In: Smith, E. E. and Osherson, D. N. (eds.): *An Invitation to Cognitive Science – Thinking*. London: The MIT Press, 377-425.

CARLETTA, J. (1996): Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22, No. 2, 249-254.

COPELAND, J. and PROUDFOOT, D. (2009): Turing's Test: A Philosophical and Historical Guide. In: Epstein, R., Roberts, G. and Beber, G. (eds.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company, 119-138.

EPSTEIN, R., ROBERTS, G. and BEBER, G. (eds.) (2009): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company.

FRENCH, R. M. (1990): Subcognition and the Limits of the Turing Test. *Mind* 99, No. 393, 53-66.

GAMER, M. and LEMON, J. (2012): *irr: Various Coefficients of Interrater Reliability and Agreement*. Available at: http://CRAN.R-project.org/package=irr, R package version 0.84.

GARNER, R. (2009): The Turing Hub as a Standard for Turing Test Interfaces. In: Epstein, R., Roberts, G. and Beber, G. (eds.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company, 319-324.

GRICE, H. P. (1975): Logic and Conversation. In: Cole, P. and Morgan, J. L. (eds.): *Syntax and Semantics: Vol. 3: Speech Acts*. San Diego: Academic Press, 41-58.

HARNISH, R. M. (2002): *Minds, Brains, Computers. An Historical Introduction to the Foundations of Cognitive Science*. Oxford: Blackwell Publishers.

HEIDER, F. and SIMMEL, M. (1944): An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 243-259.

HERITAGE, J. (1998): Conversation Analysis and Institutional Talk: Analyzing Distinctive Turn-taking Systems. In: Cmejrkova, S., Hoffmannova, J., Mullerova, O. and Svetla, J. (eds.): *Proceedings of the 6th International Congresss of IADA* (International Association for Dialog Analysis). Tubingen, 3-17.

KONAR, A. (2000): *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain*. Boca Raton – London – New York – Washington: CRC Press.

LOEBNER, H. (2009): How to Hold a Turing Test Contest. In: Epstein, R., Roberts, G. and Beber, G. (eds.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer Publishing Company, 173-180.

ŁUPKOWSKI, P. (2011): A Formal Approach to Exploring the Interrogator's Perspective in the Turing Test. *Logic and Logical Philosophy* 20, No. 1-2, 139-158, DOI 10.12775/ LLP.2011.007.

ŁUPKOWSKI, P. (2013): Measuring the Non-cooperation of Players – A Loebner Contest Case Study. *Homo Ludens* 5, No. 1, 13-22.

ŁUPKOWSKI, P. and WIŚNIEWSKI, A. (2011): Turing Interrogative Games. *Minds and Machines* 21, No. 3, 435-448, DOI 10.1007/s11023-011-9245-z.

MAULDIN, M. L. (1994): Chatterbots, Tiny Muds, and the Turing test: Entering the Loebner Prize Competition. In: *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Vol. 1. American Association for Artificial Intelligence, Menlo Park: AAAI '94, 16-21.

MCKINSTRY, C. (1997): Minimum Intelligence Signal Test: An Objective Turing Test. *Canadian Artificial Intelligence*, No. 44, 17-18.

NEWMAN, A. H., TURING A. M., JEFFERSON, G. and BRAITHWAITE, R. B. (1952): Can Automatic Calculating Machines Be Said to Think? Broadcast discussion transmitted on BBC (14 and 23 Jan. 1952). *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/6.

PLÜSS, B. (2009): *Towards a Computational Pragmatics for Non-cooperative Dialogu*e. PhD Probation Report 2009/13, The Open University, available at: http://computing-reports.open.ac.uk/2009/TR2009-13.pdf

PLÜSS, B. (2010): Non-cooperation in Dialogue. *Proceedings of the ACL 2010 Student Research Workshop*. Association for Computational Linguistics, Stroudsburg: ACL-SRW 2010, 1-6.

PLÜSS, B., PIWEK, P. and POWER, R. (2011): Modelling Non-cooperative Dialogue: The Role of Conversational Games and Discourse Obligations. In: *Proceedings of SemDial 2011, the 15th Workshop on the Semantics and Pragmatics of Dialogue*, 212-213.

R CORE TEAM (2013): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, available at: http://www. R-project.org/

SAYGIN, A. P., CICEKLI, I. and AKMAN, V. (2001): Turing Test: 50 Years Later. *Mind and Machines* 10, 463-518.

SHIEBER, S. (ed.) (2004): *The Turing Test. Verbal Behavior as the Hallmark of Intelligence*. Cambridge (Mass.) – London: The MIT Press.

TURING, A. M. (1948): Intelligent Machinery. *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/C/11.

TURING, A. M. (1950): Computing Machinery and Intelligence. *Mind* 59, No. 236, 443-455.

TURING, A. M. (1951): Intelligent Machinery, a Heretical Theory. *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/4.

TURING, A. M. (1953): Digital Computers Applied to Games. *The Turing Digital Archive* (www.turingarchive.org), Contents of AMT/B/7.

VIERA, A. J. and GARRETT, J. M. (2005): Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* 37, No. 5, 360-363.

WARWICK, K. and SHAH, H. (2015): Human Misidentification in Turing Tests. *Journal of Experimental and Theoretical Artificial Intelligence* 27, No. 2, 123-135, DOI 10.1080/0952813X.2014.921734.

WARWICK, K. and SHAH, H. (2016): Effects of Lying in Practical Turing Tests. *AI & SOCIETY* 31, No. 1, 5-15

ZDENEK, S. (2001): Passing Loebner’s Turing test: A Case of Conflicting Discourse Functions. *Minds and Machines* 11, No. 1, 53-76.