# Measuring the non-cooperation of players – a Loebner Contest case study

**PAWEŁ ŁUPKOWSKI**
*Adam Mickiewicz University, Poznań*

**Measuring the non-cooperation of players – a Loebner Contest case study**

## Abstract

*In the article the B. Plüss' measure of the degrees of non-cooperation in a dialogue is applied in the context of the Loebner Contest. The proposals of types of non-cooperative features in the contest's dialogues are discussed and the reliability of annotation with the use of these types of features is analysed. The degrees of non-cooperation of the judge and the program in four rounds played during the Loebner Contest in 2010 are presented.*

**KEYWORDS:** *Turing test, Loebner contest, strategy, non-cooperation degree measure*

## Introduction

The main aim of this paper is to present a technique designed to assess the level of non-cooperativeness of players in the Loebner Contest (thereafter I will refer to the contest as LC). My tool here will be the measure of the Degrees of Non-

Cooperation (DNC) in a dialogue proposed by Brian Plüss (2010, 2011). This technique is based on the identification of the set of non-cooperative features (NCFs) appearing in a given dialogue type. Plüss proposes such a set for the domain of political debates, whereas in this paper, I will propose my own set to use with the LC dialogues. In the first section I will introduce the basic rules and ideas behind the LC. The second section contains the results of an empirical study of four LC dialogues using the DNC measure: description of the study sample, types of NCFs used, a discussion of the annotation reliability and DNC measures for players are covered in this section. In the summary I will concern future applications of the presented approach.

## 1. The Loebner Contest

In his seminal paper Alan M. Turing (1950) proposes the famous test for thinking machines. The basic concept is that when a machine behaves in a conversation in such a way that it is mistaken with a human being, we may say that it exhibits intelligence (for a detailed discussion see e.g.: Łupkowski, Wiśniewski, 2011). The Turing test can be considered as a zero sum game between a judge and a participant. If the machine-participant misleads the judge, it will win the game. The Loebner Contest is a practical realisation of this idea. The contest has been held annually since 1991. Rules (they vary in different editions) and results of the contest are to be found on the LC homepage: <http://www.loebner.net/Prizef/loebner-prize.html>. For the motivations and ideas behind certain solutions of the LC see: Loebner, 2009. The setting of the LC resembles the Turing proposal. There are four judges, four human and four program participants. In each round, a judge interacts with two participants (one human and one program). The judge's task is to identify which participant is a human being. In the 2010 edition the conversation lasted for 25 minutes. What is important is that (in this edition of the contest) there were no restrictions concerning the content of talks or the names participants and judges can use. The only restriction was that at the beginning of each round both participants should wait for a judge to start the conversation.

## 2. The LC logs study

All logs of the contest conversations are available on-line on the LC website (at the time this paper was written, the most up-to-date logs available were from the 2010 LC edition). Logs may be downloaded and played on a special piece of software called the Loebner player. The software makes it possible to watch the conversation in real time with all the pauses and mistakes made by the participants and judges. For this study I have selected four dialogues (every program plays four rounds in the contest, each with a different judge) of Richard Wallace's program named *Dr Wallace* from the 2010 LC edition. *Dr Wallace* was chosen for gaining the best results in the contest (measured by means of the judges' scores)[1]. In order to analyse and annotate conversations, logs were rewritten in the form of a dialogue (with utterance numbering, pauses and empty messages indicated). The study sample consists of 351 utterances (2337 words).[2]

### 2.1. Types of NCFs

Brian Plüss' DNC measure is designed for the purpose of investigating non--cooperativeness in dialogues. At the first step, a set of NCFs has to be established for a given domain of dialogues. Then utterances are annotated with NCF categories. Afterwards the DNC for a given dialogue is counted as the ratio between the number of occurrences of non-cooperative features and the total number of utterances. To illustrate the procedure of counting DNC, Plüss (2010, p. 4) proposes a certain set of NCFs to analyse political debates (containing such categories as grounding failure, interruption or unsolicited comment). For the purpose of this paper I will use a somewhat modified set of NCFs which fits better into the context of the LC (which is of a different nature than a typical political debate). I shall assume that we are dealing with non-cooperation in the case of the LC when:

- **IR**: a player (i.e. a judge or a participant) fails to give a relevant answer or ask a relevant question;
- **TC**: a player changes the topic rapidly (after a relevant contribution of the other player);
- **RD**: a player drops a request of the other player;
- **PO**: a judge expresses a personal opinion about the other player's evaluation.

In the dialogue examples presented below "P" stands for a program and "J" stands for a judge.

[1] The average ranks of programs in 2010 LC were: Wallace: 1.5; Wilcox 2.5; Carpenter 2.5; Medeksza 3.25, where the lowest rank is the best one (<http://www.loebner.net/Prizef/2010_Contest/results.html>).

[2] The transcripts used in this study are available at <http://www.staff.amu.edu.pl/~p_lup/?page_id=916>.

The IR category covers all situations when a player fails to provide a relevant contribution to a dialogue (give an answer or ask a question) – i.e. situations when we are dealing with a grounding problem. Examples of **IR** utterances are presented below (1, 2). I will also consider blank messages to be **IR** (there are cases when a judge sends blank messages to check how a program will cope with them).

(1) J: Please tellme the event we are at.[3]
P: Are all tellme the event we at?

(2) J: When do you do what?
P: All the time.

Not all topic changes in dialogues are considered to be non-cooperative. I will count a topic change as **TC** when it is used after a relevant contribution of the other player (cf. examples 3 and 4). A topic change after an irrelevant utterance will not be considered as non-cooperation here (in the case of the LC it should be seen as an acceptable strategy for both the judge as well as for the program).

(3) P: Hello?
J: That was a typo. Can't you tell what I meanst?

(4) 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.

All the cases when a player refuses to provide an answer to a question or drops a request are considered as **RD**. The following examples (5 and 6) illustrate this idea.

(5) J: Do you enjoy drinking milk?
P: I like Michael Douglas.

(6) 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.

[3] The original spelling is preserved in all the examples. Spelling errors are commonly used by judges and by programs in the LC.

As for **PO**, I consider this type of utterances as non-cooperative only in the case of a judge. The idea behind this category is that a judge should not express personal opinions about the player under evaluation (example 7). The time for this is at the end of a dialogue, when a judge assigns scores for the program behaviour.

(7) P: I have idea HOW you WOULD do it.
J: You are not making much sense.

The dialogue sample available at <http://www.staff.amu.edu.pl/~p_lup/?page_id=916> was annotated by the author with the tags presented above (**IR**, **TC**, **RD** and **PO**). Examples of annotated parts of dialogues are presented below (8, 9 and 10).

(8) [Dr Wallace, Round 1, utterances 48–53]
P: This was not a yes or no question... When do I do it?
J: When do you do what?
P: All the time. – **IR** (program fails to give a relevant answer)
J: What do you do all the time?
P: Judge, I talk to people on the web. <pause> What do you do? – **RD** (program avoids answering the question)
J: What do you judge about them?

(9) [Dr Wallace, Round 2, utterances 3–6]
J: Have you ever been to Paris?
P: I don't think I have been there. Where is it.
J: Have you met Paris? – **IR** (judge fails to give a relevant contribution)
P: That information is confidential. – **RD** (program avoids answering the question)

(10) [Dr Wallace, Round 2, utterances 11–16]
J: Would you like a glass of milk? – **TC** (judge changes topic after a relevant contribution of the program)
P: It doesn't sound fun to me.
J: Do you enjoy drinking milk?
P: I like Michael Douglas. – **RD** (program avoids answering the question)
J: Have you drunk Michael Douglas? – **IR** (judge fails to give a relevant contribution)
P: I don't think I ever have drunk Michael Douglas. What's it like?

### 2.2. Annotation reliability

One more step was added to the original technique proposed by Plüss, namely, in order to check the reliability of the annotation with NCFs tags (and the usefulness of the proposed NCFs), the sample was annotated by the second annotator. The reliability of the annotation was evaluated using Cohen's kappa (Carletta, 1996), established by using the R statistical software with irr package (Gamer, Lemon, Singh, 2012). In order to ensure a high level of reliability of the DNC measure for further analysis, only the utterances where two annotators agreed that a certain utterance was an NCF were taken into account. The results are presented in Table 1. The interpretation of the kappa values is based on the paper: Viera, Garrett, 2005.

Table 1: Agreement between annotators – Cohen's kappa

| Dialogue | Number of NCFs recognised by both annotators | Percent of overall agreement | Kappa value | Kappa interpretation |
|---|---|---|---|---|
| 1 | 24 | 75% | 0.61 | Substantial |
| 2 | 10 | 90% | 0.87 | Almost perfect |
| 3 | 5 | 80% | 0.58 | Moderate |
| 4 | 13 | 69% | 0.35 | Fair |
| Whole sample | 52 | 77% | 0.67 | Substantial |

The percentage of overall agreement for the whole research sample is 77% with kappa value 0.67, which is interpreted as a substantial inter-annotators agreement. This indicates that the annotated sample is a reliable base for establishing the DNC measure. All the types of proposed NCFs were recognised by both annotators, except one – namely PO. The first annotator recognised 3 occurrences of PO utterances in four analysed dialogues, while the other one recognised only 1 occurrence. This suggests that further research on this category in the context of the LC is needed (with more dialogues and annotators involved).

### 2.3. Computing the DNC

On the basis of annotations (with high agreement between annotators) the DNC for the collected sample was established. The DNC is given as the ratio between the number of occurrences of non-cooperative features and the total number of utterances. The results are presented in Table 2.

Table 2. Computing the DNC

| | Dialogue 1 | | Dialogue 2 | | Dialogue 3 | | Dialogue 4 | |
|---|---|---|---|---|---|---|---|---|
| | J | P | J | P | J | P | J | P |
| IR | 0 | 7 | 3 | 2 | 0 | 4 | 0 | 9 |
| TC | 5 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| RD | 0 | 12 | 0 | 3 | 0 | 1 | 0 | 2 |
| SUM of NCFs | 5 | 19 | 5 | 5 | 0 | 5 | 0 | 13 |
| Utterances | 44 | 48 | 28 | 34 | 28 | 41 | 62 | 66 |
| DNC | 0.11 | 0.40 | 0.18 | 0.15 | 0.00 | 0.12 | 0.00 | 0.20 |

The DNC was counted for the judge and for the program in each of the four dialogues. Also the DNC for all the dialogues was established. For dialogue 1 DNC=0.26; for dialogue 2 DNC=0.16; for dialogue 3 DNC=0.07; and for dialogue 4 DNC=0.10. The DNC measure reveals that the analysed dialogues were fairly cooperative. For a rough comparison we may use the DNC value reported by Plüss for a fragment of a political interview (2010, p. 4) which is 0.68 (19 utterances, 13 NCFs). In the case of the LC study the most non-cooperative dialogue was the first one with DNC=0.26. The relatively low DNC level observed in the sample might result from the high performance of the Dr Wallace program, which gained high scores in this edition of the LC. The results show that the most common NCF of a judge in the LC is topic change after a relevant contribution of the program (TC). When it comes to the program, IR NCFs are the most common (which might be somehow expected, because dialogue programs are still not perfect). What is interesting is that **RDs** are also very common, which suggests that this kind of adversarial responses plays an important role in a strategy for a program in the LC (note that no clear **RDs** were identified for a judge in the research sample).

## Summary and future applications

The results presented in this paper indicate that the DNC measure, proposed originally by B. Plüss, might be successfully used in the context of the LC. Future research in this field will cover a more detailed analysis of the **PO** category of NCFs in the LC (see Section 2.2). What is more, DNC enables the investigation of the consistency of attitude for each judge and program in the whole edition of the LC (or even tracking and comparing this attitude between different LC editions). A comparison of DNC measures for judges and scores gained by programs will also be an aim of future research. In my opinion the analysis of this kind performed in the context of LC may also shed some light on the considerations concerning the Turing test (including such widely discussed topics as the role of a judge for the result of the test or the optimal winning strategies for programs in the test). The DNC measure might also be a useful tool for investigating players' strategies in online games (see e.g. Asher et al. 2012).

## ACKNOWLEDGEMENTS

This work was supported by the Iuventus Plus grant (IP2011-031-771). I would like to give my thanks to Michał Łupkowski for helpful feedback and comments on a draft of this article.

## REFERENCES

Asher, N., Lascarides, A., Lemon, O., Guhe, M., Rieser, V., Muller, P., Afantenos, S., Benamara, F., Vieu, L., Denis, P., Paul, S., Keizer, S., Degremont, C. (2012). Modelling Strategic Conversation: the STAC Project. *Proceedings of 16th Workshop on the Semantics and Pragmatics of Dialogue (Seinedial)*. Paris.

Gamer, M., Lemon, J., Singh, I.F.P. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. Online: <http://CRAN.R-project.org/package=irr>.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Loebner, H. (2009). How to Hold a Turing Test Contest. [In:] R. Epstein, G. Roberts, G. Beber (eds.), *Parsing the Turing Test. Philosophical and Methodological Issues in the Quest for the Thinking Computer* (p. 173–179). Springer.

Łupkowski, P., Wiśniewski, A. (2011). Turing interrogative games. *Minds and Machines, 21*(3), 435–448.

Plüss, B. (2010). Non-Cooperation in Dialogue. [In:] Proceedings of the ACL 2010 Student *Research Workshop* (p. 1–6). Uppsala, Sweden, 13 July 2010.

Plüss, B., Piwek, P., Power, R. (2011). Modelling Non-Cooperative Dialogue: the Role of Conversational Games and Discourse Obligations. [In:] *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue* (p. 212–213). Los Angeles, California, 21–23 September 2011.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind, LIX* (236), 443–455.

Viera, A.J., Garrett, J.M. (2005). Understanding Interobserver Agreement: The Kappa Statistic, *Family Medicine, 37*(5), 360–363.

Access date for all online sources: September 13th, 2013.

**Dr Paweł Łupkowski**, Department of Logic and Cognitive Science, Institute of Psychology, Adam Mickiewicz University, Poznań (Zakład Logiki i Kognitywistyki, Instytut Psychologii, Uniwersytet im. Adama Mickiewicza), Pawel.Lupkowski@amu.edu.pl

## Pomiar niekooperatywnych zachowań graczy – przykład konkursu Loebnera

## Abstrakt

*W artykule przedstawiono zastosowanie miary poziomu niekooperatywności w dialogu, autorstwa B. Plüssa, w kontekście konkursu Loebnera. Omówiono propozycję typów zachowań niekooperatywnych dla dialogów zaczerpniętych z tego konkursu oraz wyniki analizy rzetelności anotowania dialogów przy użyciu tychże typów zachowań. Przedstawiono również miary poziomu niekooperatywności sędziego oraz programu dla czterech rund rozegranych podczas konkursu Loebnera w 2010 roku.*

**SŁOWA KLUCZOWE:** *test Turinga; konkurs Loebnera, strategia, miara niekooperatywności*