Measuring Non-cooperation in Dialogue

Brian Plüss
Knowledge Media Institute
The Open University
Milton Keynes, UK
brian.pluss@open.ac.uk

Paul Piwek
School of Computing and Communications
The Open University
Milton Keynes, UK
paul.piwek@open.ac.uk

Abstract

This paper introduces a novel method for measuring non-cooperation in dialogue. The key idea is that linguistic non-cooperation can be measured in terms of the extent to which dialogue participants deviate from conventions regarding the proper introduction and discharging of conversational obligations (e.g., the obligation to respond to a question). Previous work on non-cooperation has focused mainly on non-linguistic task-related non-cooperation or modelled non-cooperation in terms of special rules describing non-cooperative behaviours. In contrast, we start from rules for normal/correct dialogue behaviour – i.e., a dialogue game – which in principle can be derived from a corpus of cooperative dialogues, and provide a quantitative measure for the degree to which participants comply with these rules. We evaluated the model on a corpus of political interviews, with encouraging results. The model predicts accurately the degree of cooperation for one of the two dialogue game roles (interviewer) and also the relative cooperation for both roles (i.e., which interlocutor in the conversation was most cooperative). Being able to measure cooperation has applications in many areas from the analysis – manual, semi and fully automatic – of natural language interactions to human-like virtual personal assistants, tutoring agents, sophisticated dialogue systems, and role-playing virtual humans.

1 Introduction

This paper describes a general method for measuring the degree of cooperation of dialogue participants' behaviour. Central to the method is the idea, following Traum (1994) and Matheson et al. (2000), that in dialogue obligations are continually created and resolved. Our contribution is a proposal for measuring non-cooperation in terms of the degree to which dialogue participants deviate from the obligations that they acquire during the course of the dialogue. We focus on an application of the proposed method to political interviews in order to evaluate its validity. We developed this method to extend the state-of-the-art of computational dialogue modelling to cases in which the conversational flow is compromised to some extent but without reaching complete breakdown. Shedding light on the nature of linguistic non-cooperation in dialogue promises to yield a better understanding of conversation. The method can be used for the analysis – manual, semi and fully automatic – of natural language interactions and for applications such as human-like virtual personal assistants, tutoring agents, sophisticated dialogue systems, and role-playing virtual humans.

In the remainder of this paper, we proceed as follows. In Section 2, we look at recent research in computational modelling of non-cooperative dialogue. We highlight the similarities and differences with the approach proposed in this paper. The next two sections then describe the two principal steps of our method. In Section 3, we introduce the first step. This step consists of segmentation of dialogue transcripts and coding of the speakers' contributions. We describe the segmentation and annotation schemes, and report on their reliability. In this step, the individual annotations are neutral with regards to cooperation. Section 4 introduces a fully automated method for combining the annotations from the first step with a model of the dialogue game (specific to the dialogue genre in question). The result of

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1925–1936, Osaka, Japan, December 11-17 2016.

this automatic analysis is a dialogue marked up with cooperative and non-cooperative features. These features lead to a score for each speaker that indicates the extent to which they behaved according to the obligations associated with their role in the dialogue, which we interpret as the degree of cooperation of the participant with respect to the conversational setting. The dialogue game model, in this case for political interviews, is extracted from descriptive accounts in the linguistics literature of the dialogue genre. Next, in Section 5, the validity of the method is assessed by analysing the correlation between the resulting scores and human judgement on the same set of political interview transcripts. Finally, Section 6 presents our conclusions and some suggestions for further work.

2 Related Work on Computation and Annotation of Non-cooperation in Dialogue

Possibly the earliest computational model of non-cooperation is presented by Jameson (1989). It includes an extensive study for modelling bias, individual goals, projected image and belief ascription in conversation. Jameson implemented some of these ideas, in the context of used car sales, by means of a dialogue system that can assume different roles (Jameson et al., 1994). These contributions show that user-model approaches to dialogue modelling are flexible enough to account for situations of an arbitrary degree of intricacy. However, as noted, e.g., by Taylor et al. (1996) the level of detail required in the characterisation of the user and the complexity of mechanism for reasoning about user models can lead to problems like infinite regress in nested beliefs (speaker's beliefs about the hearer's beliefs about the speaker's beliefs...).

More recently, Traum (2008) brought attention to the need for computational accounts of dialogue situations in which a broader notion of cooperation is not assumed. Traum's work on non-cooperative dialogue is mainly aimed at creating virtual humans – or embodied conversational agents (Cassell, 2001) – with abilities to engage in adversarial dialogue. Traum et al. (2005; 2008) present a model of conversation strategies for negotiation, implemented as a virtual human that can be used for teaching negotiation skills. A recent version of the system (Plüss et al., 2011; Traum, 2012) supports cooperative, neutral and deceptive behaviour, and also is able to reason in terms of secrecy in order to avoid volunteering certain pieces of information. However, their model the adversarial scenarios by means of a set of rules that the interlocutors follow. Our approach contrasts with this in that it models non-cooperation in terms of systematic deviation from the rules of the dialogue game.

Along lines similar to Traum et al., the work of Kreutel and Matheson (2001; 2003) accounts for non-cooperative behaviour at the level of the task, what the authors call *strategic acting*. At the conversational level, however, their models – as well as those of Traum and Allen (1994) and Matheson et al. (2000) – always discharge a speaker's obligations before considering their private goals. This also holds for the recent work on learning non-cooperative dialogue behaviours using statistical methods (Efstathiou and Lemon, 2014): conversational or linguistic cooperation is assumed (i.e., dialogue participants honour their discourse obligations), whereas non-linguistically, participants fail to cooperate. The method we describe in this paper is complementary to this work in that we aim to characterise, analyse and measure *conversational/linguistic* non-cooperation.

Previous research on dialogue annotation for non-cooperation is scarce. The only instances of complete research we know of are those of Davies (1997; 2006) and Cavicchio (2010) – see also Cavicchio and Poesio (2012).[1] Both are in the context of task-oriented dialogues, and more specifically the HCRC Map Task domain (Anderson et al., 1991; Carletta et al., 1997).

Davies (1994; 1997; 2006) proposes a direct approach to annotating cooperation in order to analyse its relation with effort and task success. Her annotation approach shares some characteristics with ours, but cooperation is judged directly by the annotators, as "positive codings (i.e., finding an instance of the behaviour in an utterance), and negative codings (i.e., finding an instance where we believe a particular behaviour should have been used)" (Davies, 2006, p. 43). In her doctoral thesis, Cavicchio (2010) applies Davies's coding scheme to a multi-modal corpus of the Map Task domain and studies the relation

[1] Additionally, two short papers by Asher et al. (2012) and Afantenos et al. (2012) report on ongoing data collection and preliminary annotation of negotiation dialogues surrounding a board game, following a theory of strategic conversation proposed by Asher and Lascarides (2013).

between (non-)cooperation and emotions. Her focus is not however on how to assess cooperation in dialogue, but on to what extent psychophysiological indicators of emotion (e.g., heartrate and facial expressions) correlate with cooperative behaviour.

The key difference between Davies's and our approach is that the former already includes the normative notion of dialogue game we use later in the assessment of cooperation. This reduces the flexibility of the coding scheme, as the assessment of cooperation is part of the annotation process. By detaching these steps, the method proposed here allows for assessment of cooperation of the same annotated data using different dialogue games, e.g., to explore how the same behaviour would be perceived by audiences with different cultural backgrounds.

3 Corpus Annotation

The degree of cooperation of dialogue participants is determined in two steps. The input for the process is a dialogue transcript. In the first step, this transcript is manually segmented and annotated. In this manual step, the annotators are *not* required to make any judgements about the cooperation of the interlocutors. The actual determination of the extent of cooperation takes places in the second fully automated step.

In this section, we describe the first step by briefly introducing the annotation schemes and providing our results on their reliability. The complete annotation guidelines, tool, and fully annotated corpus are available online.[2]

3.1 The Corpus

In order to test our approach, we applied it to a corpus of six political interviews with a total of 88 turns (3556 words). The number of turns and words in each fragment is shown in Table 1.

Table 1: Political interview fragments in the corpus annotation study

Interview	Turns	Words
1. Brodie and Blair	16	734
2. Green and Miliband	9	526
3. O'Reilly and Hartman	19	360
4. Paxman and Osborne	16	272
5. Pym and Osborne	10	595
6. Shaw and Thatcher	18	1069
Total	88	3556

The fragments were selected from a larger set of 15 interviews collected from publicly available sources (BBC News, CNN, Youtube, etc.). We selected this particular set with the aim of including behaviours at different levels of cooperation for both interviewer and interviewee role. At the same time, we avoided extreme cases in which the exchange broke down or turned into a dialogue of an entirely different type (e.g., confrontation or debate). A second criterion was to ensure coverage of the annotation scheme, with special attention to the dialogue act taxonomy.

3.2 Segmentation and Dialogue Act Annotation

We followed the recommendations put forward in the ISO standard proposal by Bunt et al. (2009; 2010; 2012), simplifying the terminology and some aspects of the scheme when needed. For this we drew on work by Carletta et al. (1997), Allen and Core's (1997) DAMSL, Traum and Hinkelman's (1992) Conversation Acts theory – following Poesio and Traum (1997; 1998) and proposed as a standard by the Discourse Resource Initiative (Initiative, 1997) –, and Stoyanchev and Piwek (2010a; 2010b). We consider two main classes of functions for dialogue acts: Initiating and Responsive. Initiating dialogue acts are primarily meant to provoke a response by the other speaker as opposed to being themselves responses to previous dialogue acts. Responsive dialogue acts are mainly reactions of the speaker to a previous (initiating or responsive) action of the other party. These are distinguished by the prefixes **Init** and **Resp** in Table 3. Initiating dialogue acts are further divided into information giving and information

[2] At `http://mcs.open.ac.uk/nlg/non-cooperation/`.

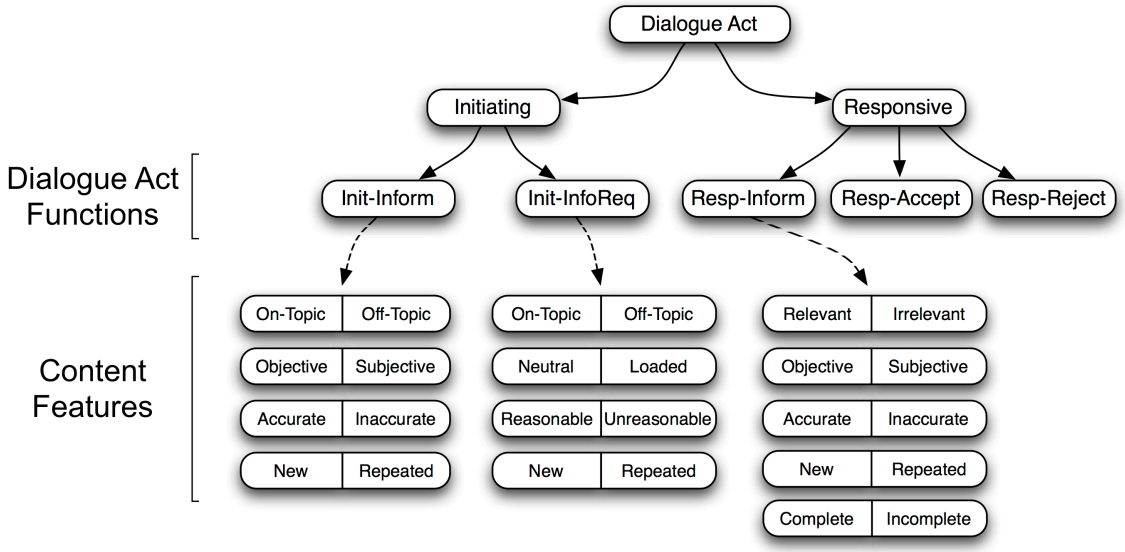

Figure 1: Annotation scheme for dialogue act functions and content features

requesting dialogue acts (**Init-Inform** and **Init-InfoReq**, respectively). Responsive dialogue acts are further divided into information giving, accepting and rejecting dialogue acts (**Resp-Inform**, **Resp-Accept**, and **Resp-Reject**, respectively). The entire annotation scheme, including dialogue act functions and content features, is shown in Figure 1.

For the segmentation and dialogue act annotation stage, four annotators (one of the authors and three native English-speaking researchers with previous experience in dialogue annotation) received transcripts of the corpus and were asked to segment the turns in each dialogue and to annotate each segment with dialogue act functions and, when applicable, with referent segments (i.e., a segment in a previous turn of the other speaker to which the current segment responds). A *segment* is defined as a stretch of a turn that can be labelled with a single dialogue act function. Stretches of a turn can belong to only one segment - i.e., segments do no overlap - and some stretches can remain unannotated. The instructions for segmenting and dialogue act functions for each turn in a dialogue are summarised as follows:

1. Segment the turn by selecting the stretches of speech that have a clear dialogue act function.

2. Assign a dialogue act function to each segment, identifying whether the dialogue act is initiating an exchange (i.e., requesting information, giving information as context for an upcoming question, etc.), or responding to a previous dialogue act (i.e., accepting a question or an answer, answering a question, rejecting a premise, providing additional information, etc.).

3. For each responsive segment, select the segment that caused the response.

Furthermore, when choosing the stretches of a turn that constitute separate segments two criteria are followed: (*a*) the stretch has to be of a length such that it can be assigned one of the available dialogue act functions, and (*b*) its contents have to request for or convey a clearly identifiable, ideally unique piece of information, or several pieces of information on the same topic.

We measured inter-annotator agreement for segmentation using Krippendor's α_U coefficient (Krippendorff, 1995), which was adapted for segmentation of transcribed dialogue. In general, agreement for segmentation, see Table 2, is high, i.e., "substantial", in terms of Landis and Koch (1977). Consistent with intuition, disagreement is greater in dialogues with longer turns.

Annotators independently segmented the turns and selected dialogue act functions for these segments in the same annotation step. This means that the units for annotation identified by one coder can differ from those identified by another coder. These differences make it possible to analyse the reliability of

Table 2: Inter-annotator agreement for segmentation (Krippendorff's α_U)

Interview	α_U	D_o	D_e
1. Brodie and Blair	0.802	3.217	16.251
2. Green and Miliband	0.618	3.276	8.565
3. O'Reilly and Hartman	0.773	4.138	18.219
4. Paxman and Osborne	0.92	0.993	12.468
5. Pym and Osborne	0.672	4.0	12.184
6. Shaw and Thatcher	0.653	7.951	22.890
Overall	0.74	23.574	90.577

Table 3: Inter-annotator agreement for dialogue act functions (Krippendorff's α)

Label	α	D_o	D_e
Init-Inform	0.409	0.040	0.068
Init-InfoReq	0.893	0.009	0.089
Resp-Inform	0.645	0.038	0.107
Resp-Accept	0.606	0.011	0.029
Resp-Reject	0.635	0.018	0.050
Overall	0.657	0.059	0.171

the original annotation data only in terms of Krippendorff's α,[3] which supports missing annotations for some of the items. The value of this coefficient for each label (i.e., regarding the rest of the categories as **Other**) and for entire dialogue act taxonomy is given in Table 3. Agreement ranges from "moderate" to "perfect", with overall agreement being "substantial".

Finally, for responsive dialogue acts, we also asked annotators to indicate which dialogue segment they were a response to. Inter-annotator agreement for referent segment annotations is "substantial" at $\alpha = 0.732$ and $(D_o, D_e) = (0.038, 0.141)$.

3.3 Content Feature Selection

For the second stage, we identified a set of dimensions on which the content of a contribution is judged (see Figure 1). These are based, in part, on Bull and Mayer's (1993) and Bull's (1994; 2003) extensive work on the micro-analysis of equivocation in political discourse.

Annotations from the previous stage were automatically aggregated to produce a single segmented and partially annotated version of each dialogue. These were used in the second stage of the study in which seven annotators (the four coders that took part in the first stage, plus another linguistic expert, with near native English, and two native English speakers with no background in linguistics or experience in dialogue analysis) were asked to select content features.

When judging the content of a segment, annotators had to consider – to the best of their knowledge – several elements of the context of the conversation (e.g., topical, political, historical), as well as common sense, world knowledge, etc. They also had to take into account previous contributions of both participants, and in some cases contributions made later on in the dialogue. Every time annotators made a judgement, they were instructed to ask themselves the following question: 'Do I have any evidence to make this choice?' If the answer was 'Yes', they could go ahead with their choice. Otherwise, they had to be *charitable*. This means that, for instance, if it is not possible to determine whether the information provided in a segment was accurate or not, the first option was chosen. Similarly, if whether a question is reasonable or not cannot be decided, then it is considered reasonable.

Table 4 shows the values of agreement for Krippendorff's α, observed and expected disagreement, observed (or average) agreement A_o, and multi-rater versions of Cohen's κ and Scott's π (or Siegel and Castellan's K) with their respective expected agreements A_e – observed agreement is the same for both coefficients and as given under A_o.[4] We report on agreement for the content features individually, aggregated for each dialogue act function, and overall for the entire corpus. Overall agreement is moderate ($\alpha = 0.454$).

[3]Krippendorff's α is a family of reliability coefficients (Krippendorff, 2003, Chapter 11) defined in terms of the ratio between the disagreement observed among the coders and the disagreement expected by chance: $\alpha = 1 - \frac{D_o}{D_e}$, where D_o and D_e are, respectively, the observed and expected disagreements.

[4]In addition to Krippendorff's α, we report reliability of the annotation of content features using multi-rater versions of Cohen's κ (Cohen, 1960; Davies and Fleiss, 1982) and Scott's π (Scott, 1955; Fleiss, 1971) – called K by Siegel and Castellan (1988). This is because these measures are often found in the literature when discussing the results of dialogue annotation exercises. The general form for both coefficients is: $\pi, \kappa = \frac{A_o - A_e}{1 - A_e}$, where A_o and A_e are, respectively, the observed – or average – agreement and the agreement expected by chance. The observed agreement A_o is the same for both coefficients and equal to the ratio between the number of instances in which any two annotators agreed in the classification of an item and the total number of pairs of annotations of each item. See discussions by Artstein and Poesio (2008) and Plüss (2014, Chapter 4) on the applications of these coefficients to studies in computational linguistics.

Table 4: Inter-annotator agreement for content features

Content Feature	$\alpha\ (D_o, D_e)$	A_o	$\kappa\ (A_e)$	$\pi \mid K\ (A_e)$
Init-Inform	0.398 (0.137, 0.227)	0.863	0.402 (0.772)	0.393 (0.775)
On-Topic | Off-Topic	0.079 (0.100, 0.109)	0.900	0.083 (0.891)	0.072 (0.892)
Objective | Subjective	0.370 (0.305, 0.483)	0.695	0.377 (0.510)	0.365 (0.520)
Accurate | Inaccurate	0.467 (0.090, 0.170)	0.910	0.467 (0.830)	0.463 (0.832)
New | Repeated	0.641 (0.052, 0.146)	0.948	0.640 (0.855)	0.638 (0.855)
Init-InfoReq	0.563 (0.081, 0.185)	0.919	0.564 (0.814)	0.560 (0.816)
On-Topic | Off-Topic	0.104 (0.022, 0.025)	0.978	0.105 (0.975)	0.100 (0.975)
Neutral | Loaded	0.481 (0.213, 0.410)	0.787	0.486 (0.586)	0.478 (0.592)
Reasonable | Unreasonable	0.514 (0.050, 0.104)	0.950	0.512 (0.897)	0.512 (0.897)
New | Repeated	0.806 (0.039, 0.202)	0.961	0.805 (0.799)	0.805 (0.799)
Resp-Inform	0.438 (0.198, 0.352)	0.802	0.443 (0.645)	0.436 (0.649)
Relevant | Irrelevant	0.407 (0.228, 0.385)	0.772	0.411 (0.613)	0.405 (0.616)
Objective | Subjective	0.316 (0.338, 0.494)	0.662	0.333 (0.493)	0.314 (0.507)
Accurate | Inaccurate	−0.014 (0.032, 0.032)	0.968	−0.014 (0.968)	−0.016 (0.968)
New | Repeated	0.763 (0.083, 0.348)	0.917	0.762 (0.652)	0.762 (0.653)
Complete | Incomplete	0.383 (0.309, 0.501)	0.691	0.385 (0.498)	0.382 (0.500)
Overall	0.454 (0.143, 0.262)	0.857	0.458 (0.736)	0.452 (0.739)

4 Computing Cooperation

4.1 From Annotations to Actions Labels

As a first step, the dialogue act functions and content features in the annotations are mapped to *action labels*. The rules of a dialogue game are formulated in terms of the actions that participants perform during a conversation. These actions are represented as labels that capture those aspects of the speakers' contributions that are necessary for applying the rules.

The mapping, see Table 5, is carried out automatically, based on rules that are tailored to a specific dialogue game and coding scheme pair. This approach allows for a separation between the prescriptive nature of the dialogue game and the descriptive character of the coding scheme. Such independence facilitates, for instance, changing the rules of the dialogue game so that it better relates to the social norms, conventions and expectations of different cultural backgrounds, while keeping the coding scheme unchanged and using the same annotated data. It is worth noting that this mapping is independent of the set of interviews in the corpus and, like the dialogue game, was devised based on the linguistics literature for political interviews (Bull and Mayer, 1993; Bull, 1994; Heritage, 1998; Clayman and Heritage, 2002; Heritage, 2005). Also, given the formalisation of the dialogue game (see Figure 2 below), the application of the rules for mapping annotated dialogue into action labels is straightforward.

Table 5: Mapping annotations to action labels in political interviews

Annotation Scheme: Dialogue Act	Content Features	Dialogue Game: Action Label
Init-Inform	**On-Topic** *and* **Objective** *and* **Accurate** *and* **New**	`valid-statement`
Init-Inform	**Off-Topic** *or* **Subjective** *or* **Inaccurate** *or* **Repeated**	`invalid-statement`
Init-InfoReq	**On-Topic** *and* **Neutral** *and* **Reasonable**	`valid-question`
Init-InfoReq	**Off-Topic** *or* **Loaded** *or* **Unreasonable**	`invalid-question`
Resp-Inform	*Any*	`invalid-reply`
Resp-Accept		`acceptance`
Resp-Reject		`rejection`

(a) Interviewer segments

Annotation Scheme: Dialogue Act	Content Features	Dialogue Game: Action Label
Init-Inform	*Any*	`invalid-statement`[a]
Init-Inform	**On-Topic** *and* **Accurate** *and* **New**	`valid-statement`[b]
Init-Inform	**Off-Topic** *or* **Inaccurate** *or* **Repeated**	`invalid-statement`[b]
Init-InfoReq	*Any*	`invalid-question`
Resp-Inform	**Relevant** *and* **Accurate** *and* **New**	`valid-reply`
Resp-Inform	**Irrelevant** *or* **Inaccurate** *or* **Repeated**	`invalid-reply`
Resp-Accept		`acceptance`
Resp-Reject		`rejection`

(b) Interviewee segments

[a] If the interview starts with a question by the interviewer.
[b] In the first turn of an interview that starts with a statement by the interviewee.

4.2 Cooperative and Non-Cooperative Feature Computation

Linguistic cooperation of a dialogue participant with respect to a conversational setting equates to the participant following the rules of the dialogue game for that conversational setting. Figure 2 shows the dialogue game of political interviews that we used for the current study, derived from the descriptive accounts in the linguistics literature (Heritage, 1998; Clayman and Heritage, 2002; Heritage, 2005). Each turn in a dialogue is associated with an amount of cooperation and an amount of non-cooperation. These are given by the number of dialogue rules that the turn, respectively, conformes with and violates. The instances in which rules are conformed with are called *cooperative features* and those in which rules are broken are called *non-cooperative features*.

Participants can break the rules of the game in two ways: (*a*) by performing a conversational action that is not allowed for their role and (*b*) by failing to perform an action they were obliged to perform. Instances of (*a*) are violations of static obligations, which we call *static non-cooperative features*. Instances of (*b*) are violations of dynamic obligations, which we call *dynamic non-cooperative features*. An analogous distinction is made for cooperative features, called, respectively, *static cooperative features* and *dynamic cooperative features*. The *degree of cooperation* of each dialogue participant is thus the ratio between the number of cooperative features – static and dynamic – and the total number of features of that participant. In general, this value can be obtained for the entire conversation and for any continuous fragments. The complete algorithms for computing these features, given an annotated transcript and dialogue game, is available online.[5]

In each turn, we check whether the actions performed by the speaker are allowed for his or her role as specified in the dialogue game. If an action is in the the speaker's set of allowed actions, then it constitutes a static cooperative feature, otherwise it becomes a static non-cooperative feature.

In each turn, we look at the speaker's obligations pending after and discharged in that turn. If an obligation on the speaker has been discharged within the turn, then it constitutes a dynamic cooperative feature, otherwise it becomes a dynamic non-cooperative feature.

Once we have computed the static and dynamic features for each turn, we can regard the proportion of these that are cooperative as an indicator of the extent to which each participant acted within the rules of the game. This is the *degree of cooperation* of a dialogue participant with respect to a dialogue game. Formally, for speaker s and dialogue $D = \langle t_1; \ldots; t_n \rangle$ this is:

$$dc_{D,s} = \frac{cf_{D,s}}{cf_{D,s} + ncf_{D,s}}$$

where $cf_{D,s}$ is the number of cooperative features – both static and dynamic – of participant s and $ncf_{D,s}$ is the analogous for non-cooperative features. This is[6]:

$$cf_{D,s} = \sum_{\substack{i=1 \\ [s_i=s]}}^{n} |sf_i(2)| + |df_i(2)| \qquad ncf_{D,s} = \sum_{\substack{i=1 \\ [s_i=s]}}^{n} |sf_i(3)| + |df_i(3)|$$

Note that, although these definitions are here expressed for the complete dialogue, the same applies to any contiguous subsequences of turns.

The *degree of non-cooperation* of a dialogue participant s in dialogue D is: $dnc_{D,s} = 1 - dc_{D,s}$.

5 Evaluation

We obtained judgements on the behaviour of participants in the political interviews in the corpus by means of an online survey constructed using SurveyMonkey.[7] Observers were shown transcripts of the dialogues and asked to rate the performance of the participants on a 5-point scale (from Incorrect to Correct), based on their intuitions on how interviewers and politicians *ought* to behave.[8]

[5] See `http://mcs.open.ac.uk/nlg/non-cooperation/`.

[6] The elements in the sequences of both static and dynamic features $SF_D = \langle sf_1; \ldots; sf_n \rangle$ and $DF_D = \langle df_1; \ldots; df_n \rangle$ are triples (s_i, C_i, NC_i), where s_i is the speaker in turn t_i, and C_i and NC_i are the associated sequences of, respectively, cooperative and non-cooperative features.

[7] `http://www.surveymonkey.com`

[8] The complete survey is available online at `http://mcs.open.ac.uk/nlg/non-cooperation/`.

$$G_{PI} = (Allow_{PI}, Introduce_{PI}, Discharge_{PI})$$

where

$$\begin{aligned}
Allow_{PI} = \{&[\texttt{ir} : \{\texttt{valid-statement}, \texttt{valid-question}, \texttt{acceptance}, \texttt{rejection}\}], && (1)\\
&[\texttt{ie} : \{\texttt{valid-statement}, \texttt{valid-reply}, \texttt{acceptance}, \texttt{rejection}\}]\} && (2)\\
Introduce_{PI} = \{&[(\texttt{ir}, (\texttt{s}) : \texttt{valid-statement}) \rightsquigarrow (\texttt{ie}, \texttt{acceptance@(s)})], && (3)\\
&[(\texttt{ir}, (\texttt{q}) : \texttt{valid-question}\ \mathbf{N}) \rightsquigarrow (\texttt{ie}, \texttt{acceptance@(q)})], && (4)\\
&[(\texttt{ie}, \texttt{acceptance@(q)}) \rightsquigarrow (\texttt{ie}, \texttt{valid-reply@(q)}\ \mathbf{C})], && (5)\\
&[(\texttt{ie}, (\texttt{s}) : \texttt{valid-statement}) \rightsquigarrow (\texttt{ir}, \texttt{acceptance@(s)})], && (6)\\
&[(\texttt{ie}, (\texttt{r}) : \texttt{valid-reply@(q)}) \rightsquigarrow (\texttt{ir}, \texttt{acceptance@(r)})], && (7)\\
&[(\texttt{ir}, \texttt{acceptance}) \rightsquigarrow (\texttt{ir}, \texttt{valid-question}\ \mathbf{N})], && (8)\\
&[(\texttt{ir}, (\texttt{s}) : \texttt{invalid-statement}) \rightsquigarrow (\texttt{ie}, \texttt{rejection@(s)})], && (9)\\
&[(\texttt{ir}, (\texttt{q}) : \texttt{invalid-question}) \rightsquigarrow (\texttt{ie}, \texttt{rejection@(q)})], && (10)\\
&[(\texttt{ir}, (\texttt{r}) : \texttt{invalid-reply}) \rightsquigarrow (\texttt{ie}, \texttt{rejection@(r)})], && (11)\\
&[(\texttt{ie}, (\texttt{s}) : \texttt{invalid-statement}) \rightsquigarrow (\texttt{ir}, \texttt{rejection@(s)})], && (12)\\
&[(\texttt{ie}, (\texttt{q}) : \texttt{invalid-question}) \rightsquigarrow (\texttt{ir}, \texttt{rejection@(q)})], && (13)\\
&[(\texttt{ie}, (\texttt{r}) : \texttt{invalid-reply}) \rightsquigarrow (\texttt{ir}, \texttt{rejection@(r)})]\} && (14)\\
Discharge_{PI} = \{&[*\text{-}\texttt{question}\ \mathbf{R} \succ \texttt{rejection}], && (15)\\
&[*\text{-}\texttt{statement} \succ \texttt{acceptance}], && (16)\\
&[*\text{-}\texttt{question}\ \mathbf{N} \succ \texttt{acceptance}], && (17)\\
&[*\text{-}\texttt{reply} \succ \texttt{acceptance}]\} && (18)
\end{aligned}$$

Figure 2: Dialogue game G_{PI} for political interviews, consisting of (*i*) $Allow_{PI}$, which specifies the actions allowed for the interviewer (ir) and interviewee (ie), respectively; (*ii*) $Introduce_{PI}$, which stipulates which actions by a specific participant give rise to obligations – e.g., (4) says that a new (**N**) valid question by the interviewer obliges the interviewee to accept that question and (5) says that after accepting a question an interviewee is obliged to provide a complete (**C**) valid reply; (*iii*) $Discharge_{PI}$ specifies how certain actions can count as other actions for (implicitly) discharging obligations – e.g., (15) says that repetition (**R**) of any questions counts as discharging the obligation for a rejection dialogue act and (16) says that a (valid or invalid) statement counts as discharging the obligation for an acceptance dialogue act; so, for instance, an obligation for acceptance by ie that has been created through application of rule (3), can be discharged by ie by producing a statement, in accordance with rule (16).

We used the six interviews in the corpus described above in Section 3.1. Judges (54 respondents in total) were shown the same context and transcript as the annotators.

We studied the relation between human judgement resulting from the survey and the degree of cooperation obtained from the method described above by means of a correlation analysis (see Figure 3).We carried out the correlation analysis separating interviewers from interviewees. The rationale for this step is that some of the rules of the dialogue game are role-specific, making the method strictly different for each participant in an interview. A similar argument applies to the way human observers are expected to judge the behaviour of interviewers and politicians. The two sets of six points are shown in Figure 4, with separate regression lines and values for Pearson's r. The results show that correlation is significantly better for interviewers ($r = 0.753$) than for interviewees ($r = 0.271$). Statistical significance is also stronger for interviewers ($p = 0.084$) indicating a trend towards positive correlation between the results of our method and human judgement. For the interviewees the correlation is not statistically significant ($p = 0.603$). With a sample of this size, correlation analysis is fairly sensitive to outliers, which could explain such a high p-value for the interviewees. Take, for instance, the interviewee in Interview 3 (O'Reilly and Hartman) which corresponds to the point furthest up from the regression line for interviewees (blue) in Figure 4. Coincidentally, Interview 3 has been described by one of the annotators as more like a debate than a political interview which could explain the unexpected value given by the method.

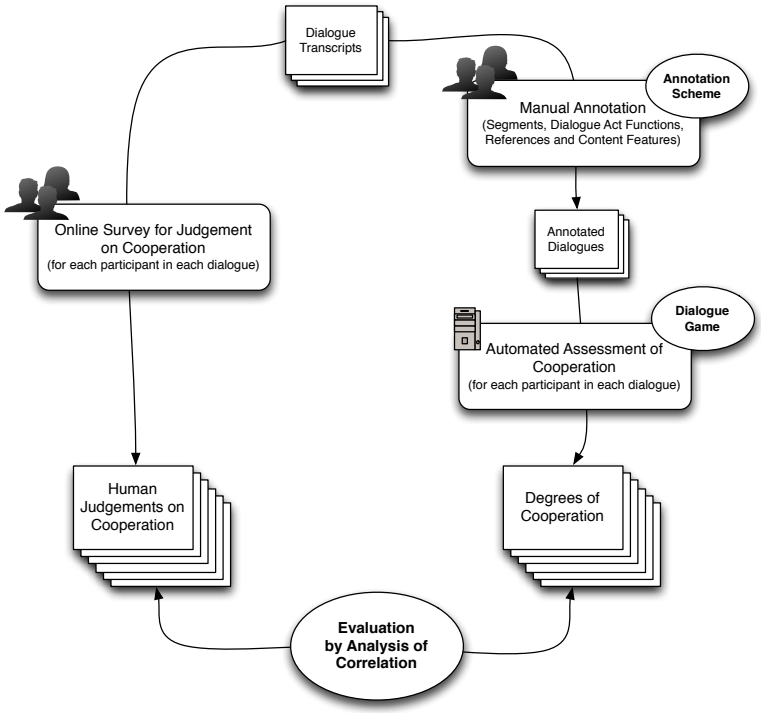

Figure 3: Evaluating the semi-automatic measure of cooperation via correlation with human judgement

6 Conclusions and Further Work

The method presented above is, to date and to our best knowledge, the most elaborate attempt at annotating and analysing naturally occurring dialogue in the light of linguistic cooperation. Also novel is the application of such an approach to a corpus of real political interviews, especially in that both speakers received the same amount of attention and that the method was subject to an extensive evaluation.

The results of the evaluation for reliability are encouraging and indicate that the method is suitable for the systematic analysis of non-cooperation. They also expose some of its weaknesses, such as the difficulties with applying some of the criteria in the manual annotation, a degree of vagueness in the definition of a few of the concepts and the inherent subjectivity of many of the judgements involved in properly characterising non-cooperation.

The evaluation of validity produced fairly good results, especially considering how little information was given to observers in the survey as to what was meant by linguistic cooperation and the total absence of a reference to the specific dialogue game adopted as part of the semi-automatic measure.

It is worth pointing out that the method, in its current form, was able to predict accurately in the six interviews of the corpus which of the participants behaved better with respect to their interlocutors. Beyond the correlation of the precise scores, the ability to determine this binary judgement without mistakes in all cases is of great interest and an indication of the adequacy of the approach.

It is unfortunate that the size of the sample in the corpus prevented from obtaining statistically significant results for each speaker role, particularly the interviewee. A larger sample, including more interview fragments would help in setting this right. Given the relative ease in collecting human judgements, the inclusion of new fragments should start with one or more surveys similar to the one described above. This would allow a decision on the choice of subset of interviews that offers the best coverage of the range of possible behaviours.

6.1 Further work

The method proposed in this paper can be extended to include further aspects of dialogue, like prosody, gestures and other multi-modal aspects of dialogue interaction, as well as sub-utterance elements such as interruptions, incomplete and overlapped speech, etc.

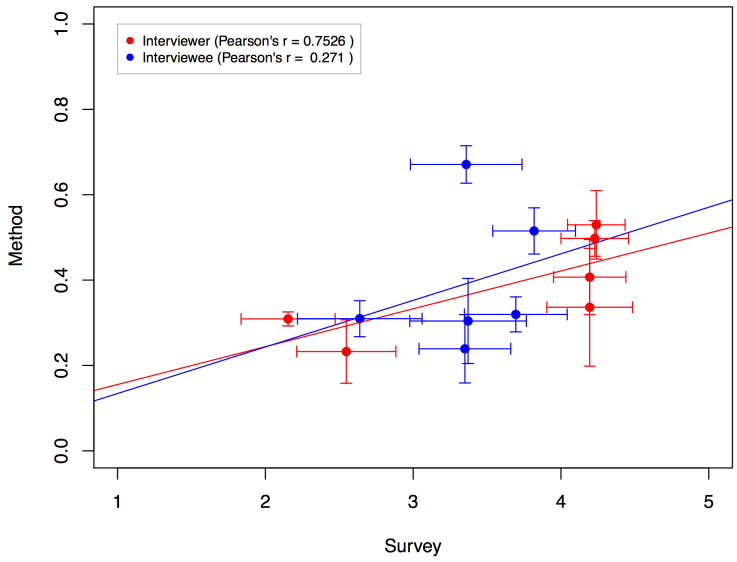

Figure 4: Survey results and the degree of cooperation for interviewers and interviewees (means with error bars, regression line and Pearson's r correlation coefficient)

A further line of work is towards full automation. Data-driven techniques using machine learning can be used to automatically annotate the dialogues with the labels needed to assess the degree of cooperation. Further, we speculate that the rules of the dialogue game could be learned from a sufficiently large corpus of interviews that are deemed conventional.

Our decoupling of the dialogue game from the annotations allows for further evaluation of the approach with participants from cultures with different conventions for political interviews (using the current corpus or a translation of it). Similarly, although the method has been described and evaluated in detail for political interviews, the approach is generally domain-independent. Applications to other conversational domains in which it is possible to identify a set of rules of expected interaction would allow further assessment of the approach. Such domains include courtroom interrogations, tutoring sessions, doctor-patient discussions, customer services, and many more.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anais Cadilhac, Cedric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Vladimir Popescu, Verena Rieser, and Laure Vieu. 2012. Modelling Strategic Conversation: model, annotation design and corpus. In *Proceedings of SemDial 2012 (SeineDial), 16th Workshop on the Semantics and Pragmatics of Dialogue*, Paris, France, September.

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. Technical report, University of Rochester.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline C. Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, and Henry S. Thompson. 1991. The HCRC Map Task Corpus. *Language and speech*, 34(4):351–366.

Ron Artstein and Massimo Poesio. 2008. Inter-coder Agreement for Computational Liguistics. *Computational Linguistics*, 34(4):555–596.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6(2):1–62, August.

Nicholas Asher, Alex Lascarides, Oliver Lemon, Markus Guhe, Verena Rieser, Philippe Muller, Stergos Afantenos, Farah Benamara, Laure Vieu, Pascal Denis, Soumya Paul, Simon Keizer, and Cedric Dégremont. 2012. Modelling Strategic Conversation: the STAC project. In *Proceedings of SemDial 2012 (SeineDial), 16th Workshop on the Semantics and Pragmatics of Dialogue*, Paris, France, September.

Peter Bull and K. Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, pages 651–666.

Peter Bull. 1994. On identifying questions, replies, and non-replies in political interviews. *Journal of Language and Social Psychology*, 13(2):115.

Peter Bull. 2003. *The Microanalysis of Political Communication: Claptrap and ambiguity*. Routledge.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Belis-Popescu, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *LREC 2012*, Istanbul, Turkey.

Harry Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Jean Carletta, Stephen Isard, G. Doherty-Sneddon, Amy Isard, J. C. Kowtko, and A. H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine*, 22(4):67–84.

F. Cavicchio and M. Poesio. 2012. (non)cooperative dialogues: the role of emotions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54:546–559.

Federica Cavicchio. 2010. *Computational Modeling of (un)Cooperation: The Role of Emotions*. Ph.D. thesis, University of Trento. CIMeC.

Steven Clayman and John Heritage. 2002. *The News Interview: Journalists and Public Figures on the Air*. Cambridge University Press.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurment*, 20(1):37–46.

Mark Davies and Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Bethan L. Davies. 1994. To cooperate or not to cooperate - is that the question? In *Proceedings of the Edinburgh Linguistics Department Conference*, pages 17–32. Citeseer.

Bethan L. Davies. 1997. *An Empirical Examination of Cooperation, Effort and Risk in Task-oriented Dialogue*. Ph.D. thesis, University of Edinburgh.

Bethan L. Davies. 2006. Testing dialogue principles in task-oriented dialogues: An exploration of cooperation, collaboration, effort and risk. *Leeds Working Papers in Linguistics and Phonetics*, 11:30–64.

Ioannis Efstathiou and Oliver Lemon. 2014. Learning non-cooperative dialogue behaviours. In *Proceedings of the SIGDIAL 2014 Conference*, pages 60–68, Philadelphia, U.S.A, 18–20 June. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

John Heritage. 1998. Conversation analysis and institutional talk. Analyzing distinctive turn-taking systems. In *Proceedings of the 6th International Congress of IADA (International Association for Dialog Analysis), Tubingen, Niemeyer*.

John Heritage. 2005. Conversation analysis and institutional talk. *Handbook of language and social interaction*, pages 103–147.

Discourse Resource Initiative. 1997. Standards for dialogue coding in natural language processing. In *Technical-Report167, Dagstuhl-Seminar*.

A. Jameson, B. Kipper, A. Ndiaye, R. Schaefer, J. Simons, T. Weis, and D. Zimmermann. 1994. Cooperating to be Noncooperative: The Dialog System PRACMA. *Lecture Notes in Computer Science*, pages 106–106.

A. Jameson. 1989. But what will the listener think? Belief ascription and image maintenance in dialog. *User Models in Dialog Systems. Springer-Verlag*, pages 255–312.

Jörn Kreutel and Colin Matheson. 2001. Cooperation and strategic acting in discussion scenarios. In *Proceedings of the Workshop on Coordination and Action*. Citeseer.

Jörn Kreutel and Colin Matheson. 2003. Incremental information state updates in an obligation-driven dialogue model. *Logic Journal of IGPL*, 11(4):485–511.

Klaus Krippendorff. 1995. *On the reliability of unitizing continuous data*. Sociological Methodology.

Klaus Krippendorff. 2003. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Incorporated, second edition edition, December.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Colin Matheson, Massimo Poesio, and David Traum. 2000. Modelling grounding and discourse obligations using update rules. In *Proceedings of the 1st conference on North American chapter of the Association for Computational Linguistics*, pages 1–8, Morgan Kaufmann Publishers Inc. Morgan Kaufmann Publishers Inc.

Brian Plüss, David DeVault, and David Traum. 2011. Toward Rapid Development of Multi-Party Virtual Human Negotiation Scenarios. In *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, November.

Brian Plüss. 2014. *A Computational Model of Non-Cooperation in Natural Language Dialogue*. Ph.D. Thesis. Department of Computing and Communications, The Open University.

Massimo Poesio and David Traum. 1997. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347.

Massimo Poesio and David Traum. 1998. Towards an Axiomatization of Dialogue Acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, pages 207–222.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, 19(3):321–325.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York, 2 edition, January.

Svetlana Stoyanchev and Paul Piwek. 2010a. Annotation Scheme for Authored Dialogues. Version 1.1. Technical Report 2010/15, Centre for Research in Computing, The Open University, July.

Svetlana Stoyanchev and Paul Piwek. 2010b. Constructing the CODA corpus: A parallel corpus of monologues and expository dialogues. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta, May.

J. A. Taylor, Jean Carletta, and C. Mellish. 1996. Requirements for belief models in cooperative dialogue. *User Modeling and User-Adapted Interaction*, 6(1):23–68.

David Traum and James Allen. 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting of ACL*, pages 1–8. Association for Computational Linguistics Morristown, NJ, USA.

David Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

David Traum, W. Swartout, S. Marsella, and J. Gratch. 2005. Fight, Flight, or Negotiate: Believable Strategies for Conversing Under Crisis. *Lecture Notes in Computer Science*, 3661:52.

David Traum, W. Swartout, J. Gratch, and S. Marsella. 2008. A virtual human dialogue model for non-team interaction. *Recent Trends in Discourse and Dialogue. Springer*.

David Traum. 2008. Extended Abstract: Computational Models of Non-cooperative dialogue. In Jonathan Ginzburg, Patrick Healey, and Yo Sato, editors, *Proceedings of LONDIAL 2008, the 12th Workshop on the Semantics and Pragmatics of Dialogue*, pages 11–14, London, UK.

David Traum. 2012. Non-cooperative and Deceptive Virtual Agents. *IEEE Intelligent Systems: Trends and Controversies: Computational Deception and Noncooperation*, 27(6):66–69.