Dialogue & Discourse 15(2) (2024) 113–144 doi: 10.5210/dad.2024.204

Common Ground inconsistencies in dialogue systems: conflict patterns implied by polar question forms

Maria Di Maro MARIA.DIMARO2@UNINA.IT
Dept. of Electrical Engineering and Information Technology,
University of Naples "Federico II", Urban/Eco Research Centre

Antonio Origlia ANTONIO.ORIGLIA@UNINA.IT
Dept. of Electrical Engineering and Information Technology,
University of Naples "Federico II", Urban/Eco Research Centre

Francesco Cutugno CUTUGNO@UNINA.IT
Dept. of Electrical Engineering and Information Technology,
University of Naples "Federico II", Urban/Eco Research Centre

Editor: David Traum

Submitted 01/2023; Accepted 11/2024; Published online 12/2024

Abstract

In linguistics, research on dialogue systems has accentuated the need to focus on various pragmatic aspects for their management and modelling. Among the most important pragma-linguistic speech acts in dialogue systems studies are Clarification Requests, corrective feedback that in some circumstances require access to the set of shared knowledge known as Common Ground. Regarding Common Ground management, pragmatic studies suggest differences in the type of polar questions that people prefer be used in Clarification Requests, where polar questions can have two possible answers: true or false. This preference appears to depend on the relationship between bias and contextual evidence. In this work, we show that varying the form of polar questions in a given pragmatic setting can influence the capability of people to track Common Ground inconsistencies. As a result, we demonstrate that using a negative polar question in Italian has functional consequences when communicating conflicting material in the Common Ground. This can improve the quality of human interactions with dialogue systems, in terms of an improved identification of the conflict. The results obtained in this work provide insights into design of error reporting approaches in natural interactions.

Keywords: Dialogue Systems, Pragmatics, Common Ground inconsistencies

1. Introduction

The wide success and the increasing popularity of conversational agents are shedding a new light on conversation analysis and on the pragmatic structure of dialogue. This includes the study of the automatic recognition of pragmatic phenomena which are common in human-human interactions (Carberry, 1985; Bunt and Black, 2000; Ammicht et al., 2003; Skantze, 2005; Hough et al., 2017). The significance of dialogue interaction in the realm of Artificial Intelligence cannot be overstated,

as exemplified by the contrasting examples of systems like Alexa and Siri and the advent of GPT-powered language models. Voice assistants such as Alexa and Siri are characterised by their limited set of interaction behaviours, which places a spotlight on the need for further advancements in conversational AI. Similarly, recent developments in GPT-powered language systems have garnered both admiration for their linguistic capabilities and concerns about their potential misuse while also having limitations in their conversational abilities. We aim to contribute to the ongoing evolution of dialogue interaction but also to provide linguistically- motivated models to help address the complex challenges and opportunities posed by the ever-expanding landscape of AI-driven communication. More specifically, our goal is to investigate theory-based strategies to identify inconsistencies in the Common Ground, the set of knowledge shared by the interlocutors in a conversation (Clark and Brennan, 1991).

Despite the documented pragmatic research in the field of conversational AI, some limitations have been highlighted. In de Wynter et al. (2023), for instance, the memorisation skill has been taken into account as an important parameter to assess the quality of the output in LLMs. The authors found out that a significant portion of generated content (about 80%, differently distributed across the models) involved memorisation. However, this memorisation often led to *factual inaccuracies*, *coherence issues*, and *logical fallacies*, where different models exhibit these problems on varying levels. Moreover, a negative correlation between factual errors and memorised content was found, suggesting that the models' ability to recall information impacted the accuracy of their output. The study confirmed the relationship between text originality and memorisation and extended this to overall discourse quality, indicating that models with higher memorisation capability tended to produce higher quality content. Pragmatic strategies adopted in such situations are, therefore, an important research subject, as they are generally concerned with *grounded* information, and thus memorisation. This must, first of all, be studied from a linguistic point of view to address the documented shortcoming of current technological models.

In the pragmatic analysis of conversation, the starting point is to define dialogues as joint activities, for which the goals of both the interlocutors, and their role in a particular interaction must be identified in order to reach the conversation targets (Macagno and Bigi, 2017). Spoken interactions are cooperative, influencing the production of utterances in a given context. As pointed out by scholars like Clark (1996), to pursue the aim to succeed in their joint activity, the interlocutors engage in a communicative process called *grounding*. In conversation analysis, *grounding* refers to the process of establishing that what we intend to say (or what has been said) can be well understood (or has been well understood) (Clark and Brennan, 1991). According to other scholars, such as Allwood et al. (2000), grounding can refer to the determination of what level of perception and comprehension is deemed acceptable. This can vary depending on the type of activity and how critical the information is to that activity. Additionally, the analysis introduces another factor influencing *grounding behaviour*, intended as the evaluation or assessment of how and why grounding (mutual understanding) happens. This holds even in scenarios like casual conversations or conflicts, where there isn't a clear purpose beyond social interaction or disagreement related to the information being conveyed, similar to what is described in this work. To verify the establishment of a *Common Ground* during a dialogic interaction, different linguistic or para-linguistic feedback analysis strategies (Traum, 1999) can be exploited. From a linguistic point of view, dialogue efficiency can rely on the analysis of communicative feedback, whose relevance was pointed out by Allwood et al. (1992) and which continues to be considered a fundamental characteristic in dialogue modelling (Buschmeier and Kopp, 2018). In this work, specific attention is dedicated to computational

pragmatics, with respect to the systems' use of pragmatic tools in specific inconsistent contexts and their impact on the human capability to solve them. The significance of the grounding process in conversational systems has been emphasised in various ways, ranging from signalling uncertainty (Fernández et al., 2007; Hough and Schlangen, 2017) to exploring different levels of grounding (Roque and Traum, 2008; Roque, 2009; Petukhova et al., 2015), and even to the application in the evaluation of dialogue systems (Curry et al., 2017; Zou, 2020). Specifically for the problems concerning inconsistency check and signalling, data we collected in previous studies (Di Maro et al., 2020, 2021a,b), summarised in Section 2, motivate the experiments we carried out in this work.

Various scholars have highlighted the importance of including dialogue exchanges that make use of corrective utterances in their systems to improve the communication process (Bousquet-Vernhettes et al., 2003; Bohus, 2007). This resulted from the users' need to interact with an agent capable of cooperating through communicative actions. Human interlocutors always contribute with questions, answers, and feedback (Beun and van Eijk, 2004). A corrective dialogue is a particular type of dialogue made up of sequential acts which occur when: i) the user notices an error in the system and corrects it; or ii) the user changes their mind; or iii) the user's beliefs are in contradiction with the system's beliefs and expectations. In the first two cases, the corrective dialogue is initiated by the user, whereas, in the last case, it is initiated by the system (Bousquet-Vernhettes et al., 2003).

One example of corrective dialogue in human-machine interaction is the one presented in Beun and van Eijk (2004). The authors focused on a particular communicative problem related to conceptual discrepancies between a computer system and its user. There are two potential scenarios to consider:

- the system may detect when the user mistakenly applies an incorrect action to a specific object;
- the user's dialogue contribution might include inaccurately chosen words, words placed in the wrong order, or an incorrect combination of words.

Concerning specifically the second scenario, the authors reported the following example: if a user asks the system to 'edit process 308', the system could infer that, contrary to its prior assumption, the user believes that processes are editable. Starting from the hypothesis that both the computer and its user have a mental representation of a domain, the mental representation on the computer side, also referred to as ontology, contains conceptualisations that are made explicit in a formal language. Although these conceptualisations are usually incomplete and inaccurate, they can be used to trace the system's reasoning about the concepts, items, and their properties. This representation also allows the detection of conceptual discrepancies, for example when the system observes that the user applies an incorrect action to a particular object. The authors also stated that, although feedback is now used in such systems, there is still no accurate *'mathematical theory'* for natural communicative behaviours and their computational model to human-machine interaction, especially as far as conceptual discrepancies are concerned (Beun and van Eijk, 2004, p. 2). According to the authors, what is still missing is, therefore, a reference model guiding the adoption of a specific type, content, and form of the feedback that has to be generated in a particular situation (Beun and van Eijk, 2004), a gap that we investigate in this work.

General mathematical models describing a linguistic theory are only partially implemented in technological systems. In Prakken (2018), for the case of argumentation-based dialogue, it is highlighted that ad-hoc solutions are often presented for the task at hand. Moreover, the most recent

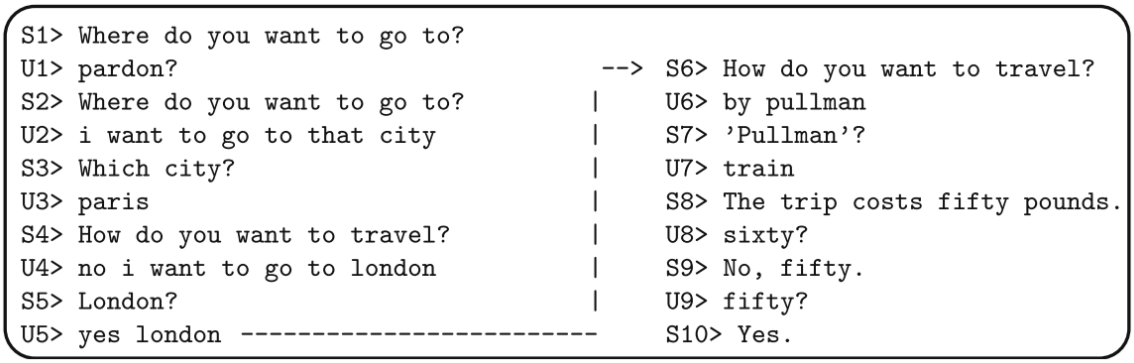


Figure 1: Example of the CLARIE system's capabilities to use Clarification Requests (Source: Purver (2004, p.277).

approaches relinquish the task to machine learning models, which, however, build statistical models, uninformed by underlying reasons.

For example, the choice of the best feedback to use could depend on different factors: i) the domain knowledge in both the system and its user: more specifically, the system's knowledge about the user's conceptualisation; ii) the role played by the system in the interaction (i.e., whether the system is a the expert or not), a parameter which might affect the definition of some of the the ontology features. An example of corrective dialogue, and the forms adopted, in conversational agents is given, for instance, in Purver (2004, 2006), where the system (CLARIE) is capable of handling Clarification Requests uttered by human users (Figure 1). Here, confirmation requests in the form of polar questions are also used as shown in S5, S7, U8, and U9.

In this work, a type of corrective dialogue is investigated, in which a simulated system has a non-expert role and initiates repair strategies for its grounded knowledge, when conceptual discrepancies, or inconsistencies, in the sequence of actions uttered by the user occur. The types and forms of feedback are investigated here, not only as far as the appropriateness is concerned, but especially for the practical effects that act on the interaction itself. We are particularly interested in polar questions, which are defined as questions that make relevant affirmation/confirmation or disconfirmation (Stivers and Enfield, 2010). Polar questions can have two possible binary answers: true versus false. More in detail, the following research questions will be investigated:

- Main question: Does linguistic feedback in the form of different polar questions influence the capability of people to identify the cause of reported Common Ground inconsistencies?
- Secondary question: Do different polar question forms influence the speed with which people take decisions when searching for the cause of reported Common Ground inconsistencies?

The study on inconsistencies within the Common Ground is driven by the need to delve into the practical aspects of this phenomenon. Specifically, when communication is framed in the context of a collaborative effort to accomplish an operational task, the different use of language forms may have a direct impact of task performance, beyond perceived naturalness. In light of this, we have undertaken a comprehensive examination, primarily concentrating on its operational implications. A previous linguistically-informed experiment, which involved the collection of introspective data

(Di Maro et al., 2021c), has paved the way for the investigation outlined here. The main objective of this follow-up experiment is to measure how the capability of human subjects to solve inconsistencies, signalled by a dialogue system, changes when, in the case of positive bias/negative evidence conflicts, a negative polar question (i.e., *Should I not have added salt?*) is used, compared to the use of a positive polar question (i.e., *Should I have added salt?*). By doing so, we aim to provide valuable insights into how inconsistencies in the Common Ground can be managed and leveraged effectively in various linguistic and communicative contexts. This research is not only significant in advancing our understanding of computational pragmatics but also holds the potential to inform real-world applications and improve machine-to-human communication strategies.

The methodology we propose in this work has been developed specifically to answer these questions and it has been designed to artificially elicit inconsistent instructions in the participants. Also, the experimental procedure aims to create the illusion of misinterpretation by the subjects to represent the situation of interest as faithfully as possible.

The paper is organised as follows: in Section 2 the background theory on polar questions and their pragmatic functions is summarised along with the effects on usability in human-machine interaction we expect to observe; Section 3 describes a set of experiments designed to answer the above-mentioned research questions, as far as conflict detection and speed of the detection itself are concerned; Section 4 provides the results of those experiments, showing that the use of different polar questions can influence the capability to identify previously-undetected conflicts with a statistically significant difference: concerning the main question, high negative polar questions resulted in a higher percentage of conflict detection, whereas for the secondary question the speed conflict detection is considered; in Section 5 the theoretical consequences of the obtained results are discussed and formalised; Section 6 draws the conclusions.

2. Background theory

Among the linguistic feedback used in corrective dialogues to convey specific epistemic meanings, polar questions are frequently explored. Polar questions usually encode in themselves not only a mere request but also presuppositions, agendas and preferences. Furthermore, the use of a polar question can also implicate a disaffiliation, meaning the act of opposing something a co-participant has said or done (Steensig and Drew, 2008). In this case of the reference to the informational content, we can, therefore, also speak of epistemically biased questions. According to the literature, one way of expressing disaffiliation is through the use of *Reversed Polarity Questions*, which are questions that convey bias towards the opposite valence than the utterance (Koshik, 2002, 2005). For example, negative interrogatives can also function as positive assertions challenging the recipient's position and vice versa (i.e., *Could I be more wrong?*) (Heritage, 2002). Criticisms and challenges can also be expressed through declaratives (i.e., *You shouldn't have done that*), imperatives (i.e., *Don't do that to me again*), or exclamations (i.e., *How dare you?*), which are perceived more confrontational and explicit and can be therefore face-threatening (Hayano, 2013; Sidnell and Stivers, 2012). Among non-standard communications, conflicting representations (Huang, 2017) are listed as interactions taking place when a discrepancy between what is communicated and what is believed by the agent occurs. In these scenarios, polar questions can, therefore, serve as a knowledge challenging tool (i.e., *You sure about that?*).

Various authors have pointed out how either the original bias of the speaker or the contextual evidence bias could influence the syntactic form of polar questions. For example, according to

		Bias		
		positive	**neutral**	**negative**
Evidence	**positive**		PPQ/RPQ	RPQ
	neutral	HNPQ (*outer*)	PPQ	
	negative	HNPQ (*outer/inner*)	LNPQ	

Table 1: Results for preferred polar question forms per pragmatic cell in English and German, redrafted by Domaneschi et al. (2017); HNPQ refers to high negative polar questions, LNPQ to low negative polar questions, PPQ to positive polar questions, RPQ to really-positive polar questions.

Ladd (1981), in English, high negative polar questions (i.e., *Isn't there a good restaurant nearby?*) is mandatorily used to express the original speaker bias. Whereas a positive bias for a specific contextual evidence is highlighted with a positive polar question (i.e., *Is it raining?*). We define these basic concepts as follows:

Original speaker bias "Belief or expectation of the speaker that p is true, based on his epistemic state prior to the current situational context and conversational exchange" (Ladd, 1981, p. 166).

Contextual evidence bias "Expectation that p is true (possibly contradicting a prior belief of the speaker) induced by evidence that has just become mutually available to the participants in the current discourse situation" (Buring and Gunlogson, 2000, p. 7).

In Domaneschi et al. (2017), possible combinations of original bias of the speaker and contextual evidence were investigated, to point out the influence they may have on the choice of polar question forms. This contrast represents, indeed, the conflict existing between the presupposed knowledge of the questioner and that of the answerer. The experiment was carried out by the authors for English and German but the same observations, with small variations, also appear to be valid for Italian (Di Maro et al., 2021c).

The result of the reference study, in Table 1, shows that both the original bias and the bias derived from the contextual evidence interact in the selection of the appropriate question: in both languages positive polar questions (**PPQ**) are typically selected when there is no original speaker belief and positive or non-informative contextual evidence is provided; low negation questions (**LNPQ**, i.e., *Do you not...?*) are most frequently chosen when no original belief meets negative contextual evidence; high negation questions (**HNPQ**, i.e., *Don't you...?*) are prompted when a positive original speaker belief is followed by negative or non-informative contextual evidence; positive questions with *really* (**RPQ**) are produced most frequently when a negative original bias is combined with positive contextual evidence. Regarding HNPQ, we can distinguish two readings in the column with positive bias and neutral or negative evidence. Ladd (1981) referred to the *outer negation reading* when the speaker wants to double check p, and the *inner negation reading* in which the speaker wants to double check $\neg p$. In the inner reading, negation is part of the proposition being checked, whereas in the outer reading it is not. The two readings can be distinguished by the presence of positive polarity items (i.e. *some*, *already* or *too*), and negative polarity items (i.e. *any*, *yet*, *either*) (Domaneschi et al., 2017).

The results provided in Domaneschi et al. (2017) and in Di Maro et al. (2021c) indicate a preference for specific forms of polar questions to signal different kinds of conflict. Nevertheless, it is important to remember that these are intended as tendencies, as the authors specify that differ-

ent forms are indeed possible in the described scenarios. For example, in a positive bias/negative evidence scenario, PPQs can be adopted alongside HNPQs, according to the strength of the bias, the role of the interlocutors, and their intentionality. From an interaction design point of view, this has the potential to inform about the form in which to present a confirmation and/or Clarification Request as a polar question, depending on the pragmatic context. The experiment resulted in the identification of the most appropriate form according to the type of conflict. This result was used as a starting point for our study which is conversely focused on the resulting functionalities of the appropriate Italian forms, in terms of robustness. Starting from some human-computer interaction usability principles, we can point out some properties which can also be investigated to evaluate the effect of confirmation requests form on interaction quality. In Dix et al. (2003), three main interaction design principles are listed, namely learnability, flexibility, and robustness. In this work, we focus on system robustness. This is defined as the level of support that the system provides to the user in completing and assessing a task successfully; this can also be ensured by the ability a system can have to check message understanding and correcting alleged errors in order to successfully complete the required tasks; the following principles are applied to support system robustness:

1. *observability*, which refers to the possibility of observing the internal state of the system; this can be further represented by five different principles: i) browsability allows the user to explore the internal state without modifying it, ii) default suggests the user possible actions, iii) reachability enables the navigation through observable states, iv) persistence refers to the duration of an observable state, v) task performance includes the services supporting all the possible tasks;

2. *recoverability*, which is the ability a system has to recover in case of errors; error recovery can act forward and backward: i) forward error recovery refers to errors in the current state causing a negotiation from that state to the desired one, ii) backward error recovery aims at correcting the effects of previous states in order to return to a preceding state; Common Ground Clarification Requests function, indeed, as backward inconsistencies recovery of grounded information;

3. *responsiveness*, which is the time the system need to give feedback and communicate with the user;

4. *task conformance*, which refers to the level of support a system offers when a task is executed in an expected way.

Related to the robustness principle, systems can make their internal states observable through verbal or non-verbal interaction. Specifically, when problems occur in information processing, the observable characteristics of such states can be utilised to recover from the problems. From an interaction design point of view, the results obtained in Domaneschi et al. (2017) and Di Maro et al. (2021c) have a potential effect on the principle of robustness. In particular, what can actually take advantage of the correct use of Clarification Requests in terms of consistency between the form used and the type of the Common Ground problem, namely Common Ground inconsistencies, that can be related to *observability* (here, the main question) and *recoverability* (here, the secondary question) features. We, therefore, investigate if different forms of polar questions used to signal an artificially created inconsistency, in series of commands from a human user, lead to a better handling of problematic situations characterised by conflicting statements.

2.1 Common Ground Inconsistencies

Information stored in the Common Ground may be subject to revision when inconsistencies occur. Starting from a formal definition of the possible conflict situations, we focus on the specific problem of bias/evidence conflict.

Given a domain D containing domain items (e.g. cutlery, ingredients, etc.), we define an ordered sequence of actions A on items belonging to D. Since every action has different pre-conditions and consequences, each $a_i \in A$ is associated with a set S_i composed of verifiable pre-conditions $[pre(a_1, pr_1), \ldots, pre(a_k, pr_q)]$ and post-conditions $[post(a_1, po_1), \ldots, post(a_k, po_s)]$. Note that some actions may have multiple pre- or post- conditions (i.e. $[pre(a_1, pr_1), pre(a_1, pr_2)]$. Also, some actions may not even have pre- or post-conditions. In these, po_i and pr_i parameters represent *propositions* that must, in the case of pre-conditions, be verified and, in the case of post-conditions, immediately become true. We will refer to generic propositions using p in the rest of the paper. During the course of the dialogue, applying post-conditions updates a set of propositions P, containing what is true and what is false about items in D. Verifying that actions can actually be performed implies verifying that target propositions in P hold. Consequently, executable actions automatically introduce or remove propositions in P. Conflicts may, therefore, occur in the following conditions:

i) a $pre(a, p)$ is incompatible with the rules of the *Communal Common Ground* (CCG)[1], including Common Sense, resulting in an *impossible* action; for instance, *grind the milk* is impossible, since the pre-condition of the action [Grinding] is to have an ingredient which is solid and not liquid or powder;

ii) a $pre(a_i, \neg p) \in S_i$ is not verified because of a $post(a_j, p) \in S_j$ resulting from a preceding a_j, saved in the set of shared knowledge - the *Personal Common Ground* (PCG)[2]. This causes what we call an *inconsistency*.

Concerning the second point, which is the main interest of this work, we call this type of conflict *Common Ground Inconsistency*, with reference to the incompatibility between the listener belief and the new evidence provided by the speaker. For instance, as above-specified, the action of [Grinding] requires a solid ingredient. The action of [Grinding] results in the ingredient becoming powder as post-condition. This means that, after this action, an action like [Cutting] cannot be performed on the same ingredient as the previous post-condition causes the pre-condition of [Cutting] (i.e., solid) not to be verified.

Clarification Requests can be in this case adopted as a corrective feedback. It is worth noting that, in this work, we are not concerned with the solution to the problem, which may vary (i.e., cancel the previous action, insert corrective actions, etc...). We only concentrate on the detection and signalling of the problem between users.

In Figure 2, the scenario eliciting a Common Ground Clarification Requests is displayed. In the representation of the female agent A, the CCG is stored to guide the process of accumulating information in the PCG. The information (i_1, i_2, i_3, ..., i_n) are communicated by the male agent B to A, and sequentially stored in her PCG. When B utters a new information i_z, this is represented as a new item candidate to be part of the PCG. This representation has *disastrous* results, in that the presence of the new item i_z in the PCG clashes with the presence of another item i_3, whose validity is now

1. The amount of information shared with people that belong to the same community (Clark, 2015)
2. The amount of information collected over time through communicative exchanges with an interlocutor (Clark, 2015)

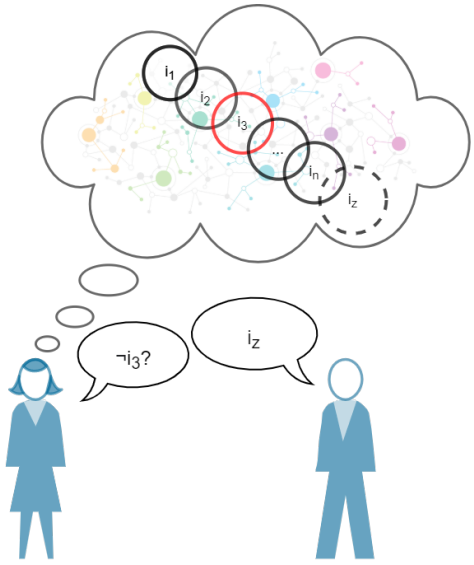


Figure 2: Representation of the Common Ground Clarification Requests elicitation scenario.

questioned. This conflict represents a Common Ground Inconsistency and is translated in the Common Ground Clarification Request $\neg i_3?$, whose form, function and illocutive effect are analysed in the next chapters. As already highlighted, it can be pointed out how important polar questions are to Common Ground Inconsistencies, in that their epistemic stance (or presuppositional stance) is clearly expressed compared to other types of questions. Finally, Common Ground Clarification Requests do not necessary refer to the immediately previous utterance, but to previously, alleged wrongly, grounded information.

In this work, we, therefore, investigate different polar questions in Common Ground inconsistencies scenarios and how they can influence the capability of people to identify the cause of the conflict.

3. Experimental setup

The type of dialogue that we concentrate on to investigate the problem at hand can be divided in two sub-dialogues:

- **Instructing phase**: the participant produces a sequence of utterances representing instructions for the interlocutor to execute;
- **Repair phase**: when a logical inconsistency occurs, the participant has to re-examine the sequence of instructions they gave, according to the indication given by the interlocutor. Re-examination is intended here only in terms of conflict identification and not resolution.

To evaluate the effect of different question forms, we needed to replicate a situation as close as possible to a natural one. Specifically, we create a situation where people unintentionally give a conflicting sequence of instructions, and verbalise them while not realising the mistake until this is signalled to them by the interlocutor. For our case, it is, therefore, important to elicit the incorrect sequence of instructions from participants by making them verbalise it and by artificially creating

the impression of having committed a mistake. In the following subsections, we will describe the experiment in terms of its objectives, reference domain, expected results and hypotheses, and methods adopted in the development of the setup.

3.1 Rationale

To our knowledge, the most widely used experimental procedure to elicit dialogues that include conflict resolution is the map task (Anderson et al., 1991). We draw inspiration from this particular setting to elicit conflict resolution in our domain with two main different aspects to consider:

- The map task presents a global misalignment represented by the different maps. In our case, we concentrate on specific elements in the instructing phase;
- In the map task, people are allowed to realise that the maps are different. In our case, after artificially creating the error, we also hide it during the repair phase.

We propose an experimental procedure to create an apparently consistent sequence of instructions that is later contradicted by another instruction. Furthermore, we aim to create the impression in the subjects of having committed an interpretation mistake when verbalising a target instruction.

The goal of the experiments presented in this work is to evaluate how, given a specific pragmatic situation, different Italian polar question forms can impact the quality of the interaction. Specifically, we concentrated on conflicting situations in which a statement is accepted but, in a later stage of the dialogue, makes another statement unacceptable. Both statements, *per se*, must be acceptable, with conflicts arising only from their combination. This is consistent with the stimuli used in Domaneschi et al. (2017) and Di Maro et al. (2021c) to investigate the appropriateness of the different question forms in each possible conflict category, as explained in Section 2. To avoid considering dialogue situations influenced by personal interests, we focus on a particular type of dialogue, namely the *deliberation* dialogue (Walton and Krabbe, 1995; Prakken, 2018), characterised by the process by which two or more agents reach a consensus on a course of action in a collaborative way. This type of dialogue takes advantage of argumentative capabilities, thus belonging to the category of argumentation-based dialogues. Argumentation-based dialogue refers to the modelling of the verbal interaction aimed at the resolution of conflicts of opinions via the adoption of specific strategies. This field of study consists of a variety of different approaches and individual systems, with few unifying accounts or general frameworks (Prakken, 2018).

More in detail, starting from Domaneschi et al. (2017), we take a different angle on two points:

1. The bias is constructed through an actual sequence of utterances given by the subject according to the semantic content shown on a sequence of slides the subject is looking at;
2. The subject is asked to perform a task on the basis of the information that can be extracted from an error prompt using different forms of polar questions.

In fact, while in Domaneschi et al. (2017) and Di Maro et al. (2021c), the appropriateness is considered, here the functionalities of the forms are taken into account, as a task is depending on the type of question. The considered situation is particularly interesting to guide dialogue systems design, as it corresponds to cases in which the machine's internal representation of the Common Ground is made inconsistent not because of the last incoming command but because of previous actions. In such cases, observability and recoverability are influenced by the capability of the machine to communicate information concerning the problem in a natural and synthetic way.

3.2 Domain description

Let's consider a set of cooking recipes $R = \{r_1, \ldots, r_n\}$, $\forall r_i = \{S_i, A_i, G_i, T_i\}$, where S_i is the series of slides representing r_i, A_i the actions, G_i the ingredients, and T_i the cooking tools for each r_i. Each recipe has a different $\{S_i, A_i, G_i, T_i\}$ tuple coherent with it. The experiment stimuli were composed of series of slides $S_i = \{s_1, \ldots, s_n\}$ and, for each of them, participants were asked to elaborate spoken commands. The slide series defines the recipe. Each slide s_i was designed to represent an action in the recipe r_i belonging to the cooking domain. This domain was chosen for three important reasons: i) the familiarity with this domain is presumably high among speakers, being part of everyday life; ii) similarly to the map-task (Baker and Hazan, 2011), this domain could be applied in a *deliberation* dialogue; iii) contrary to the traditional map-task, the number of different actions is higher, making the tasks more varied and slightly more articulated; moreover, single actions, although atomic, are often linked to each other, in the sense that an action can affect a consequent one in ways that may not be immediately evident.

The use of visual stimuli was adopted to avoid influencing the participants' production with linguistic material. Also, to make the task less imposing, the slides' structure was kept coherent and the representation strategy was designed in such a way that the same action would always be represented by the same image. Hence, each s_i was represented using a fixed structure: given $a_i \in A_i$, the action involved in s_i, a_i was represented on the left side of the slide through an animated image[3]; given the set of ingredients G_i and the set of cooking tools T_i, the set of parameters represented on the right side of the slide through static images was taken from the set $P = G_i \cup T_i$. Figure 3 shows an example of visual stimulus for the action *grinding* applied to the ingredient *nutmeg*.

To simulate the occurrence of conflicting situations, we replaced a slide s'_x representing a correct action in the original recipe with a slide s_x introducing an inconsistent action in S. As already specified, the action depicted in s_x is acceptable, and therefore not impossible, so the subject can continue providing commands with no error prompts when s_x is presented. The inconsistency emerges when the last action in sequence, s_k, cannot be performed because of s_x. This inconsistency, in the form of a contrast between positive bias (having to do X) and negative evidence (having to do something that is prevented by the consequences of doing X), was determined by the opposition of some aspects of s_x and some aspects of s_k. Following s_k, a further slide s_q was presented, containing an error message formulated in different ways, thus creating the experimental conditions considered in this work.

The system prompt message represented in s_q, other than presenting the different forms of Clarification Requests considered in this work, also instructed the participants to go back through the recipe to look for the conflict, so that it was possible to observe the participants' behaviour to evaluate the effect of using the different question forms on the task of error identification. Once s_q is presented and the subject has understood the new directives, they were free to use voice instructions to move in S.

The conflicts introduced in s_x were of two different types: quantity-related or ingredient-related. Quantity-related conflicts refer to the situation where an ingredient is used without specifying the quantity; in fact, when this specification is missing, the interlocutor presupposes that after the action is processed the ingredient is no longer available. On the other hand, ingredient-related conflicts refer to ingredients which have been used in a preceding action instead of the correct ingredient. This makes them no longer available, although they would have been if the correct ingredient was

3. Gifs were generated from video recipes taken from *GialloZafferano* `https://www.giallozafferano.it/`

Figure 3: An example slide from the experiment; the represented action, on the left, elicits a command for the action *grinding* (more specifically, this action is generally represented by showing a cook using a grinder to grind lemon peel) applied to the ingredient *nutmeg* represented in the parameters' slot on the right; the corresponding action to be uttered is, therefore, *grind the nutmeg*. This slide was translated from Italian for the reader's convenience to help describe the setting. In the following Figures, the original slides in Italian are shown.

Recipe	**Code**	**Conflict Type**	**# Slides**
Béchamel	R01	Quantity	9
Carbonara	R02	Ingredient	16
Oat yoghurt baskets	R03	Ingredient	16
Potato croquettes	R04	Quantity	14
Pancakes	R05	Quantity	13
Baked potatoes	R06	Quantity	9
Piadina	R07	Quantity	8
Tuna meatballs	R08	Quantity	11
Tiramisù	R09	Ingredient	11
Small pizzas	R10	Ingredient	12

Table 2: Recipes tested within the experiment with their conflict type description and the number of slides/actions used from the original recipe.

used before. The relationship between conflict types and the 10 recipes used in the experiment is summarised in Table 2.

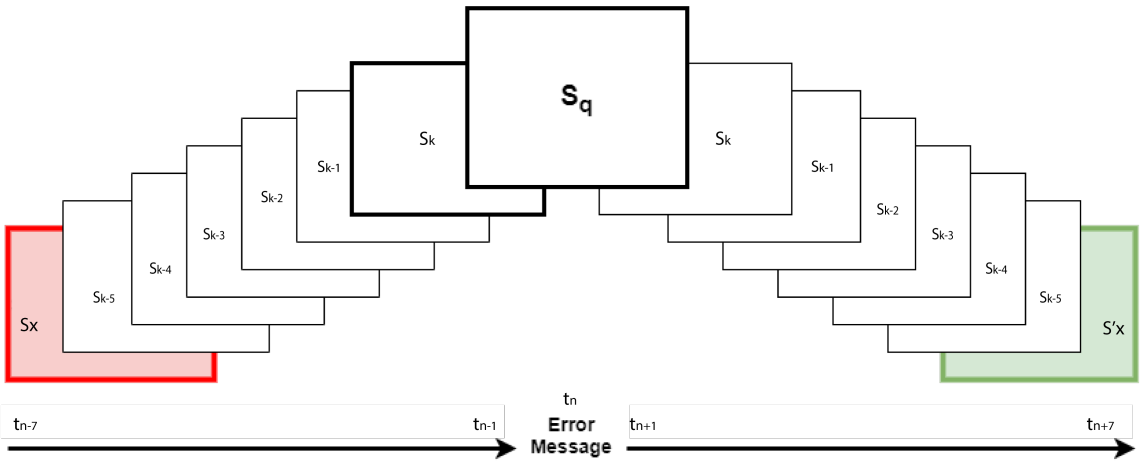


Figure 4: Experiment structure. People go through the actions sequence containing the slide with the error s_x. Then, they see the error message in slide s_q and are instructed to go backwards to find the error. While they instruct the experimenter to do so, the experimenter actually goes forwards in the presentation, showing the reversed sequence containing the correct slide s'_x. This creates the illusion to go backwards in the same sequence while they are actually moving thorough a reversed one.

The slides were organised in such a way that, after the s_q slide, the reversed S sequence was found. Also, in the reversed sequence the original action in the recipe replaced s_x, thus becoming *consistent* and noted as s'_x. The experiment evolved in the following way (as illustrated in Figure 4):

- from the start time t_0, people saw each slide of the first part of the sequence, from s_0 to s_k up to t_{n-1}, seeing the incorrect slide s_x at time t_x;
- at time t_n people saw the error message slide s_q and were instructed to go backwards to find the incoherence;
- as time passed from t_{n+1} onwards, people thought they were going backwards in the sequence, while the experimenter actually moved the presentation forward, through the reversed sequence from s_k backwards;
- at time $t_{n+k-x+1}$, people saw the correct slide s'_x, as if it had always been the one included in the original sequence.

For example, suppose that a recipe has 9 slides and that the fourth slide s_3 contains the error. The presentation contains, therefore, 19 slides: the 9 slides representing the recipe with the error, the error message slides, and the same 9 slides representing the recipe mounted backwards, with the correct slide substituting the wrong one. People go through the sequence from s_0 to s_8 from time t_0 to t_8 and they see the wrong slide s_3 at time t_3. At t_9, they are shown the error message slide s_q.

From time t_{10} to t_{17}, they believe they are going backwards in the sequence, while the experimenter actually moves forward, through the reversed sequence, so that they see the correct slide s'_3 at time t_{17}.

Summarising, after the error message, the time index keeps increasing, while the slide index decreases. This created the effect of showing the reversed sequence to the subject, as if they were going back, while *substituting* the incorrect s_x slide with the correct slide s_y, as shown in Figure 4. This is intended to create a situation as close as possible to the one where people actually pronounce the incorrect command while believing it is correct and proceed with the task. When the error is made clear, they can attempt to find to the command causing the problem. Since it is known that verbalisation can reinforce the memorisation of visual stimuli (Weatherford et al., 2021), the experiment induces the subjects in verbalising wrong commands to recreate a believable situation, in which the possibility to detect the issue independently from the type of confirmation request is still present. This is especially important to obtain a fair baseline for the presented comparisons. The task ends when people renounce or indicate a slide as the one containing the error. The task ends regardless of whether people are correct or not. Furthermore, to better represent a situation in which an error occurs, participating subjects were not informed that there was the possibility that an elicited command could give rise to an error. This way, they would not force themselves into remembering precisely the previous commands they gave. Furthermore, to reduce the possibility that the s_x was in the subject's short term memory, s_x was introduced at a minimum distance of 5 slides from s_k.

Finally, to better exemplify the experiment, in Figure 5 and Figure 6, we reported the sequence of slides used for the recipe *piadina*, highlighting s_0 as the source of quantity-related incoherence. The recipe used comprises the following actions:

s_0 **Put the flour in the bowl**

s_1 Mix salt, lard, baking soda, and water to the flour

s_2 Stir the mixture

s_3 Add water to the mixture

s_4 Stir the mixture

s_5 Add water to the mixture

s_6 Knead the dough

s_7 Put the flour on the counter

s_q Error message

[...]

s'_0 Put part of the flour in the bowl

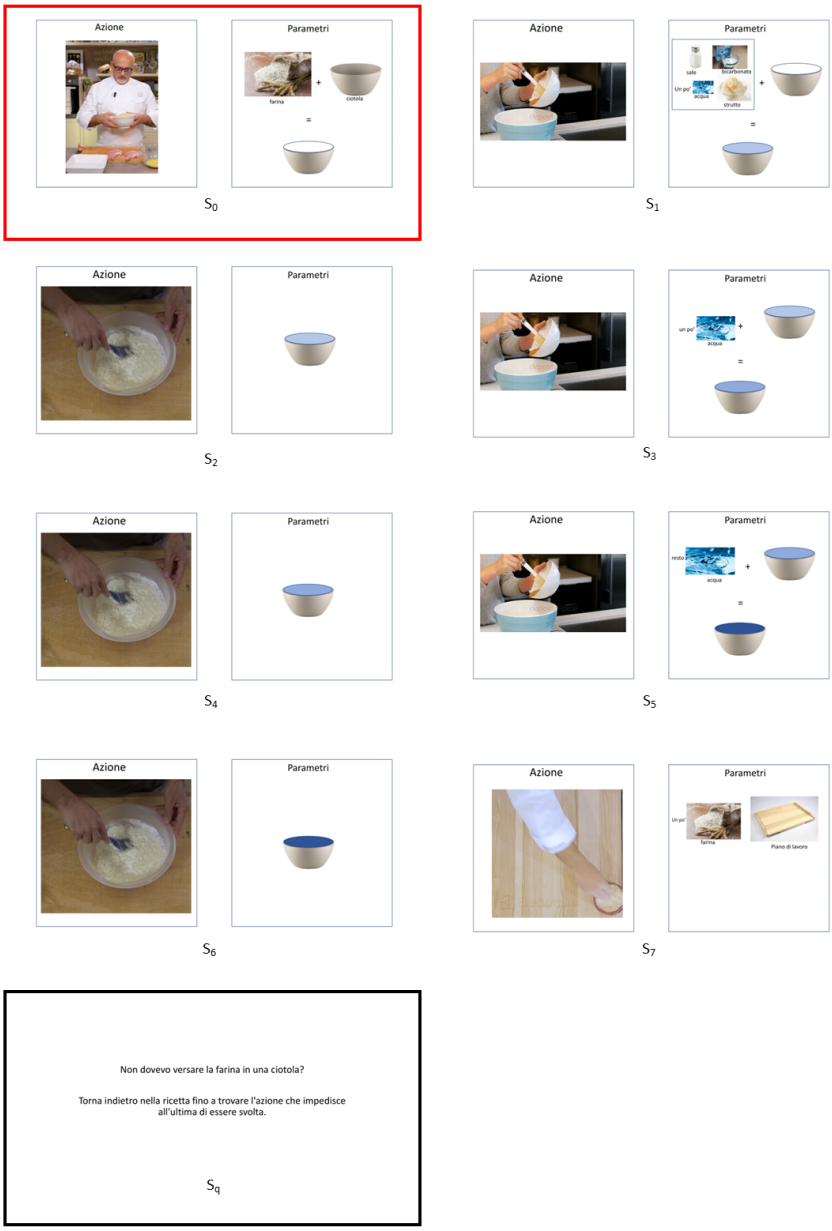


Figure 5: The actions sequence in the recipe during the first phase of the experiment, up to the error message slide s_q (English translation of the above reported message: *Should I not have added the flour to the bowl? Go back through the recipe until you find the action that prevents the last one from being carried out.*), and containing the error slide s_0.

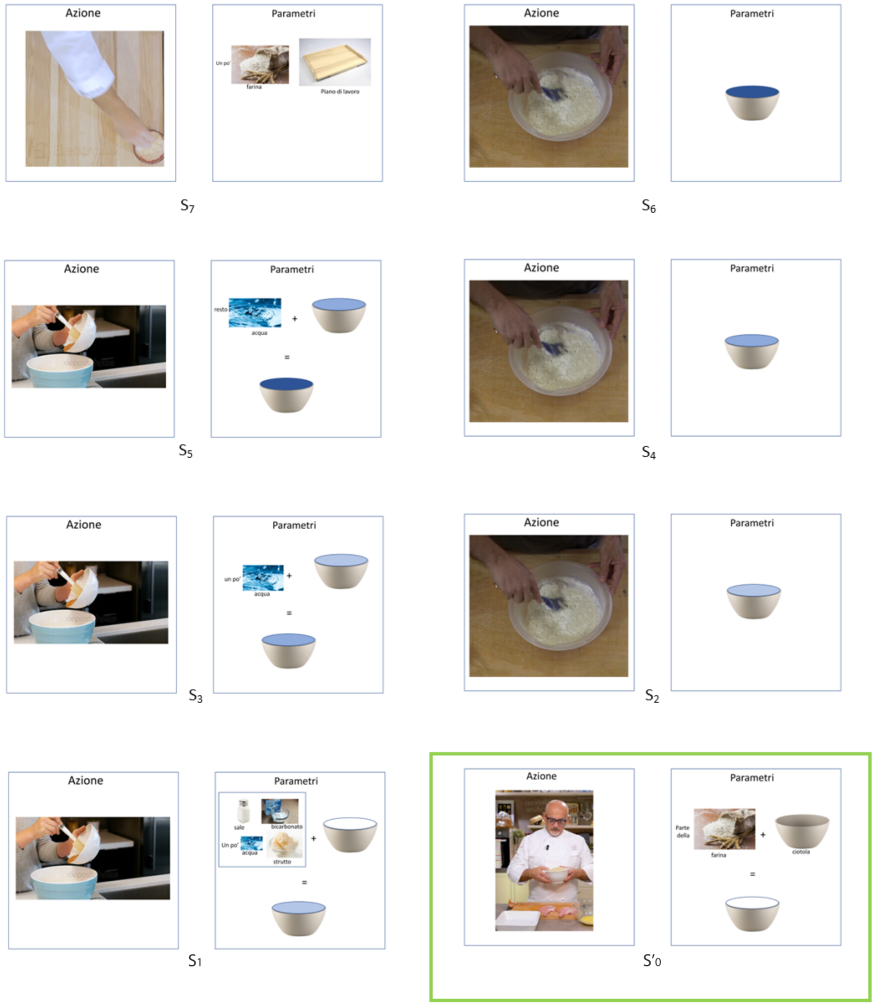


Figure 6: The reversed actions sequence in the recipe, shown during the second part of the experiment, containing the correct slide s'_0

3.3 Hypothesis

As previously mentioned, the goal of the experiment was to check if different system error messages, shown in s_q, were more or less efficient in signalling to the user the existence of a conflict arising from s_x and its details in a succinct, natural way. Our initial hypothesis is that, similar to what has been observed in other production and introspective studies (Domaneschi et al., 2017; Di Maro et al., 2021c), in the conflicting situation presented here, the adoption of an HNPQ results perceptually and operationally in a more effective conflict detection. More specifically, we expected that in expressing conflicts between previous beliefs (positive bias) and opposing contextual obser-

	1st Recipe	2nd Recipe
P1	R09	R08
P2	R07	R03
P3	R03	R02
P4	R01	R07
P5	R06	R01
P6	R04	R10
P7	R02	R05
P8	R05	R06
P9	R10	R09
P10	R08	R04
P11	R07	R05
P12	R05	R07

Table 3: Recipes' distribution for each experimental session.

vations (negative evidence), the use of a negative polar question would: a) improve the capability of the participants to identify the error, suggesting that the use of such questions would improve the observability degree of a dialogue system, i.e., main question, and b) decrease the required effort to intervene on the system to address the inconsistency, i.e., secondary question (Section 1).

3.4 Methods

Instructions given to the participants were presented in the first slide of the experiment and contained the following text:

> In this experiment, we ask you to look at a series of slides describing a recipe and tell the experimenter what to do to make it. Each slide is made up of actions and parameters: actions generally describe what to do, while parameters show which objects are involved in the action. In case of problems, you are free to move back and forth in the recipe as you like by asking the experimenter in which direction to move the slides.
> Take your time to think about what to say and give your instructions when you are sure they are correct.
> We first start with a training recipe, in which you are free to ask any questions you want to the experimenter. Once the test starts, the experimenter will no longer be able to answer you, except for carrying out your instructions until the end of the experiment.

A total of 36 participants was recruited for the experiment, each of which had to perform 2 tasks (72 tasks were collected in total). Participants were divided into three, gender balanced, groups, one for each experimental condition:

- Control group: the first experimental condition consisted of two tasks combining two different recipes as shown in Table 3; this condition was used both as validation for the experimental setup, in order to understand if the slides and the task were understandable for the participants,

and as analysis of the general error message which was used to signal the conflict to the participant (i.e., *This action is not possible because it clashes with a previous one*). The resulting collected values were, therefore, used as a term of comparison for statistical analysis.

- PPQ group: the second experimental condition differed from the previous one just for the typology of error message presented to the participant, where a positive polar question was instead used (i.e, *Should I have added the flour to the container?*). The use of the most frequent polar question form was useful to test its appropriateness in bias-evidence conflicts in simulated human-machine interactions.
- HNPQ group: the third experimental condition, similarly as the previous one, made use of a negative polar question, and more specifically of a high negation polar question in the past tense (i.e, *Should I not have added the flour to the container?*), whose appropriateness in the positive bias versus negative evidence scenarios was confirmed in the experiments described in (Domaneschi et al., 2017).

For the control group, the average age was of 25.5, with an average self-evaluation of their cooking skills equal to 2.33 (on a scale from 1 to 5). For the PPQ group, the average age was of 27.25 with an average of 2.67 self-evaluated cooking skills. Finally, for the HNPQ group, the participants were on average 26.08 years old and their average self-evaluated cooking skills was 2.92 points. No significant differences were found in the comparison of the self-evaluated capabilities among the groups. Because of Covid-19 restrictions, the experiment was carried out online.

Before presenting the two recipes used for the experiment, a training recipe was used to let participants familiarize themselves with the setting, to learn how to interpret the slides and to ask questions. This did not contain a conflict, but was just used to explain the structure of the experiment. After this training session, participants did not report significant challenges in interpreting the slides. An interaction example is reported below:

```
Recipes: Zucchini "alla scapece"

U1: Cut zucchini;
U2: Add salt;
U3: Put the oil in a bowl;
--- ERROR ---
U4: Add salt and mint to the bowl;
-------------
-- CORRECT --
U4*: Add salt and chili to the bowl;
-------------
U5: Mix the mixture;
U6: Cut the garlic;
U7: Add garlic and mint to the mixture;

Error message:

Control group: "This action is not possible because it clashes
with a previous one";

PPQ group: "Should I have added mint to the bowl?"
```

```
HNPQ group: "Should I not have added mint to the bowl?"
```

In this example[4], we reported the sequence of utterances for a recipe. The erroneous action (U4) states to add the *mint* to the *bowl*, instead of *chili* (U4*). At the time U7 is uttered, *mint* is no longer available, as it was already used. At this point, the error message is prompted and the recipe can not proceed. The correct action, as reported in U4*, suggested, instead, the addition of another ingredient, i.e., *chili* instead of *mint*, which, instead, will be required later. The error message is different according to the experimental group, as previously described.

4. Results

At the end of the data collection phase, about 372 minutes (122 minutes for the Control group, 145 minutes for the PPQ group, 105 minutes for the HNPQ group) of audiovisual recording were collected and annotated using ELAN (Wittenburg et al., 2006). Annotations, provided by the authors, were used to compute the results. They marked the slides boundaries appearing in the videos and the speech fragments containing subject-uttered commands. For the rest of the presented analyses, slide change times will be considered to investigate the proposed research questions. First of all, we verified that the error inserted in each recipe was found at least one time and that the error was never systematically found. This indicates that the considered recipes and the kind of error introduced were neither too difficult nor too easy. A summary of the percentage of times the error was correctly identified, in each recipe, is shown in Figure 7. Next, we consider the number of times in which participants were able to identify the slide containing the action that caused the conflict, according to the type of error message. A summary of the performance of each participant, for each experimental setting (control, PPQ, HNPQ), is shown in Table 4, where the conflicts found are ticked, while aggregate counts are reported in Table 5.

Results for our main research question show that, as expected, the slide causing the conflict was found most often in the HNPQ configuration. Since both the PPQ and the HNPQ configurations improved the capability of the participants to identify the error, the significance of the effect was checked using the binomial test, a non-parametric test for binary variables (Wagner-Menghin, 2014). The test showed that, when using HNPQs, the conflict was found more frequently, with respect to the Control group, in a statistically significant way ($p = 0.005$). This result is also consistent with HNPQs being the preferred form of Clarification Requests for the considered conflicts in Domaneschi et al. (2017) and Di Maro et al. (2021c). On the other hand, when using PPQs, the difference with the Control group resulted not to be statistically significant ($p = 0.4$). Although PPQs do provide more information about the problem than a generic error message does, since the counts of conflicts found is higher in PPQ than in Control group, its use is not statistically different from that of the generic error message. The interpretation of the results leads us to the following observations:

- PPQs and HNPQs are not expected to be mutually exclusive in the considered situation: as a matter of fact, they are both reported to be acceptable but there is a difference in the ability of the subjects to detect errors. This is consistent with what was reported in previous experiments Domaneschi et al. (2017); Di Maro et al. (2021c), where the frequency of occurrence of

4. Translated from Italian
PPQ: *Dovevo aggiungere la menta alla ciotola*
HNPQ: *Non dovevo aggiungere la menta alla ciotola?*

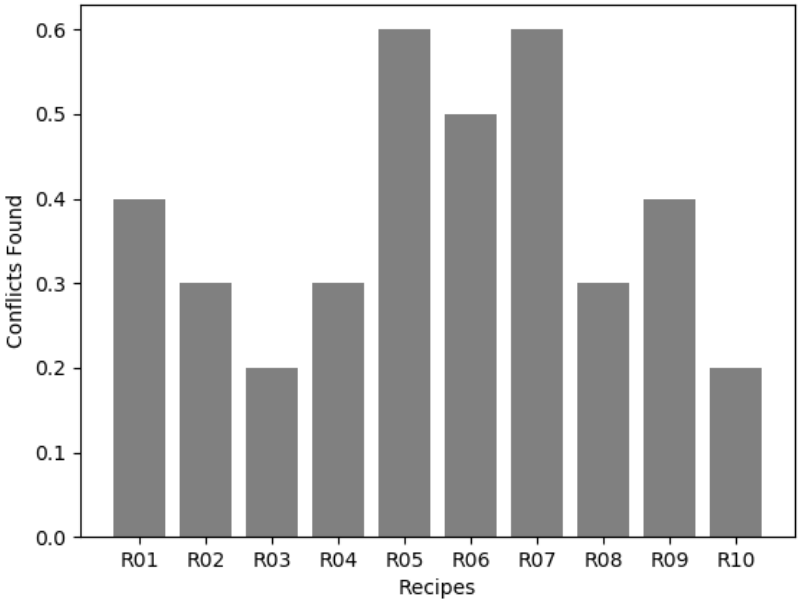


Figure 7: Percentage of times conflicts found per recipe.

	Control		PPQ		HNPQ	
	1st Recipe	**2nd Recipe**	**1st Recipe**	**2nd Recipe**	**1st Recipe**	**2nd Recipe**
P1	✓	✓	✓			✓
P2			✓			
P3			✓	✓	✓	
P4	✓	✓	✓		✓	
P5			✓	✓	✓	
P6			✓		✓	✓
P7	✓				✓	
P8	✓	✓			✓	✓
P9				✓	✓	✓
P10				✓	✓	
P11	✓	✓			✓	✓
P12			✓		✓	✓

Table 4: Distribution of conflicts found by each participant: for each experimental group (control, PPQ, and HNPQ), found conflicts are ticked per each participant for each corresponding recipe (1st and 2nd Recipe).

	1st Recipe	2nd Recipe	Total	Percentage
Control	5	4	9	37.5
PPQ	7	4	11	45.83
HNPQ	10	6	16	66.67

Table 5: Number of conflicts found in the three experimental setups (per recipe, in total, and in percentage).

HNPQs was higher than PPQs in the considered type of conflict. PPQs, however, were not completely unobserved;

- Since the only difference between the two forms is the negation element *non*, both convey the same amount of informative content more than the baseline but only the HNPQ form outperforms it in a statistically significant way, supporting a view that HNPQs do provide an advantage in functional terms;
- The acceptability of the two forms in previous experiments also explains the non statistically significant difference between HNPQs and PPQs: while they can convey the intended message, in general, being used more frequently to communicate other kinds of problem (Section 2), also causes them to be sometimes misinterpreted in this pragmatic condition.

Support to the third point was partially provided by the feedback of one participant, who spontaneously reported that the PPQ led him to think that the question was referring to the last presented slide, rather than to a previous one, as if it was a different kind of confirmation. In fact, while PPQs can be considered as a *grounding act*, which verify the correctness of the previous discourse unit, HNPQs function more as an *argumentation act*, referring to a global rather than local problem (Traum, 1999; Di Maro, 2021a).

This suggests that using the appropriate syntactic form to convey pragmatic meaning when signalling bias/evidence conflicts may not only be a generic preference or a mere matter of appropriateness but it may actually improve the quality of the conveyed message, as people are indeed able to identify problems more easily. From the previous analysis, it was found that people tend to make the right choice, in addressing a $p/\neg p$ conflict, more frequently when this is signalled using a HNPQ. A performance decrease is observed during the second round of the experiment, in all the conditions: it seems reasonable to attribute this to fatigue effects rather than factors due to the stimuli because it is consistent among the participants and is not caused by the recipes themselves because they all appeared both as the first sequence and as the second one.

To answer to our secondary research question we needed to evaluate how *efficiently* people reach a conclusion after being presented the error prompt. For this reason, the sequence of steps followed when searching for the conflict was considered. As participants may take a quick decision but indicate the wrong slide or they may simply rapidly decide they are not able to find the conflict, it is not possible to simply consider the time spent looking for the conflict. In the indicated cases, for example, a short amount of time to reach a wrong conclusion or to quit the task would be considered equally to cases in which people rapidly got to the right choice. For this reason, the sequence of steps followed by participants during the search phase was compared with the ideal sequence that would lead a person who has understood the problem to the problematic slide. For the comparison, the Dynamic Time Warping (DTW) (Müller, 2007) algorithm was used and, specifically, warping

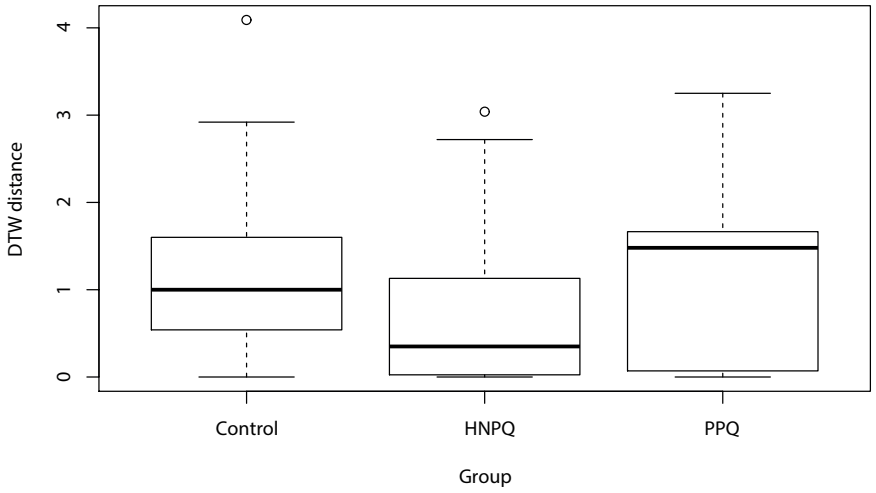
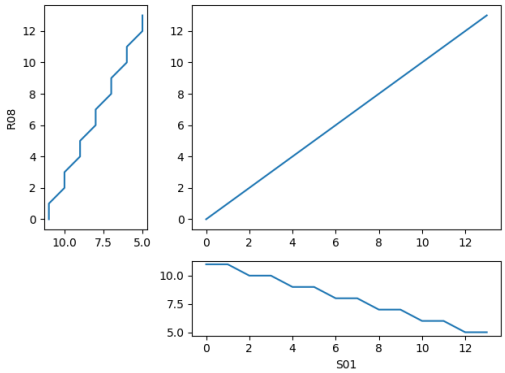


Figure 8: Box plots representing distances distribution in the three experimental conditions.

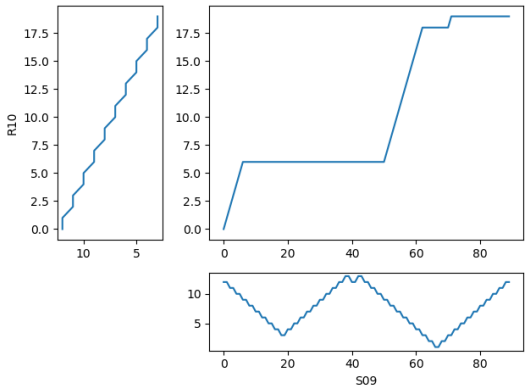
	Control	**HNPQ**
HNPQ	0.089	-
PPQ	0.75	0.39

Table 6: Pairwise comparisons using Wilcoxon rank sum test with Hölm adjustment.

distances were considered as an indicator of how efficient the observed sequences of steps were with respect to the ideal one, which directly reaches the problematic slide. An example of a user who goes directly to the conflicting slide is given in in Figure 9a, whereas an example of a user who is not sure of which slide caused the problem and goes back and forth in the sequence is shown in Figure 9b. In absolute terms, users explored the sequence in a closer way, on average, to the reference curve when the HNPQ was used, when compared with a general error message and a PPQ. To check whether the average differences were significantly different, the Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to check for normality. Since the distributions were not normal, the Kruskal-Wallis test (Kruskal and Wallis, 1952) was used to check the differences shown in Figure 8. In this Figure, the distributions of Dynamic Time Warping distances, for each experimental condition, are shown. The difference was found not to be statistically significant in any case. Pairwise comparisons were performed using the Wilcoxon rank-sum test with the Hölm adjustment to further detail the situation, shown in Table 6. Therefore, while it is confirmed that, by using HNPQs, participants identify the problem more frequently, when they understand the kind of mistake they should look for, we could not find evidence, in our data, that they find it in statistically different amounts of time.



(a) An observed sequence (lower graph) from a user who understood the problem after reading the prompt compared with the optimal sequence (leftmost graph). The number of turns needed to reach the conclusion is exactly 13 as expected in the reference graph. The user directly goes to the wrong slide in the numbered sequence (5), leading to perfect alignment. In this case, the dynamic time warping algorithm reports a warping distance of 0 (no alignment effort needed).



(b) An observed sequence (lower graph) from a user who repeatedly moved through the slides sequence looking for the problem after reading the prompt compared with the optimal sequence (leftmost graph). The number of turns needed to reach the conclusions is much higher than expected and the user keeps going back and forth in the numbered slides sequence. In this case, the dynamic time warping algorithm reports a warping distance of 1.93 (significant effort).

Figure 9: Evaluations examples with Dynamic Time Warping. For each Figure, the reference graph, on the left, represents the optimal sequence of steps through the recipe to get to the error slide and report the error, thus terminating the exploration. The lower graph represents the observed sequence of steps taken by an example subject before reporting the error. In the first case (a), the user immediately found the error, producing the ideal sequence of steps. In the second case (b), the user went back and forth through the sequence of slides multiple times before identifying the error. The central graph shows the alignment between the two sequences, produced using the Dynamic Programming algorithm for classic DTW, as reported in Müller (2007). The alignment effort represents the DTW distance between the reference sequence and the observed one. In the reference graph, the x axis represents the slide number and the y axis represents the step. The axes are inverted, in the lower graph, with the y axis representing the slide number and the X axis the number of steps. The central graph x and y axes correspond, respectively to the x axis of the observed sequence and to the y axis of the reference sequence. This represents the optimal alignment between the two sequences.

5. Discussion

This Section describes how the application of the results obtained from an experimental methodology on the operation of certain linguistic forms in specific pragmatic contexts can be formalised for the purpose of framing them in a developing theoretical framework for argumentation-based dialogue. Our discussion will now cover how these findings contribute to the design of real dialogue management systems. We believe our results have an impact in the following ways:

- the kind of conflict we studied has real-world implications for designers of dialogue systems, as the syntactic form in which the error is presented, depending on the nature of the error itself, does have an impact in the users understanding it;
- the conflict we analysed can be represented in formal terms so that dialogue systems can do the necessary inference to recognise and respond appropriately. We will show in the rest of this Section how the use of graph databases enables this.

The presented approach is designed to be compatible with the system architecture assumed by the Framework for Advanced Natural Tools and Applications with Social Interactive Agents (FANTASIA) (Origlia et al., 2019, 2022). FANTASIA[5] is a plugin for the Unreal Engine designed to support the development of Embodied Conversational Agents. From a terminological point of view, we will adopt some of the concepts presented in FANTASIA.

First, the type of system in which the pragmatic strategies described here can be applied will be specified. Then, details will be provided on the structuring of the knowledge it deals with, with its formalisation and conflict identification strategy. Proving the efficiency of such a form in a defined pragmatic situation, like the one tested in the experiment presented in this work, can be used to good advantage in dialogue systems aimed at learning sequences of actions uttered by a human interlocutor. These applications require the user to have a leading role, and therefore a higher knowledge (K+), and the machine to have a subordinate one, corresponding to a lower knowledge (K-). This type of task can be considered as a sub-type of *User-initiative tasks*. In addition to the characteristics typical of a User-initiative system, in our model, the system checks for consistency based on shared rules. In such situations, the domain information given by the users builds the PCG, that is the set of information collected over time through communicative exchanges with an interlocutor. In other words, it can be considered as a record of shared experiences new to the receiving system, although the general knowledge of the domain are conversely already shared in the CCG, that is the amount of information shared with people that belong to the same community, that is to say, people that share general knowledge, knowledge about social background, education (schools attended, levels of education attained), religion, nationality, and language(s) (Clark, 2015) (Section 2.1). The user has, therefore, more knowledge (K+ position) of the domain with respect to the system, as the desired goal is known by the user. Conversely, the system does not have the same K+ position. Nonetheless, the structured PCG is used to build presuppositions with strong confidence, that make the system closer to the K+ position to the point that, in case of inconsistencies, the system can assume the role of questioner. In such a scenario, each set of actions $A = \{a_1, \ldots, a_n\}$ contains both pre- and post-conditions $a_i = \{pre\_con, post\_con\}$. Pre- and post-conditions-based inconsistencies between two uttered actions occur when the post-condition resulted from a previous action is not compatible with the pre-condition of a new action, based on the rules of the CCG. The system

5. github.com/antori82/FANTASIA

can use a knowledge representation module, in the form of a graph, to verify the compatibility with the PCG, as shown in Di Maro et al. (2021b).

In this case a confirmation request is used. On the one hand, if an inconsistency occurs, the problem is recognised and signalled by using a HNPQ, which resulted in an improvement in the efficiency of conflict detection with respect to the baseline. In fact, as already pointed out in Domaneschi et al. (2017) for German and English, this polar question form is suitable to the type of conflict arising when a positive bias clashes with a negative contextual evidence. This is because a preceding action, part of the PCG, becomes the system's presupposition, whereas the new uttered action is made impossible because of an unverified pre-condition, representing the negative evidence. If, on the other hand, the inconsistency is between a pre-condition and the rules of the CCG, the corrective feedback, would be of another type, i.e., explanation.

Previous experiments showed that PPQs and HNPQs can be both adopted in negative bias/positive evidence conflicts, but with a significant difference in the frequency of occurrence. However, PPQs are more frequent in other pragmatic situations (i.e., positive bias/neutral evidence), with a confirmation function, so that their use in the context studied in this work, while acceptable, may be easily misleading for the participants. While the use of PPQs, therefore, does not implicitly prevent the problem identification, it makes it harder to identify it, not producing a significant improvement with respect to the baseline. Furthermore, it has to be remembered that the questions were presented in the written form. The interpretation of bias in polar questions can, conversely, be also affected by intonation (Asher and Reese, 2007; Savino, 2012). Participants could, therefore, have interpreted the questions with different intonations. Further investigations will also be directed towards this level of analysis. Concluding, our results show that, while it is possible for people to identify the problem when a PPQ is presented, it is easier for the subjects to find the solution when the expected form, HNPQ, is presented. The adoption of HNPQ together with its details coming from conflict detection represents the argumentative capabilities needed in this type of dialogue. Therefore, conflict detection can constitute a first building block to create a formal theory of argumentation-based dialogue centred on conflict patterns.

Formally, a PCG is inconsistent if a post-condition of a certain action a_j introduces a proposition p_j such that p_j is in conflict with a subsequent pre-condition for an action a_i. This can be formalised as:

$$inconsistent(PCG) \implies \exists a_i \in A, \exists p_j \in P | pre(a_i, \neg p_j) \tag{1}$$

Given that the set of propositions P is generated during the dialogue, this corresponds to the PCG. Therefore, when a pre-condition $pre(a_j, \neg p)$ is not verified because of a post-condition $post(a_i, p)$, it is possible to report the *cause* of the inconsistency, beyond the mere existence of the problem, by reporting a_i.

In this case, consistently with Domaneschi et al. (2017), a HNPQ is generated. From the user's point of view, this model represents our proposal that a human interacting with a dialogue system, when presented a Clarification Request in the form of a HNPQ, is implicitly led to look for specific conflict patterns. This implies a revision of previous beliefs, incrementally accepted in the PCG, to identify conflicts. These would be characterised by the post-conditions of an action a_i, in the sequence, causing the violation of the pre-conditions posed by the last requested action a_n, by establishing the truthfulness of a given proposition p that should not be verified, in order for a_n to be accepted.

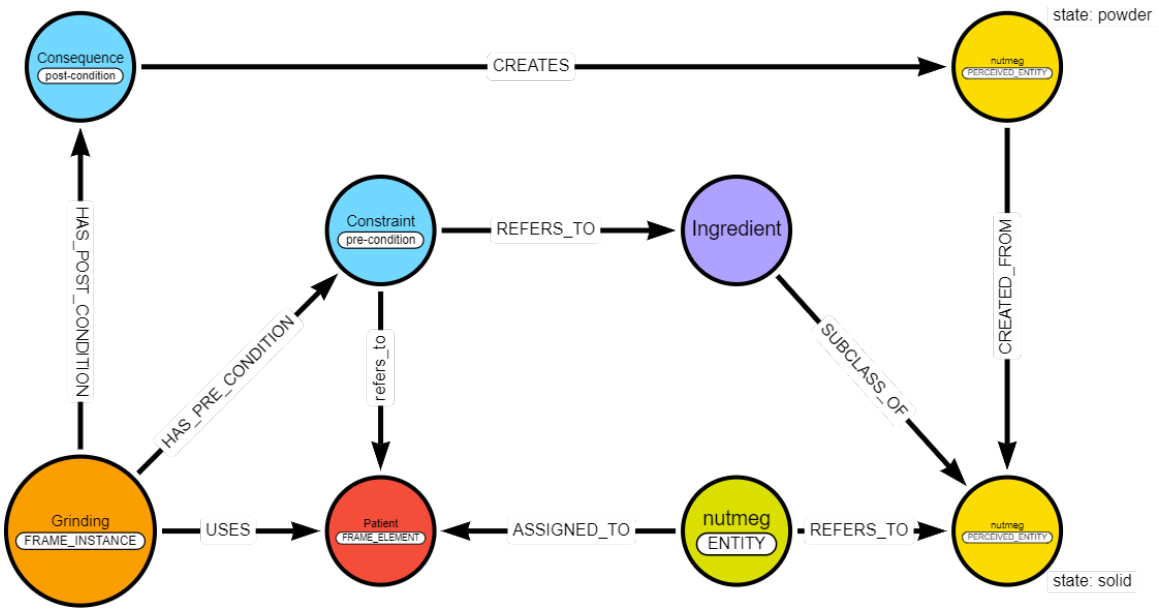


Figure 10: An example of the graph-based interpretation for the instruction *Grind the nutmeg*. After verifying that the pre-condition of the state of the referent is *solid*, the consequence of the action is that the *powder* state is acquired. Note that in this way the sequence of transformation is preserved to support Clarification Requests when necessary.

Given the previous formal representations of a consistent/¬ consistent PCG, the negative polar question implies a specific error pattern as follows:

$$HNPQ(a_j) \implies pre(a_i, \neg p) \wedge post(a_j, p) \tag{2}$$

This can be represented in the form of a graph, as suggested in Di Maro et al. (2021b), connecting actions and entities involved in the actions with their specific properties, as pre- and post-conditions. Specifically, Di Maro et al. (2021b) made use of Neo4j (Webber, 2012). Neo4j is an open source graph database manager that has been developed over the last 16 years and applied to a high number of tasks related to data representation (Dietze et al., 2016), exploration (Drakopoulos et al., 2015) visualisation (Jiménez et al., 2016) and dialogue management (Di Bratto et al., 2024). In Neo4j, nodes and relationships may be assigned *labels* that describe the type of object they are associated with. Neo4j is characterised by high scalability, ease of use and its proprietary query language: Cypher. Cypher is designed to be a *declarative* language that highlights patterns' structure using an SQL-inspired *ASCII-art syntax*.

In Figure 10, an example showing how the domain of the interaction is represented using graph-based formalism is presented. In such a graph, using Cypher, the conflict can be detected as follows:

```
MATCH (a1:ACTION)-[:REFERS_TO]->(e:ENTITY)-[:ASSIGNED_TO]->(fe:
    FRAME_ELEMENT {name: 'Patient'}),
(e)-[:REFERS_TO]->(pe1:PERCEIVED_ENTITY)
WHERE NOT (a1)-[:IS_FOLLOWED_BY]->() AND 'POWDER' IN labels(pe1)
```

Listing 1: Cypher query checking the pre-conditions of the Grinding frame for which a perceived entity cannot have the POWDER label in order to make the action possible

This query lets the system identify a possible inconsistency, as the action refers to an entity, in turn, referring to a specific perceived entity in the interaction. For the action of Grinding, reported in the query, the perceived entity must not have the label POWDER, according to the pre-conditions of the CCG, to not cause the inconsistency to occur. In case of inconsistency, with the use of another query, the action that caused the perceived entity to acquire the label POWDER, if present as a post-condition, is returned. In this way, the Clarification Request can be structured with the retrieved information. For more details, the reader is referred to Di Maro et al. (2021b).

The practical consequence of such a model and its future extensions, considering other types of conflict, is a more efficient generation of error messages using the appropriate form of polar questions depending on the reported situation.

5.1 Threats to validity

The limitations of the present study in the context of dialogue and argumentation can be multi-faceted. In this work, we limited the experimental context to a very specific situation to evaluate the impact of the using specific syntactic forms to signal the conflict of interest. Specifically, we intended to measure the practical advantage of using patterns observed in linguistic studies to improve the quality of human-machine interaction. Nevertheless, we are aware that dialogue and argumentation deal with many other important aspects which are not addressed, such as social dynamics, contextual factors, and user initiated questions. It's important to consider these limitations to understand the relevance of the study and the possibility to generalise, as well as its implications for different types of dialogue, tasks, and objectives. One significant limitation of the study is its applicability to a specific type of collaborative argumentation-based dialogues. Nevertheless, as also reported by other studies (Domaneschi et al., 2017; Di Maro et al., 2021c), the use of HNPQ in the described pragmatic scenario should be considered appropriate regardless of the domain and dialogue type. Most dialogue systems, including AI-driven ones, can have imperfect processing and interpretation capabilities, which can also lead to inconsistencies. These problems were addressed in a previous study that classified Clarification Requests based on the type of problem, considering Contact, Perception, Understanding, and Intention as different communication levels (Di Maro, 2021b). Common Ground Inconsistencies are a subset of problems at the Understanding level. The exploration of other types of inconsistencies or communicative problems is left to future studies. Moreover, since participants were asked to read the error message, one further limitation may lie in the absence of prosodic features which may have disambiguate the interpretation of PPQs.

6. Conclusions

Subtle changes in human communication, also depending on the contextual situation, may lead to different interpretation of the communicative act. Recent pragmatic studies have highlighted how the interaction between knowledge that was previously added to the Common Ground and different kinds of contextual evidence lead to different forms of Clarification Requests. For the specific case of positive knowledge negated by incoming contextual evidence, the use of high negative polar questions appears to be perceived as more appropriate by human evaluators. In this work, we have shown that this is not simply a matter of preference or naturalness: adopting the most expressive syntactic form to signal specific kinds of conflict between bias and evidence indeed has a potential impact on interaction quality. More specifically, we investigated the following research questions:

- Main question: Does linguistic feedback in the form of different polar questions influence the capability of people to identify the cause of reported common ground inconsistencies?
- Secondary question: Do different polar question forms influence the speed with which people take decisions when searching for the cause of reported Common Ground inconsistencies?

For the case of error signalling in the $p/\neg p$ case, we have shown that the use of HNPQs significantly improves the capability of human subjects to understand where the cause of the conflicts considered in this work is, with respect to the baseline. This is coherent with the expectations coming from the reference pragmatic study and it has relevant implications for the design of dialogue systems concerning the observability feature, thus providing a positive answer for our main research question. The use of HNPQs in bias/evidence conflicts is not just a matter of appropriateness but it influences the capability of the subjects to complete the assigned task. Concerning the secondary research question, about the efficiency with which a decision is reached by the human participants, results do not show a statistically significant positive effect, in our data. This may suggest that, once the kind of problem is understood by the participant, the time needed to identify the inconsistent action does not vary significantly between HNPQ and PPQ. The main issue may lie, therefore, in the participants being able to actually find the inconsistent action, at all. However, this claim needs further investigation as it is solely based on the analysis of speech response times. Given the time needed, for example, for response planning, it is possible that appreciable differences in reaction times may be observable on other signals, like for example eye tracking and pupil dilation. This is left for future work.

The theoretical formalisation presented in Section 5 adds a further perspective on possible future applications of our findings. We have shown that a formal representation of conflicts can be related to the superficial form of Clarification Requests, which are a characteristic element of argumentation-based dialogues. We have also shown that human subjects react positively when presented with a Clarification Request form coherent with type of conflict that was considered in the experiment. This is consistent with the theory presented in Domaneschi et al. (2017) and Di Maro et al. (2021c) which also describe a number of other types of conflicts. This suggests that a larger number of cases in which argumentation-based capabilities are necessary can be described through the use of conflict patterns, which would become a foundational element of an encompassing theory for argumentation-based dialogue.

Moreover, to overcome the reported limitations (Section 5.1), future plans foresee the implementation of a dialogue system based on these observations to confirm the observed effects in human-machine dialogues. Also, the formalisation effort for argumentation-based dialogue will be extended to cover all the cases mentioned literature.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26, 1992.

Jens Allwood, David Traum, and Kristiina Jokinen. Cooperation, dialogue and ethics. *International Journal of Human-Computer Studies*, 53(6):871–914, 2000.

Egbert Ammicht, J Fosler-Lussier, and Alexandros Potamianos. System and method for representing and resolving ambiguity in spoken dialogue systems, 2003. US Patent App. 10/170,510.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The HCRC map task corpus. *Language and speech*, 34(4):351–366, 1991.

Nicholas Asher and Brian Reese. Intonation and discourse: Biased questions. *Interdisciplinary studies on information structure*, 8:1–38, 2007.

Rachel Baker and Valerie Hazan. Diapixuk: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods*, 43(3):761–770, 2011.

Robbert-Jan Beun and Rogier M van Eijk. Conceptual discrepancies and feedback in human-computer interaction. In *Proceedings of the conference on Dutch directions in HCI*, page 13, 2004.

Dan Bohus. Error awareness and recovery in task-oriented spoken dialogue systems. *Ph.D. Thesis, Computer Science Department, Carnegie Mellon University*, 2007.

Caroline Bousquet-Vernhettes, Régis Privat, and Nadine Vigouroux. Error handling in spoken dialogue systems: toward corrective dialogue. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

Harry Bunt and William Black. *Abduction, belief and context in dialogue: studies in computational pragmatics*, volume 1. John Benjamins Publishing, 2000.

Daniel Buring and Christine Gunlogson. Aren't positive and negative polar questions the same? *Working paper*, 2000. URL `http://hdl.handle.net/1802/1432`.

Hendrik Buschmeier and Stefan Kopp. Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 1213–1221, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems. URL `http://dl.acm.org/citation.cfm?id=3237383.3237880`.

Mary Sandra Carberry. *Pragmatic Modeling in Information System Interfaces (Goals, Dialogue, Plans, Ill-formedness)*. PhD thesis, Newark, DE, USA, 1985.

Eve V. Clark. Common ground. In *The Handbook of Language Emergence*, page 328–353. Wiley, Chichester, UK, 2015. doi: 10.1002/9781118346136.ch15.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D., editors, *Perspectives on Socially Shared Cognition*, pages 13–1991. American Psychological Association, 1991.

Amanda Cercas Curry, Helen Hastie, and Verena Rieser. A review of evaluation techniques for social dialogue systems. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, pages 25–26, 2017.

Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. An evaluation on large language model outputs: Discourse and memorization. *arXiv preprint arXiv:2304.08637*, 2023.

Martina Di Bratto, Antonio Origlia, Maria Di Maro, and Sabrina Mennella. Linguistics-based dialogue simulations to evaluate argumentative conversational recommender systems. *User Modelling And User Adapted Interaction, Special Issue on Conversational Recommender Systems: Theory, Models, Evaluations, and Trends*, 2024.

Maria Di Maro. Computational grounding: An overview of common ground applications in conversational agents. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):133–156, 2021a.

Maria Di Maro. "Shouldn't I use a polar question?" Proper Question Forms Disentangling Inconsistencies in Dialogue Systems. *Ph.D. Thesis, Università degli Studi di Napoli Federico II*, 2021b.

Maria Di Maro, Mohamed Diaoulé Diallo, and Francesco Cutugno. Information-processing machines and the access-conscious recognition of common ground inconsistencies: a proposal. In *PSYCHOBIT*, 2020.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. Conflict search graph for common ground consistency checks in dialogue systems. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*, 2021a.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. Cutting melted butter? common ground inconsistencies management in dialogue systems using graph databases. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):157–190, 2021b.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. Polarexpress: Polar question forms expressing bias-evidence conflicts in italian. *International Journal of Linguistics*, pages 14–35, 2021c.

Felix Dietze, Johannes Karoff, André Calero Valdez, Martina Ziefle, Christoph Greven, and Ulrik Schroeder. An open-source object-graph-mapping framework for neo4j and scala: Renesca. In *International Conference on Availability, Reliability, and Security*, pages 204–218. Springer, 2016.

Alan Dix, Alan John Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. *Human-computer interaction*. Pearson Education, 2003.

Filippo Domaneschi, Maribel Romero, and Bettina Braun. Bias in polar questions: Evidence from English and German production experiments. *Glossa: A Journal of General Linguistics*, 2(1), 2017. doi: 10.5334/gjgl.27.

Georgios Drakopoulos, Andreas Kanavos, Christos Makris, and Vasileios Megalooikonomou. On converting community detection algorithms for fuzzy graphs in neo4j. In *Proceedings of the 5th International Workshop on Combinations of Intelligent Methods and Applications, CIMA*, 2015.

Raquel Fernández, Andrea Corradini, David Schlangen, and Manfred Stede. Towards reducing and managing uncertainty in spoken dialogue systems. In *Proceedings of the 7th International Workshop on Computational Semantics (IWCS07)*, 2007.

Kaoru Hayano. 19 question design in conversation. *The handbook of conversation analysis*, page 395, 2013.

John Heritage. The limits of questioning: Negative interrogatives and hostile question content. *Journal of Pragmatics*, 34(10-11):1427–1446, 2002.

Julian Hough and David Schlangen. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 274–282. IEEE, 2017.

Julian Hough, Sina Zarrieß, and David Schlangen. Grounding Imperatives to Actions is Not Enough: A Challenge for Grounded NLU for Robots from Human-Human data. In *GLU 2017 International Workshop on Grounding Language Understanding*, pages 88–91, 08 2017. doi: 10.21437/GLU.2017-18.

Yan Huang. *The Oxford Handbook of Pragmatics*. Oxford University Press, 2017.

Pablo Jiménez, Javier Villalba Diez, and Joaquin Ordieres-Mere. Hoshin kanri visualization with neo4j. empowering leaders to operationalize lean structural networks. *Procedia CIRP*, 55:284–289, 2016.

Irene Koshik. A conversation analytic study of yes/no questions which convey reversed polarity assertions. *Journal of Pragmatics*, 34(12):1851–1877, 2002.

Irene Koshik. *Beyond rhetorical questions: Assertive questions in everyday interaction*, volume 16. John Benjamins Publishing, 2005.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

D Robert Ladd. A first look at the semantics and pragmatics of negative questions and tag questions. In *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill*, number 17, pages 164–171, 1981.

Fabrizio Macagno and Sarah Bigi. Analyzing the pragmatic structure of dialogues. *Discourse Studies*, 19(2):148–168, 2017.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

Antonio Origlia, Francesco Cutugno, Antonio Rodà, Piero Cosi, and Claudio Zmarich. Fantasia: a framework for advanced natural tools and applications in social, interactive approaches. *Multimedia Tools and Applications*, 78:13613–13648, 2019.

Antonio Origlia, Martina Di Bratto, Maria Di Maro, and Sabrina Mennella. Developing embodied conversational agents in the unreal engine: The fantasia plugin. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6950–6951, 2022.

Volha Petukhova, Harry Bunt, Andrei Malchanau, and Ramkumar Aruchamy. Experimenting with grounding strategies in dialogue. *SEMDIAL 2015 goDIAL*, page 198, 2015.

Henry Prakken. *Historical overview of formal argumentation*, volume 1. College Publications, 2018.

Matthew Purver. Clarie: The clarification engine. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 77–84. Citeseer, 2004.

Matthew Purver. Clarie: Handling clarification requests in a dialogue system. *Research on Language and Computation*, 4(2-3):259–288, 2006.

Antonio Roque. *Dialogue management in spoken dialogue systems with degrees of grounding*. University of Southern California, 2009.

Antonio Roque and David Traum. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, 2008.

Michelina Savino. The intonation of polar questions in Italian: Where is the rise? *Journal of the International Phonetic Association*, 42(1):23–48, 2012.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Jack Sidnell and Tanya Stivers. *The Handbook of Conversation Analysis*, volume 121. John Wiley & Sons, 2012.

Gabriel Skantze. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341, 2005.

Jakob Steensig and Paul Drew. *Questioning and Affiliation/Disaffiliation in Interaction: Special Issue of Discourse Processes*. SAGE Publications, 2008.

Tanya Stivers and Nick J Enfield. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626, 2010.

David R Traum. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*, pages 124–131, 1999.

Michaela M Wagner-Menghin. Binomial test. *Wiley StatsRef: Statistics Reference Online*, 2014.

Douglas Walton and Erik CW Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995.

Dawn R Weatherford, Mitchell A Meltzer, Curt A Carlson, and James C Bartlett. Never forget a face: Verbalization facilitates recollection as evidenced by flexible responding to contrasting recognition memory tests. *Memory & Cognition*, 49(2):323–339, 2021.

Jim Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218, 2012.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, 2006.

Yiqian Zou. An experimental evaluation of grounding strategies for conversational agents. *Master Thesis, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg*, 2020.