An empirical evaluation of argumentation in explaining inconsistency tolerant query answering

Abdelraouf Hecham[1], Abdallah Arioua[2], Gem Stapleton[3], and Madalina Croitoru[1]

[1] University of Montpellier
[2] University of Claude Bernard Lyon 1
[3] University of Brighton

Abstract. In this paper we answer empirically the following research question: "Are dialectical explanation methods more effective than one-shot explanation methods for Intersection of Closed Repairs inconsistency tolerant semantics in existential rules knowledge bases?" We ran two experiments with 84 and respectively 38 participants and showed that under certain conditions dialectical approaches are significantly more effective than one-shot explanations.

1 Introduction

We place ourselves in a logical based setting where we consider inconsistent knowledge bases expressed using existential rules [9]. Existential rules have been recently extensively studied in the knowledge representation and reasoning community due to their expressiveness: existential rules generalise many Semantic Web commonly employed languages [19] [9] [13][4]. Reasoning in presence of inconsistency has been another challenge to be addressed due to the uselessness of existing reasoners when inconsistency arises. To address reasoning with inconsistency for existential rules, numerous inconsistency tolerant semantics have been proposed [7, 5, 14, 16]. The intuition behind most of these semantics is to consider maximal consistent subsets of knowledge bases as a support for reasoning. Unfortunately, explaining such reasoning techniques to an user is challenging - with only a few approaches currently practically available. Amongst the approaches for explanation we distinguish two types of methods: one-shot explanation methods and interactive explanation methods. In the one shot explanation methods we include both provenance based methods [6] and argument based notions [3] as the two explanation methods are equivalent for certain semantics. This work will focus on one of such semantics, the ICR (Intersection of Closed Repairs) semantics. Interactive explanation methods rely on dialectical approaches [2].

In this paper we are asking the following research question: "*Are dialectical approaches more effective than one shot explanations for ICR inconsistent tolerant semantics?*". We ran two experiments where the participants are exposed with seven inconsistent knowledge bases. To avoid unwanted effects of a priori knowledge used by the participants, the knowledge bases were completely fictitious. For each knowledge base and for given query, each participant was presented in a random manner the query's ICR explanation (one-shot or dialectical). Next, the participants were invited to answer a new query on the knowledge base. We measured the effectiveness of an explanation

based on the user's (1) answer correctness and (2) answer time as the goal of one-shot or dialogue explanations is to help users understand ICR semantics. Our studies show that the dialectical approaches are significantly more effective than one-shot explanation only as long as the *intent of the explainer is clearly conveyed*. In our case the expressing the intent of the explainer was achieved by using the word "possibly" during the dialectical phase.

The significance of our work is two fold. First, to the best of our knowledge, we conduct the first empirical study in the literature that address the problem of explanation effectiveness for ICR semantics in existential rules knowledge bases. Second, and more broadly, we align ourselves to the recent line of work around investigating the added value of argumentation via explanation [18].

After giving the main theoretical background notions in Section 2 (existential rules, inconsistent tolerant semantics, explanation) we detail and discuss the results of the experimentation in Section 3. The raw data of the experimentation is publicly available for reproducibility reasons[4]. We conclude the paper with Section 5.

2 Background notions

After describing the logical language used in this paper, existential rules, we lay out the problem of inconsistency tolerant query answering. We present two methods for addressing this problem: either using the so called inconsistency tolerant semantics [15] or using logic based argumentation [10]. We give the basic notions of logical argumentation and its instantiation using existential rules. We then present the state of the art with respect to explanation of query answering in this setting using two methods: one shot explanations and dialectical explanations.

2.1 Existential rules

The existential rules language has attracted much interest recently in the Semantic Web and Knowledge Representation community for its suitability of representing knowledge in a distributed context (such as Ontology Based Data Access (OBDA) applications where the domain knowledge is represented by an ontology facilitating query answering over existing data) [15] [19]. The language [9] extends plain Datalog with *existential variables* in the rule conclusion. A subset of this language, also known as $Datalog^{\pm}$, refers to identified decidable existential rule fragments [13][4]. The existential rule language is composed of formulas built with the usual quantifiers $(\exists, \forall)$ and *only* two connectors, implication $(\rightarrow)$ and conjunction $(\wedge)$. It contains the following elements:

- A *fact* is an existentially closed atom (of the form $p(t_1, \ldots, t_k)$ where p is a predicate of arity k and $t_i, i \in [1, \ldots, k]$ are terms, i.e. variables or constants [5]).
- An existential *rule* is of the form $\forall \overrightarrow{X}, \overrightarrow{Y}\ H[\overrightarrow{X}, \overrightarrow{Y}] \rightarrow \exists \overrightarrow{Z} C[\overrightarrow{Z}, \overrightarrow{X}]$ where H and C are facts or conjunctions of facts and $\overrightarrow{X}, \overrightarrow{Y}, \overrightarrow{Z}$ their respective sets of variables.

[4] `https://github.com/anonIJCAI/ExplanationExperiment`

[5] The unique name assumption is made for constants.

– A *negative constraint* is a particular kind of rule where H is a conjunction of atoms and C is $\bot$ (*absurdum*). It implements *weak negation*.
– A knowledge base $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ is composed of a set of facts $\mathcal{F}$, a set of rules $\mathcal{R}$ and a set of negative constraints $\mathcal{N}$. We denote by $\texttt{Cl}_{\mathcal{R}}(\mathcal{F})$ the *closure* of $\mathcal{F}$ by $\mathcal{R}$ (computed by all possible applications of the rules in $\mathcal{R}$ over $\mathcal{F}$ until a fixed point is reached). $\texttt{Cl}_{\mathcal{R}}(\mathcal{F})$ is said to be $\mathcal{R}$*-consistent* if no negative constraint hypothesis can be deduced from it. Otherwise $\texttt{Cl}_{\mathcal{R}}(\mathcal{F})$ is $\mathcal{R}$*-inconsistent*. A knowledge base $(\mathcal{F}, \mathcal{R}, \mathcal{N})$ is said to be *inconsistent* iff $\mathcal{F}$ is $\mathcal{R}$-inconsistent. When considering consistent facts, entailment implicitly considers rules application (i.e. $\mathcal{F} \models Q$ is equivalent to $\texttt{Cl}_{\mathcal{R}}(\mathcal{F}) \models q$).

Example 1. Consider the following knowledge base $\mathcal{K}$: Victor is a rabbit. Victor has a delta badge. Victor is short sighted. All rabbits have long ears. Everyone with a delta badge has access to the quarantine ward. Everyone that has access to the quarantine ward wears protective glasses. If one is short-sighted then it must wear eye glasses. One cannot wear protective glasses and eye glasses in the same time. Formally, $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$, where:

$$\begin{aligned}
\mathcal{F} =& \{rabbit(Victor), hasDbadge(Victor),\\
& shortsighted(Victor)\}.\\
\mathcal{R} =& \{\forall x\, rabbit(x) \rightarrow longEars(x),\\
& \forall x hasDbadge(x) \rightarrow hasAccessQuarant(x),\\
& \forall x\, hasAccessQuarant(x) \rightarrow wearPGlasses(x),\\
& \forall x\, shortsighted(x) \rightarrow wearEGlasses(x)\}.\\
\mathcal{N} =& \{\forall x\, wearPGlasses(x) \wedge wearEGlasses(x) \rightarrow \bot)\}.
\end{aligned}$$

$$\begin{aligned}
\texttt{Cl}_{\mathcal{R}}(\mathcal{F}) =& \mathcal{F} \cup \{longEars(Victor), hasAccessQuarant(Victor),\\
& wearPGlasses(Victor), wearEGlasses(Victor)\}.
\end{aligned}$$

Since $\texttt{Cl}_{\mathcal{R}}(\mathcal{F})$ is $\mathcal{R}$-inconsistent (it entails the hypothesis of the negative constraint) then $\mathcal{K}$ is inconsistent. Classical entailment will allow to deduce anything out of an inconsistent knowledge base.

In practical OBDA systems involving large amounts of data and multiple data sources, data inconsistency commonly occurs [14]. In this setting, classical reasoners cannot be employed. Luckily, inconsistency tolerant semantics address this problem.

2.2 Inconsistent tolerant semantics

Inconsistency-tolerant semantics [7, 5, 14, 16] have been proposed in the literature to address the problem of reasoning in the presence of inconsistency in OBDA. These semantics rely on the notion of data repairs. A repair is a maximal (with respect to set inclusion) consistent subset of $\mathcal{F}$. The set of all repairs of a knowledge base is denoted $\mathcal{R}epair(\mathcal{K})$. Once the repairs are computed, different semantics can be used for query answering over the knowledge base. In this paper we focus on (**I**ntersection of **C**losed **R**epairs semantics) [5]. The semantics considers the repairs enriched with extra information obtained by rule application (i.e. closed) and then intersects them. The obtained

(consistent) set is then used for classical entailment. Formally, the query Q is ICR-entailed from $\mathcal{K}$, written $\mathcal{K} \models_{ICR} Q$, iff:

$$\bigcap_{\mathcal{A} \in \mathcal{R}epair(\mathcal{K})} \mathtt{Cl}_{\mathcal{R}}(\mathcal{A}) \models Q$$

Example 2 (Cont'd Example 1). The repairs of $\mathcal{K}$ are $\mathcal{R}epair(\mathcal{K}) = \{\mathcal{A}_1, \mathcal{A}_2\}$. $\mathcal{A}_1$ states that Victor is a rabbit and it has a delta badge. $\mathcal{A}_2$ states that Victor is a rabbit and it is short sighted. The closure of A_1 by $\mathcal{R}$ adds the information that Victor has access to the quarantine ward and Victor wears protective glasses. Similarly, the closure of A_2 by $\mathcal{R}$ adds the information that Victor must wear eye glasses. Formally:

$$\begin{aligned}
\mathcal{A}_1 =& \{rabbit(Victor), hasDbadge(Victor)\},\\
\mathtt{Cl}_{\mathcal{R}}(A_1) =& \{rabbit(Victor), hasDbadge(Victor),\\
& longEars(Victor), hasAccessQuarant(Victor),\\
& wearPGlasses(Victor)\},\\
\mathcal{A}_2 =& \{rabbit(Victor), shortsighted(Victor)\},\\
\mathtt{Cl}_{\mathcal{R}}(A_2) =& \{rabbit(Victor), shortsighted(Victor),\\
& longEars(Victor), wearEGlasses(Victor)\}.
\end{aligned}$$

It follows that $\mathtt{Cl}_{\mathcal{R}}(A_1) \cap \mathtt{Cl}_{\mathcal{R}}(A_2) = \{rabbit(Victor), longEars(Victor)\}$.

The query $Q_1 : longEars(Victor)$ is ICR-entailed and the query $Q_2 : wearEGlasses(Victor)$ is not ICR entailed.

Argumentation for existential rules. A semantically equivalent reasoning method with ICR entailment existential rules is defined in [11]. The authors instantiate an argumentation framework over the inconsistent knowledge base and prove sceptically preferred semantics over this argumentation framework to be equivalent to ICR entailment over the inconsistent knowledge base.

More precisely, given a knowledge base $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$, the *corresponding argumentation framework* $\mathcal{AF}_{\mathcal{K}}$ is a pair $(\mathtt{Arg}, \mathtt{Att})$ where $\mathtt{Arg}$ is the set of arguments that can be constructed from $\mathcal{F}$ and $\mathtt{Att}$ is the *attack* relation defined over $\mathtt{Arg} \times \mathtt{Arg}$. An argument $a = (H, C)$ is a pair with H the minimal support of the argument (also denoted $Supp(a)$) and C its conclusion (denoted $Conc(a)$) satisfying $H \models C$ [6]. An argument a attacks an argument b iff there exists a fact $f \in Supp(b)$ such that the set $\{Conc(a), f\}$ is $\mathcal{R}$-inconsistent. We say that $\mathcal{E} \subseteq \mathtt{Arg}$ is *conflict free* iff there exist no arguments $a, b \in \mathcal{E}$ such that $(a, b) \in \mathtt{Att}$ and that $\mathcal{E}$ *defends* argument a iff, for every argument $b \in \mathtt{Arg}$, if $(b, a) \in \mathtt{Att}$ then there exists $c \in \mathcal{E}$ such that $(c, b) \in \mathtt{Att}$. $\mathcal{E}$ is a *preferred extension* iff it is a maximal conflict free set defending all its arguments (please see [12] for other types of semantics). An argument is sceptically (preferred) accepted if it is in all (preferred) extensions. [11] shows the equivalence between sceptically acceptance under preferred semantics and ICR-entailment: $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N}) \models_{ICR} Q$ iff Q is sceptically preferred entailed from $\mathcal{AF}_{\mathcal{K}}$.

[6] The finiteness of the argumentation framework follows from the chase reducer employed by the entailment.

2.3 Explanation notions

The equivalence result of the previous section means one can employ argumentation inspired explanation techniques for entailment under ICR semantics. Two such notions have been investigated in the literature: *one-shot arguments* and *dialectical explanations*.

One-shot argument explanations. Inspired from provenance-based explanations in databases [8], in [6] the authors introduce the notion of *one-shot provenance explanation* explanation. Their explanation is semantically and syntactically equivalent to the explanation introduced by [3] that considers that a query explanation is a *one-shot argument* supporting the query. In the rest of the paper, we denote such explanation as *one-shot argument explanations*.

Example 3 (Cont'd Example 1). For example, a *one-shot argument explanation* for Q_1 : $longEars(Victor)$ is $(rabbit(Victor), longEars(Victor))$. A *one-shot argument explanation* for the query $Q_2 : wearEGlasses(Victor)$ is $(shortsighted(Victor), wearEGlasses(Victor))$.

Dialectical explanations. Introduced by [2], the dialectical explanation is interactive. It build upon the notion of argument in [3] and takes the form of an *explanation dialogue* [1] in which an *explainer* aims to make an *explainee* understand why Q is or is not ICR-entailed. Intuitively, for a query that is ICR-entailed there will be an argument supporting it in every repair (since, by definition, the query is in the intersection of the closed repairs). For a query that is not ICR-entailed one can eventually find a repair in which the query is not entailed. In the following we will briefly give the basic formal notions underlying the dialectical explanation for ICR entailment.

An explanation dialogue $\mathcal{D}_n = (m_0, m_1, \ldots, m_n)$ over $\mathcal{AF}_\mathcal{K}$ is a sequence of moves exchanged between an explainer (called EXPR) and an explainee (called EXPE) about a query Q (i.e. the subject of the dialogue, denoted $\texttt{Subject}(\mathcal{D}_n)$). The moves are started by the explainee $\texttt{Part}(m_0) = \texttt{EXPE}$ (where $\texttt{Part}(m_i)$ denotes the participant who plays the move m_i). The participants take turns advancing one move at a time: for all $m_i \in \mathcal{D}_n$, $i > 0$, $\texttt{Part}(m_i) = \texttt{EXPR}$ iff i is odd otherwise $\texttt{Part}(m_i) = \texttt{EXPE}$. m_n is called the most recent move in $\mathcal{D}_n$. Each move m_i has a locution $\texttt{loc}(m_i) \in \{$EXPLAIN, ATTEMPT, CLARIFY, CLARIFICATION, DEEPEN, DEEPENING, POSITIVE, NEGATIVE$\}$ and a content a that is an argument in $\mathcal{AF}_\mathcal{K}$ or a well-formed syntactical entity.

The dialogue is asymmetric in the sense that the participants do not use the same locutions. All in all, the explainee is allowed to use the locutions $\mathcal{C}_{\texttt{EXPE}} = \{$EXPLAIN, CLARIFY, DEEPEN, POSITIVE, NEGATIVE$\}$, which are respectively an explanation request, a clarification request, a deepening request and either a declaration of understanding or declaration of inability of understanding. The explainer is allowed to use the locutions $\mathcal{C}_{\texttt{EXPR}} = \{$ATTEMPT, CLARIFICATION, DEEPENING$\}$ which are the corresponding answers respectively.

In what follows we describe each move and we give its semantics with respect to the underlying argumentation framework $\mathcal{AF}_\mathcal{K} = (\texttt{Arg}, \texttt{Att})$. We first introduce the

following concepts and then present the semantics. A **clarification** of an argument a is a sequence of rules and facts that starts form $Supp(a)$ and ends by $Conc(a)$. It represents the line of reasoning from the support to the conclusion. A **deepening** between two arguments a and b such that b attacks a intends to explain the conflict between a and b by showing the set of violated constraints over $\{Conc(b), Supp(a)\}$ [7].

- EXPLAIN(a). The explainee asks for an explanation of a query Q. $a \in \texttt{Arg}$ is the argument such that $Supp(a) = Conc(a) = Q$. The argument a is referred to *the subject argument*.
- ATTEMPT(a). The explainer advances the argument a that **explains** Q. That is, an argument whose conclusion entails the query, i.e. $Conc(a) \models Q$. The next possible replying moves are one of the followings: CLARIFY(a), DEEPEN(a), POSITIVE(a) or NEGATIVE(a).
- CLARIFY(a). It is a request made by the explainee for a clarification of a.
- CLARIFICATION(a). A clarification of a advanced by the explainer. If the explainee has not asked before for a deepening then it is allowed to advance DEEPEN(a).
- DEEPEN(a). It is a request made by the explainee for a deepening of the conflict between a and the subject argument.
- DEEPENING(a). A deepening of a advanced by the explainer. If the explainee has not asked before for a clarification then it is allowed to advance CLARIFY(a).
- POSITIVE(a). The explainee confirms his understanding of the subject of the dialogue where a is the subject argument. No move can be played afterwards.
- NEGATIVE(a). The explainee declares his inability to understand the explanation. The explainer can advance another ATTEMPT(a') such that a' is another explanation of the subject of the dialogue.

Let us take the example of a real dialogue that occurred during the experimentation based on the knowledge base presented in Example 1.

Example 4 (Cont'd Example 1). Consider Example 2 and the following query:

Query: "Is Victor wearing protective glasses?"

The user wants to know why it is not ICR-entailed. There are two possibilities, the one-shot explanation or the dialogue (dialectical explanation). They are presented hereafter:

One-shot explanation: "Victor is short-sighted."

Dialectical explanation takes form of the dialogue depicted by the top right table in the next column.

[7] Please note that in what follows we may use the word explanation and argument exchangeably to mean an argument because an argument is an explanation in our case.

Part	Text	Formal
EXPE	Explain why the answer is negative?	EXPLAIN(a)
EXPR	Because it is possible that he is short-sighted.	ATTEMPT(b)
EXPE	Clarify the explanation, I don't see how this could be a problem.	CLARIFY(b)
EXPR	If Victor is short-sighted then he should wear eyeglasses. Consequently, he cannot wear protective glasses.	CLARIFICATION(b)
EXPE	Deepen please, how is that a problem?	DEEPEN(b)
EXPR	Because a person cannot wear eyeglasses and protective glasses in the same time.	DEEPENING(b)
EXPE	I understand.	POSITIVE(a)

3 Experiment Method

The goal of one-shot or dialogue explanations is to help users understand query entailment in inconsistent knowledge bases. However, each explanation might achieve this goal to a different degree. In this section, we describe the experiment protocol we used to compare the explanations and to determine -if possible- *which one is most effective*.

3.1 Experiment Design

As understanding is a vague concept and sometimes subjective, we consider that *one explanation is more effective than another* if it is *significantly* more likely for a user to give the correct answer for a query after being exposed to the more effective explanation. If the difference is not significant, we consider more effective the one where *correct answers are provided significantly more quickly*. Otherwise, we consider both explanation to have the same efficacy.

Our experiment protocol to test the effect of an explanation is to present a user with different descriptions of situations (knowledge bases) containing inconsistencies. To avoid unwanted effects of a priori knowledge used by the participants, these situations are completely fictitious. For each situation, the user is presented with a textual description of the inconsistent knowledge base that is as faithful as possible to the underlying logical formalism. Then we provide a query and the answer for that query, along with an explanation (either one-shot or dialogue). We assume that if the user is able to correctly answer another query *on the same inconsistent knowledge base*, then the *explanation had a positive effect towards understanding query answering under ICR-semantics*. An example of a situation is described in the following Example 5.

Example 5. The participant is shown the following inconsistent situation:

> "Jude is a snake. Jude is a puma. All snakes wear sunglasses. All pumas wear running shoes. One cannot be a snake and a puma at the same time. There is only one Jude in the forest."

Then, he is presented with a query and its answer:

Query: "Does Jude wear sunglasses?"
Answer: "No."

Alongside the answer, the participant is provided with an explanation (**either** one-shot or dialogue).

One-shot explanation: "No, because Jude is a puma"

Dialogue explanation (with Alice as the explainee and Bob as the explainer):

`Alice` : Does Jude wear sunglasses?
`Bob` : No, she does not.
`Alice` : Why not? Jude is a snake, therefore she wears sunglasses.
`Bob` : I don't agree, Jude is a puma.
`Alice` : I don't see how this could be a problem.
`Bob` : Jude cannot be a puma and a snake in the same time.
`Alice` : I understand.

Then, the participant is asked to answer the query: "Does Jude wear running shoes?"

We ran a first experiment (referred to as Experiment 1) with a between-group design, to which we recruited 84 participants split into two groups depending on the type of explanation with which they were presented. The first group was presented with one-shot explanations, and the second group received the dialogue explanations. All participants were first year university students in computer science; they were not familiar with logic and argumentation and none were members of the authors' research group.

Some of the participants who received the dialogue explanation reported that they had difficulties agreeing with the explanation provided by Bob (the explainer) as he seemed to assert claims they felt were not exactly true. For example, in the dialogue explanation of Example 5, Bob argues that 'Jude is a puma', but the participants considered that it was not necessarily true as 'Jude is a snake'.

In reality, the aim of the explainer (Bob) is not to assert that Jude is *definitely* a puma, but to state that there exists some information (an argument) supporting the fact that Jude is a puma, i.e. Jude is *possibly* a puma, therefore we cannot derive any conclusion from the fact it contradicts, namely, that Jade is a snake, thus, under ICR-semantics, the answer to the query "Does Jude wear sunglasses?" is 'No'.

Fearing that this misunderstanding might have had an effect on the result of the experiment, we ran a second experiment in which we changed the wording of the dialogue explanation to include the word 'possibly' and to better convey the intent of the explainer. The following example describes how the wording of the dialogue explanation of Example 5 was changed.

Example 6. The new dialogue explanation for the situation in Example 5 is as follows:

`Alice` : Does Jude wear sunglasses?
`Bob` : No, she is not.
`Alice` : Why not? Jude is a snake, therefore she wears sunglasses.
`Bob` : I don't agree. Jude is possibly a puma.
`Alice` : I don't see how this could be a problem.
`Bob` : Jude cannot be a puma and a snake at the same time. Since both are possible, it is safer to assume that Jude is not wearing sunglasses.
`Alice` : I understand.

In this second experiment (we refer to it by Experiment 2), 38 participants were recruited with the same in between-group design and the same inconsistent situations as in Experiment 1. The participants were drawn from the same pool, and none had taken part in the first study.

3.2 Experiment Execution

Both experiments 1 and 2 were performed on weekdays between 9am and 5pm. Participants took part in the study in a classroom that provided a quiet environment and which was free from interruption. A dedicated website (called the research vehicle) was used to perform the experiments and gather performance data (i.e. it recorded the answer provided to each question and the time taken to provide the answer). Participants were accompanied, during the experiment, by an experimental facilitator who ran the study. Furthermore, the participants were requested not to discuss any of the details with other people after they had taken part. The participants were informed that they could withdraw at any time and they all completed the experiment in under an hour.

Each experiment of the study had two main phases: training phase, and the main data collection phase. In training phase participants were introduced to the research vehicle and to inconsistent situations using an example. This example was not later used in the main study.

The second phase is where we collected performance data (accuracy and time). Participants were presented with 7 inconsistent knowledge bases in random order. For each knowledge base, the participant was presented with (i) a set of sentences describing the knowledge base, (ii) a query on that knowledge base and (iii) the answer for that query along with an explanation (either a one-shot explanation or a dialogue one depending on the group of the participant) as described in Example 5. When the participant is ready, he would click to display a question consisting in another query on the same previously presented knowledge base. The time taken to answer a question is determined from the instant the question was presented until the instant the participant had selected an answer. The participant is then asked to justify his answer (the justification was later used to make sure that participants answered seriously to the test query). Afterwards, the research vehicle would ask them to indicate when they were ready to proceed to the next question, thus allowing a pause between questions. There was a maximum time limit of two minutes to answer a question to ensure that the experiment did not continue indefinitely. From the 7 presented questions (queries), 4 were not ICR-entailed (the correct

answer the query is false) and 3 where ICR-entailed (the correct answer for the query is true).

4 Analysis and Results

In the following two subsections, we present the statistical analysis and results from our two experiments. The method employed to analyse the accuracy data was a Mann-Whitney test and, for the time data, an ANOVA was performed. Regarding the time performance indicator, we only analyzed the data from questions for which a correct answer was provided, consistent with previous research such as [17]. When we determined which treatment (explanation) most effectively supported task performance, we viewed accuracy as the most important indicator. That is, one treatment was taken to be more effective than another if it was significantly more likely to yield a correct answer. Otherwise, one treatment was taken to be more effective than another if correct answers were provided significantly more quickly; in any case, we present the time analysis for completeness. For each test, we used a 5% significance level to call statistically significant results.

4.1 Experiment 1

The results from experiment 1 are based on data collected from 84 people, each answering seven questions. Of the 588 responses, there were a total of 401 correct answers giving an overall accuracy rate of 68.2% and, thus, an error rate of 31.8%. The one-shot group's accuracy rate ($N = 41$) was 69.34% and, for the dialogue group ($N = 43$) was 67.11%. Subjecting the data to a Mann-Whitney test revealed no significant differences between the one-shot and dialogue treatments with respect to accuracy, with $p = 0.6523$ (adjusted for ties).

Regarding time, the mean response time for the 401 correct responses was 13.3 seconds (sd: 13.3). The mean time taken to answer questions correctly by the one-shot group was 13.4 seconds (sd: 11.2) and, for the dialogue group was 13.1 seconds (sd: 15.1). The time data were not normal so a log transformation was applied, which yielded a skewness of 0.23. It was, therefore, robust to proceed with an ANOVA on the transformed data, which yielded a p-value of 0.827. Therefore, there was no statistically significant difference between the one-shot and dialogue treatments with respect to time. In summary, taking into account both accuracy and time as performance indicators, we may suggest that there was no significant difference between and the one-shot and dialogue explanations when the word 'possibly' was omitted from the phrasing of the dialogue explanation.

4.2 Experiment 2

The results from experiment 2 are based on data collected from 38 people, each answering seven questions. Of the 266 responses, there were a total of 178 correct answers giving an overall accuracy rate of 66.9% and, thus, an error rate of 33.1%. The one-shot group's accuracy rate ($N = 21$) was 58.5% and, for the dialogue group ($N = 17$) was 77.3%. Subjecting the data to a Mann-Whitney test revealed significant differences between the two treatments, with $p = 0.0012$ (adjusted for ties). Therefore, the dialogue

treatment supported significantly better task performance, in terms of accuracy, than the one-shot treatment.

For completeness, we also include the time analysis for this experiment. The mean response time for the 178 correct responses was 13.7 seconds (sd: 19.1). The mean time taken to answer questions correctly by the one-shot group was 16.6 seconds (sd: 24.9) and, for the dialogue group was 11.0 seconds (sd: 16.7). As with experiment 1, the time data were not normal so a log transformation was applied, which yielded a skewness of 0.45. Conducting an ANOVA on the transformed data yielded a p-value of 0.248. Therefore, there was no statistically significant difference between the one-shot and dialogue treatments with respect to time.

In summary, we may suggest that *the dialogue treatment supports significantly better task performance, relative to the one-shot treatment, when the word 'possibly' was included in the phrasing*. To provide an indication of the practical effect size, we saw approximately 19 more correct answers, for every 100 questions, from the dialogue group compared to the one-shot group. Given these results, we can also suggest that the dialogue treatment better supports accuracy without there being a significant time-penalty. Therefore, *our study supports the use of the dialogue explanation* as long as the intent of the explainer is clearly conveyed (by including the word 'possibly' in the phrasing for example).

5 Discussion

In this paper we empirically studied the effectiveness of available explanation methods for query answering in presence of inconsistency and showed that under certain conditions dialectical approaches are significantly more effective than one shot ones. Please note that we showed the effect of dialogue without it being interactive. In future work we consider running an experiment to take advantage of the added value of interaction.

We conclude the paper by discussing potential threats to validity of our experiments. The threats to the validity of our results can either be internal or external. Internal validity considers the flaws related to the experimental setting and whether there is sufficient evidence to substantiate the claim. External validity considers the extent to which we can generalise the results.

With regards to internal validity, the between group design allows us to avoid the carry-over effect where one treatment might affect another if applied on the same participant. The random order in which situations are presented along with the fact that they were completely fictitious prevents the use of a priori knowledge. To minimize false positives (i.e. the participant answers randomly), we used the justification provided by the participants to make sure they took the time to read and try to understand the explanation provided. No false positives were reported.

As for the external validity, participants where representative of a wider audience as they had no previous experience with query answering in inconsistent knowledge bases, the tested queries were yes or no questions (4 of them were not ICR-entailed, while the remaining 3 were ICR-entailed).

Acknowledgements. This work was partially supported by the French ANR project ASPIQ (ANR-12-BS02-0003).

References

1. A. Arioua and M. Croitoru. Formalizing explanatory dialogues. In *Proceedings of the 9th International Conference on Scalable Uncertainty Management (SUM'15)*, pages 282–297, 2015.
2. A. Arioua and M. Croitoru. A dialectical proof theory for universal acceptance in coherent logic-based argumentation frameworks. In *Proceedings of the 22nd European Conference on Artificial Intelligence, (ECAI'16)*, 2016.
3. A. Arioua, N. Tamani, and M. Croitoru. Query answering explanation in inconsistent datalog +/- knowledge bases. In *Proceedings of the 26th International Conference Database and Expert Systems Applications (DEXA'15), Part I*, pages 203–219, 2015.
4. J.-F. Baget, M.-L. Mugnier, S. Rudolph, and M. Thomazo. Walking the complexity lines for generalized guarded existential rules. In *Proceedings of IJCAI'11*, pages 712–717, 2011.
5. M. Bienvenu. On the complexity of consistent query answering in the presence of simple ontologies. In *Proc of AAAI*, 2012.
6. M. Bienvenu, C. Bourgaux, and F. Goasdoué. Explaining inconsistency-tolerant query answering over description logic knowledge bases. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pages 900–906, 2016.
7. M. Bienvenu and R. Rosati. Tractable approximations of consistent query answering for robust ontology-based data access. In *Proceedings of IJCAI'13*, pages 775–781. AAAI Press, 2013.
8. P. Buneman and W.-C. Tan. Provenance in databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1171–1173. ACM, 2007.
9. A. Calì, G. Gottlob, and T. Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. In *Proceedings of the Twenty-Eigth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'09)*, pages 77–86, 2009.
10. M. Croitoru, R. Thomopoulos, and S. Vesic. Introducing Preference-Based Argumentation to Inconsistent Ontological Knowledge Bases. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems - 18th International Conference, Bertinoro, Italy, October 26-30, 2015, Proceedings*, pages 594–602, 2015.
11. M. Croitoru and S. Vesic. What Can Argumentation Do for Inconsistent Ontology Query Answering? In *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*, pages 15–29, 2013.
12. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.
13. G. Gottlob, T. Lukasiewicz, and A. Pieris. Datalog+/-: Questions and Answers. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*, 2014.
14. D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, and D. F. Savo. Inconsistency-tolerant semantics for description logics. In *Proceedings of the Fourth International Conference on Web Reasoning and Rule Systems*, RR'10, pages 103–117, Berlin, Heidelberg, 2010. Springer-Verlag.
15. M. Lenzerini. Data integration: A theoretical perspective. In *Proc. of PODS 2002*, 2002.
16. T. Lukasiewicz, M. V. Martinez, and G. I. Simari. Complexity of inconsistency-tolerant query answering in datalog+/-. In R. Meersman, H. Panetto, T. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. D. Leenheer, and D. Dou, editors, *Proceedings of the 12th International Conference on Ontologies, Databases, and Applications of Semantics, September 10-11, 2013*, volume 8185 of *LNCS*, pages 488–500. Springer, 2013.

17. W. Meulemans, N. Henry Riche, B. Speckmann, B. Alper, and T. Dwyer. Kelpfusion: A hybrid set visualization technique. *IEEE Transactions on Visualization and Computer Graphics*, 19(11):1846–1858, 2013.
18. S. Modgil and M. Caminada. *Argumentation in Artificial Intelligence*, chapter The added value of argumentation, pages 105–129. Springer US, 2009.
19. A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. In *Journal on data semantics X*, pages 133–173. Springer, 2008.