Axiomathes
DOI 10.1007/s10516-017-9339-1

ORIGINAL PAPER

Solving Smullyan Puzzles with Formal Systems

José Félix Costa[1,2] · **Diogo Poças**[3]

Received: 6 January 2017 / Accepted: 18 April 2017

Abstract Solving numeric, logic and language puzzles and paradoxes is common within a wide community of high school and university students, fact witnessed by the increasing number of books published by mathematicians such as Martin Gardner (popular books as old as Gardner in Aha! insight. W. H. Freeman & Co., London, 1978, Wheels, life and other mathematical amusements. W H Freeman & Co., London, 1985), Douglas Hofstadter [in one of the best popular science books on paradoxes (Hofstadter in Godel, escher, bach: an eternal golden braid, Penguin, London, 2000)], inspired by Gödel's incompleteness theorems), Patrick Hughes and George Brecht (see Hughes and Brecht in Vicious circles and infinity, an anthology of paradoxes. Penguin Books, London, 1993) and Raymond M. Smullyan (the most well known being Smullyan in Forever undecided, puzzle guide to godel. Oxford Paperbacks, Oxford 1988, To Mock a Mockingbird and other logic puzzles. Oxford Paperbacks, Oxford 2000, The lady or the tiger? And other logic puzzles. Dover Publications Inc., Mineola 2009), inter alia. Books by Smullyan (such as Smullyan 1988, 2000) are, however, much more involved, since *they introduce learning trajectories and strategies* across several subjects of mathematical logic, as difficult as combinatorial logic (see, e.g., Smullyan 2000), computability theory (see Smullyan 1988), and proof theory (see Smullyan 1988, 2009). These books provide solutions to their suggested exercises. Both statements and their solutions are written in the natural language, introducing some informal algorithms. As an

✉ José Félix Costa
fgc@math.tecnico.ulisboa.pt

Diogo Poças
pocasd@math.mcmaster.ca

[1] Department of Mathematics, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal

[2] Centro de Filosofia das Ciências da Universidade de Lisboa, Lisbon, Portugal

[3] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada

Published online: 28 April 2017

exercise in Mathematics we wonder if an easy proof system could be devised to solve the amusing equations proposed by Smullyan in his books. Moreover, university students of logic could well train themselves in constructing deductive systems to solve puzzles instead of a non-uniform treatment one by one. In this paper, addressing students, we introduce one such formal systems, a tableaux approach able to provide the solutions to the puzzles involving either propositional logic, first order logic, or aspect logic. Let the reader amuse herself or himself!

Keywords Aspect logic · Automatic reasoning · Paradoxes · Proof systems · Puzzles · Tableaux

1 Introduction

Popular puzzle books centered in natural language appear grouped in different classes. E.g., Jim Sukach (see Sukach 1995, 1997) is known for his police stories where the solution to each mystery is rooted in contradiction and Occam's razor arguments; Paul Sloane and Des MacHale in Sloane and MacHale (1992) developed training strategies for lateral thinking based on partial information and possible different meanings words can get. Raymond M. Smullyan explores the logic of natural language with the flavour of philosophical logic with self reference (see Smullyan 1988, 2000, 2008, 2009, 2010, 2011, 2012a, b, 2013).

Modeling discourse and argumentation is basic to automatic reasoning or to analyse natural language. Modeling natural language is an extraordinary useful exercise in understanding the expressive power of logic connectives, quantifiers, tense and modal aspects. Natural language is so tricky that with a few simple syntactical assertions we can build paradoxes and very difficult puzzles. Raymond M. Smullyan used this notorious fact to develop a training program in logic, distributed in a variety of books that we can map between popular science and mathematical logic textbooks, where the concepts and operating rules are digested by the readers in sequences of puzzles, interesting notes, all prefaced by short accounts on the basic ideas.

In this paper, directed to the students in the transition from high school to university, we investigate the potentialities of two of such books, namely Smullyan (1988, 2009). In Smullyan (1988), Raymond M. Smullyan, the famous mathematical logician, pianist and magician, addresses the logical world of Knights and Knaves and proceeds with proof theory. In the other book (Smullyan 2009), Smullyan addresses mainly combinatorial algebras, although he revisits the same world of Knights and Knaves.

We start by introducing formal language to model the universe of discourse. At any new feature of the natural language, we keep adding formal language through new logical operators and rules to deal with the new situations.

The basic constituents of the discourse are the agents, either Knights or truth tellers, those that speak the truth in any occasion, and Knaves, the liars, those that always lie: $T(X)$ means that agent X is a truth teller and $\neg T(X)$ means that agent X is

liar. Note that $T(X)$ asserts that agent X is a *permanent* truth teller and $\neg T(X)$ asserts that agent X is a permanent liar. Thus agents are invariant with respect to falsehood.

The next ingredient, discussed in Sects. 4 and 5, is the **claim** operator $\mathbf{C}_X$ that reads *agent X claims that*. In this way, $\mathbf{C}_X\varphi$ reads *agent X claims that* φ. Thus $\mathbf{C}_Y\neg T(D)$ is a formal statement that reads *agent Y* **claims** *that agent D is a liar*, or just, in a more common language, *Arthur* **claims** *that Dan is a liar*, or *Arthur once* **claimed** *that Dan is a liar*, together with all possible modalities of natural language. Assertion $\mathbf{C}_Y\neg T(D)$ can then be seen as *Arthur York once said that Dan* **was** *a liar*. Section 4 illustrates first the expressive power of the language with more simple cases. Of course, we find limitations, e.g., *Dan might have been a liar at that particular trial, but he is not lying this time*. However, since our agents are invariant with respect to the truth, the language cannot capture time dependencies. It would require more hard work on language operators and other complexities. Some weak form of truthhood time dependency is nevertheless discussed in Sect. 6, where agents have a time table for truth and falsehood.

We can expand expressive power if we allow quantification. With quantification in Sect. 7 we can model assertions of the kind *for all agent x, we can find another agent y such that y will claim that both x and y are constant liars*.

More operators and subtleties can be introduced and modeled, respectively, by the Smullyan reader. Besides getting acquainted with mathematical logic and the tableaux system, the reader may acquire skills to model natural language and learning how to retrieve meaning from ambiguous sets of sentences.

To introduce the main theme of Sects. 4, 5, 6, and 7, we start with a more rigorous specification of the language and the calculus in Sect. 2, followed by some further notation in Sect. 3.

2 Signature, Syntax and Calculus

We define a system to express and prove propositions in the Smullyan's universe of Knights and Knaves. We adopt the notation and style of Sernadas and Sernadas (2008). We consider:

- *Variables*, also called agent variables, represented by lowercase letters such as $x, y, z, \ldots$
- *Constant symbols or names*, also called agent symbols, represented by uppercase letters such as $A, B, C, \ldots, X, \ldots$
- Two important *predicate symbols*, the truthhood symbol T depending on the agent and the equality simbol $\equiv$ depending on two agents.

Agent variables together with agent symbols constitute the so-called *terms*. To define the *sentences* (we prefer here sentence to formula for we are analising natural language), we use the constructors $\neg$ (negation), $\vee$ (disjunction) and $\mathbf{C}$ (claim):

- $T(t)$ is a sentence whenever t is a term;
- $t_1 \equiv t_2$ is a sentence whenever t_1, t_2 are terms;
- $(\neg\varphi)$ is a sentence whenever φ is a sentence;

- $(\varphi_1 \vee \varphi_2)$ is a sentence whenever φ_1 and φ_2 are sentences;
- $(\mathbf{C}_t\varphi)$ is a sentence whenever t is a term and φ is a sentence.

Notice that $T(t)$ reads 'agent t is a truth speaker'. The statement $t_1 \equiv t_2$ reads 'agent t_1 is agent t_2' in situations such like 'the defendant (D) is Arthur York (Y). The last constructor is used to express agent claims; that is, the sentence $\mathbf{C}_t\varphi$ should be read as 'agent t claims that φ is the case'. In subsequent sections we will require new constructors and symbols.

The last step is the calculus of sentences, that is, a collection of *inference rules* that allow the logician to obtain sentences that are logic consequences of the hypotheses. Our rules of inference will take the form

$$\frac{\Gamma}{\Delta} \qquad\qquad \frac{\Gamma}{\Delta_1 \qquad \Delta_2}$$

where Γ is a set of sentences and Δ, Δ_1 and Δ_2 are non-empty sets of sentences. When $\Gamma = \emptyset$ the rule of inference is called an axiom. The second rule type above (more involved) intuitively means 'whatever is derivable both from Δ_1 and Δ_2 is derivable from Γ'. This is a top-down system of inference branching downwards.

For a given set (collection) of sentences Γ, the set of sentences derivable (we will say provable) from Γ is defined as follows:

- *Hypotheses*: every sentence in Γ is provable from Γ;
- *Principle of excluded middle*: every statement φ is such that $\varphi \vee \neg\varphi$ is provable from Γ;
- *Disjunction*: every statement provable both from $\Gamma \cup \{\varphi_1\}$ and $\Gamma \cup \{\varphi_2\}$ is (considered) provable from $\{\varphi_1 \vee \varphi_2\} \cup \Gamma$;
- *Negative disjunction*: if the statement $\neg(\varphi_1 \vee \varphi_2)$ is provable from Γ, then $\neg\varphi_1$ and $\neg\varphi_2$ are both provable from Γ;
- *True claim*: for any sentence φ and term t, if $\mathbf{C}_t\varphi$ and $T(t)$ are both provable from Γ, then φ is also (considered) provable from Γ;
- *False claim*: for any sentence φ and term t, if $\mathbf{C}_t\varphi$ and $\neg T(t)$ are both provable from Γ, then $\neg\varphi$ is also (considered) provable from Γ;
- *Absurd*: for any sentences φ and ψ, if φ and $\neg\varphi$ are both provable from Γ, then ψ is provable from Γ;
- *Reflexivity*: for any term t, we have that $t \equiv t$ is provable from Γ;
- *Symmetry*: for any terms t_1 and t_2, if $t_1 \equiv t_2$ is provable from Γ, then also $t_2 \equiv t_1$ is provable from Γ;
- *Transitivity*: for any terms t_1, t_2, and t_3, if both $t_1 \equiv t_2$ and $t_2 \equiv t_3$ are provable from Γ, then so is provable from Γ the sentence $t_1 \equiv t_3$;
- *Congruence for T*: for any terms t_1 and t_2, if both $t_1 \equiv t_2$ and $T(t_1)$ are provable from Γ, then $T(t_2)$ is provable from Γ;
- *Congruence for* $\mathbf{C}$: for any terms t_1 and t_2, for any sentence φ, if both statements $t_1 \equiv t_2$ and $\mathbf{C}_{t_1}\varphi$ are provable from Γ, then $\mathbf{C}_{t_2}\varphi$ is also provable from Γ;

Notice that by *provable* in the previous list of case rules we mean *in one step of reasoning we can infer that*. Of course, this concept of provability is rooted in the rules of thought that *we might accept or not*. We may invest some time in analising the rules one by one to see if any of them troubles our mind.

The above rules can be represented in a *proof tree* notation. For example, the *true claim* can be represented as

$$\frac{T(t), \mathbf{C}_t\varphi}{\varphi}$$

and the *disjunction* can be represented as

$$\frac{\varphi_1 \vee \varphi_2}{\varphi_1 \qquad \varphi_2}$$

We now illustrate how to prove a statement that derives (is provable) from a set of hypotheses or premises. For example, consider any sentences φ_1 and φ_2 and let the hypotheses be $\Gamma = \{\varphi_1 \vee \varphi_2, \neg\varphi_1 \vee \varphi_2\}$, i.e., the premises or hypotheses are two: $\varphi_1 \vee \varphi_2$ and $\neg\varphi_1 \vee \varphi_2$, meaning 'hypothesis 1: either φ_1 is the case or φ_2 is the case' and 'hypothesis 2: either $\neg\varphi_1$ is the case or φ_2 is the case'. We now prove that φ_2 can be proved from Γ, *statement that the reader surely accepts without proof since it seems quite obvious*. The reader should analise each line on top of the RULES of provability introduced above:

1. $\varphi_1 \vee \varphi_2$ and $\neg\varphi_1 \vee \varphi_2$ are the hypotheses Γ to start with (*hypothesis rule*);
2. Every statement that can be proved both from $\Gamma \cup \{\varphi_1\}$ and $\Gamma \cup \{\varphi_2\}$ can be proved from Γ since $\varphi_1 \vee \varphi_2 \in \Gamma$ (*disjunction rule*);
3. φ_2 is provable from $\Gamma \cup \{\varphi_2\}$, since φ_2 is itself an hypothesis in $\Gamma \cup \{\varphi_2\}$ (*hypothesis rule*);
4. Every statement that can be proved both from $\Gamma \cup \{\varphi_1\} \cup \{\neg\varphi_1\}$ and $\Gamma \cup \{\varphi_1\} \cup \{\varphi_2\}$ is provable from $\Gamma \cup \{\varphi_1\}$, since $\neg\varphi_1 \vee \varphi_2 \in \Gamma$ (again *disjunction rule*);
5. φ_2 is provable from $\Gamma \cup \{\varphi_1\} \cup \{\varphi_2\}$, since φ_2 is itself an hypothesis in $\Gamma \cup \{\varphi_1\} \cup \{\varphi_2\}$ (*hypothesis rule*);
6. φ_2 is provable from $\Gamma \cup \{\varphi_1\} \cup \{\neg\varphi_1\}$, since the context contains a contradiction, φ_1 and $\neg\varphi_1$ (*absurd rule*);
7. φ_2 is provable from $\Gamma \cup \{\varphi_1\}$ by steps 4, 5 and 6;
8. φ_2 is provable from Γ by steps 2, 3 and 7.

It is useful to present the above reasoning in the compact form of a proof tree as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\varphi_1 \vee \varphi_2, \neg\varphi_1 \vee \varphi_2}}{\varphi_1 \vee \varphi_2, \neg\varphi_1 \vee \varphi_2, \varphi_1} \qquad \dfrac{}{\varphi_1 \vee \varphi_2, \neg\varphi_1 \vee \varphi_2, \varphi_2}}{\varphi_1 \vee \varphi_2, \neg\varphi_1 \vee \varphi_2, \varphi_1, \neg\varphi_1} \qquad \dfrac{}{\varphi_1 \vee \varphi_2, \neg\varphi_1 \vee \varphi_2, \varphi_1, \varphi_2}}{\varphi_2}$$

The reader can follow the steps of the proof from 1 to 8 above in the proof tree, and he or she should do it before starting reading the next sections. Reading proof trees help in assigning meaning to their branches and understanding sentence decomposition. The reader may look to a proof tree such as that one above as a dense and illegible pile of symbols. We can guaranty to the reader that all those symbols are easy to read, it will be enough a slight effort in the beginning.

This method is due to Raymond M. Smullyan himself and put in practice by Richard C. Jeffrey. Deductive systems can be deeply studied in several old and new good books such as Bell and Machover (1977), Leblanc and Wisdom (1972) or Sernadas and Sernadas (2008).

We can rigorously define a proof tree as a finite rooted tree with a map that assigns to each node a set of statements. Moreover, each node that is not a leaf represents an instantiation of some rule of inference, and its children are the possible conclusions of that rule. Using this notion, we can prove an equivalence between proving that a sentence is derivable from a set of hypotheses and providing a proof tree.

Proposition 1 *Let Γ be a set of sentences and γ a sentence. Then we write $\Gamma \vdash \gamma$ if and only if there is a proof tree such that*

- *Γ is the set of sentences assigned to the root node;*
- *γ is a member of every set of sentences assigned to a leaf node.*

In a proof tree, we use branching to introduce additional hypotheses (the only use of branching is via the *disjunction rule*). When we say that γ is a member of every set of sentences assigned to a leaf node, this means that in any branch of the tree (that is, in every choice of additional hypotheses) we can derive γ, which intuitively means that γ is derivable from the root node.

3 Useful Rules and Constructors

We have listed in the previous section some rules of inference, which form the core of our deduction system. However, we can prove additional rules (obtainable by the core rules) that can be helpful in reducing the size of the proof trees. For example, an useful rule is obtained by combining the principle of excluded middle and disjunction.

Proposition 2 *(Cut rule)*

$$\frac{}{\varphi \qquad \neg\varphi}$$

Proof The above is a simplification of

$$\dfrac{\dfrac{}{\varphi \vee \neg\varphi}}{\varphi \qquad \neg\varphi}$$

□

Proposition 3 *(Double negation)*

$$\frac{\neg\neg\varphi}{\varphi}$$

Proof To prove the above rule simply apply the cut rule in φ:

$$\dfrac{\neg\neg\varphi}{\neg\neg\varphi,\varphi \qquad \dfrac{\neg\neg\varphi,\neg\varphi}{\varphi}}$$

We also want to express sentences using constructors $\wedge$ (conjunction), $\rightarrow$ (implication), and $\leftrightarrow$ (equivalence). These can be expressed as **abbr**eviations, in terms of $\vee$ and $\neg$:

- $(\varphi_1 \wedge \varphi_2) \overset{\text{abbr}}{=} \neg(\neg\varphi_1 \vee \neg\varphi_2)$;
- $(\varphi_1 \rightarrow \varphi_2) \overset{\text{abbr}}{=} \neg\varphi_1 \vee \varphi_2$.
- $(\varphi_1 \leftrightarrow \varphi_2) \overset{\text{abbr}}{=} (\neg\varphi_1 \vee \varphi_2) \wedge (\varphi_1 \vee \neg\varphi_2)$.

We can also state and prove the correspondent rules for *conjunction*, *negative conjunction*, *implication* and *negative implication*, *equivalence* and *negative equivalence*. We shall present only one of these rules. □

Proposition 4 *(Conjunction)*

$$\frac{\varphi_1 \wedge \varphi_2}{\varphi_1, \varphi_2}$$

Proof To prove the above rule simply apply the negative disjunction and the double negation:

$$\dfrac{\varphi_1 \wedge \varphi_2 \overset{\text{abbr}}{=} \neg(\neg\varphi_1 \vee \neg\varphi_2)}{\dfrac{\neg\neg\varphi_1}{\varphi_1} \qquad \dfrac{\neg\neg\varphi_2}{\varphi_2}}$$

Before moving to the problems, we present a useful technique to simplify a proof tree. Each time we use the *absurd* rule, we can write the symbol $*$ to denote any sentence (since from *absurd* any sentence is derivable). Then, we only need to look at the leaves of the proof tree that do not have this symbol. For instance, here is an alternate proof of the double negation:

$$\dfrac{\neg\neg\varphi}{\neg\neg\varphi,\varphi \qquad \dfrac{\neg\neg\varphi,\neg\varphi}{*}}$$

□

4 The Census Taker—*I Claimed (Level I)*

We start with the most simple modeling exercises. The puzzles from this section are taken from Smullyan (1988).

> The census taker Mr. McGregor once did some fieldwork on the Island of Knights and Knaves. On this island, women are also called Knights and Knaves. McGregor decided on this visit to interview married couples only.
>
> Smullyan in (1988)

4.1 And

> McGregor knocked on one door; the husband partly opened it and asked McGregor his business. 'I am a census taker', replied McGregor, 'and I need information about you and your wife, Which, if either, is a Knight, and which, if either, is a Knave?'
> 'We are both Knaves!' said the husband angrily as he slammed the door.
> What type is the husband [H] and what type is the wife [W]?
>
> Smullyan in (1988)

That the husband is a Knave is denoted by $\neg T(H)$ and the same for his wife $\neg T(W)$. The conjunction of both potential facts is denoted by $\neg T(H) \wedge \neg T(W)$. The complexity is added when we realize that $\neg T(H) \wedge \neg T(W)$ is the statement of the husband. Thus the husband claims that $\neg T(H) \wedge \neg T(W)$ or, in simple symbolic terms, $\mathbf{C}_H(\neg T(H) \wedge \neg T(W))$. Now, we apply the calculus to see if we can say something about the nature of husband and wife. Consider the following proof tree:

$$\dfrac{\mathbf{C}_H(\neg T(H) \wedge \neg T(W))}{\dfrac{\dfrac{\dfrac{\mathbf{C}_H(\neg T(H) \wedge \neg T(W)), T(H)}{T(H), \neg T(H) \wedge \neg T(W)}}{T(H), \neg T(H), \neg T(W)}}{*} \qquad \dfrac{\dfrac{\dfrac{\mathbf{C}_H(\neg T(H) \wedge \neg T(W)), \neg T(H)}{\neg T(H), \neg(\neg T(H) \wedge \neg T(W))}}{\neg T(H), \neg\neg T(H) \vee \neg\neg T(W)}}{\dfrac{\neg T(H), \neg\neg T(H)}{*} \qquad \dfrac{\neg T(H), \neg\neg T(W)}{\neg T(H), T(W)}}$$

We conclude that $\{\mathbf{C}_H(\neg T(H) \wedge \neg T(W))\} \vdash \neg T(H), T(W)$, hence the husband ($H$) is a Knave and his wife (W) is a Knight.

4.2 If-Then

> The next home visited by McGregor proved more of a puzzler. The door was opened timidly by a rather shy man. After McGregor asked him to say something about himself and his wife, all the husband said was: 'If I am a Knight, then so is my wife.'
> McGregor walked away none too pleased. 'How can I tell anything about either of them from such a noncommittal response?' he thought. He was about

> to write down 'Husband and wife both unknown', when he suddenly recalled an old logic lesson from his Oxford undergraduate days. 'Of course,' he realised, 'I can tell *both* their types!'
> What type is the husband [*H*] and what type is the wife [*W*]?
>
> Smullyan in (1988)

'If I am a Knight, then so is my wife' is denoted just by $T(H) \rightarrow T(W)$, i.e., 'I am a Knight', $T(H)$, implies 'my wife is a Knight', $T(W)$. Consider the following proof tree:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_H(T(H) \rightarrow T(W))}{T(H), \mathbf{C}_H(T(H) \rightarrow T(W))}}{T(H), T(H) \rightarrow T(W)}}{T(H), \neg T(H) \vee T(W)}}{T(H), \neg T(H) \quad T(H), T(W)}}{*} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_H(T(H) \rightarrow T(W))}{\neg T(H), \mathbf{C}_H(T(H) \rightarrow T(W))}}{\neg T(H), \neg(T(H) \rightarrow T(W))}}{\neg T(H), T(H), \neg T(W)}}{*}$$

Notice that, in the proof tree, we used the fact that $\neg(T(H) \rightarrow T(W))$ is equivalent to $T(H) \wedge \neg T(W)$ ($T(H), \neg T(W)$ in the bottom right of the tree, just before the asterisk). We conclude that $\{\mathbf{C}_H(T(H) \rightarrow T(W))\} \vdash T(H), T(W)$, hence both the husband and his wife are Knights.

4.3 If-and-Only-If

> When the census taker interviewed the third couple, the husband said: 'My wife and I are of the same type: we are either both Knights or both Knaves.' What can be deduced about the husband and what can be deduced about the wife?
>
> Smullyan in (1988)

'My wife and I are of the same type' can be denoted by $T(H) \leftrightarrow T(W)$. Consider the following proof tree:

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_H(T(H) \leftrightarrow T(W))}{T(H), \mathbf{C}_H(T(H) \leftrightarrow T(W))}}{T(H), T(H) \leftrightarrow T(W)}}{T(H), T(W)} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_H(T(H) \leftrightarrow T(W))}{\neg T(H), \mathbf{C}_H(T(H) \leftrightarrow T(W))}}{\neg T(H), \neg(T(H) \leftrightarrow T(W))}}{\neg T(H), (T(H) \wedge \neg T(W)) \vee (\neg T(H) \wedge T(W))}}{\neg T(H), T(H) \wedge \neg T(W) \quad \neg T(H), \neg T(H) \wedge T(W)}}{T(H), \neg T(H), \neg T(W) \qquad \neg T(H), T(W)}}{*}$$

Notice that, in the proof tree, we used the fact that $\neg(T(H) \leftrightarrow T(W))$ is equivalent to $(T(H) \wedge \neg T(W)) \vee (\neg T(H) \wedge T(W))$. We conclude that $\{\mathbf{C}_H(T(H) \leftrightarrow T(W))\} \vdash T(W)$, hence the wife is a Knight. However, we are not able to conclude whether the husband is a Knight or a Knave, one leaf says yes and the other says no.

5 The trials of Arthur York—*I Claim that I Claimed (Level II)*

On the following subsections we analyse several puzzles introduced in Smullyan (2000) and then solve them using the calculus introduced in Sects. 2, 3 and 4.

We begin with the trials of Arthur York.

> Inspector Craig of Scotland Yard was called to the Island of Knights and Knaves to help find a criminal named Arthur York. What made the process difficult was that it was not known whether Arthur York was a Knight or a Knave.
>
> Smullyan in (2000)

5.1 The First Trial of Arthur York

> One suspect was arrested and brought to trial. Inspector Craig was the presiding judge. Here is a transcript of the trial:
> **Craig**: What do you know about Arthur York?
> **Defendant**: Arthur York once claimed that I was a Knave.
> **Craig**: Are you by any chance Arthur York?
> **Defendant**: Yes.
> Is the defendant [D] Arthur York [Y]?
>
> Smullyan in (2000)

To start with the set of sentences, we have first to write them in the formal language. It should be straightforward: (a) that Y stands for Arthur York and D for the defendant, (b) that the last answer should be modeled by *the defendant claims that he is Arthur York*, or D *claims that* $D \equiv Y$, or simply $\mathbf{C}_D D \equiv Y$, (c) the first answer corresponds to D *claims that* Y *claimed that* $\neg T(D)$, i.e., $\mathbf{C}_D\mathbf{C}_Y\neg T(D)$. We join the two sentences in the same set $\Gamma = \{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \mathbf{C}_D D \equiv Y\}$, and apply to it our calculus. Consider the following proof tree:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \mathbf{C}_D D \equiv Y}{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \mathbf{C}_D D \equiv Y, T(D)}}{\mathbf{C}_D\mathbf{C}_Y\neg T(D), T(D), D \equiv Y}}{\mathbf{C}_Y\neg T(D), T(D), D \equiv Y}}{\mathbf{C}_Y\neg T(D), T(D), D \equiv Y, T(Y)}}{T(D), D \equiv Y, T(Y), \neg T(D)}}{*} \qquad \dfrac{\dfrac{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \mathbf{C}_D D \equiv Y}{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \mathbf{C}_D D \equiv Y, \neg T(D)}}{\dfrac{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \neg T(D), \neg D \equiv Y}{\neg T(D), \neg D \equiv Y, \neg\mathbf{C}_Y\neg T(D)}}$$

We provide the explanation of a proof tree once in this section. All the other proof trees read in the same way.

The branching results from one application of the *cut*: either the defendant speaks the truth, or what he says is false, that is, either $T(D)$ or $\neg T(D)$. In the left branch, we conclude first that, since the defendant speaks the truth and claims that he is Arthur York, the defendant is indeed Arthur York; and then what he first claims is true, that is, that Arthur York claimed that the defendant is a liar; then we conclude

that Arthur York says the truth, i.e., $T(Y)$; and finally that the defendant is a liar; the contradiction between $T(D)$ and $\neg T(D)$ is marked with asterisk. The branch is closed and no information was obtained! In the right branch, we first conclude that what the defendant said in second place is false, i.e., that the defendant is not Arthur York; finally, since the defendant is a liar, we conclude that it is not true that Arthur York once claimed that the defendant is a liar.

We conclude, looking at the right leaf, that $\{\mathbf{C}_D\mathbf{C}_Y\neg T(D), \mathbf{C}_D D \equiv Y\} \vdash \neg T(D)$, $\neg D \equiv Y$, hence the defendant is not Arthur York. He is also a Knave.

5.2 The Second Trial of Arthur York

> Another suspect was arrested and brought to trial. Here is a transcript of the trial:
> **Craig**: The last suspect was a queer bird; he actually claimed to be Arthur York! Did you ever claim to be Arthur York?
> **Defendant**: No.
> **Craig**: Did you ever claim that you are *not* Arthur York?
> **Defendant**: Yes.
> Craig's first guess was that the defendant [D] was not Arthur York [Y] , but are there really sufficient grounds for acquitting him?
>
> Smullyan in (2000)

The first claim of the defendant is straightforward. The second statement is that *the defendant claims that the defendant claimed that* $\neg D \equiv Y$, i.e., $\mathbf{C}_D\mathbf{C}_D\neg D \equiv Y$. Consider the following proof tree:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\mathbf{C}_D\neg\mathbf{C}_D D \equiv Y, \mathbf{C}_D\mathbf{C}_D\neg D \equiv Y}}{T(D), \mathbf{C}_D\neg\mathbf{C}_D D \equiv Y, \mathbf{C}_D\mathbf{C}_D\neg D \equiv Y}}{T(D), \neg\mathbf{C}_D D \equiv Y, \mathbf{C}_D\neg D \equiv Y}}{T(D), \neg\mathbf{C}_D D \equiv Y, \neg D \equiv Y} \qquad \dfrac{\dfrac{\dfrac{\dfrac{}{\mathbf{C}_D\neg\mathbf{C}_D D \equiv Y, \mathbf{C}_D\mathbf{C}_D\neg D \equiv Y}}{\neg T(D), \mathbf{C}_D\neg\mathbf{C}_D D \equiv Y, \mathbf{C}_D\mathbf{C}_D\neg D \equiv Y}}{\neg T(D), \neg\neg\mathbf{C}_D D \equiv Y, \neg\mathbf{C}_D\neg D \equiv Y}}{\neg T(D), \mathbf{C}_D D \equiv Y, \neg\mathbf{C}_D\neg D \equiv Y}}{\neg T(D), \neg D \equiv Y, \neg\mathbf{C}_D\neg D \equiv Y}$$

We conclude that $\{\mathbf{C}_D\neg\mathbf{C}_D D \equiv Y, \mathbf{C}_D\mathbf{C}_D\neg D \equiv Y\} \vdash \neg D \equiv Y$ (notice that the statement $\neg D \equiv Y$ is *the unique logic consequence* of the premises that can be inferred in this proof tree), hence the defendant is not Arthur York. We are not able to conclude whether the defendant is a Knight or a Knave, one leaf says yes and the other says no.

5.3 The Third Trial of Arthur York

> 'Don't despair,' said Craig to the chief of the island police, 'we may find our man yet!'
> Well, a third suspect [D] was arrested and brought to trial. He brought with him his defense attorney [A] , and the two made the following statements in court.
> **Defense Attorney**: My client is indeed a Knave, but he is not Arthur York.

Defendant: My attorney always tells the truth!
Is there enough evidence either to acquit or convict the defendant?

Smullyan in (2000)

That 'My attorney always tells the truth' is denoted by $\mathbf{C}_D T(A)$. The statement of the defense attorney can be encoded into $\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y)$. Consider the following proof tree:

$$\dfrac{\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y), \mathbf{C}_D T(A)}{\dfrac{\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y), \mathbf{C}_D T(A), T(D)}{\dfrac{\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y), T(D), T(A)}{\dfrac{T(D), T(A), \neg T(D) \wedge \neg D \equiv Y}{\dfrac{T(D), T(A), \neg T(D), \neg D \equiv Y}{*}}}} \quad \dfrac{\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y), \mathbf{C}_D T(A), \neg T(D)}{\dfrac{\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y), \neg T(D), \neg T(A)}{\dfrac{\neg T(D), \neg T(A), \neg(\neg T(D) \wedge \neg D \equiv Y)}{\dfrac{\neg T(D), \neg T(A), \neg\neg T(D) \vee \neg\neg D \equiv Y}{\mathbf{R}}}}}$$

Right subtree (**R**):

$$\dfrac{\neg T(D), \neg T(A), \neg\neg T(D) \vee \neg\neg D \equiv Y}{\dfrac{\neg T(D), \neg T(A), \neg\neg T(D)}{\dfrac{\neg T(D), \neg T(A), T(D)}{*}}} \quad \dfrac{\neg T(D), \neg T(A), \neg\neg D \equiv Y}{\neg T(D), \neg T(A), D \equiv Y}$$

We conclude that $\{\mathbf{C}_A(\neg T(D) \wedge \neg D \equiv Y), \mathbf{C}_D T(A)\} \vdash \neg T(D), \neg T(A),\ D \equiv Y$, hence the defendant is Arthur York. We also conclude that the defendant and his attorney are both Knaves.

5.4 The Father

On the next day Inspector Craig came across a native who said: 'My father once said that he and I are different types, one a Knight and one a Knave.'
Is it possible that his father really said that?
Consider the following proof tree:

Smullyan in (2000)

$$\dfrac{\mathbf{C}_S\mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\dfrac{T(S), \mathbf{C}_S\mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\dfrac{T(S), \mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\mathbf{L}}}} \quad \dfrac{\neg T(S), \mathbf{C}_S\mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\neg T(S), \neg\mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}$$

with the left tree L:

$$\dfrac{T(S), \mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\dfrac{T(S), T(F), \mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\dfrac{T(S), T(F), T(S) \leftrightarrow \neg T(F)}{*}}} \quad \dfrac{T(S), \neg T(F), \mathbf{C}_F(T(S) \leftrightarrow \neg T(F))}{\dfrac{T(S), \neg T(F), \neg(T(S) \leftrightarrow \neg T(F))}{\dfrac{T(S), \neg T(F), T(S) \leftrightarrow T(F)}{*}}}$$

We conclude that $\{\mathbf{C}_S\mathbf{C}_F(T(S) \leftrightarrow \neg T(F))\} \vdash \neg T(S)$, hence the son is lying.

6 Day-Knights and Night-Knights—*I Claimed at Night (Level III)*

> There is a strange place called Subterranea. It is a city completely underground; the inhabitants have never seen the light of the day. Clocks, watches and all other timepieces are strictly forbidden. Yet the inhabitants have an uncanny sense of time; they always know when it is day and when it is night. Each inhabitant is of one of two types — *day-Knights* and *night-Knights*. The day-Knights tell the truth during the day and lie during the night; the night-knights tell the truth during the night and lie during the day.
> Visitors to the city are allowed, but of course they may not bring any timepieces with them. Any visitor to the city is bound to become disoriented; after a few days she or he loses all sense of when it is day and when it is night.
> Smullyan in (2000)

In this section, and only in this section, we modify the semantics of the claim operator in order to solve time dependencies. Let us add language to model the novelty.

To reason about inhabitants of Subterranea, we need to adapt our logic. We remove the truthhood symbol T and replace it with the symbol D ($D(X)$ should be read as 'X is a day-Knight'). We also introduce the propositional constant symbol **DAY** that reads 'now (that I did my claim) it is day'. Of course $\neg$**DAY** reads 'now (that I did my claim) it is night'. Thus, the set of sentences or statements is now augmented in order to contain:

- $D(t)$ is a statement whenever t is a term;
- **DAY** is a statement.

Notice that $D(t)$ should be interpreted as 'agent t is a daytime truth speaker', i.e., 'agent t speaks the truth during the day and is a liar during the night'. We also introduce five new very intuitive rules of inference:

- *Congruence for D*: for any terms t_1 and t_2, if both $t_1 \equiv t_2$ and $D(t_1)$ are provable from Γ, then $D(t_2)$ is provable from Γ;
- For any statement φ and agent X if we can prove $\mathbf{C}_X\varphi$, $D(X)$, and **DAY**, then we have a proof of statement φ;
- For any statement φ and agent X if we can prove $\mathbf{C}_X\varphi$, $\neg D(X)$, and **DAY**, then we have a proof of statement $\neg\varphi$;
- For any statement φ and agent X if we can prove $\mathbf{C}_X\varphi$, $D(X)$, and $\neg$**DAY**, then we have a proof of statement $\neg\varphi$;
- For any statement φ and agent X if we can prove $\mathbf{C}_X\varphi$, $\neg D(X)$, and $\neg$**DAY**, then we have a proof of statement φ;

To see that these rules are self-evident, just translate them into the natural language. E.g., for the first: if Arthur claims that φ is the case, Arthur speaks the truth during the day and it is daytime, then surely φ is the case. E.g., for the last: if

Arthur claims that φ is the case, Arthur is a liar at daytime but it is at night, then surely φ is the case.

6.1 The Night-Knight

> On one occasion an inhabitant said: 'I am a night-Knight and it is now day.' Was he a day-Knight or a night-Knight? Was it then day or night?
>
> Smullyan in (2000)

The sentence translates into the premise $\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})$. Consider the following proof tree:

$$\dfrac{\dfrac{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})}{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY}), \mathbf{DAY}}\,\mathbf{L} \qquad \dfrac{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})}{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY}), \neg\mathbf{DAY}}\,\mathbf{R}}{}$$

Left subtree (**L**):

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY}), \mathbf{DAY}}{\mathbf{DAY}, D(X), \mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})}}{\mathbf{DAY}, D(X), \neg D(X) \wedge \mathbf{DAY}}}{\mathbf{DAY}, D(X), \neg D(X)}}{*} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY}), \mathbf{DAY}}{\mathbf{DAY}, \neg D(X), \mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})}}{\mathbf{DAY}, \neg D(X), \neg(\neg D(X) \wedge \mathbf{DAY})}}{\mathbf{DAY}, \neg D(X), D(X) \vee \neg\mathbf{DAY})}}{\mathbf{DAY}, D(X), \neg D(X)} \quad \dfrac{}{\mathbf{DAY}, \neg\mathbf{DAY}, \neg D(X)}$$

Right subtree (**R**):

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY}), \neg\mathbf{DAY}}{\neg\mathbf{DAY}, D(X), \mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})}}{\neg\mathbf{DAY}, D(X), \neg(\neg D(X) \wedge \mathbf{DAY})}}{\neg\mathbf{DAY}, D(X), D(X) \vee \neg\mathbf{DAY}}}{\neg\mathbf{DAY}, D(X) \quad \neg\mathbf{DAY}, D(X)} \qquad \dfrac{\dfrac{\dfrac{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY}), \neg\mathbf{DAY}}{\neg\mathbf{DAY}, \neg D(X), \mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})}}{\neg\mathbf{DAY}, \neg D(X), \neg D(X) \wedge \mathbf{DAY}}}{\neg\mathbf{DAY}, \neg D(X), \mathbf{DAY}}}{*}$$

We conclude that $\{\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})\} \vdash \neg\mathbf{DAY}, D(X)$, hence the inhabitant was a day-Knight speaking during the night. Notice that, as soon as the cut between $D(X)$ and $\neg D(X)$ is done, we applied one of the four rules of inference: if $\mathbf{C}_X(\neg D(X) \wedge \mathbf{DAY})$ and $D(X)$ and **DAY**, then we conclude that $\neg D(X) \wedge \mathbf{DAY}$, and so on...

6.2 Two Simultaneous Questions

> On another occasion I asked an inhabitant two questions: 'Are you a day-Knight?' and 'Is it now day?' He replied: 'Yes is the correct answer to at least one of your questions.' Was he [X] a day-Knight or a night-Knight? Was it then day or night?
>
> Smullyan in (2000)

The sentence should be read as 'I am a day-Knight or it is now day', which translates into the premise $\mathbf{C}_X(D(X) \vee \mathbf{DAY})$. Consider the following simplified proof tree:

$$\dfrac{\mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\mathbf{L}} \quad \dfrac{\neg D(X), \mathbf{C}_X(D(X \vee \mathbf{DAY})}{\mathbf{R}}}$$

Left subtree (**L**):

$$\dfrac{D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{\mathbf{DAY}, D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{\mathbf{DAY}, D(X), D(X) \vee \mathbf{DAY}}{\mathbf{DAY}, D(X) \quad \mathbf{DAY}, D(X)}} \quad \dfrac{\neg\mathbf{DAY}, D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{\neg\mathbf{DAY}, D(X), \neg D(X) \wedge \neg\mathbf{DAY}}{\dfrac{\neg\mathbf{DAY}, D(X), \neg D(X)}{*}}}}$$

Right subtree (**R**):

$$\dfrac{\neg D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{\mathbf{DAY}, \neg D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{\mathbf{DAY}, \neg D(X), \neg D(X) \wedge \neg\mathbf{DAY}}{\dfrac{\mathbf{DAY}, \neg D(X), \neg\mathbf{DAY}}{*}}} \quad \dfrac{\neg\mathbf{DAY}, \neg D(X), \mathbf{C}_X(D(X) \vee \mathbf{DAY})}{\dfrac{\neg\mathbf{DAY}, \neg D(X), D(X) \vee \mathbf{DAY}}{\dfrac{\neg\mathbf{DAY}, \neg D(X), D(X)}{*} \quad \dfrac{\neg\mathbf{DAY}, \neg D(X), \mathbf{DAY}}{*}}}}$$

We conclude that $\{\mathbf{C}_X(D(X) \vee \mathbf{DAY})\} \vdash \mathbf{DAY}, D(X)$, hence the inhabitant was a day-Knight speaking during the day.

6.3 The Two Brothers

> I once came across two brothers *A* and *B* and did not know the type of either, nor did I even know whether they were of the same type. I also did not know whether it was day or night at the time. Here is what they said:
> **A**: At least one of us is a day-night.
> **B**: A is a night-Knight.
> I then knew the type of each and whether it was day or night.
>
> Smullyan in (2000)

The first sentence translates into the premise $\mathbf{C}_A(D(A) \vee D(B))$ and the second sentence translates into the other premise $\mathbf{C}_B\neg D(A)$. Consider the following proof tree:

$$\dfrac{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{C}_B\neg D(A)}{\dfrac{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{C}_B\neg D(A), \mathbf{DAY}}{\mathbf{L}} \quad \dfrac{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{C}_B\neg D(A), \neg\mathbf{DAY}}{\mathbf{R}}}$$

Left subtree (**L**):

$$\mathbf{C}_A(D(A) \vee D(B)), \mathbf{C}_B\neg D(A), \mathbf{DAY}$$
$$\overline{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{DAY}, D(B), \mathbf{C}_B\neg D(A)} \qquad \overline{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{DAY}, \neg D(B), \mathbf{C}_B\neg D(A)}$$
$$\overline{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{DAY}, D(B), \neg D(A)} \qquad \overline{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{DAY}, \neg D(B), \neg\neg D(A)}$$
$$\overline{\mathbf{DAY}, D(B), \neg D(A), \neg(D(A) \vee D(B))} \qquad \overline{\mathbf{DAY}, \neg D(B), D(A), D(A) \vee D(B)}$$
$$\overline{\mathbf{DAY}, D(B), \neg D(A), \neg D(A) \wedge \neg D(B)} \qquad \overline{\mathbf{DAY}, \neg D(B), D(A) \qquad \mathbf{DAY}, \neg D(B), D(A), D(B)}$$
$$\overline{\mathbf{DAY}, D(B), \neg D(A), \neg D(B)} \qquad \qquad \qquad \qquad \qquad \qquad * \qquad$$
$$* \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad$$

Right subtree (**R**):

$$\mathbf{C}_A(D(A) \vee D(B)), \mathbf{C}_B\neg D(A), \neg\mathbf{DAY}$$
$$\overline{\mathbf{C}_A(D(A) \vee D(B)), \neg\mathbf{DAY}, D(B), \mathbf{C}_B\neg D(A)} \qquad \overline{\mathbf{C}_A(D(A) \vee D(B)), \neg\mathbf{DAY}, \neg D(B), \mathbf{C}_B\neg D(A)}$$
$$\overline{\mathbf{C}_A(D(A) \vee D(B)), \neg\mathbf{DAY}, D(B), \neg\neg D(A)} \qquad \overline{\mathbf{C}_A(D(A) \vee D(B)), \neg\mathbf{DAY}, \neg D(B), \neg D(A)}$$
$$\overline{\neg\mathbf{DAY}, D(B), D(A), \neg(D(A) \vee D(B))} \qquad \overline{\neg\mathbf{DAY}, \neg D(B), \neg D(A), D(A) \vee D(B)}$$
$$\overline{\neg\mathbf{DAY}, D(B), D(A), \neg D(A) \wedge \neg D(B)} \qquad \overline{\neg\mathbf{DAY}, \neg D(B), \neg D(A), D(A)} \qquad \overline{\neg\mathbf{DAY}, \neg D(B), \neg D(A), D(B)}$$
$$\overline{\neg\mathbf{DAY}, D(B), D(A), \neg D(A), \neg D(B)} \qquad * \qquad \qquad *$$
$$* \qquad \qquad \qquad \qquad \qquad \qquad \qquad \qquad$$

We conclude that $\{\mathbf{C}_A(D(A) \vee D(B)), \mathbf{C}_B\neg D(A)\} \vdash \mathbf{DAY}, D(A), \neg D(B)$, hence A was a day-Knight, B was a night-Knight and it was then day.

6.4 The Two Inhabitants

> On another occasion I came across two inhabitants A and B who made the following statements:
> **A**: Both of us are day-Knights.
> **B**: That is not true!
> Which one should be believed?
>
> Smullyan in (2000)

The first sentence translates into the premise $\mathbf{C}_A(D(A) \wedge D(B))$ and the second sentence translates into the other premise $\mathbf{C}_B\neg(D(A) \wedge D(B))$. Consider the following proof tree:

$$\mathbf{C}_A(D(A) \wedge D(B)), \mathbf{C}_B\neg(D(A) \wedge D(B))$$
$$\overline{\mathbf{C}_A(D(A) \wedge D(B)), \mathbf{C}_B\neg(D(A) \wedge D(B)), \mathbf{DAY}} \qquad \overline{\mathbf{C}_A(D(A) \wedge D(B)), \mathbf{C}_B\neg(D(A) \wedge D(B)), \neg\mathbf{DAY}}$$
$$\mathbf{L} \qquad \qquad \qquad \qquad \qquad \qquad \mathbf{R}$$

Left subtree (**L**):

$$\mathbf{C}_A(D(A) \wedge D(B)), \mathbf{C}_B\neg(D(A) \wedge D(B)), \mathbf{DAY}$$
$$\overline{\mathbf{C}_A(D(A) \wedge D(B)), \mathbf{DAY}, D(A), \mathbf{C}_B\neg(D(A) \wedge D(B))} \qquad \overline{\mathbf{C}_A(D(A) \wedge D(B)), \mathbf{DAY}, \neg D(A), \mathbf{C}_B\neg(D(A) \wedge D(B))}$$
$$\overline{D(A) \wedge D(B), \mathbf{DAY}, D(A), \mathbf{C}_B\neg(D(A) \wedge D(B))} \qquad \overline{\neg(D(A) \wedge D(B)), \mathbf{DAY}, \neg D(A), \mathbf{C}_B\neg(D(A) \wedge D(B))}$$
$$\overline{\mathbf{DAY}, D(A), D(B), \mathbf{C}_B\neg(D(A) \wedge D(B))}$$
$$\overline{\mathbf{DAY}, D(A), D(B), \neg(D(A) \wedge D(B))}$$
$$\overline{\mathbf{DAY}, D(A), D(B), \neg D(A) \vee \neg D(B)}$$
$$\overline{\mathbf{DAY}, D(A), \neg D(A), D(B)} \qquad \overline{\mathbf{DAY}, D(A), D(B), \neg D(B)}$$
$$* \qquad \qquad \qquad *$$

Right subtree (**R**):

$$\dfrac{\dfrac{\dfrac{\mathbf{C}_A(D(A)\wedge D(B)),\mathbf{C}_B\neg(D(A)\wedge D(B)),\neg\mathbf{DAY}}{\mathbf{C}_B\neg(D(A)\wedge D(B)),\neg\mathbf{DAY},D(A),\mathbf{C}_A(D(A)\wedge D(B))}}{\mathbf{C}_B\neg(D(A)\wedge D(B)),\neg\mathbf{DAY},D(A),\neg(D(A)\wedge D(B))} \quad \dfrac{\dfrac{\dfrac{\mathbf{C}_B\neg(D(A)\wedge D(B)),\neg\mathbf{DAY},\neg D(A),\mathbf{C}_A(D(A)\wedge D(B))}{\mathbf{C}_B\neg(D(A)\wedge D(B)),\neg\mathbf{DAY},\neg D(A),D(A)\wedge D(B)}}{\mathbf{C}_B\neg(D(A)\wedge D(B)),\neg\mathbf{DAY},\neg D(A),D(A),D(B)}}{*}$$

We conclude that $\{\mathbf{C}_A(D(A)\wedge D(B)),\mathbf{C}_B\neg(D(A)\wedge D(B))\}\vdash\neg(D(A)\wedge D(B))$, hence the second inhabitant was telling the truth, they are not of the same type! We are not able to conclude the type of any brother nor whether it was day or night.

7 The Sociologist—*for Every One I Claimed (Level IV)*

In this last section we return to the same interpretation of the claim operator as in Sects. 3, 4 and 5 and introduce quantified sentences. To that end, we use the constructor $\forall$. Agents are again either constant truth tellers or constant liars.

We update the inductive definition of sentences of Sect. 2 with:

- $(\forall x\varphi)$ is a sentence whenever x is an agent variable and φ is a sentence.

We can also talk about the *bounded* and *free* variables of a sentence φ, defined as usual. We can also denote by $[\varphi]^x_t$ the sentence obtained by replacing in sentence φ all free occurences of variable x by term t. We also say that a variable is *fresh* with respect to a sentence φ if it does not occur (either free or bounded) in φ. Finally, we can introduce new rules of inferences for these connectives:

- *universal*: for any sentence φ, variable x and term t such that $\forall x\varphi$ is a provable sentence, then $[\varphi]^x_t$ is also a provable sentence, as long as t is an agent symbol or t is an agent variable that does not become bound after the substitution;
- *negative universal*: for any sentence φ and variables x, z such that $\neg(\forall x\varphi)$ is provable, then we have that $\neg[\varphi]^x_z$ is also provable, as long as z is a fresh variable in φ.

We can also express the constructor $\exists$ as an abbreviation:

- $\exists x\varphi \overset{\text{abbr}}{=} \neg(\forall x\neg\varphi)$.

Proposition 5 (Existential quantifier)

$$\frac{\exists x\varphi}{[\varphi]^x_z}\ z \text{ is a fresh variable with respect to } \varphi$$

Proof The proof is left to the reader. □

7.1 Inspector Craig Meets the Sociologist

> On the final day, Inspector Craig met a sociologist who was visiting the island. He gave the following report:
> 'I have interviewed all the inhabitants of this island and I have observed a curious thing: For every native *x*, there is at least one native *y* such that *y* claims that *x* and *y* are both Knaves.'
> Does this report hold water?
>
> Smullyan in (2000)

That '*x* and *y* are both liars' is easy to denote by $\neg T(x) \wedge \neg T(y)$. That *y* himself claims that fact is represented by $\mathbf{C}_y(\neg T(x) \wedge \neg T(y))$. The full sentence resumes to the use of quantifiers: $\forall x \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y))$. The derivation uses the rules above, together with the abbreviation relative to the existential quantifier. Notice in the following subtree that the existential quantifier is replaced by substituting the fresh variable *z* for *y*:

$$\dfrac{\dfrac{\dfrac{\forall x \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y))}{\forall x \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y)), \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y))}}{\forall x \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y)), \mathbf{C}_z(\neg T(x) \wedge \neg T(z))}}{\mathbf{L} \qquad\qquad \mathbf{R}}$$

The subtrees **L** and **R** are just the result of a cut relative to $T(z)$ (and $\neg T(z)$). Left subtree (**L**):

$$\dfrac{\dfrac{\dfrac{T(z), \mathbf{C}_z(\neg T(x) \wedge \neg T(z))}{T(z), \neg T(x) \wedge \neg T(z)}}{T(z), \neg T(x), \neg T(z)}}{*}$$

Notice in the following subtree that the existential quantifier is again replaced by substituting the fresh variable *w* for *y*. Then the fresh variable refers to a new agent *w* and a cut relative to $T(w)$ solves the problem. Right subtree (**R**):

$$\dfrac{\dfrac{\dfrac{\dfrac{\forall x \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y)), \neg T(z)}{\neg T(z), \exists y \mathbf{C}_y(\neg T(z) \wedge \neg T(y))}}{\neg T(z), \mathbf{C}_w(\neg T(z) \wedge \neg T(w))}}{\dfrac{\dfrac{\neg T(z), T(w), \mathbf{C}_w(\neg T(z) \wedge \neg T(w))}{\neg T(z), T(w), \neg T(z) \wedge \neg T(w)}}{\neg T(z), T(w), \neg T(z), \neg T(w)}}{*} \qquad \dfrac{\dfrac{\dfrac{\neg T(z), \neg T(w), \mathbf{C}_w(\neg T(z) \wedge \neg T(w))}{\neg T(z), \neg T(w), \neg(\neg T(z) \wedge \neg T(w))}}{\neg T(z), \neg T(w), T(z) \vee T(w)}}{\dfrac{\neg T(z), \neg T(w), T(z)}{*} \qquad \dfrac{\neg T(z), \neg T(w), T(w)}{*}}$$

We conclude that from $\forall x \exists y \mathbf{C}_y(\neg T(x) \wedge \neg T(y))$ we can derive any sentence, hence the sociologist gave a false report. Indeed the report holds water!

8 Concluding Remarks

We introduced logic operators guided by the need of modeling successively more complex features of the natural language. To each logic operator we associated a semantics based on derivation rules to reason about sets of sentences of the language — a calculus of sentences. Finally, we considered proof trees as a means to solve in a uniform way puzzles and language paradoxes from a few popular science books on logic by Raymond M. Smullyan.

The idea is to endow the student with the capacity of solving language problems with the help of automatic reasoning. This method reduces the difficulty of solving some classes of Smullyan puzzles to a modeling issue.

The extension of the method to further puzzles is left to the interested student.

References

Bell JL, Machover M (1977) A course in mathematical logic. North-Holland Publishing Co., Amsterdam
Gardner M (1978) Aha! insight. W. H. Freeman & Co., London
Gardner M (1985) Wheels, life and other mathematical amusements. W H Freeman & Co., London
Hofstadter DR (2000) Godel, escher, bach: an eternal golden braid, 20 Anniversary edn. Penguin, London
Hughes P, Brecht G (1993) Vicious circles and infinity, an anthology of paradoxes. Penguin Books, London
Leblanc H, Wisdom WA (1972) Deductive logic. Prentice Hall, Inc., Upper Saddle River
Sernadas A, Sernadas C (2008) Foundations of logic and theory of computation. College Publications, London
Sloane P, MacHale D (1992) Challenging lateral thinking puzzles. Sterling Publishing Company Inc., New York
Smullyan RM (1988) Forever undecided, puzzle guide to godel. Oxford Paperbacks, Oxford
Smullyan RM (2000) To Mock a Mockingbird and other logic puzzles. Oxford Paperbacks, Oxford
Smullyan RM (2008) Logical labyrinths. A. K. Peters/CRC Press, Natick
Smullyan RM (2009) The lady or the tiger? And other logic puzzles. Dover Publications Inc., Mineola
Smullyan RM (2010) King Arthur in search of his dog and other curious puzzles. Dover Publications Inc., Mineola
Smullyan RM (2011) What is the name of this book? The riddle of dracula and other logical puzzles. Dover Publications Inc., Mineola
Smullyan RM (2012) Alice in puzzle-land: a Carrollian tale for children under eighty. Dover Publications Inc., Mineola
Smullyan RM (2012) Satan, cantor & infinity: mind-boggling puzzles, New edition edn. Dover Publications Inc., Mineola
Smullyan RM (2013) The Gödelian puzzle book: puzzles, paradoxes and proofs. Dover Publications Inc., Mineola
Sukach J (1995) Quicksolve whodunit puzzles. Sterling Publishing Company Inc., New York
Sukach J (1997) Quicksolve whodunit puzzles. Sterling Publishing Company Inc., New York