Proceedings of the Workshop
“Logic & Cognition”

European Summer School in Logic, Language, and Information
Opole, Poland
13-17 August, 2012

edited by Jakub Szymanik and Rineke Verbrugge

Preface

1 The workshop theme

The roots of logic go back to antiquity, where it was mostly used as a tool for analyzing human argumentation. In the 19th century Gottlob Frege, one of the founders of modern logic and analytic philosophy, introduced anti-psychologism in the philosophy of mathematics. In the following years anti-psychologism, the view that the nature of mathematical truth is independent of human ideas, was one of the philosophical driving forces behind the success of mathematical logic. During the same period in the 19th century, also modern psychology (Helmholtz, Wundt) was born. However, the notion of anti-psychologism often stood in the way of a potential merge of the disciplines and led to a significant separation between logic and psychology research agendas and methods. Only since the 1960s, together with the growth of cognitive science inspired by the 'mind as computer' metaphor, the two disciplines have started to interact more and more. Today, we finally observe an increase in the collaborative effort between logicians, computer scientists, linguists, cognitive scientists, philosophers, and psychologists.

Topics of the workshop revolve around empirical research motivated by logical theories as well as logics inspired by experimental studies reflecting an increasing collaboration between logicians and cognitive scientists.

2 Invited talks at the workshop

In addition to the contributed talks, of which the articles are gathered in this volume, the workshop also presents two invited talks:

- Paul Egré and David Ripley *Vagueness and hysteresis: a case study in color categorization*
- Iris van Rooij *Rationality, intractability and the prospects of "as if" explanations*

The abstracts of both presentations may be found in this volume.

3 Best papers

The Programme Committee of the workshop decided to award the *Best Paper Prize* to: Nina Gierasimczuk, Han van der Maas, and Maartje Raijmakers for the paper *Logical and Psychological Analysis of Deductive Mastermind*.

The *Best Student Paper Prize* was awarded to Fabian Schlotterbeck and Oliver Bott for the paper *Easy Solutions for a Hard Problem? The Computational Complexity of Reciprocals with Quantificational Antecedents.*

4 Acknowledgements

We would like to thank all the people who helped to bring about the workshop *Logic and Cognition*. First of all, we thank all invited speakers and contributed speakers for ensuring an interesting conference.

Special thanks are due to the members of the program committee for their professionalism and their dedication to select papers of quality and to provide authors with useful, constructive feedback during the in-depth reviewing process:

Program Committee:

Leon de Bruin
Eve Clark
Robin Clark
Paul Egré
Fritz Hamm
Alice ter Meulen
Marcin Mostowski
Maartje Raijmakers
Iris van Rooij
Keith Stenning
Marcin Zajenkowski

In addition, Catarina Dutilh Novaes and Raphael van Riel shared their expertise to help review the papers. Thanks! We would also like to thank the organizers of the European Summer School in Logic, Language and Information, Opole, 2012, that which hosts our workshop.

Finally, we would like to express our gratitude to NWO for largely financing this workshop through Vici grant NWO 227-80-00, Cognitive Systems in Interaction: Logical and Computational Models of Higher-Order Social Cognition.

Jakub Szymanik
Rineke Verbrugge
Groningen, July 2012

Vagueness and hysteresis: a case study in color categorization

Paul Egré and David Ripley

Institut Jean-Nicod (CNRS), Paris

Abstract. This paper presents the first results of an experimental study concerning the semantic status of borderline cases of vague predicates. Our focus is on the particular case of color predicates (such as 'yellow', 'orange', 'blue', 'green'), and on the influence of context in the categorization of color shades at the border between two color categories. In an unpublished study, D. Raffman and colleagues found that subjects have no difficulty in categorizing the same color shade as 'blue' or 'green' depending on the direction of the transition between the two categories, suggesting a phenomenon of hysteresis or persistence of the initial category. Hysteresis is a particularly intriguing phenomenon for vagueness for two reasons: i) it seems to comport with the tolerance principle, which says that once applied, a category can be extended to cases that differ very little ii) it seems to suggest that borderline cases are cases of overlap, rather than underlap between semantic categories (see Raffman 2009, Egré 2011, Ripley 2012). In our first study, we probed for hysteresis in two different tasks: in the first, subjects had to perform a task of color matching, namely to decide of each shade in a series between yellow and orange (respectively blue and green) whether it was more similar to the most yellow or to the most orange kept on the display. In the second task, subjects had to decide which of the two color labels 'yellow' or 'orange' was the most suitable. Shades were presented in three different orders, random, ascending from yellow to orange, and descending. While we found no order effect in the perceptual matching task, we found an effect of negative hysteresis in the linguistic task in each color set, namely subjects switched category at a smaller position rather than at a later position depending on the order. In a second study, we used the same design but asked subjects to report agreement or disagreement with various sentential descriptions of the shade (viz. 'the shade is yellow/not yellow/yellow and not yellow'). No order effect was found in that task. These findings raise two particular issues concerning the boundaries of vague semantic categories, which we discuss in turn: the first concerns the interpretation of negative, as opposed to positive hysteresis. Another concerns the sensitivity of order effects to the task.

This is a joint work with Vincent de Gardelle.

Rationality, intractability and the prospects of 'as if' explanations

Iris van Rooij

Radboud University Nijmegen
Donders Institute for Brain, Cognition and Behavior

Abstract. Proponents of a probabilistic (Bayesian) turn in the study of human cognition have used the intractability of (non-monotonic) logics to argue against the feasibility of logicist characterizations of human rationality. It is known, however, that probabilistic computations are generally intractable as well. Bayesians have argued that, in their own case, this is merely as pseudoproblem. Their argument is that humans do not really perform the probabilistic calculations prescribed by probability theory, but only act as if they do–much like the planets do not calculate their own orbits, and birds fly without any knowledge of the theory of aerodynamics.

The prospects of such an 'as if' explanation dissolving the intractability problem depends inter alia on what is meant by 'as if'. I analyze some of the most plausible meanings that are compatible with various statements in the literature, and argue that none of them circumvents the problem of intractability.

The analysis will show that, even though the constraints imposed by tractability may prove pivotal for determining adequate characterizations of human rationality, these constraints do not directly favor one type of formalism over another. Cognitive science would be better off realizing this and putting efforts into dealing with the problem of intractability head-on, rather than playing a shell game.

This is joint work with:
Cory Wright (University of California, Long Beach, USA),
Johan Kwisthout (Radboud University Nijmegen, The Netherlands),
Todd Wareham (Memorial University of Newfoundland, Canada).

Logical and Psychological Analysis of Deductive Mastermind

Nina Gierasimczuk[1], Han van der Maas[2], and Maartje Raijmakers[2]

[1] Institute for Logic, Language and Computation, University of Amsterdam,
Nina.Gierasimczuk@gmail.com,
[2] Department of Psychology, University of Amsterdam

Abstract. The paper proposes a way to analyze logical reasoning in a deductive version of the Mastermind game implemented within the Math Garden educational system. Our main goal is to derive predictions about the cognitive difficulty of game-plays, e.g., the number of steps needed for solving the logical tasks or the working memory load. Our model is based on the analytic tableaux method, known from proof theory. We associate the difficulty of the Deductive Mastermind game-items with the size of the corresponding logical tree derived by the tableau method. We discuss possible empirical hypotheses based on this model, and preliminary results that prove the relevance of our theory.

Keywords: Deductive Mastermind, Mastermind game, deductive reasoning, analytic tableaux, Math Garden, educational tools

1 Introduction and Background

Computational and logical analysis has already proven useful for the investigations into the cognitive difficulty of linguistic and communicative tasks (see [1–5]). We follow this line of research by adapting formal logical tools to directly analyze the difficulty of non-linguistic logical reasoning. Our object of study is a deductive version of the *Mastermind* game. Although the game has been used to investigate the acquisition of complex skills and strategies in the domain of reasoning about others [6], as far as we know, it has not yet been studied for the difficulty of the accompanying deductive reasoning. Our model of reasoning is based on the proof-theoretical method of analytic tableaux for propositional logic, and it gives predictions about the empirical difficulty of game-items. In the remaining part of this section we give the background of our work: we explain the classical and static versions of the Mastermind game, and describe the main principles of the online Math Garden system. In Section 2 we introduce Deductive Mastermind as implemented within Math Garden. Section 3 gives a logical analysis of Deductive Mastermind game-items using the tableau method. Finally, Section 4 discusses some hypotheses drawn on the basis of our model, and gives preliminary results. In the end we briefly discuss the directions for further work.

1.1 Mastermind Game

Mastermind is a code-breaking game for two players. The modern game with pegs was invented in 1970 by Mordecai Meirowitz, but the game resembles an earlier pen and paper game called Bulls and Cows. The Mastermind game, as known today, consists of a decoding board, code pegs of k colors, and feedback pegs of black and white. There are two players, the code-maker, who chooses a secret pattern of ℓ code pegs (color duplicates are allowed), and the code-breaker, who guesses the pattern, in a given n rounds. Each round consists of code-breaker making a guess by placing a row of ℓ code pegs, and of code-maker providing the feedback of zero to ℓ key pegs: a black key for each code peg of correct color and position, and a white key for each peg of correct color but wrong position. After that another guess is made. Guesses and feedbacks continue to alternate until either the code-breaker guesses correctly, or n incorrect guesses have been made. The code-breaker wins if he obtains the solution within n rounds; the code-maker wins otherwise. Mastermind is an *inductive inquiry* game that involves *trials of experimentation and evaluation.* As such it leads to the interesting question of the underlying logical reasoning and its difficulty. Existing mathematical results on Mastermind do not provide any answer to this problem—most focus has been directed at finding a strategy that allows winning the game in the smallest number of rounds (see [7–10]).

Static Mastermind [11] is a version of the Mastermind game in which the goal is to find the minimum number of guesses the code-breaker can make all at once at the beginning of the game (without waiting for the individual feedbacks), and upon receiving them all at once completely determine the code in the next guess. In the case of this game some strategy analysis has been conducted [12], but, more importantly, Static Mastermind has been given a computational complexity analysis [13]. The corresponding Static Mastermind Satisfiability Decision Problem has been defined in the following way.

Definition 1 (Mastermind Satisfiability Decision Problem).

Input *A set of guesses G and their corresponding scores.*
Question *Is there at least one valid solution?*

Theorem 1. *Mastermind Satisfiability Decision Problem is NP-complete.*

This result gives an objective computational measure of the difficulty of the task. NP-complete problems are believed to be cognitively hard [14–16], hence this result has been claimed to indicate why Mastermind is an engaging and popular game. It does not however give much insight into the difficulty of reasoning that might take place while playing the game, and constructing the solution.

1.2 Math Garden

This work has been triggered by the idea of introducing a dedicated logical reasoning training in primary schools through the online Math Garden system

(Rekentuin.nl or MathsGarden.com, see [17]). Math Garden is an adaptive training environment in which students can play various educational games especially designed to facilitate the development of their abstract thinking. Currently, it consists of 15 arithmetic games and 2 complex reasoning games. Students play the game-items suited for their level. A difficultly level is appropriate for a student if she is able to solve 70% items of this level correctly. The difficulty of tasks and the level of the students' play are being continuously estimated according to the Elo rating system, which is used for calculating the relative skill levels of players in two-player games such as chess [18]. Here, the relative skill level is computed on the basis of student v. game-item opposition: students are rated by playing, and items are rated by getting played. The rating depends on accuracy and speed of problem solving [19]. The rating scales go from − infinity to + infinity. If a child has the same rating as an item this means that the child can solve the item with probability .5. In general, the absolute values of the ratings have no straightforward meaning. Hence, one result of children playing within Math Garden is a rating of all items, which gives item difficulty parameters. At the same time every child gets a rating that reflects her or his reasoning ability. One of the goals of this project is to understand the empirically established item difficulty parameters by means of a logical analysis of the items. Figure 5 in the Appendix depicts the educational and research context of Math Garden.

2 Deductive Mastermind

Now let us turn to Deductive Mastermind, the game that we designed for the Math Garden system (its corresponding name within Math Garden is *Flowercode*). Figure 1 shows an example item, revealing the basic setting of the game. Each item consists of a decoding board (1), short feedback instruction (2), the domain of flowers to choose from while constructing the solution (3) and the timer in the form of disappearing coins (4). The goal of the game is to guess the right sequence of flowers on the basis of the clues given on the decoding board. Each row of flowers forms a conjecture that is accompanied by a small board on the right side of it. The dots on the board code the feedback about the conjecture: one green dot for each flower of correct color and position, one orange dot for each flower of correct color but wrong position, and one red dot for each flower that does not appear in the secret sequence at all. In order to win, the player is supposed to pick the right flowers from the given domain (3), place them in the right order on the decoding board, right under the clues, and press the OK button. She must do so before the time is up, i.e., before all the coins (4) disappear. If the guess is correct, the number of coins that were left as she made the guess is added to her overall score. If her guess is wrong, the same number is subtracted from it. In this way making fast guesses is punished, but making a fast correct response is encouraged [19].

Unlike the classical version of the Mastermind game, Deductive Mastermind does not require the player to come up with the trial conjectures. Instead, Flowercode gives the clues directly, ensuring that they allow *exactly one correct solution.*

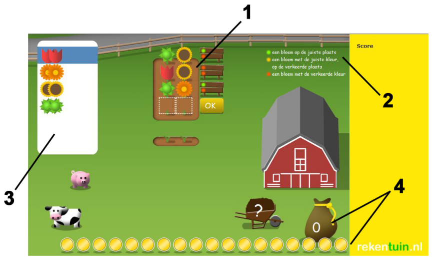

1
2
3
4
Score
● een bloem op de juiste plaats
● een bloem met de juiste kleur, op de verkeerde plaats
● een bloem met de verkeerde kleur
OK
?
0
rekentuin.nl

Fig. 1. Deductive Mastermind in the Flowercode setting

Hence, Deductive Mastermind reduces the complexity of classical Mastermind by changing from an inductive inference game into a very basic logical-reasoning game. On the other hand, when compared to Static Mastermind, Deductive Mastermind differs with respect to the goal. In fact, by guaranteeing the existence of exactly one correct solution, Deductive Mastermind collapses the postulated complexity from Theorem 1, since the question of the Static Mastermind Satisfiability Problem becomes void. It is fair to say that Deductive Mastermind is a combination of the classical Mastermind game (the goal of the game is the same: finding a secret code) and the Static Mastermind game (it does not involve the trial-and-error inductive inference experimentation). Its very basic setting allows access to atomic logical steps of non-linguistic logical reasoning. Moreover, Deductive Mastermind is easily adaptable as a single-player game, and hence suitable for the Math Garden system. The simple setting provides educational justification, as the game trains very basic logical skills.

The game has been running within the Math Garden system since November 2010. It includes 321 game-items, with conjectures of various lengths (1-5 flowers) and number of colors (from 2 to 5). By January 2012, 2,187,354 items had been played by 28,247 primary school students (grades 1-6, age: 6-12 years) in over 400 locations (schools and family homes). This extensive data-collecting process allows analyzing various aspects of training, e.g., we can access the individual progress of individual players on a single game, or the most frequent mistakes with respect to a game-item. Most importantly, due to the student-item rating system mentioned in Section 1.2, we can infer the relative difficulty of game-items. Within this hierarchy we observed that the present game-item domain contains certain "gaps in difficulty"—it turns out that our initial difficulty estimation in terms of non-logical aspects (e.g., number of flowers, number of colors, number of lines, the rate of the hypotheses elimination, etc.) is not precise, and hence the domain of items that we generated does not cover the whole difficulty space. Providing a logical apparatus that predicts and explains the difficulty of Deductive Mastermind game-items can help solving this problem and hence also facilitate the training effect of Flowercode (see Appendix, Figure 7).

3 A Logical Analysis

Each Deductive Mastermind game-item consists of a sequence of conjectures.

Definition 2. *A* conjecture *of length l over k colors is any sequence given by a total assignment, $h : \{1,\dots,\ell\} \to \{c_1,\dots,c_k\}$. The* goal sequence *is a distinguished conjecture, $goal : \{1,\dots,\ell\} \to \{c_1,\dots,c_k\}$.*

Every non-goal conjecture is accompanied by a feedback that indicates how similar h is to the given goal assignment. The three feedback colors: green, orange, and red, described in Section 2, will be represented by letters g, o, r.

Definition 3. *Let h be a conjecture and let goal be the goal sequence, both of length l over k colors. The* feedback f for h with respect to *goal is a sequence*

$$\overbrace{g\dots g}^{a}\;\overbrace{o\dots o}^{b}\;\overbrace{r\dots r}^{c} = g^a o^b r^c,$$

where $a,b,c \in \{0,1,2,3,\dots\}$ and $a+b+c=\ell$. The feedback consists of:

- *exactly one g for each $i \in G$, where $G = \{i \in \{1,\dots\ell\} \mid h(i) = goal(i)\}$.*
- *exactly one o for every $i \in O$, where $O = \{i \in \{1,\dots,\ell\}\backslash G \mid$ there is a $j \in \{1,\dots,\ell\}\backslash G,$ such that $i \neq j$and $h(i) = goal(j)\}$.*
- *exactly one r for every $i \in \{1,\dots,\ell\}\backslash(G \cup O)$.*

3.1 The informational content of the feedback

How to logically express the information carried by each pair (h,f)? To shape the intuitions let us first give a second-order logic formula that encodes any feedback sequence $g^a o^b r^c$ for any h with respect to any *goal*:

$$\begin{aligned}\exists G{\subseteq}\{1,\dots\ell\}(card(G){=}a \wedge \forall i{\in}G\; h(i){=}goal(i) \wedge \forall i{\notin}G\; h(i){\neq}goal(i)\\ \wedge\; \exists O{\subseteq}\{1,\dots\ell\}\backslash G\; (card(O){=}b \wedge \forall i{\in}O\; \exists j{\in}\{1,\dots\ell\}\backslash G(j{\neq}i \wedge h(i){=}goal(j))\\ \wedge\; \forall i{\in}\{1,\dots\ell\}\backslash(G{\cup}O)\; \forall j{\in}\{1,\dots\ell\}\backslash G\; h(i){\neq}goal(j))).\end{aligned}$$

Since the conjecture length, ℓ, is fixed for any game-item, it seems sensible to give a general method of providing a less engaging, propositional formula for any instance of (h,f). As literals of our Boolean formulae we take $h(i) = goal(j)$, where $i,j \in \{1,\dots\ell\}$ (they might be viewed as propositional variables $p_{i,j}$, for $i,j \in \{1,\dots\ell\}$). With respect to sets G, O, and R that induce a partition of $\{1,\dots\ell\}$, we define $\varphi^g_G, \varphi^o_{G,O}, \varphi^r_{G,O}$, the propositional formulae that correspond to different parts of feedback, in the following way:

- $\varphi^g_G := \bigwedge_{i\in G} h(i){=}goal(i) \wedge \bigwedge_{j\in\{1,\dots,\ell\}\backslash G} h(j){\neq}goal(j)$,
- $\varphi^o_{G,O} := \bigwedge_{i\in O}(\bigvee_{j\in\{1,\dots,\ell\}\backslash G, i\neq j} h(i) = goal(j))$,
- $\varphi^r_{G,O} := \bigwedge_{i\in\{1,\dots\ell\}\backslash(G\cup O), j\in\{1,\dots\ell\}\backslash G, i\neq j} h(i) \neq goal(j)$.

Observe that there will be as many substitutions of each of the above schemes of formulae, as there are ways to choose the corresponding sets G and O. To fix the domain of those possibilities we set $\mathbb{G} := \{G | G \subseteq \{1, \dots, \ell\} \wedge card(G){=}a\}$, and, if $G \subseteq \{1, \dots, \ell\}$, then $\mathbb{O}^G = \{O | O \subseteq \{1, \dots, \ell\} \backslash G \wedge card(O){=}b\}$. Finally, we can set $Bt(h, f)$, the Boolean translation of (h, f), to be given by:

$$Bt(h,f) := \bigvee_{G \in \mathbb{G}} (\varphi^g_G \wedge \bigvee_{O \in \mathbb{O}^G} (\varphi^o_{G,O} \wedge \varphi^r_{G,O})).$$

Example 1. Let us take $\ell = 2$ and (h, f) such that: $h(1){:=}c_1$, $h(2){:=}c_2$; $f{:=}or$. Then $\mathbb{G}{=}\{\emptyset\}$, $\mathbb{O}^{\{\emptyset\}}{=}\{\{1\},\{2\}\}$. The corresponding formula, $Bt(h, f)$, is:

$$(goal(1){\neq}c_1 \wedge goal(2){\neq}c_2) \wedge ((goal(1){=}c_2 \wedge goal(2){\neq}c_1) \vee (goal(2){=}c_1 \wedge goal(1){\neq}c_2))$$

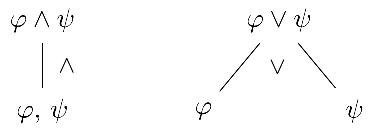
Each Deductive Mastermind game-item consists of a sequence of conjectures together with their respective feedbacks. Let us define it formally.

Definition 4. A Deductive Mastermind game-item *over ℓ positions, k colors and n lines, $DM(l, k, n)$, is a set $\{(h_1, f_1), \dots, (h_n, f_n)\}$ of pairs, each consisting of a single conjecture together with its corresponding feedback. Respectively, $Bt(DM(l, k, n)) = Bt(\{(h_1, f_1), \dots, (h_n, f_n)\}) = \{Bt(h_1, f_1), \dots, Bt(h_n, f_n)\}$.*

Hence, each Deductive Mastermind game-item can be viewed as a set of Boolean formulae. Moreover, by the construction of the game-items we have that this set is satisfiable, and that there is a unique valuation that satisfies it. Now let us focus on a method of finding this valuation.

3.2 Analytic Tableaux for Deductive Mastermind

In proof theory, the analytic tableau is a decision procedure for propositional logic [20–22]. The tableau method can determine the satisfiability of finite sets of formulas of propositional logic by giving an adequate valuation. The method builds a formulae-labeled tree rooted at the given set of formulae and unfolding breaks these formulae into smaller formulae until contradictory pairs of literals are produced or no further reduction is possible. The rules of analytic tableaux for propositional logic, that are relevant for our considerations are as follows.[3]



By our considerations from Section 3 we can now conclude that applying the analytic tableaux method to the Boolean translation of a Deductive Mastermind game-item will give the unique missing assignment *goal*. In the rest of the paper we will focus on 2-pin Deductive Mastermind game-items (where $\ell = 2$), in particular, we will explain the tableau method in more detail on those simple examples.

[3] We do not need the rule for negation because in our formulae only literals are negated.

2-pin Deductive Mastermind Game-items Since the possible feedbacks consist of letters g (green), o (orange), and r (red), in principle for the 2-pin Deductive Mastermind game-items we get six possible feedbacks: gg, oo, rr, go, gr, or. From those: gg is excluded as non-applicable; go is excluded because there are only two positions. Let us take a pair (h, f), where $h(1){=}c_i$, $h(2){=}c_j$, then, depending on the feedback, the corresponding boolean formulae are given in Figure 2. We can compare the complexity of those feedbacks just by looking at their tree representations created from the boolean translations via the tableau method. Those representations clearly indicate that the order of the tree-difficulty for the four possible feedbacks is: $oo{<}rr{<}gr{<}or$. As the feedbacks oo, rr are conjunctions, they do not require branching (the other two include disjunctions, and as such demand reasoning by cases). Unlike rr, the feedback oo in fact gives the solution immediately. Within the two remaining rules, gr requires less memory to store the information in each branch.

Feedback	**Boolean translation**
oo	$goal(1) \neq c_i \wedge goal(2) \neq c_j \wedge goal(1) = c_j \wedge goal(2) = c_i$
rr	$goal(1) \neq c_i \wedge goal(2) \neq c_j \wedge goal(1) \neq c_j \wedge goal(2) \neq c_i$
gr	$(goal(1) = c_i \wedge goal(2) \neq c_j) \vee (goal(2) = c_j \wedge goal(1) \neq c_i)$
or	$(goal(1) \neq c_i \wedge goal(2) \neq c_j) \wedge$ $((goal(1) = c_j \wedge goal(2) \neq c_i) \vee (goal(2) = c_i \wedge goal(1) \neq c_j))$

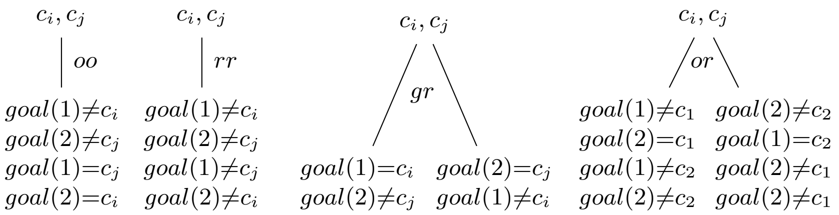


Fig. 2. Formulae and their trees for 2-pin Deductive Mastermind feedbacks

We will now briefly discuss the tableau method on an example. Let us consider the following Deductive Mastermind game-item (Figure 3). The tree on the left stands for the reasoning which corresponds to analyzing the conjectures as given, from top to bottom. The first branching gives the result of applying the gr feedback. In the next level of the tree we apply the oo feedback to the second conjecture. We must first do so assuming the left branch of the first conjecture to be true. This leads to a contradiction—on this branch we get that $goal(2){=}c_1$ and $goal(2){\neq}c_1$. Then we move to the right branch of the first conjecture. This assumption leads to discovering the right assignment, $goal(1){=}c_2$ and $goal(2){=}c_1$, there is no contradiction on this branch. This tableau procedure is required to build the whole tree for the game-item. That is not always necessary. The right-most part of Figure 3 shows what happens if you chose to start the analysis from the second conjecture. We first apply the feedback oo to the second conjecture. We immediately get: $goal(1){=}c_2$ $goal(2){=}c_1$. The full

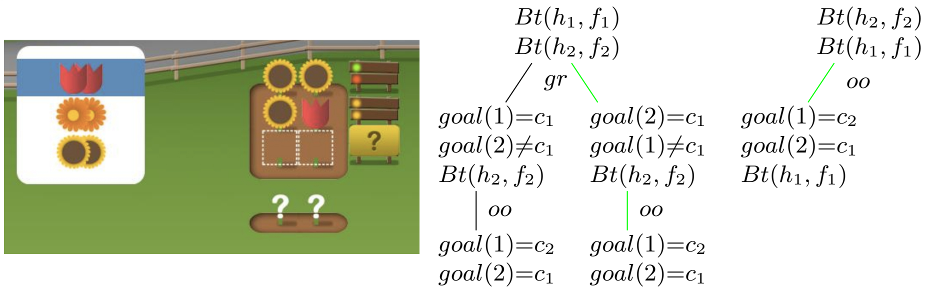


Fig. 3. Comparison of two different trees for one item. The formalization is as follows: c_1 stands for sunflower, c_2 for tulip; $h_1(1)=c_1, h_1(2)=c_1, f_1=gr$, etc. Green branch gives the right valuation. The tree on the right hand-side analyzes feedback *oo* first.

unique assignment with no contradiction. We can stop the computation at this point—if the other conjecture contradicted this assignment, then it would mean that the two conjectures must be inconsistent and hence not satisfiable. This contradicts the setting of our game.

The tree might not always give us the complete valuation explicitly. In some game-items it is required to use a flower that did not appear in the conjectures. This is so in the example in Figure 4; the right-most branch does not give a contradiction, it does assign color 1 to the second position of the goal conjecture. In such case we draw the remaining color, c_3 (a tulip in the picture), as the missing value of the first position of the goal conjecture, i.e., $goal(1)=c_3$.

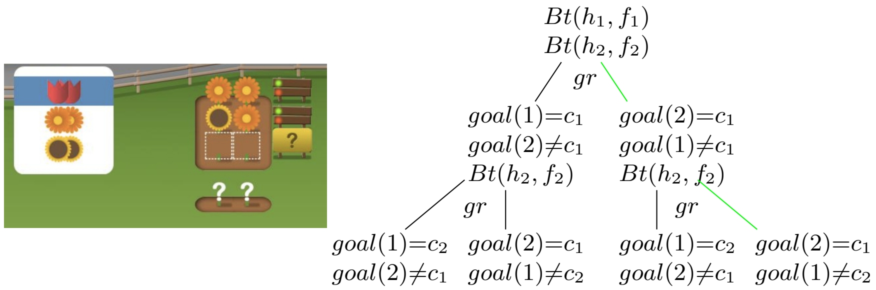


Fig. 4. Comparison of the trees of two different items, while processing the conjectures from top to bottom. Here, c_1 stands for marguerite, c_2 for sunflower, c_3 for tulip. This item rating is 4.2 (see Figure 6).

4 Hypotheses and Preliminary Results

Normatively speaking, the full tree generated by the tableau method for the set of formulae corresponding to a Deductive Mastermind game-item gives its ideal reasoning scheme and thus the size of the tree can be thought of as an abstract complexity measure. Obviously, the shape and the size of the tree for each Deductive Mastermind item depends to some extent on the order in which the formulae are analyzed (see Figure 3). Hence, using the tableau method it is even possible to analyze whether and in what way the students apply reasoning strategies, i.e., how they manipulate with the elements of the task in order to optimize the size of the reasoning tree (i.e., the length of the computation). In this way, items' logical difficulty can be expressed via the size of their minimal trees. The empirical data, resulting from children playing the Deductive Mastermind game in Math Garden, includes item ratings. In the first analysis of this data we aimed at relating the item ratings to the parameters of the trees. To this end we define a computational method based on the tableau method. The computational method makes two assumptions. First, the formulae are not processed from top to bottom, but instead the order depends on the length of the rule that is associated with the feedback. That is, feedback is processed in the following order: oo, rr, gr, or. Ties are solved by processing the top formula first. Second, the computational method is stopped once a consistent solution is found, assuming that there exists at least one solution. Based on these principles and the tableau method we programmed a recursive algorithm for calculating the type and number of steps until solution for each item is reached. We define 4 measures of theoretically derived item difficulties; the required number of oo, rr, gr, and rr steps, which together might predict item difficulty. Below we will describe the empirically derived item ratings of the 2-pin items and we will show how these relate to the theoretically derived measures of item difficulties.

Method Participants were 28,247 students from grades 1-6, of age: 6-12 years. Together, they played 2,187,354 items between November 2010 and January 2012. From the total of 321 items in the Master Mind game, 100 items have two pins. From these 100 items, 10 items involved 2 colors, 30 items involved 3 colors, 30 items involved 4 colors and 30 items involved 5 colors.

Results To test the relation between empirically derived item ratings and theoretically derived measures of item difficulty we did a regression analysis that includes the basic features of the items (number of colors and number of guesses) and the required number of *oo*, *rr*, *gr*, and *or* analysis steps as predicted by the tableau-based computational algorithm ($F(6,93)$=33, p < .0001, $R2$=.66). All these factors but one (i.e., number of *gr* evaluations) were significant in predicting item difficulties: number of colours (β=1.07, p=.02), number of hypotheses (β=1.75, p<.01), number of *oo* feedbacks ($\beta = -5.1$, p<.001), number of *rr* feedbacks ($\beta = -3.19, p$<.0001), number of *gr* evaluations (ns), number of *or* evaluations (β=1.6, p<.0001). Note that among the rules, only the required number of or steps increases item difficulty. A second aspect of the observed item ratings to explain is the shape of its distribution. The distribution of the item difficulties shows a remarkable property for the 2-pin items (see Appendix, Figure 6). The distribution is bimodal, meaning that there is a cluster of more difficult items (item ratings >0) and a cluster of easier items (item ratings <0). In the cluster of easy items, there are items with 2, 3, 4, and 5 colors. The cluster of difficult items consists of items with 3, 4 and 5 colors. The two clusters could be expressed fully in terms of model-relevant features. It appeared that items are easy in the following two cases: (1) no *or* feedback and no *gr* feedback; (2) no *or* feedback, at least one *gr* feedback, and all colors are included in at least one of the conjectures. Items are difficult otherwise. This shows that or steps make the item relatively difficult, but only if or is required to solve the item. A second aspect that makes an item difficult is the exclusion of one of the colors in the solution from the hypotheses rows.

5 Conclusions and Future Work

In this paper we proposed a way to use a proof-theoretical concept to analyze the cognitive difficulty of logical reasoning. The first hypotheses that we have drawn from the model gave a reasonably good prediction of item-difficulties, but the fit is not perfect. However, it must be noted that several non-logical factors may play a role in the item difficulty as well. For example, the motor requirements to give the answer also introduce some variation in item difficulty, which depends not only on accuracy but also on speed. The minimal number of clicks required to give an answer varies between items. We did not take these aspects into account so far. In particular, the two difficulty clusters observed in the empirical study can be explained with the use of our method.

Our further work can be split into two main parts. We will continue this study on the theoretical level by analyzing various complexity measures that can be obtained on the basis of the tableau system. We also plan to compare the fit with empirical data of the tableau-derived measures and other possible logical formalizations of level difficulty (e.g., a resolution-based model). On the empirical side we first would like to extend our analysis to game-items that consist of conjectures of higher length. This will allow comparing difficulty of tasks of different size and measure the trade-off between the size and the structure of the tasks. We will also study this model in a broader context of other Math Garden games. This would allow comparing individual abilities within Deductive Mastermind and the abilities within other games that have been correlated for instance with the working memory load. Finally, it would be interesting to see whether the children are really using the proposed difficulty order of feedbacks in finding an optimal tableau—this could be checked in an eye-tracking experiment (see [23, 24]).

References

1. Szymanik, J.: Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language. PhD thesis, Universiteit Van Amsterdam (2009)
2. Gierasimczuk, N., Szymanik, J.: Branching quantification v. two-way quantification. Journal of Semantics **26**(4) (2009) 367–392
3. Szymanik, J., Zajenkowski, M.: Comprehension of simple quantifiers. Empirical evaluation of a computational model. Cognitive Science **34**(3) (2010) 521–532
4. Zajenkowski, M., Styła, R., Szymanik, J.: A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. Journal of Communication Disorders **44**(6) (2011) 595–600

5. Van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., Toni, I.: Intentional communication: Computationally easy or difficult? Frontiers in Human Neuroscience **5** (2011) 1–18
6. Verbrugge, R., Mol, L.: Learning to apply theory of mind. Journal of Logic, Language and Information **17**(4) (2008) 489–511
7. Knuth, D.E.: The computer as master mind. Journal of Recreational Mathematics **9**(1) (1977) 1–6
8. Irving, R.W.: Towards an optimum Mastermind strategy. Journal of Recreational Mathematics **11** (1978-79) 81–87
9. Koyama, M., Lai, T.: An optimal Mastermind strategy. Journal of Recreational Mathematics **25** (1993) 251—256
10. Kooi, B.: Yet another Mastermind strategy. ICGA Journal **28**(1) (2005) 13–20
11. Chvatal, V.: Mastermind. Combinatorica **325–329** (1983)
12. Greenwell, D.L.: Mastermind. Journal of Recreational Mathematics **30** (1999-2000) 191–192
13. Stuckman, J., Zhang, G.: Mastermind is NP-complete. INFOCOMP Journal of Computer Science **5** (2006) 25–28
14. Van Rooij, I.: The tractable cognition thesis. Cognitive Science **32** (2008) 939–984
15. Szymanik, J.: Computational complexity of polyadic lifts of generalized quantifiers in natural language. Linguistics and Philosophy **33**(3) (2010) 215–250
16. Ristad, E.S.: The Language Complexity Game. The MIT Press (1993)
17. Van der Maas, H.L.J., Klinkenberg, S., Straatemeier, M.: Rekentuin.nl: combinatie Van oefenen en toetsen. Examens (2010) 10–14
18. Elo, A.: The Rating of Chessplayers, Past and Present. Arco (1978)
19. Klinkenberg, S., Straatemeier, M., Van der Maas, H.L.J.: Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. Computers in Education **57** (2011) 1813–1824
20. Beth, E.W.: Semantic entailment and formal derivability. Mededelingen Van de Koninklijke Nederlandse Akademie Van Wetenschappen. Afdeling Letterkunde **18(13)** (1955) 309–342
21. Smullyan, R.: First-order logic. Springer-Verlag, Berlin (1968)
22. Van Benthem, J.: Semantic tableaus. Nieuw Archief voor Wiskunde **22** (1974) 44–59
23. Ghosh, S., Meijering, B., Verbrugge, R.: Logic meets cognition: Empirical reasoning in games. In Boissier, O., Fallah-Seghrouchni, A.E., Hassas, S., Maudet, N., eds.: MALLOW, CUER Workshop Proceedings (2010)
24. Ghosh, S., Meijering, B.: On combining cognitive and formal modeling: a case study involving strategic reasoning. In Van Eijck, J., Verbrugge, R., eds.: Proceedings of the Workshop on Reasoning About Other Minds: Logical and Cognitive Perspectives (RAOM-2011), Groningen, CEUR Workshop Proceedings (2011)

Appendix

The appendix contains three figures. The first one illustrates the educational and research context of the Math Garden system (Figure 5); the second shows the distribution of the empirical difficulty of all items in Deductive Mastermind (Figure 6); and the third gives the frequency of players and game-items with respect to the overall rating (Figure 7). We refer to those figures in the proper text of the article.

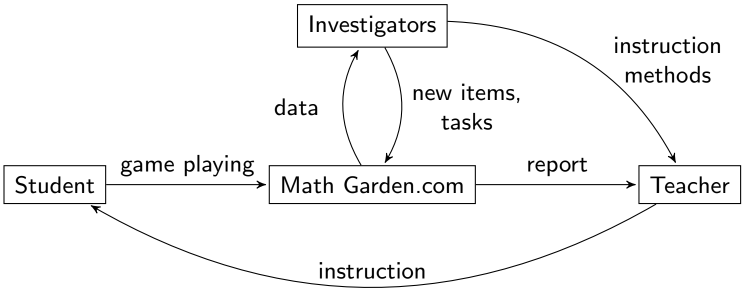


Fig. 5. Math Garden educational and research context

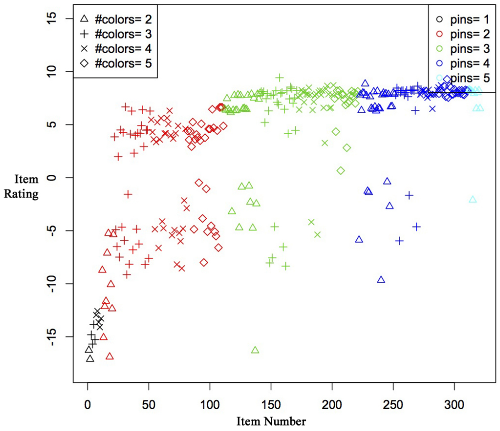


Fig. 6. The distribution of the empirically established game-item difficulties for all 321 game-items in the Deductive Mastermind game. The item number (x-axis) is some arbitrary number of the game-item. The y-axis shows the item ratings. For example, the item presented in Figure 4 is number 23 and its rating is 4.2.

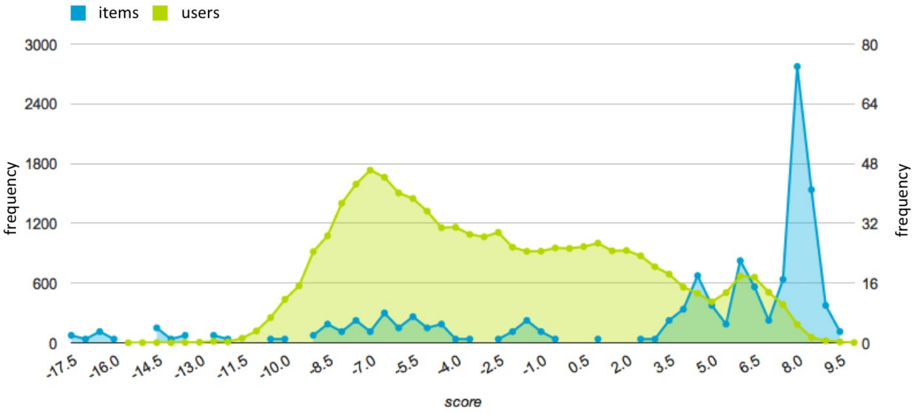


Fig. 7. The frequency of players (green, y-axis on the left) and game-items (blue, y-axis on the left) with respect to the overall rating (x-axis).

Euclid's Diagrammatic Logic and Cognitive Science

Yacin Hamami[1] and John Mumma[2]

[1] Centre for Logic and Philosophy of Science, Vrije Universiteit Brussel, Brussels
yacin.hamami@gmail.com
[2] Max Planck Institute for the History of Science, Berlin
john.mumma@gmail.com

Abstract. For more than two millennia, Euclid's *Elements* set the standard for rigorous mathematical reasoning. The reasoning practice the text embodied is essentially diagrammatic, and this aspect of it has been captured formally in a logical system termed **Eu** [2, 3]. In this paper, we review empirical and theoretical works in mathematical cognition and the psychology of reasoning in the light of **Eu**. We argue that cognitive intuitions of Euclidean geometry might play a role in the interpretation of diagrams, and we show that neither the mental rules nor the mental models approaches to reasoning constitutes by itself a good candidate for investigating geometrical reasoning. We conclude that a cognitive framework for investigating geometrical reasoning empirically will have to account for both the interpretation of diagrams and the reasoning with diagrammatic information. The framework developed by Stenning and van Lambalgen [1] is a good candidate for this purpose.

1 Introduction

A distinctive feature of elementary Euclidean geometry is the natural and intuitive character of its proofs. The prominent place of the subject within the history and education of mathematics attests to this. It was taken to be the foundational starting point for mathematics from the time of its birth in ancient Greece up until the 19th century. And it remains within mathematics education as a subject that serves to initiate students in the method of deductive proof. No other species of mathematical reasoning seem as basic and transparent as that which concerns the properties of figures in Euclidean space.

One may not expect a formal analysis of the reasoning to shed much light on this distinctive feature of it, as the formal and the intuitive are typically thought to oppose one another. Recently, however, a formal analysis, termed **Eu**, has been developed which challenges this opposition [2, 3]. **Eu** is a formal proof system designed to show that a systematic method underlies the use of diagrams in Euclid's *Elements*, the representative text of the classical synthetic tradition of geometry. As diagrams seem to be closely connected with the way we call upon our intuition in the proofs of the tradition, **Eu** holds the promise of contributing to our understanding of what exactly makes the proofs natural.

In this paper, we explore the potential **Eu** has in this respect by confronting it with empirical and theoretical works in the fields of mathematical cognition and the psychology of reasoning. Our investigation is organized around the two following issues:

1. What are the interpretative processes on diagrams involved in the reasoning practice of Euclidean geometry and what are their possible cognitive roots?
2. What would be an appropriate cognitive framework to represent and investigate the constructive and deductive aspects of the reasoning practice of Euclidean geometry?

By providing a formal model of the reasoning practice of Euclidean geometry, **Eu** provides us with a tool to address these two issues. We proceed as follows. To address the first issue, we first state the interpretative capacities that according to the norms fixed by **Eu** are necessary to extract, from diagrams, information for geometrical reasoning. We then present empirical works on the cognitive bases of Euclidean geometry, and suggest that cognitive intuitions might play a role in the interpretative aspects of diagrams in geometrical reasoning. To address the second issue, we compare the construction and inference rules of **Eu** with two major frameworks in the psychology of reasoning—the mental rules and the mental models theories. We argue that both have strengths and weaknesses as a cognitive account of geometrical reasoning as analyzed by **Eu**, but that one will need to go beyond them to provide a framework for investigating geometrical reasoning empirically.

The two main issues are of course intimately related. In a last section, we argue that the framework developed by Stenning and van Lambalgen in [1], which connects interpretative and deductive aspects of reasoning, might provide the right cognitive framework for investigating the relation between them.

2 A Logical Analysis of the Reasoning Practice in Euclid's *Elements*: The Formal System Eu

Eu is based on the seminal paper [4] by Ken Manders. In [4] Manders challenges the received view that the *Elements* is flawed because its proofs sometimes call upon geometric intuition rather than logic. What is left unexplained by the received view is the extraordinary longevity of the *Elements* as a foundational text within mathematics. For over two thousand years there were no serious challenges to its foundational status. Mathematicians through the centuries understood it to display what the basic propositions of geometry are grounded on. The deductive gaps that exist according to modern standards of logic story were simply not seen.

According to Manders, Euclid is not relying on geometric intuition illicitly in his proofs; he is rather employing a systematic method of diagrammatic proof. His analysis reveals that diagrams serve a principled, theoretical role in Euclid's mathematics. Only a restricted range of a diagram's spatial properties are permitted to justify inferences for Euclid, and these self-imposed restrictions can be explained as serving the purpose of mathematical control. **Eu** [2, 3] was designed to build on Manders' insights, and precisely characterize the mathematical significance of Euclid's diagrams in a formal system of geometric proof.

Eu has two symbol types: diagrammatic symbols $\varDelta$ and sentential symbols A. The sentential symbols A are defined as they are in first-order logic. They express relations between geometric magnitudes in a configuration. The diagrammatic symbols are defined as $n \times n$ arrays for any n. Rules for a well-formed diagram specify how points, lines and circles can be distinguished within such arrays. The points, lines and circles of Euclid's diagrams thus have formal analogues in **Eu** diagrams. The positions the elements of Euclid's diagrams can have to one another are modeled directly by the position their formal analogues can have within a **Eu** diagram.

The content of a diagram within a derivation is fixed via a relation of diagram equivalence. Roughly, two **Eu** diagrams $\varDelta_1$ and $\varDelta_2$ are equivalent if there is a bijection between its elements which preserves their non-metric positional relations.[3] The equivalence relation is intended to capture what Manders terms the *co-exact* properties of a Euclidean diagram. A close examination of the *Elements* shows that Euclid refrains from making any inferences that depend on a diagram's metric properties. At the same time, Euclid does rely on diagrams for the non-metric positional relations they exhibit—or in Manders' terms, the co-exact relations they exhibit. Diagrams, it turns out, can serve as adequate representations of such relations in proofs.

Eu exhibits this by depicting geometric proof as running on two tracks: a diagrammatic one, and a sentential one. The role of the sentential one is to record metric information about the figure and provide a means for inferring this kind of information. The role of the diagrammatic track is to record non-metric positional information of the figure, and to provide a means for inferring this kind of information about it. Rules for building and transforming diagrams in derivations are sensitive only to properties invariant under diagram equivalence. It is in this way that the relation of diagram equivalence fixes the content of diagrams in derivations.

What is derived in **Eu** are expressions of the form

$$\varDelta_1, A_1 \longrightarrow \varDelta_2, A_2$$

where $\varDelta_1$ and $\varDelta_2$ are diagrams, and A_1 and A_2 are sentences. The geometric claim this is stipulated to express is the following:

> $\star$*Given a configuration satisfying the non-metric positional relations depicted in $\varDelta_1$ and the metric relations expressed in A_1, then one can obtain a configuration satisfying the positional relations depicted by $\varDelta_2$ and metric relations specified by A_2.*

[3] For a more detailed discussion of **Eu** diagrams and diagram equivalence, we refer the reader respectively to sections A and B of the appendix.

3 Interpretative Aspects of Geometrical Reasoning with Diagrams

Diagrams in Euclid's *Elements* are mere pictures on a piece of paper. How can a visual experience triggered by looking at a picture lead to a cognitive representation that can play a role in geometrical reasoning? From a cognitive perspective, taking seriously the use of diagrams in reasoning requires an account of the way diagrams are *interpreted* in order to play a role in geometrical reasoning. The formal system **Eu** provides a theoretical answer to this question. Here we compare this theoretical account with recent works in mathematical cognition probing the existence of cognitive intuitions of Euclidean geometry. We will argue that the **Eu** analysis of geometrical reasoning suggests a possible role for these cognitive intuitions in the interpretation of diagrams.

3.1 Interpretation of Diagrams in Eu

As described in section 2, the key insight behind the **Eu** analysis of the diagrammatic reasoning practice of Euclid's *Elements* is the observation that appeals to diagrams in proofs are highly controlled: proofs in Euclid's *Elements* only make use of the *co-exact* properties of diagrams. From a cognitive perspective, the practice of extracting the co-exact properties from a visual diagram is far from a trivial one. According to **Eu**, the use of diagrams presupposes the two following cognitive abilities:

1. The capacity to categorize elements of the diagram using normative concepts: points, linear elements and circles.
2. The capacity to abstract away irrelevant information from the visual-perceptual experience of the diagram.

The formal syntax of **Eu**, combined with the equivalence relation between **Eu** diagrams, can be interpreted as specifying these capacities precisely. More specifically:

- The first capacity amounts to the ability to see in the diagram the elements p, l, c of a particular **Eu** diagram $\langle n, p, l, c \rangle$, which respectively denote the sets of points, lines and circles.
- The second capacity amounts to the ability to see in the particular diagram positional relations that are invariant under diagram equivalence.

Thus, the interpretation of a visual diagram according to **Eu** results in a formal object which consists in an equivalence class of **Eu** diagrams. It is precisely on these formal objects, the interpreted diagrams, that inference rules operate in the **Eu** formalization of reasoning in elementary geometry.

3.2 Intuitions of Euclidean Geometry in Human Cognition

Recently, several empirical studies in mathematical cognition [5–7] have directly addressed the issue of the cognitive roots of Euclidean geometry. These studies might be classified in two categories: one approach consists in providing empirical evidence for the existence of *abstract geometric concepts* [5, 7], the other approach consists in providing empirical evidence for the perception of *abstract geometric features* [6]. Here we successively report some of the empirical findings provided by these two approaches.

The two studies [5] and [7] have been conducted on an Amazonian Indigene Group called the Mundurucu, with no previous education in formal Euclidean geometry. In the first study [5], the experiment consisted in identifying the presence, or the absence, of several topological and geometric concepts in Mundurucu participants. To this end, the experimenter proposed, for each concept under investigation, six slides in which five of the images displayed the given geometric concept (e.g., parallelism), while the last one lacked the considered property (non-parallel lines). The test shows that several basic geometrical concepts are present in Mundurucu conceptual systems, such as the concepts of straight line, parallel line, right angle, parallelogram, equilateral triangle, circle, center of circle. Nevertheless, the study [5] does not address geometric concepts that go beyond perceptual experience. Such concepts are the topic of another study [7]. In [7], empirical evidence is provided for the capacity of Mundurucu to reason about the properties of dots and infinite lines. In particular, most of the Mundurucu participants consider that, given a straight line and a dot, we can always position another straight line on the dot which will never cross the initial line.

The second approach for detecting the presence of intuitions of Euclidean geometry is based on the framework of *transformational geometry* [8]. According to this framework, a geometric theory is identified with respect to the transformations that preserve its theorems. Euclidean geometry is then identified by its four types of transformations—translation, rotation, symmetry and homothecy—leaving then *angle* and *length proportions* as the main defining features of figures in Euclidean geometry. The experimental approach based on this framework consists in investigating the capacity of participants to perceive abstract geometric features in configurations where irrelevant features are varied. The experiments reported in [6] adopt such an approach. The empirical results show that both children and adults are able to use angle and size to classify shapes, but only adults are able to discriminate shapes with respect to sense, i.e., the property distinguishing two figures that are mirror images of each other.

The two approaches reported here to investigate cognitive intuitions of Euclidean geometry bring out cognitive abilities which seem related to the abilities for the interpretation of diagrams postulated by **Eu**. This observation suggests a possible role for the cognitive intuitions of Euclidean geometry in the reasoning practice of elementary geometry, as we will now see.

3.3 Cognitive Intuitions and the Interpretation of Diagrams

How might the cognitive intuitions of Euclidean geometry relate to the deductive aspects of Euclid's theory of geometry? In a correspondence in *Science* [9] on the study reported in [5], Karl Wulff has challenged the claim that Dehaene *et* *al.* address Euclidean geometry by denying a role for these cognitive intuitions in the demonstrative aspects of Euclid's theory of geometry. An opposing position can be found in [10]:

> The axioms of geometry introduced by Euclid [. . .] define concepts with spatial content, such that any theorem or demonstration of Euclidean geometry can be realized in the construction of a figure. Just as intuitions about numerosity may have inspired the early mathematicians to develop mathematical theories of number and arithmetic, Euclid may have appealed to universal intuitions of space when constructing his theory of geometry. [10, p. 320]

Interestingly, stating precisely the postulated cognitive abilities involved in the interpretation of diagrams according to **Eu** might suggest a role for the cognitive intuitions of Euclidean geometry in the reasoning practice of Euclid's *Elements*.

Firstly, we have noticed that one of the key abilities in the interpretation of diagrams, according to **Eu**, is the capacity to *categorize* or *type* the different figures of the visual diagram using the normative concepts of geometry. This ability seems to be universal, according to the empirical data that we reported in the previous section, as the Mundurucu people, without previous formal education in Euclidean geometry, seem to possess an abstract conceptual systems which contains the key normative concepts of elementary geometry, and are able to use it to categorize elements of visual diagrams.

Secondly, interpretation of diagrams in **Eu** requires an important capacity of *abstraction* which is formally represented by an equivalence relation between diagrams. This equivalence relation aims to capture the idea that some features of the diagram are not relevant for reasoning: for instance, the same diagram rotated, translated or widened, would play exactly the same role in a geometrical proof of the *Elements*. This capacity of abstraction seems to connect with the cognitive ability of perceiving abstract geometric features, such as angle and length proportions, while abstracting away irrelevant information from the point of view of Euclidean geometry.

We now turn to the deductive and constructive aspects of geometrical reasoning. In section 5, we argue that a plausible cognitive framework for an empirical investigation of geometrical reasoning will have to bring together the interpretative, deductive and constructive aspects of geometrical reasoning to be faithful to the mathematical practice of Euclidean geometry.

4 Deductive and Constructive Aspects of Geometrical Reasoning with Diagrams

In this section, we begin by presenting the **Eu** formalization of constructive and deductive steps in Elementary geometry, and then discuss the capacity of the mental rules and the mental models theories to represent these reasoning steps. We conclude that an adequate framework for an empirical investigation of geometrical reasoning will have to go beyond these two theories.

4.1 Eu Construction and Inference Rules

The **Eu** proof rules specify how to derive $\Delta_1, A_1 \longrightarrow \Delta_2, A_2$ expressions. They specify, in particular, the operations that can be performed on the pair Δ_1, A_1 to produce a new pair Δ_2, A_2. Given the intended interpretation of $\Delta_1, A_1 \longrightarrow \Delta_2, A_2$ given by $\star$, the fundamental restriction on these rules is that they be geometrically sound. In other words, if Δ_2, A_2 is derivable from $\Delta_, A_1$, then with *any* configuration satisfying the geometric conditions represented by $\Delta_1 A_1$ one must either have a configuration satisfying the conditions represented by Δ_2, A_2 (if $\Delta_1 = \Delta_2$), or be able to construct a configuration satisfying the conditions represented by Δ_2, A_2 (if Δ_2 contains objects not in Δ_1).

The formal process whereby Δ_2, A_2 is derived has two stages: a construction and demonstration. Hence **Eu** has two kinds of proof rules: construction rules and inference rules. The construction stage models the application of a proof's construction on a given diagram. The demonstration stage models the inferences made from the assumed metric relations and the particular diagram produced by the construction. The soundness restriction thus applies only to the inference rules. The construction rules together codify a method for producing a representation that can serve as a vehicle of inference. The demonstration rules codify the inferences that can be made from such a vehicle.[4]

4.2 Mental Rules Theory

The mental rules theory [11, 12] represents reasoning as the application of *formal rules*, rules which are akin to those of natural deduction. According to this theory, reasoning is conceived as a syntactic process whereby sentences are transformed. These transformations are made according to specific rules defined precisely in terms of the syntactic structures of sentences. Deduction of one sentence from a set of other sentences (premisses) is seen as the *search* for a *mental proof*, which consists precisely in the production of the conclusion from the premisses by application of the rules a finite number of times. Consequently, the mental rules theory represents reasoning as consisting in syntactic operations on the *logical forms* of sentences.

The formalization of geometrical reasoning provided by **Eu** shares one important feature with the mental rules theory: **Eu** represents geometrical reasoning in terms of syntactic rules of inference. This is made possible by considering diagrams as a kind of *syntax*, and then by stipulating rules that control inferences that are drawn from diagrams. Thus, **Eu** diagrams might be seen as representing something like the *logical form* of concrete visual diagrams. One could perhaps say that **Eu** diagrams represent their *geometric form*, and that **Eu** inference rules operate on these forms. From this point of view, the formalization of geometrical reasoning provided by **Eu** appears to be in direct line with the the mental rules theory of reasoning.

[4] For the formal details, see sections 1.4.1 and 1.4.2 of [2], available online at `www.johnmumma.org`.

Nevertheless, the mental rules theory seems to run into troubles when we consider the *construction* operations on diagrams, which are central to the reasoning practice of Euclidean geometry, and which are formalized by **Eu**'s construction rules. One might still argue that **Eu** construction rules are syntactic operations on **Eu** diagrams, since diagrams are included its syntax, and so **Eu** construction rules fit the framework of mental rules theory. However, this does not seem correct as the mental rules are considered to be *deduction* rules; the soundness restriction applies to them. They are thereby of a very different nature than **Eu**'s construction rules. Consequently, even though the mental rules theory could suitably represent inferential steps in geometrical reasoning, it seems that the theory lacks the necessary resources to account for the construction operations on diagrams, an aspect fundamental to the reasoning practice in Euclidean geometry.

4.3 Mental Models Theory

The mental models theory [13, 14] postulates that reasoning depends on envisaging *possibilities*. When given a set of premisses, an individual constructs mental models which correspond to the possibilities elicited by the premisses. Different reasoning strategies are then available to extract information from the mental models: one may represent several possibilities in a diagram and draw a conclusion from the diagram, one may make an inferential step from a single possibility, or one can use a possibility as a counter-example for falsifying a conclusion. One interesting feature of the theory is that mental models are supposed to be *iconic*: the structure of a mental model is supposed to reflect the structure of the possibility that it represents. This feature is nicely illustrated when the mental models theory is applied to account for reasoning with visual-spatial information [15].

Contrary to the mental rules theory, the mental models theory seems suited to provide an account of the representation of the different construction operations on diagrams: visual diagrams used in geometrical proofs can be conceived as mental models entertained by the reasoner. Mental models associated to diagrams would then be of a visual-spatial nature, reflecting the spatial relations between the different elements of the diagram. This just seems to be a description of **Eu** diagrams, which encode the information that can be legitimately used in geometrical reasoning.

Accordingly, if **Eu** diagrams are conceived as mental models, the construction operations on diagrams would be explained in terms of the ways mental models are constructed. Nevertheless, one might worry about the capacity of the mental models theory to account for the specific use of diagrams in geometrical reasoning. One of the main points of **Eu** is to exhibit precisely that diagrammatic information enters legitimate mathematical inferences in a very controlled way. In order to account for the use of diagrams in geometrical reasoning, the mental models theory would have to be supplemented with a regimentation of the information that can be legitimately extracted from diagrams conceived as mental models. The formal system **Eu** shows that this can be done using syntactic rules that operate on diagrams represented as syntactic objects.

4.4 Beyond the Mental Rules vs Mental Models Debate

Geometrical reasoning, as practiced in Euclid's *Elements*, constitutes an interesting test for current theories in the psychology of reasoning. Our previous comparison predicts that the mental rules and mental models theories present both advantages and disadvantages as an account of the reasoning practice of geometrical reasoning with diagrams. Indeed, it seems that these two theories are actually complementary in their capacity to account for geometrical reasoning as described by **Eu**: the mental rules theory seems adequate to represent *inferences* in geometrical reasoning, while the mental models theory seems adequate to represent the *diagrammatic constructions* that support such inferences. Thus, it seems that to provide a framework for an empirical investigation of geometrical reasoning with diagrams one will need to go *beyond* these two theories. Such a move is also suggested, although for different reasons, in [16, 17]. In this respect, **Eu** constitutes a possible starting point for developing a cognitive framework for geometrical reasoning which would be faithful to both deductive and constructive aspects of the mathematical practice of Euclidean geometry.

5 Interpretation and Reasoning in Elementary Geometry

According to the **Eu** analysis of geometrical reasoning, interpretation and reasoning with diagrams are intimately related. This observation, originating in the study of the reasoning practice in Euclid's *Elements* [4], is directly in line with a recent approach to the psychology of reasoning developed by Stenning and van Lambalgen [1] which attributes a central role to interpretative processes in human reasoning:

> We [. . .] view reasoning as consisting of two stages: first one has to establish the domain about which one reasons and its formal properties (what we will call 'reasoning *to* an interpretation') and only after this initial step has been taken can ones reasoning be guided by formal laws (what we will call 'reasoning *from* an interpretation'). [1, p. 28]

Geometrical reasoning with diagrams, as formalized by **Eu**, precisely fits within this framework: reasoning *to* an interpretation corresponds to the process of interpreting a visual diagram along with an associated metric assertion, resulting in **Eu** into a pair Δ, A; reasoning *from* an interpretation is then represented as the application of the formal rules of **Eu** to prove geometric claims.

This perspective unifies the two main issues addressed in this paper. Our main conclusions can then be restated as follows: (i) intuitions of Euclidean geometry as studied by mathematical cognition are likely to play a role in reasoning *to* an interpretation of a diagrams, (ii) the mental rules and mental models theories of reasoning are inadequate for representing reasoning *from* an interpretation, as none of them is able to account for both deductive and constructive aspects of geometrical reasoning. In the perspective of Stenning and van Lambalgen [1], **Eu** appears as a good candidate for representing the process of reasoning *from* an interpretation in elementary geometry.

6 Conclusion

The empirical investigation of the cognitive bases of Euclidean geometry is a multi-disciplinary enterprise involving both mathematical cognition and the psychology of reasoning, and which shall benefit from works in formal logic and the nature of mathematical practice. In this paper, we used the formal system **Eu** to review existing empirical and theoretical works in cognitive science with respect to this enterprise. Our investigation shows: (i) the necessity of dealing jointly with interpretation and reasoning, (ii) the relevance of works on the cognitive bases of Euclidean geometry for the interpretation of diagrams and (iii) the necessity to go beyond the mental rules vs mental models distinction for accounting for both constructive and deductive aspects of geometrical reasoning.

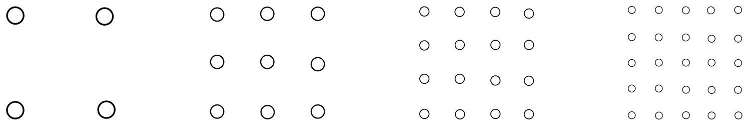
References

1. Stenning, K., Van Lambalgen, M.: Human Reasoning and Cognitive Science. The MIT Press, Cambridge, MA (2008)
2. Mumma, J.: Intuition Formalized: Ancient and Modern Methods of Proof in Elementary Geometry. PhD thesis, Carnegie Mellon University (2006)
3. Mumma, J.: Proofs, pictures, and Euclid. Synthese **175**(2) (2010) 255–287
4. Manders, K.: The Euclidean diagram. In Mancosu, P., ed.: Philosophy of Mathematical Practice. Clarendon Press, Oxford (2008)
5. Dehaene, S., Izard, V., Pica, P., Spelke, E.: Core knowledge of geometry in an Amazonian indigene group. Science **311**(5759) (2006) 381–384
6. Izard, V., Spelke, E.: Development of sensitivity to geometry in visual forms. Human Evolution **23**(3) (2009) 213–248
7. Izard, V., Pica, P., Spelke, E., Dehaene, S.: Flexible intuitions of Euclidean geometry in an Amazonian indigene group. Proceedings of the National Academy of Sciences **108**(24) (2011) 9782–9787
8. Klein, F.: A comparative review of recent researches in geometry. Bulletin of the New York Mathematical Society **2** (1893) 215–249
9. Wulff, K.: Examining knowledge of geometry. Science **312**(5778) (2006) 1309
10. Izard, V., Pica, P., Dehaene, S., Hinchey, D., Spelke, E.: Geometry as a universal mental construction. In Brannon, E., Dehaene, S., eds.: Space, Time and Number in the Brain: Searching for the Foundations of Mathematical Thought. Number XXIV in Attention & Performance. Oxford University Press, Oxford (2011) 319–332
11. Rips, L.: Cognitive processes in propositional reasoning. Psychological Review **90**(1) (1983) 38–71
12. Rips, L.: The Psychology of Proof: Deductive Reasoning in Human Thinking. The MIT Press, Cambridge, MA (1994)
13. Johnson-Laird, P.: Mental Models. Cambridge University Press, Cambridge (1983)
14. Johnson-Laird, P.: Mental models and human reasoning. Proceedings of the National Academy of Sciences **107**(43) (2010) 18243–18250
15. Knauff, M.: A neuro-cognitive theory of deductive relational reasoning with mental models and visual images. Spatial Cognition & Computation **9**(2) (2009) 109–137
16. Knauff, M.: How our brains reason logically. Topoi **26**(1) (2007) 19–36
17. Goel, V.: Anatomy of deductive reasoning. Trends in Cognitive Sciences **11**(10) (2007) 435–441

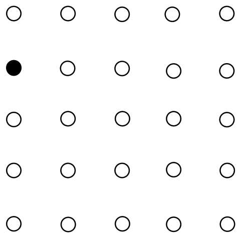
Appendix

A Eu diagrams

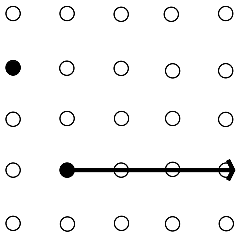
The syntactic structure of diagrams in **Eu** has no natural analogue in standard logic. Their underlying form is a square array of dots of arbitrary finite dimension.



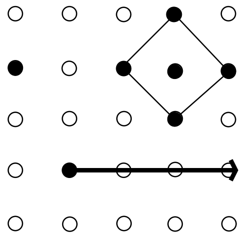
The arrays provide the planar background for an **Eu** diagram. Within them geometric elements—points, linear elements, and circles— are distinguished. A point is simply a single array entry. An example of a diagram with a single point in it is



Linear elements are subsets of array entries defined by linear equations expressed in terms of the array entries. (The equation can be bounded. If it is bounded on one side, the geometric element is a ray. If it is bounded on two sides, the geometric element is a segment.) An example of a diagram with a point and linear element is



Finally, a circle is a convex polygon within the array, along with a point inside it distinguished as its center. An example of a diagram with a point, linear element and a circle is

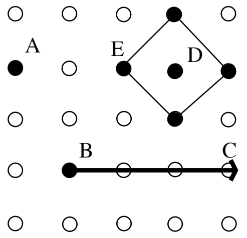


The size of a diagram's underlying array and the geometric elements distinguished within it, comprise a diagram's identity. Accordingly, a diagram in **Eu** is a tuple

$$\langle n, p, l, c \rangle$$

where n, a natural number, is the length of the underlying array's sides, and p, l and c are the sets of points, linear elements and circles of the diagram, respectively.

Like the relation symbols which comprise metric assertions, the diagrams have slots for variables. A diagram in which the slots are filled is termed a *labeled diagram.* The slots a diagram has depends on the geometric elements constituting it. In particular, there is a place for a variable beside a point, beside the end of a linear element (which can be an endpoint or endarrow), and beside a circle. One possible labeling for the above diagram is thus



Having labeled diagrams within **Eu** is essential, for otherwise it would be impossible for diagrams and metric assertions to interact in the course of a proof. We can notate any labeled diagram as

$$\langle n, p, l, c \rangle [\boldsymbol{A}, R]$$

where $\boldsymbol{A}$ denotes a sequence of variables and R a rule matching each variable to each slot in the diagram.

B Diagram equivalence

The definition of diagram equivalence is based on the notions of a diagram's *completion* and that of *co-exact map*. The completion $\varDelta'$ of a diagram $\varDelta$ is simply the diagram obtained by adding to $\varDelta$ intersection points to the intersections present in it. (See page 38-40 of [2]) Two diagrams are then equivalent if there is a co-exact map between their completions.

The key idea behind the definition of diagram equivalence, then, is that of of a co-exact map. A necessary condition for their to be such a map between diagrams $\varDelta_1$ and $\varDelta_2$ is that both diagrams contain the same number of points, linear elements and circles, and these objects are labeled by the same variables and variable pairs.

Suppose $\varDelta_1$ and $\varDelta_2$ are two such diagrams, and $\phi_{\boldsymbol{x}}$ is the bijection that maps an element in $\varDelta_1$ to the element in $\varDelta_2$ with the same label. In virtue of their position with respect to the underlying array of $\varDelta_1$, the elements of $\varDelta_1$ satisfy various geometric relations. The bijection $\phi_{\boldsymbol{x}}$ is a co-exact map if and only if it preserves a certain sub-set of these relations. Precisely $\phi_{\boldsymbol{x}}$ is a co-exact map if and only if it satisfies the following nine conditions, where P and Q are points of $\varDelta_1$, N and M are linear elements of $\varDelta_1$ (i.e. segments, rays, or lines) and C_1 and C_2 are circles of $\varDelta_1$.

- P lies on M in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(P)$ lies on $\phi_{\boldsymbol{x}}(M)$ in $\varDelta_2$.
- P and Q lie on a given side M in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(P)$ and $\phi_{\boldsymbol{x}}(Q)$ lie on the same side of $\phi_{\boldsymbol{x}}(M)$ in $\varDelta_2$. (The same side of $\phi_{\boldsymbol{x}}(M)$ in is determined via the orientation specified by the two variable label of M and $\phi_{\boldsymbol{x}}(M)$.)
- P lies inside/on/outside circle C_1 in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(P)$ lies inside/on/outside $\phi_{\boldsymbol{x}}(C_1)$ in $\varDelta_2$.
- M intersects N at a point/along a segment in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(M)$ intersects $\phi_{\boldsymbol{x}}(N)$ at a point/along a segment in $\varDelta_2$.
- M cuts in one point/cuts in two points/is tangent at one point to/is tangent along a segment to C_1 in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(M)$ cuts in one point/cuts in two points/is tangent at one point to/is tangent along a segment to $\phi_{\boldsymbol{x}}(C_1)$ in $\varDelta_2$.
- C_1 has the same intersection signature with respect to C_2 in $\varDelta_1$ as $\phi_{\boldsymbol{x}}(C_1)$ has with respect to $\phi_{\boldsymbol{x}}(C_2)$ in $\varDelta_2$. (The intersection signature classifies the way C_1 and C_2 intersect each other as convex polygons. For its precise definition, see appendix A in the thesis.)
- Circle C_1 lies outside/ is contained by circle C_2 in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(C_1)$ lies outside/is contained by circle $\phi_{\boldsymbol{x}}(C_2)$ in $\varDelta_2$.
- Circle C_1 lies within a given side of M in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(C_1)$ lies within the same side of $\phi_{\boldsymbol{x}}(M)$ in $\varDelta_2$.
- End-arrow P is on a given side of M in $\varDelta_1 \longleftrightarrow \phi_{\boldsymbol{x}}(P)$ is on the same side of $\phi_{\boldsymbol{x}}(M)$ in $\varDelta_2$.

The Double Disjunction Task as a Coordination Problem

Justine Jacot

University of Lund, Department of Philosophy
justine.jacot@fil.lu.se

Abstract. In this paper I present the double disjunction task as introduced by Johnson-Laird. This experiment is meant to show how mental model theory explains the discrepancy between logical competence and logical performance of individuals in deductive reasoning. I review the results of the task and identify three problems in the way the task is designed, that all fall under a lack of coordination between the subject and the experimenter, and an insufficient representation of the semantic/pragmatic interface. I then propose a reformulation of the task, that makes explicit the underlying semantic reasoning and emphasizes the difference of interpretation of the DDT between the experimenter and the subjects.

1 Introductory Remarks

1.1 The Double Disjunction Task

The *double disjunction task* (DDT), introduced by Johnson-Laird et al. (1992), is an inferential task where subjects are asked to say "What, if anything, follows" from a set of two premises, typically of the form:

Alice is in India or Barbara is in Pakistan [or both/but not bot] (P_1)

Barbara is in Pakistan or Cheryl is in Afghanistan [or both/but not bot] (P_2)

The optional parameters (in square brackets) are always covariant in premises. In the terminology introduced by Johnson-Laird et al. (1992), the authors determine an *inclusive* and an *exclusive* variant, when (resp.) the first and second values are chosen. These variants are called *affirmative*, as opposed to those obtained when (P_2) is substituted with:

Barbara is in Bangladesh or Cheryl is in Afghanistan [or both/but not bot] (P'_2)

which are called *negative*, because of the incompatibility between (P_1) and (P'_2), that gives rise to a contradiction. With subscript for inclusive or exclusive, and superscripts for positive and negative, one obtains four variants, DDT^+_I, DDT^+_E, DDT^-_I and DDT^-_E.

The task was aimed at enforcing some predictions by the mental model theory (see below) according to which the more models an agent has to represent the less she is able to make correct inferences from the premises. Inclusive disjunctions should be therefore more difficult to cope with than exclusive ones, as well as disjunctions involving a contradiction in the premises. For matters of clarity, here is the truth table for the inclusive (noted $\vee$) and exclusive (noted $\underline{\vee}$) disjunctions:

P	Q	$P \vee Q$	$P \veebar Q$
T	T	T	F
T	F	T	T
F	T	T	T
F	F	F	F

Fig. 1. Truth table for inclusive and exclusive disjunctions

1.2 The Mental Model Theory

The *mental model theory* (MMT) has been developed in reaction against the idea of a 'mental logic' according to which individuals have logical rules in their mind and draw logical inferences in line with these rules. According to the MMT, people make logical inferences by following neither the rules of natural deduction nor those of truth tables, but by constructing mental models of the premises from which they derive a model of the conclusion. Deductive reasoning does not depend on syntactic rules but on a three-steps semantic process. First, semantic information is mentally construed as a representation of situations given by the premises. Second, an attempt to draw a conclusion from the mental models of the premises is made. Third, the validity of the conclusion is checked by ensuring that no model of the premises renders it false. For instance, a disjunction such as:

There is a club or there is a spade

calls for two models:

♣
    ♠

When the following premise is added:

There is no club

the model representing the club is discarded and the informational content of the new premise is added to the remaining model:

¬♣ ♠

The conclusion 'There is a spade' can be drawn, since there is no model of the premises that makes this conclusion false.

Relying on the assumption that individuals are fundamentally rational but make errors in practice, Johnson-Laird identifies three principles for deductions made by 'logically untutored individuals':

> [They] eschew conclusions that contain less semantic information than premises. [...] Similarly, they seek conclusions that are more parsimonious than the premises. [...] They try to draw conclusions that make explicit some information only implicit in the premises. In short, to deduce is to maintain semantic information, to simplify, and to reach a new conclusion. (Johnson-Laird et al. 1992, p. 419)

These constraints on deductive reasoning are close to Grice's maxims of conversation (Grice 1989). These maxims ask for agents involved in a communicative situation to be as informative as is required (maxim of quantity), to try to make their contribution true (maxim of quality), to be relevant (maxim of relevance), and to be perspicuous (maxim of manner). Although Grice is mentioned only in passing by Johnson-Laird et al. (1992, p. 419), their analysis of deduction is recovered in these maxims: maintaining semantic information pertains to the maxim of quantity, non redundancy and simplification to the maxim of manner, and reaching a new conclusion to the maxim of relevance, while deduction guarantees the quality of the conclusion, conditional on the quality of the premises.

Grice stresses that conversations where the maxims apply are a special case of cooperative situations, and that these maxims are instantiations of a more general Cooperation Principle (CP). An experiment is, after all, a cooperative situation, and therefore the effect of CP should be evaluated at the beginning of the task, because deductive reasoning of lay people is constrained by semantic principles as well as pragmatic principles.

The remainder of this paper is organized as follows. The next section firstly discusses the results of the DDT, secondly identifies the main issues at stake in the interpretation of the results and the way the task is formulated. Section 3 shows that most of these problems arise because of a lack of coordination between the subject and the experimenter in the interpretation of the task. Section 4 suggests some possible reformulations that could improve the base rate of originally expected answer, by eliciting both the adequate modal representation, and the selection of relevant information, through pragmatic principles.

2 Results and Interpretation of the DDT

2.1 Results

Johnson-Laird et al. (1992) discuss only the selection made for the 'positive' DDT, consisting in the following two statements:

Alice is in India and Cheryl is in Afghanistan, or Barbara is in Pakistan, or both (C_i^+)

Alice is in India and Cheryl is in Afghanistan, or Barbara is in Pakistan, but not both (C_e^+)

The valid conclusion for the negative disjunctions were predicted to be more difficult to make because they suppose the ability to first detect the contradiction in the premises. The percentages of valid conclusions made by the subjects for the four kind of disjunctions are displayed in Fig. 2.

	Affirmative	Negative
Exclusive	21%	8%
Inclusive	6%	2%

Fig. 2. Percentage of valid conclusions for the DDT

As predicted by the MMT, the percentage of success decreases when the number of models to build increases, coupled with a floor effect: when a deduction needs three or more models, it is almost impossible for the subjects to make a correct inference, or even to decide whether a conclusion follows from the premises.

But while Johnson-Laird is interested in the link between one particular case of error with one particular subset of the models of the premises (Johnson-Laird et al. 1992, p. 433-434), and although the prediction of MMT for relative rates of success are confirmed, the low base rates remain only imperfectly explained. Besides, a non negligible number of subjects shows an ability to perform disjunctive reasoning. The results reported by Toplak and Stanovich (2002) show a higher success rate for the DDT (37%), because the authors collapsed in one 'super-category' all the conclusions that displayed a "disjunctive pattern" of reasoning.[1]

The two categories collapsed into the disjunctive responses were what Toplak and Stanovich (2002) call 'the *partial contingency*' response and the 'Alice-Cheryl response', which were both given by 12% of the subjects. The partial contingency response, for instance, 'If Alice is in India and Cheryl is in Afghanistan, Barbara is not in Pakistan', is given when people draw "the implications of one of the disjunctive path but not the other." (Toplak and Stanovich 2002, p. 203). According to Toplak and Stanovich (2002), it seems in this case that subjects are able to assign hypothetic truth values to each disjunct in the premises but unable to represent situations where all states of affairs hold, comprising those where Alice is not in India and Cheryl is not in Afghanistan.

2.2 Identifying the Problems in the Design of the DDT

MMT explains failure in terms of the number of mental models to be construed in order to solve the task. Reasoning being semantic, the semantic content affects the resolution of the task. The more models individuals have to process, the less they succeed in making correct inferences, due to the difficulty to store in the working memory the totality of the models of the premises needed to draw the correct inferences, particularly when different situations have to be processed in parallel.

While MMT assumes that "naive individuals have a modicum of deductive competence" (Johnson-Laird 1999, Johnson-Laird and Byrne 1991) but that they (sometimes) fail at deductive performance, things have to be viewed the other way round. The theory predicts that, faced with a deductive inference, individuals apply some filtering constraints depending on the kind of inference they have to make. I suggest that they should be viewed as applying some relevance constraints because of the form of the reasoning they have to make, and then draw the conclusions that seem relevant.

Changing the perspective preserves some insights from MMT (i.e. why 'trivial' conclusions are not even considered), but sheds light on other aspects left unexplained by the theory. For instance, answers in disjunctive form are more common than in conditional form (see Toplak and Stanovich 2002), which seems to indicate that people tend to preserve the syntactic form of the premises.

[1] What Toplak and Stanovich (2002) call 'disjunctive reasoning' is the construction of mental models for the premises and at least a partial combination of them in the conclusion. The normative answer (fully disjunctive) was given by only 13% of the subjects, which is lower than in Johnson-Laird et al. (1992).

One may then identify three problematic categories in the formulation of the DDT. The first category of problems concerns the logical form of the premises. First, the conjunction of the two premises is implicit, but this conjunction must become explicit in order for the intended (correct) conclusion to be drawn. Second, it is not clear whether the disjunct ‘Barbara is in Pakistan’ should be at the same place in the two premises, or at the first place in the first premise and at the second place in the second premise. Indeed, it is not sure that not logically trained individuals always interpret the disjunction as commutative in the DDT. But at the same time, it seems that people want to preserve the syntactic form of the premises: answers in disjunctive form are more common than in conditional form. Although Johnson-Laird et al. (1992, p. 433) explained subjects the difference between an inclusive and an exclusive disjunction, guaranteeing a common interpretation of “. . . or . . . or both” and “. . . or . . . but not both” (between participants and experimenter), they give little detail about this explanation, save for the fact that it did not rely on truth-tables or natural deduction rules.

The second category pertains to the linguistic context opened by the DDT. First, it seems that, in designing the task, Johnson-Laird and his colleagues wanted to avoid the problem of an abstract task by giving the subjects a thematic task. But it is unclear whether the difference would negatively affect the results. Indeed, the premises in the DDT are rather abstract since they present names and places out of the blue. No context or situation supports the story. Second, nothing in the task says that ‘Barbara’ in both premises is the same individual, for the reasons mentioned above.

The third category of problems relates to the way the DDT is framed in general, and particularly the way the question is asked: ‘what, *if anything*, follows from these premises?’ Either classical logic is assumed and an infinite number of valid conclusions follows from the premises, and then subjects have no time or no space to give the full answer; or the answer is intended to be the conclusion that preserves the most semantic information from the premises and ‘if anything’ can be dropped from the question.

In what follows I will show that these issues can be addressed if we reconsider the DDT as a coordination problem between subject and experimenter. I will focus on the second and third category of problems, the first one being answered in the new formulation of the task I propose next.

3 A Coordination Problem

3.1 Narrowing Down the Context

Experimenters take often for granted that only one possible interpretation of the logical vocabulary is correct, and think that departure from this interpretation is an error. But, as shown by Stenning and van Lambalgen (2008, chap. 3), the problem is that experimenters also leave under-determined this interpretation for the subjects, to such an extent that subjects sometimes complete the under-determined semantic content in an unintended way. If MMT is correct and logical reasoning is indeed semantic, then the DDT should be able to provide tests such that either the semantic content is determined enough to yield a single interpretation, or the normative answer allows for different interpretations.

As noted by Shafir (1994), one problem with the DDT is that it demands subjects to "think through" the disjuncts, and to draw a conclusion from both the premises. In order to do so, subjects must process all the premises at the same time, but they appear on paper as two separate sentences, and the conjunction of the premises remains implicit. While the disjunctive form may trigger some ability to reason by cases (even though the performance's rate is low), what in fact is needed is to reason by cases *together with* the ability to keep track of the different possibilities, as e.g. answers to different questions regarding the premises.

Moreover, as put forward by relevance theory (Sperber et al. 1995, Sperber and Wilson 1995, Wilson and Sperber 2004) there is a discrepancy between the laboratory task and the real life situations in which the subjects might have to perform these kind of inferences. The lack of relevance of the task at stake, or the lack of communicative interaction can yield some pragmatic 'disabilities.' On a linguistic level, some linguistic forms produce some kind of relevance 'filters' content- and context- dependent which, in turn, induce a preference ordering over possible interpretations (Girotto et al. 2001, p. 70). On one hand, in the type of task presented to the subjects, logically relevant and logically irrelevant content are mixed, whereas it is asked to the individuals to pay attention only to the logically relevant content. On the other hand, in order to achieve the task, individuals have to interpret it correctly. They must therefore pay attention to the logically relevant *and* irrelevant content. But the logically irrelevant content sometimes prompts individuals to a certain interpretation, which is not always the interpretation the experimenter has in mind.[2]

Further discussion of the relevance theory explanation of the success and failure of logical reasoning is not necessary, since my argument needs no other agreement than retaining its central insight. The DDT is content- and context-dependent and it seems reasonable to think that subjects interpret it as such, in a way that intermingles semantic and pragmatic inferences in reasoning.

3.2 Semantics vs. Pragmatics and the Principle of Cooperation

The way the question is asked in the DDT is disputable, for reasons that are admitted by Johnson-Laird himself, but that he does not seem to factor in the formulation of the question. Indeed, Johnson-Laird acknowledges that "infinitely many valid conclusions follow from any premises" (Johnson-Laird 1999, p. 113) but, by saying 'if anything', nevertheless implies that it could be the case that nothing follows from them. This is obviously triggered by the three principles of deductive reasoning already mentioned in section 1.2. Assuming these facts, it is highly improbable that subjects will draw conclusions such as any conjunction of the premises.

If it is true that logically untrained individuals do not draw 'trivial' conclusions from a set of premises, there is no need to try to lure them in thinking that there might be

[2] Sperber et al. (1995, p. 44): "Past discussions of subjects' performance have tended to focus on the task as already interpreted (by the experimenter). [...] But interpreting the task is part and parcel of performing it, and obeys criteria of rationality in its own right. The study of 'content effects' is the study of sound cognitive processes that are by no means out of place in subjects' performance."

some. Besides, the formulation of the question must indicate that there is some relevant information that can be inferred from the premises. Indeed, the Gricean CP would ask for a speaker's intentions to be recognized by the hearer in order for the speaker's utterance to be correctly interpreted. Why not assuming such a process in the DDT? As previously mentioned, a psychological experiment involves a certain level of cooperation between subjects and the experimenter. Certainly CP governs inferences to intentions and goals of others, and reconstructing them in DDT inevitably involves speculation. But it is not impossible to reconstruct a possible route to C_i^+ and C_e^+ that takes CP more seriously than the original theory.

4 Two Versions of the DDT

4.1 The Intended Interpretation

The theory predicts that subjects will: (*a*) compute a representation of three (two) mental scenarios, or 'mental models', for each inclusive (resp.: exclusive) premise; (*b*) combine the representations, eliminating those incompatible with the information; and: (*c*) 'read off' some appropriate conclusion and answer the question.

In this general setting, the intended interpretation requires an abstraction step. In step (*a*), subjects are supposed to replace the premises with a variable $X \in \{A, B, C\}$ where A, B, C stand for each of the individual in the disjunct, to which a value is assigned, equivalent to 'I know/I do not know where is *X*'. Step (*b*), for each variant of the DDT, is tantamount to assigning a value to each variable, compatible with the constraints embodied in the models of the premises.

Comparatively, subjects should be expected to have greater difficulties to solve the 'negative' tasks. Johnson-Laird et al. explain this difficulty by the necessity of obtaining an additional (implicit) premise, which follows from subjects background knowledge:

> Of course, if [Barbara] is in [Bangladesh], then [she] is not in [Pakistan]. According to the model theory, affirmative deductions should be easier than such negative deductions because the latter call for the detection of the inconsistency between the contrary constituents. (Johnson-Laird et al. 1992, p. 433)

The detection of inconsistency is an additional (inferential) step, left unanalyzed by Johnson-Laird et al. This step is equivalent to impose additional constraints to compute the composition of the representations associated with (P_1) and (P'_2), using background knowledge. The additional complexity of tracking combinations of two values for parameter *B* in both DDT_{I}^- DDT_{E}^-, explains the cost of "detection of inconsistency" by the stress imposed on computational abilities. In addition, DDT_{E}^- outputs an additional mental model, as compared to DDT_{E}^+, inducing a heavier load on working memory.

So far, this reconstruction of steps (*a*) and (*b*) of DDT agrees with Johnson-Laird et al. However, reconstruction of step (*c*) raises serious issues with both the design of the experiment, and the interpretation of its results. Johnson-Laird et al. discuss only the selection made for the 'positive' DDT. Their justification for this selection is the "support" mental models yield to these conclusions.

Indeed, treating *A*, *B* and *C* as propositions, and substituting known (asserted) and unknown (excluded) locations with 1 and 0, respectively, one obtains distributions of truth-values that satisfy (C_i^+) and (C_e^+). However, this explanation is at odds with the claim that reasoning based on mental models does not proceed by building and reading equivalent of truth tables, since “truth tables [. . .] are too bulky to be mentally manipulated” (Johnson-Laird et al. 1992, p. 421).

Considering the representation Johnson-Laird et al. gives for the set of final mental models in $\mathrm{DDT}_{\mathrm{I}}^{+}$ and $\mathrm{DDT}_{\mathrm{E}}^{+}$, reproduced Fig. 3, (C_e^+) is the most natural of the answers considered. If the mental construction is actually carried as hypothesized by Johnson-Laird et al., (C_e^+) can be ‘read off’ neglecting excluded locations, and using ‘or. . . but not both’ to express the existence of (two) alternative scenarios. This method does not rely on explicit substitution with truth-values, but has no straightforward equivalent in $\mathrm{DDT}_{\mathrm{I}}^{+}$ that would read (C_i^+) off.

Alice	Barbara	Cheryl	Alice	Barbara	Cheryl
[I]		[A]	[I]	[P]	[A]
	[P]		[I]	[P]	
			[I]		[A]
				[P]	[A]
				[P]	

Fig. 3. Mental Models for $\mathrm{DDT}_{\mathrm{E}}^{+}$ and $\mathrm{DDT}_{\mathrm{I}}^{+}$

One could hypothesize that subjects approximate $\mathrm{DDT}_{\mathrm{I}}^{+}$ by addressing it *as if* it were $\mathrm{DDT}_{\mathrm{E}}^{+}$, and then simply switch the proviso from “but not both” to “or both.” However, this rationale for the choice of (C_i^+), is not consistent with the data, in particular the relative difference between the rate of success in $\mathrm{DDT}_{\mathrm{E}}^{+}$ and $\mathrm{DDT}_{\mathrm{I}}^{+}$. Johnson-Laird et al. may have selected (C_i^+) by the above reasoning, or simply by judging (C_e^+) intuitively natural enough, and, switching the proviso, assuming that (C_i^+) was natural, too. However they give no hint has to what their selection is based on, and rest content with the relative frequencies, that validate the predictions of MMT.

In the absence of an analysis of the ‘reading off’ method that outputs (C_i^+), it seems difficult to argue whether the answer corresponds to some natural method. Likewise, the absence of discussion of answers expected to be given to $\mathrm{DDT}_{\mathrm{I}}^{-}$ and $\mathrm{DDT}_{\mathrm{E}}^{-}$ makes the assessment of the task more complicated than Johnson-Laird et al. seem to assume.

4.2 The Abstraction Problem

The naturalness of (C_e^+) depends essentially on the representation of the problem that neglects *specific* locations, and in which one checks only whether a location is known, or not. If one assumes with Johnson-Laird et al. that reasoners somehow automatically select representations that are as implicit as possible, the inescapable conclusion is that the underlying automatisms are not very good at selecting the level of abstraction that makes the task solvable.

Yet, abstraction in variants of the DDT may not be as simple as assumed by Johnson-Laird et al. Reading (or hearing) (P_1), (P_2) and (P_2'), in either their inclusive or exclusive variants, may elicit a very different representation, if the range of values is taken to be the set of locations. Then, upon reading (hearing) (P_1), a subject may consider that the set of relevant values is:[3]

$$v_{(\mathrm{P}_1)} = \{i, p, \overline{ip}\} \tag{$v_{(\mathrm{P}_1)}$}$$

(where $\overline{ip}$ corresponds to any other possible location than India and Pakistan). However, reading (P_2) produces a new set of values for that premise, as well as an update of ($v_{(??)}$), as follows:

$$v'_{(\mathrm{P}_1)} = \{a, i, p, \overline{aip}\} \tag{$v'_{(\mathrm{P}_1)}$}$$

$$v_{(\mathrm{P}_2)} = \{a, i, p, \overline{aip}\} \tag{$v_{(\mathrm{P}_2)}$}$$

The effect is even more striking with (P'_2), where the number of seemingly relevant values (and therefore of their combinations) increases even more, with:

$$v''_{(\mathrm{P}_1)} = \{a, b, i, p, \overline{abip}\} \tag{$v''_{(\mathrm{P}_1)}$}$$

$$v_{(\mathrm{P}'_2)} = \{a, b, i, p, \overline{abip}\} \tag{$v_{(\mathrm{P}'_2)}$}$$

Given the complexity of the task, one will typically not complete step (*a*) before the time the answer has to be given. Again, when "unable to construct a set of models or [. . .] to discern what holds over all of them" subject will respond that *no valid conclusion follows*.

5 Conclusion

DDT as it is designed in psychology settings, is not a bad test for reasoning skills in the sense that it would mistake the competence to test. The problem comes from the way experimenters interpret the task itself. The difference in interpretation between experimenters and subjects is what leads to a misinterpretation of the results, and shows the role of relevance and computational complexity in the answers of the subjects. Reconsidering the DDT as a coordination situation brings to light the reasoning process of the subjects in drawing inferences from a double disjunction. This process, as shown in the sketch of the formal reformulation of the task above, is a sequential procedure which starts from a step-by-step construction of representations task, and outputs a global conclusion. Our reformulation emphasizes the discrepancy between the intended interpretation of the DDT by the experimenters, and the task as it may be understood by the subjects.

What remains to do in a further research is, from a theoretical perspective, building a complete formal representation of the DDT displaying the intended interpretation by the experimenter and the possible interpretation(s) by the subjects. This reformulation of the task relies on the difference made by Stenning and van Lambalgen (2008) between reasoning *to* an interpretation and reasoning *from* an interpretation: the authors argue that subjects must always first *reason to* the experimenter's (intended) interpretation before they can *reason from* it to a solution.

[3] A symmetric argument can be given for the case where (P_2) or (P'_2) are presented first.

From an empirical perspective, we need to design an experiment based on the explanatory hypothesis argued for in this paper, especially concerning the level of abstraction required from the subjects in order to solve the task. To this effect, it seems desirable to handle both sides of the problem by testing subjects' performance in a fairly abstract task where the disjuncts are displayed as propositional variables such as A, B, C, and compare the results with a DDT where the premises are contextually fleshed out by a story to fulfill the need for a thematic task.

Bibliography

Girotto, V., Kemmelmeier, M., Sperber, D., and van der Henst, J.-B. (2001). Inept reasoners or pragmatic virtuosos? Relevance and the deontic selection task. *Cognition*, 81:69–76.

Grice, H. P. (1989). Logic and conversation. In *Studies in the Way of Words*, pages 22–40. Harvard University Press.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, 50:109–135.

Johnson-Laird, P. N. and Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., Byrne, R. M. J., and Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, 99:418–439.

Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition*, 50:403–430.

Sperber, D., Cara, F., and Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57:31–95.

Sperber, D. and Wilson, D. (1995). *Relevance. Communication and Cognition.* Blackwell, 2nd edition.

Stenning, K. and van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. The MIT Press.

Toplak, M. E. and Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 94:197–209.

Wilson, D. and Sperber, D. (2004). Relevance theory. In Horn, L. R. and Ward, G., editors, *Handbook of Pragmatics*, pages 607–632. Blackwell.

Conditionals, Inference, and Evidentiality

Karolina Krzyżanowska[1], Sylvia Wenmackers[1], Igor Douven[1], and Sara Verbrugge[2]

[1] Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands
[2] Department of Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium

Abstract. At least many conditionals seem to convey the existence of a link between their antecedent and consequent. We draw on a recently proposed typology of conditionals to revive an old philosophical idea according to which the link is inferential in nature. We show that the proposal has explanatory force by presenting empirical results on two Dutch linguistic markers.

1 Introduction

Although various theories of conditional sentences have been proposed, none of them seems to account for all the empirical data concerning how people use and interpret such sentences.[3] It is almost universally agreed that no theory of conditionals counts as empirically adequate if it validates sentences like these:

(1) a. If badgers are cute, then $2 + 2 = 4$.
b. If weasels are vegetables, then unicorns hold Frege in particularly high esteem.

It is easy to understand why people are reluctant to accept conditionals like (1a) and (1b): the antecedents of those conditionals have nothing to do with their consequents. And it seems that using a conditional construction is meant to convey the existence of some sort of link between the content of the if-clause and the content of the main clause.

But what kind of link might this be? According to an old philosophical idea, the link is inferential in nature. If this idea is currently unpopular, at least in philosophy, that may be due to the fact that theorists have failed to recognize that the inferential connection need not be of one and the same type. We draw on a recently proposed typology of conditionals to reinstate the old idea, at least for a large class of conditionals that linguists have aptly termed "inferential conditionals." We buttress our proposal by presenting experimental results concerning two Dutch linguistic markers.

[3] For a survey of various accounts of conditionals and problems they are facing see, for instance, Edgington [1] or Bennett [2].

2 Inferential Conditionals

Probably the most general distinction to be made when it comes to conditionals is that between indicative and subjunctive conditionals. In this paper, we will only be concerned with indicative conditionals.[4] For many theorists this is only the beginning of a typology, though there is little unanimity as to what the typology should further look like. What has become common in linguistics, but is rarely encountered in the philosophical literature, is to classify indicative conditionals as *inferential* and *content* conditionals.[5] The class of content conditionals is not particularly well defined—its members are sometimes loosely said to describe relations between states of affairs or events as they happen in reality—but those will not concern us here. We are going to limit our attention to inferential conditionals, that is conditional expressing a reasoning process, having the conditional's antecedent as a premise and its consequent as the conclusion, for example: "If she has not had much sleep recently, she will perform poorly on her exams" or "If he owns a Bentley, he must be rich." Arguably, these constitute a common, if not the most common, type among the conditionals we encounter in natural language.

That a conditional sentence can be considered as a kind of "condensed argument" [8, p. 15] is not altogether new to philosophy; it can be traced back at least to Chrysippus, a stoic logician from the third century BC. He is believed to have held the view that a conditional is true if it corresponds to a valid argument [9]. Obviously, if we limit our understanding of a valid argument only to the classical deductive inference, we can easily find counterexamples to the above claim. Yet deduction is not the only type of reasoning people employ, and a plausible theory of inferential conditionals should not neglect this fact.

Although linguists have proposed various finer-grained typologies of inferential conditionals (see, e.g., Declerck and Reed [6]), most of these stem from grammatical distinctions. However, we are interested in a specific typology recently presented in Douven and Verbrugge [10] who acknowledge the variety of inferential relations between a conditional's antecedent and its consequent. A first distinction these authors make is that between *certain* and *uncertain* inferences, where certain inferences guarantee the truth of the conclusion given the truth of the premises while uncertain inferences only tend to make the truth of the conclusion likely given the truth of the premises; the former are standardly referred to as "deductive inferences." Douven and Verbrugge further follow standard philosophical usage in grouping uncertain inferences into *abductive* and *inductive* ones, where the former are inferences based on explanatory considerations and the latter are inferences based on statistical information. More exactly, in an abductive inference we infer a conclusion from a set of premises because

[4] The distinction between indicative and subjunctive conditionals is not as clear-cut as one might wish, but the conditionals we used in our materials were all uncontroversial cases of indicative conditionals. Henceforth, indicative conditionals will be referred to simply as "conditionals."

[5] See, among others, Dancygier [3,4], Dancygier and Sweetser [5], Declerck and Reed [6], and Haegeman [7].

the conclusion provides the best explanation for those premises; for example, we may infer that Sally failed her exam from the premises that Sally had an exam this morning and that she was just seen crying and apparently deeply unhappy: that she failed the exam is the best explanation for her apparent unhappiness. Inductive inferences rely on information about frequencies (which may be more or less precisely specified); for instance, we infer that Jim is rich from the premise that he owns a Bentley because we know that by far the most Bentley owners are rich. The validity of both abductive and inductive inferences is a matter of controversy and ongoing debate. It is largely uncontested, however, that people do engage in these types of inferences on a routine basis.

Douven and Verbrugge's typology of inferential conditionals follows the aforesaid typology of inference. That is to say, they distinguish between certain (or deductive) and uncertain inferential conditionals, and then divide the latter class further into abductive and inductive inferential conditionals. More specifically, they propose the following:

Definition 1. *"If p, then q" is a deductive inferential (DI, for short) / inductive inferential (II) / abductive inferential (AI) conditional if and only if q is a deductive / inductive / abductive consequence of p.*

(Various formalizations of the inductive and abductive consequence relations have been offered in the literature, though, like Douven and Verbrugge in their paper, we refrain from committing to any in particular.) Douven and Verbrugge note also that, often, the inference may rely on the antecedent p together with background assumptions that are salient in the context in which the conditional is asserted or evaluated. Such conditionals are called *contextual* DI, AI, or II conditionals, depending on the type of inference involved.[6]

Definition 2. *"If p, then q" is a contextual DI / II / AI conditional if and only if q is a deductive / inductive / abductive consequence of $\{p, p_1, \ldots, p_n\}$, with $p_1, \ldots, p_n$ being background premises salient in the context in which "If p, then q" is asserted or evaluated.*

Douven and Verbrugge do not claim that their typology is *correct* and the ones that so far have been propounded by other theorists are *incorrect.* Indeed, it might even be odd to think that there are *natural kinds* of conditionals that a typology should try to chart. What they do claim is that their typology is exceedingly simple and that it is non-ad hoc in that it relies on a time-tested distinction between types of inference. More importantly still, they show in their 2010 paper that the typology has considerable explanatory force by recruiting it in service of testing a thesis, first proposed by Adams [11] and championed by many since, according to which the acceptability of a conditional is measured by the probability of its consequent conditional on its antecedent.

[6] As Douven and Verbrugge [10, p. 304] note, in contextual AI conditionals, the consequent need not always be the best explanation of the antecedent. It may also be that the consequent is, *in light of the antecedent*, the best explanation of one of the background assumptions.

We expand here on the main experiment of Douven and Verbrugge's paper, because it served as the starting point for our own empirical work. In their experiment, Douven and Verbrugge divided the participants into two groups, asking one group to judge the acceptability of ten DI, ten AI, and ten II conditionals and the other group to judge the corresponding conditional probabilities. This is an example of a question asking for the acceptability of an AI conditional:[7]

> Context: Judy is waiting for a train. She is looking for her iPod. It is not in her coat. She suddenly sees that the zipper of her bag is open. She cannot remember having opened it. It is announced that there are pickpockets active in the train station.
>
> Conditional: If Judy's iPod is not in her bag, then someone has stolen it.
>
> Indicate how acceptable you find this conditional in the given context:
>
> Highly unacceptable 1 2 3 4 5 6 7 Highly acceptable

The corresponding question asking for the conditional probability is this:

> Context: Judy is waiting for a train. She is looking for her iPod. It is not ain her coat. She suddenly sees that the zipper of her bag is open. She cannot remember having opened it. It is announced that there are pickpockets active in the train station. Suppose that Judy's iPod is not in her bag.
>
> Sentence: Someone has stolen Judy's iPod.
>
> Indicate how probable you find this sentence in the given context:
>
> Highly improbable 1 2 3 4 5 6 7 Highly probable

The results obtained in this experiment show that Adams' thesis holds for DI conditionals at best, and that for AI conditionals the most that can be said is that acceptability and conditional probability are highly correlated; for II conditionals not even that much is true.

This already shows that the typology of inferential conditionals proposed by Douven and Verbrugge has considerable explanatory force. Here, we aim to extend the case for this typology by relating it to two putative evidential markers. Before we move on to report our experimental results on these markers (in section 4), we briefly discuss the concept of an evidential marker.

3 Evidential Markers

In some languages, it is common for speakers to communicate information about the evidential grounds for the contents of their assertions. Some languages also possess a rich arsenal of prefixes, suffixes, particles, and other linguistic items for

[7] See Appendix A of Douven and Verbrugge [10] for the full materials used in this experiment.

this purpose.[8] European languages do not encode evidentiality grammatically, but this is not to say that speakers of European languages do not have the resources to indicate evidential grounds, or that they never use those resources. Sentences like:

(2) Adam works hard.

do not indicate the speaker's source of information at all—the speaker can actually share an office with Adam and see him working hard on an everyday basis, but she could also have inferred (2) from Adam's work output or simply heard someone else say so.

But other sentences do suggest the speaker's evidential grounds. For instance, when we are wondering about the translation of a phrase in Latin and we know that Susan studied classical languages for a number of years, we might say

(3) Susan should be able to translate this phrase.

This assertion would seem odd if we knew that (say) the phrase is from a text which Susan recently published in English translation. Similarly, when an English speaker tries to call her friend but does not get an answer, she may infer that her friend is out and express the resulting belief by saying:

(4) She must be out.

Were she to see her friend walking on the street, her assertion of (4) would again seem odd or even inappropriate.

Some authors have argued that the modal auxiliary verb "must" makes the assertion weaker than the one without it.[9] But this is not generally true. For sometimes "must" seems to indicate the necessity of what has been asserted: if one knows that Mary has put a bottle of wine either in the fridge or in the cupboard, and one has checked that it is not in the cupboard, it seems natural for one to conclude:

(5) It must be in the fridge.

As noticed by von Fintel and Gillies [17,18], what (4) and (5) have in common in the first place is that they signal the presence of an inference. Specifically, the verb "must" indicates that the speakers' grounds for their assertions are inferential, and hence indirect, but as they also argue, this need not mean that

[8] On the basis of data from 32 languages, Willett [12] proposed a taxonomy of markers encoding main types of sources of information. The main distinction is between direct (perceptual) and indirect access, and the latter can be further divided into other speaker's reports and inference. Since then, various aspects of evidentiality in language have been investigated, like for instance a developmental study by Papafragou and colleagues [13] on Korean speaking children's learning of evidential morphology and their ability to reason about sources of information.

[9] See e.g. Karttunen [14, p. 12], Groenendijk and Stokhof [15, p. 69] or Kratzer [16, p. 645]

these grounds are weak or inconclusive. Again, some confusion could have been avoided if the variety of inference relations had been attended to.

Of course, sometimes we convey information about our evidential grounds in more direct ways, as when we say that we *saw* that John crossed the street, or that it *seems* to us that Harriet is worried, or by the use of such words as "probably," "presumably," "possibly," "apparently," "allegedly," "putatively," and so on. But in this paper we focus on "must" and "should" in their roles as evidential markers, or rather, we focus on their Dutch counterparts "moet wel" ("must"[10]) and "zal wel" ("will," "should"). Evidential markers can serve a number of purposes. For instance, they may be used to indicate the *source* of the speaker's evidence: whether it is perceptual evidence, or evidence from testimony, or evidence from some third type of source still. They may also be used to indicate the *quality* of the evidence (e.g., indicate how reliable the source was). We will be mainly interested in the question of whether "moet wel" and "zal wel" play any distinctive role in signaling the *kind* of inference that is involved in making whatever evidence the speaker has bear on the content of her assertion. Can anything systematic be said about whether these markers go better with some type or types of inference than with others?

To clarify this question, note that the inference underlying the assertion of (4) in our example is most plausibly thought of as being abductive, that is, as an inference to the best explanation: that the friend is out is the best explanation for the evidence that the speaker has, to wit, that her friend does not answer the phone. In the example of Susan, it rather seems to be some form of inductive reasoning that warrants the assertion of (3): the people we met in our lives who had studied classical languages for a number of years were typically able to translate Latin phrases; given that Susan studied classical languages for a number of years, we expect her to be able to translate the designated phrase. Naturally, the inferential connection between evidence and grounds for assertion may also be deductive, as in the case of (5). The question we are interested in is whether the use of "moet wel" and "zal wel" gives us any indication as to what kind of inference (if any at all) led the speaker to feel warranted in making the assertion she did on the basis of the evidence she had.

4 Experiment: Linguistic Markers

Our experiment makes use of the typology of inferential conditionals discussed above. We look at a number of instances of the various types whose degrees of acceptability have been ascertained in previous research and we look whether these degrees are affected by inserting "moet wel" or "zal wel" into the sentences. We encountered the putative English counterparts of these markers already in our examples involving (4) and (3). "Must" and "zal wel" have also been described as inferential markers in the literature. As for the latter, Verbrugge [20] established a close connection between "zal wel" and inferential conditionals.

[10] "Wel" is a positive polar marker which has no counterpart in English; see Nuyts and Vonk [19, p. 701]. In German, "bestimmt" comes close.

Specifically, in an elicitation task in which they were requested to complete conditionals whose antecedents were given, participants tended to come up with a significantly higher number of inferential conditionals (as opposed to content conditionals) when they were in addition requested to use "zal wel" in the consequents than when they were not. As for "must," Dietz [21, p. 246] notes that in "It must be raining," the auxiliary indicates that the speaker only has (what he calls) "inferential evidence," and no direct observational evidence, that it is raining; see also the papers by von Fintel and Gillies cited earlier as well as Anderson [22], Papafragou [23], and Nuyts and Vonk [19]. However, so far researchers have not considered differentiating between the various types of inference by means of the said markers. Might not one marker go better with the expression of one type of inference and the other with the expression of another type of inference? Even more fundamentally, is "must" really an inferential marker? We are not aware of any empirical evidence that warrants a positive answer. Philosophers may be convinced that "must" can serve as an evidential marker, but philosophers were also convinced that the acceptability of a conditional is equal to the corresponding conditional probability, and—as Douven and Verbrugge showed—empirical evidence gives the lie to that thought.

To investigate the aforementioned questions, we used the materials of the experiment described in the previous section. We inserted "moet wel" and "zal wel" into the consequents of the conditionals used in Douven and Verbrugge's main experiment and checked whether this made a difference to acceptability ratings and their correlations with probability ratings. We were particularly interested in the effect the presence of the auxiliaries has on these correlations for different types of conditionals.

4.1 Method

Participants
Fifty seven students of the University of Leuven took part in the experiment.

Design
The type of conditional (DI / AI / II) was manipulated within subjects. The different lexical markers were manipulated between subjects.

Materials and Procedure
All materials were in Dutch, the participants' mother tongue. Thirty items were presented in a booklet. Every participant had to evaluate ten abductive, ten inductive, and ten deductive items. Items were randomized per booklet and the booklets were randomly distributed in the lecture hall. The items consisted of the same context–sentence pairs that were presented to the participants in the main experiment of Douven and Verbrugge [2010] who were asked to judge the acceptability of conditionals, except that the conditionals now contained the markers "moet wel" and, respectively, "zal wel." For instance, to the AI conditional "Als Judy's iPod niet in haar tas zit, dan heeft iemand die gestolen" ("If Judy's iPod is not in her bag, then someone has stolen it"), whose acceptability participants in Douven and Verbrugge's experiment had been asked to grade, corresponded

in our experiment the AI conditionals "Als Judy's iPod niet in haar tas zit, dan moet iemand die wel gestolen hebben" ("If Judy's iPod is not in her bag, then someone must have stolen it") and "Als Judy's iPod niet in haar tas zit, dan zal iemand die wel gestolen" ("If Judy's iPod is not in her bag, then someone will have stolen it").

Participants ($N = 30$) in the condition "moet wel" were asked to judge the acceptability of the conditionals containing "moet wel." Participants ($N = 27$) in the condition "zal wel" were asked to judge the acceptability of the conditionals containing "zal wel." The instructions were the same as the ones used in the acceptability condition of Douven and Verbrugge's experiment.

Results

Comparisons with the condition investigating the probability were set up. We computed the mean per sentence (ten abductive sentences, ten inductive sentences, and ten deductive sentences) over the participants. We thus obtained a mean for each of the thirty sentences. Then we computed the correlations between the condition with marker ("zal wel" / "moet wel") and the mean probabilities per item obtained on the basis of Douven and Verbrugge's experiment.

For the thirty sentences in the experiment, probabilities as obtained in Douven and Verbrugge's experiment and acceptability of the sentences with "zal wel" were highly correlated: $N = 30$, Spearman $R = .837712$, $t(N - 2) = 8.116914$, $p < .0001$. For the thirty sentences in the experiment, probabilities as obtained in Douven and Verbrugge's experiment and acceptability of the sentences with "moet wel" were highly correlated: $N = 30$, Spearman $R = .855871$, $t(N - 2) = 8.756641$, $p < .0001$.

We next considered different types of conditionals. For the abductive sentences, probability and "moet wel" were highly correlated: $N = 10$, Spearman $R = .993902$, $t(N - 2) = 25.49522$, $p < .00001$. We obtained similar results for "zal wel": $N = 10$, Spearman $R = .936175$, $t(N - 2) = 7.532386$, $p = .0001$. For the inductive sentences, correlations did not reach significance level; for "moet wel" and "zal wel": $N = 10$, Spearman $R = .612121$, $t(N - 2) = 2.189453$, $p = .059972$. For the deductive sentences, probability was highly correlated with "moet wel": $N = 10$, Spearman $R = .814593$, $t(N - 2) = 3.972223$, $p < .01$. For probability and "zal wel," the correlation did not reach significance level: $N = 10$, Spearman $R = .613985$, $t(N - 2) = 2.200141$, $p = .058981$.

Comparison with the results from Douven and Verbrugge's experiment showed that, overall, the markers had little effect on the perceived acceptability of the conditionals as well as, correspondingly, on the correlation between acceptability and probability (see Table 1). However, splitting the results for the various types of conditionals was more revealing. It appeared that, while for II conditionals, adding either of our two markers had virtually no effect on the correlation between acceptability and probability, adding them to the AI conditionals did increase the said correlation, even to the extent of yielding a near-to-perfect correlation for the AI conditionals containing "moet wel" (for no marker, $R = .8997$; for "zal wel," $R = .936175$; and for "moet wel," $R = .993902$). For DI conditionals,

Table 1. Correlations with and without markers (* = marginally significant; for all other results $p < .01$)

	All	DI	II	AI
No marker	.851102	.818182	.620064*	.899700
"zal wel"	.837712	.613985	.612121*	.936175
"moet wel"	.855871	.814593	.612121*	.993902

adding "moet wel" had almost no effect, but adding "zal wel" led to a considerable decrease of the correlation between acceptability and probability.

4.2 Discussion

These results confirm that "moet wel" and "zal wel" are inferential markers, given that (i) inserting either of them in the AI conditionals has the effect of increasing the correlation between acceptability and probability, and (ii) inserting "zal wel" in the DI conditionals has the effect of decreasing that correlation. It is no surprise that inserting "moet wel" in the DI conditionals does not have a similar effect: like "must" in English, "moet wel" can also serve as an alethic modality, and may thus be naturally interpreted in a DI conditional as underlining the necessity of the inference.

5 Concluding Remarks

The typology proposed in Douven and Verbrugge [10] helps to explain why adding inferential markers to conditionals makes the systematic kind of difference that we found in our experiment. That is further support for the thought that this typology is of theoretical significance. The experimental work reported here is part of a larger project. Experiments concerning the English "must," "should," and "will" are currently being undertaken, and the results are to be compared to the results of the experiment reported above. A further avenue for future research concerns applications of these markers. For instance, if some markers can be used as a kind of litmus test for distinguishing between various types of conditionals, then testing for the effect that adding these markers to conditionals has may help us in classifying conditionals whose type is controversial. Perhaps the most important part of the project is to see whether the current typology of conditionals can serve to ground a new semantics and/or pragmatics of conditionals. Once it is recognized that various types of inferential connection may be involved, it becomes quite plausible to claim that at least many conditionals require for their acceptability the existence of an inferential link between antecedent and consequent. Whether this is then to be taken as a brute pragmatic fact, or whether it has a deeper explanation in terms of truth conditions, is the question we ultimately hope to answer. The investigations reported here are meant as a first step toward that answer.

References

1. Edgington, D.: On conditionals. Mind **104**(414) (1995) 235–329
2. Bennett, J.: A Philosophical Guide to Conditionals. Oxford University Press, Oxford (2003)
3. Dancygier, B.: Conditionals and Predictions: Time, Knowledge and Causation in Conditional Constructions. Cambridge University Press, Cambridge (1998)
4. Dancygier, B.: Classyfying conditionals: Form and function. English Language and Linguistics **7** (2003) 309–323
5. Dancygier, B., Sweetser, E.: Mental Spaces in Grammar. Conditional Constructions. Cambridge University Press, Cambridge (2005)
6. Declerck, R., Reed, S.: Conditionals: A Comprehensive Empirical Analysis. Mouton de Gruyter, Berlin/New York (2001)
7. Haegeman, L.: Conditional clauses: External and internal syntax. Mind & Language **18**(4) (2003) 317–339
8. Woods, M.: Conditionals. Oxford University Press, Oxford (2003)
9. Sanford, D.H.: If P, Then Q: Conditionals and the Foundations of Reasoning. Routledge, London (1989)
10. Douven, I., Verbrugge, S.: The Adams family. Cognition **117** (2010) 302–318
11. Adams, E.W.: The logic of conditionals. Inquiry **8** (1965) 166–197
12. Willett, T.: A cross-linguistic survey of the grammaticization of evidentiality. Studies in Language **12**(1) (1988) 51–97
13. Papafragou, A., Li, P., Choi, Y., Han, C.: Evidentiality in language and cognition. Cognition **103** (2007) 253–299
14. Karttunen, L.: "Possible" and "Must". In: Syntax and Semantics. Academic Press, New York (1972)
15. Groenendijk, J.A., Stokhof, M.J.: Modality and conversational information. Theoretical Linguistics **2** (1975) 61–112
16. Kratzer, A.: Modality. In von Stechow, A., Wunderlich, D., eds.: Semantics: An International Handbook of Contemporary Research. de Gruyter, Berlin (1991) 38–74
17. von Fintel, K., Gillies, A.: An opinionated guide to epistemic modality. Oxford Studies in Epistemology **2** (2007) 32–63
18. von Fintel, K., Gillies, A.S.: *Must* . . . stay . . . strong! Natural Language Semantics **18**(4) (2010) 351–383
19. Nuyts, J., Vonk, W.: Epistemic modality and focus in Dutch. Linguistics **37**(4) (1999) 699–737
20. Verbrugge, S.: Frequency and lexical marking of different types of conditionals by children and adults in an elicitation task. In: Proceedings of the EuroCogSci 07 Conference. (2007)
21. Dietz, R.: Epistemic modals and correct disagreement. In García-Carpintero, M., Kölbel, M., eds.: Relative Truth. Oxford University Press, Oxford (2008) 23–262
22. Anderson, L.B.: Evidentials, paths of change and mental maps: Typologically regular asymmetries. In Chage, W.L., Nichols, J., eds.: Evidentiality: The Linguistic Coding of Epistemology. Ablex, Norwood NJ (1986)
23. Papafragou, A.: Inference and word meaning: The case of modal auxiliaries. Lingua **105** (1998) 1–47

Using Quantified Epistemic Logic as a Modeling Tool in Cognitive Neuropsychology

Rasmus K. Rendsvig
Philosophy and Science Studies, Roskilde University

1 Introduction

The classic *modus operandi* for model construction in cognitive neuropsychology (CN) is by use of box-and-arrow diagrams[1] to capture the functional architecture of the mental information-processing system under consideration. In such diagrams, of which Figure 1 is an example, boxes represent particular components of the system and arrows represent pathways of information flow. Along with a suitable interpretation, box-and-arrow diagrams represent typical CN theories, consisting of statements regarding what specific modules are included in the system as well as statements regarding how information may flow between these components.

Recently, it has been argued that this methodology should be augmented by the use of computational models, allowing for the realization of CN theories in the form of executable computer programs, structurally isomorphic to the box-and-arrow version the theory, [3, p. 166]:

> This way of doing cognitive psychology is called computational cognitive psychology, and its virtues are sufficiently extensive that one might argue that all theorizing in cognitive psychology should be accompanied by computational modeling.

Though *epistemic logics* are not by themselves executable programs, as a modeling tool they do however possess many of the features and virtues required by [3]. It is the purpose of this paper to discuss the viability of using epistemic logics as a modeling tool in cognitive neuropsychology.

Benefits of epistemic logic. There are three fields that could benefit from epistemic logical modeling of theories from cognitive neuropsychology. First, a motivation for constructing formal logical models aimed at cognitive neuropsychology is that precision and logical entailment can provide explanatory force and working hypotheses. Further, epistemic logics can be used to express higher cognitive functions such as knowledge and belief in straightforward languages. This further allows the formulation of derived principles, such as object recognition. That epistemic logics can straightforwardly express such functions further makes it easy to read off predictions from logical theories, allowing for simple comparisons with empirical observations.

Second, logicians interested in modeling information flow and communication acts could gain more realistic models of the internal parts of these processes, if they took to modeling the functional architecture underlying our abilities to perform such actions.

[1] The "universal notation in modern cognitive neuropsychology", according to [4].

Finally, if logicians and cognitive neuropsychologists merged formal tools and empirical insight, philosophers would stand to gain. Having empirical theories couched in flexible modal logical frameworks would allow precise analyses of philosophical problems using tools with which many philosophers are already familiar. To exemplify, the model introduced below has been used in [14] to give novel analyses of Frege's Puzzle about Identity, based on formal notions of semantic competence.

Plan of attack. As argued in [3], a proper modeling of a CN theory should be true to that theory in an isomorphic sense: the formal model should include exactly the modules and pathways of the CN theory, while maintaining their separate and joint functions. However, while many CN theories include a *semantic system* module, it is far from clear where such a collection of concepts can be found in, e.g., a quantified S5 logic. To overcome this difficulty, we here take the perspective that a logical model of a CN theory is composed of *both* a formal logic as well as semantics for this logic. More specifically, then the modules of a functional architecture is represented by model-theoretic structures over which agents' capabilities can then be expressed using formal logical syntax. A complete logic for the model-theoretic structure may then be seen as a theory detailing the capabilities of an agent with such a mental makeup.

The present approach differs from the computational cognitive neuropsychological (CCN) way outlined in [3] in an important aspect. The CCN approach there discussed construct a model of normal behavior, which can the be 'lesioned' to simulate brain damage, and thereby compare to experimental observations. In the present, we reverse this approach. Instead of constructing first a stronger logic for normal behavior from which we can then remove, we construct a weaker logic for the completely damaged, to which abilities can then be added. As such, the logic is generic: in order to produce a subject-specific logic, further assumptions regarding specific knowledge and abilities must be made.

Evaluating logical models. Given that a motivation for logical modeling of CN theories was a promise of working hypotheses, one would expect such hypotheses to be empirically testable. As the observables in CN tests are subjects' performance based on given input, and these abilities are described by formal logical statements over the model-theoretic structure modeling the functional architecture, it is thus natural to take such hypotheses as constituted by the formulas satisfied in the actual world of the semantic model. As will be shown below, this is a feasible way of comparing formal model with empirical observations. However, there is no hope that *any* normal epistemic logical model can produce hypotheses all of which are consistent with made observations. *Qua* the problem of logical omniscience, any modeled agent will always know all logical consequences of her knowledge [7], which *no* subject can do. This results in a problem for the present approach in relation to evaluating, rejecting and refining the formal models, for every hypothesis may now be wrong for two reasons: it may be the case that the modeled functional architecture does not represent reality, or it may be that the chosen hypothesis requires reasoning skills beyond normal, human capabilities. In the latter case, this will result in a non-answer, and the danger is that the correct functional architecture is rejected. We offer no solution here, but note that this is a problem for which one needs to control when performing theory evaluation.

Structure. The structure of the paper is as follows. In the ensuing section, we introduce a simplified version of a CN theory of the structure of lexical, semantic competence (SLC) from [12]. This is to act as a toy CN theory, which will be modeled using a two-sorted quantified epistemic logic (QEL) introduced in section 3. Most weight is on section 4, where the connections between the QEL and the SLC are drawn. It is first argued that the modules of the SLC are represented in the two-sorted model-theory. Secondly, it is argued that the three distinct competence types of the SLC are expressible in the formal language, and that their dissociations are preserved in the two-sorted logic.[2] Finally, we consider studies which show the shortcomings of the present model, and thereby falsify the presented model. We then conclude.

2 The Structure of Lexical Competence

In [12], a box-and-arrow theory of the *s*tructure of semantic, *l*exical *c*ompetence (SLC) is constructed on the basis of studies from cognitive neuropsychology.[3] The elements of the SLC consist of three competence types defined over four ontologies, two of these being mental modules, see Figure 1. The three competence types are *inferential competence* and two types of *referential competence*, being *naming* and *application*. The four ontologies include one of *external objects*, one of *external words* (e.g., spoken or written words) and two mental modules: a *word lexicon* and the *semantic system*. This structure is illustrated in Figure 1.

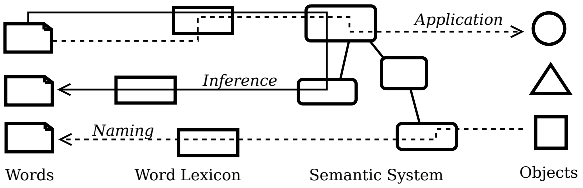


Fig. 1. A simplified illustration of the SLC. Elements in the WL are not connected, only elements in the SS are. Inferential competence requires connecting two items from the word lexicon *through* the semantic lexicon.

Word Lexicon, Semantic Lexicon and Inferential Competence. *Inferential competence* lies between three ontologies: *external word*, *word lexicon* and the *semantic system*. On input, the external word is first analyzed and related to an mental representation from the word lexicon (WL). In [12], two word lexica are included for different input, a phonetic and a graphical. Here, attention is restricted to a simplified structure with only one arbitrary such, consisting only

[2] Due to space restrictions, proof theory will not be considered, but a complete axiomatization can easily be constructed based on the general completeness result for many-sorted modal logics from [13].

[3] For the review of these studies, arguments for the structure and references to relevant literature, the reader is referred to [12]. The presentation here differs slightly, but in-essentially, from that.

of proper names,[4] as illustrated in Figure 1. Using a graphical lexicon as an example, the word lexicon consists of the words an agent is able to recognize in writing. Secondly, the mental representation of the word is related to a mental concept in the semantic system (SS). The SS is a collection of non-linguistic, mental concepts possessed by the agent, *distinct* from the WL. The semantic system reflects the agent's mental model of the world, and the items in this system stand in various relations to one another.[5] In contrast, in the WL connections between the various items *do not exist.* Such only exist *via* the SS. The third step is exactly a connection between two entries in the SS. Finally, the latter of these are connected to an entry in the WL and output can be performed.[6] *Inferential competence* is the ability to correctly connect lexical items via the SS, e.g., connecting 'dog' to 'animal'. This ability underlies performance such as stating definitions, paraphrasing and finding synonyms.

Referential Competence and External Objects. *Referential competence* is "the ability to map lexical items onto the world" [12, p. 60]. This is an ability involving all four ontologies, the last being *external objects.* It consists of two distinct subsystems. The first is *naming.* This is the act of retrieving a lexical item from the WL when presented with an object. It is a two-step process, where first the external object is connected to an suitable concept in the SS, which is then connected to a WL item for output. The second subsystem is that of *application.* Application is the act of identifying an object when presented with a word. Again, this is a two-stage process, where first the WL item is connected to an SS item, which is then connected to an external object. A naming or application deficit can occur if either stage is affected: if, e.g., either an object is not mapped to a suitable concept due to lack of recognition, or a suitable concept is not mapped to the correct (or any) word, then a naming procedure will not be successfully completed.

Empirical Backing for Multiple Modules and Competence Types. The SLC may seem overly complex. It may be questioned why one should distinguish between word and semantic type modules, or why referential competence is composed of two separate competence types, instead of one bi-directional. The reasons for these distinctions are based on empirical studies from cognitive neuropsychology where reviews of subjects with various brain-injuries indicate that these modules of human cognition are separate (see [12] for references).

The distinction between WL and SS is further supported in [8] by cases where patients are able to recognize various objects, but are unable to name them (they cannot access the WL from the SS). In the opposite direction, cases are reported where patients are able to reason about objects and their relations when shown

[4] To only include proper names is technically motivated, as the modeling would otherwise require second-order expressivity. This is returned to below.

[5] Marconi uses the term '*semantic lexicon*', but to keep this presentation in line with standard cognitive neuropsychological terminology, 'semantic system' is used instead.

[6] For simplicity, a distinction between input and output lexica will not be made. See [15] for discussion.

objects, yet unable to do the same when prompted by their names (i.e., the patients cannot access the `SS` from the `WL`). The latter indicates that reasoning is done with elements from the `SS`, rather than with items the `WL`.

Regarding competence types, it is stressed in [12] that inferential and referential competence are distinct abilities. Specifically, it is argued that the ability to name objects does not imply inferential competence with the used name, and, *vice versa*, that inferential knowledge about a name does not imply the ability to use it in naming tasks. No conclusions are drawn with respect to the relationship between inferential competence and application. Further, application is dissociated from naming, in the sense that application can be preserved while naming is lost. No evidence is presented for the opposite dissociation, i.e. that application can be lost, but naming maintained.

In the following section, a model will be constructed which include the mentioned ontologies over which the competence types can be defined, and in which these are appropriately dissociated.

3 Modeling the Structure of Lexical Competence

To construct a toy model of the SLC, a two-sorted first-order epistemic logic will be used. A very limited syntax is used, though the syntax and semantics could easily be extended to include more agents, sorts, function- and relation symbols, cf. [13].

A two-sorted language is used to ensure that the model respects the dissociation of word lexicon and semantic system. The first sort, σ_{OBJ}, is used to represent external objects and the semantic system entries. As such, these are *non-linguistic* in nature. The second sort, σ_{LEX}, is used to represent the external words from the agent's language and entries in the word lexicon. Had terms been used to represent both simultaneously, the model would be in contradiction with empirical evidence.

The choice of quantified epistemic logic (QEL) fits well with the SLC, if one assumes the competence types to be (perhaps implicitly) knowledge-based.[7] The notions of object identification required for application is well-understood as modeled in the quantified S5 framework, cf. [5]. The 'knowing who/what' constructions using *de re*-constructions in QEL from [10] captures nicely the knowledge required for object identification by the subjects reviewed in [12]. This is returned to in the following section.

Syntax. Define language $\mathcal{L}$ to include two sorts, σ_{OBJ} and σ_{LEX}. For sort σ_{OBJ}, include 1) a countable set of *object constant symbols*, $OBJ = \{a, b, c, ...\}$, and 2) a countably infinite set of *object variables* $VAR = \{x_1, x_2, ...\}$. The set of terms of sort σ_{OBJ} is $TER_{OBJ} = OBJ \cup VAR$. For sort σ_{LEX}, include 1) a countable set of *name constant symbols*, $LEX = \{n_1, n_2, ...\}$, and 2) a countably infinite set of *name variables*, $VAR_{LEX} = \{\dot{x}_1, \dot{x}_2, ...\}$. The set of terms of sort σ_{LEX} is $TER_{LEX} = LEX \cup VAR_{LEX}$.

[7] In [3], visual recognition tasks are explicitly referred to in terms of knowledge (see, e.g., p. 149). Which type of knowledge is however not discussed.

Include further in $\mathcal{L}$ a unary function symbol, μ, of arity $TER_{LEX} \longrightarrow TER_{OBJ}$. The set of all *terms, TER,* of $\mathcal{L}$ are $OBJ \cup VAR \cup LEX \cup VAR_{LEX} \cup \{\mu(t)\}$, for all $t \in LEX \cup VAR_{LEX}$. Finally, include the binary relation symbol for identity, $=$. The well-formed formulas of $\mathcal{L}$ are given by

$$\varphi ::= (t_1 = t_2) \,|\, \neg\varphi \,|\, \varphi \wedge \psi \,|\, \forall x\varphi \,|\, K_i\varphi$$

The definitions of the remaining boolean connectives, the dual operator of K_i, $\hat{K_i}$, the existential quantifier and free/bound variables and sentences are all defined as usual. Though a mono-agent system, the operators are indexed by i to allow third-person reference to agent i.

Semantics. Define a 2QEL *model* to be a quadruple $M = \langle W, \sim, Dom, \mathcal{I}\rangle$ where

1. $W = \{w, w_1, w_2, ...\}$ is a set of *epistemic alternatives* to actual world w.
2. $\sim$ is an *indistinguishability (equivalence) relation* on $W \times W$.
3. $Dom = Obj \cup Nam$ is the *(constant) domain of quantification*, where $Obj = \{d_1, d_2, ...\}$ is a non-empty set of *objects*, and $Nam = \{\dot{n}_1, \dot{n}_2, ..., \dot{n}_k\}$ is a finite, non-empty set of *names.*
4. $\mathcal{I}$ is an *interpretation function* such that
$$\mathcal{I} : OBJ \times W \longrightarrow Obj \,|\, \mathcal{I} : LEX \longrightarrow Nam \,|\, \mathcal{I} : \{\mu\} \times W \longrightarrow Obj^{Nam}$$

To assign values to variables, define a *valuation function*, v, by

$$v : VAR \longrightarrow Obj \,|\, v : VAR_{LEX} \longrightarrow Nam$$

and a *x-variant of* v as a valuation v' such that $v'(y) = v(y)$ for all $y \in VAR_{(LEX)}/\{x\}$.

Truth conditions for formulas of $\mathcal{L}$ are now defined as follows:

$$\begin{array}{ll}
M, w \models_v (t_1 = t_2) & \text{iff } d_1 = d_2, \text{ where } d_i = \begin{cases} v\,(t_i) & \text{if } t_i \in VAR \cup VAR_{LEX} \\ \mathcal{I}\,(w, t_i) & \text{if } t_i \in OBJ \\ \mathcal{I}\,(t_i) & \text{if } t_i \in LEX \end{cases} \\
M, w \models_v \varphi \wedge \psi & \text{iff } M, w \models_v \varphi \text{ and } M, w \models_v \psi \\
M, w \models_v \neg\varphi & \text{iff not } M, w \models_v \varphi \\
M, w \models_v K_i\varphi & \text{iff for all } w' \text{ such that } w \sim w',\ M, w' \models_v \varphi \\
M, w \models_v \forall x\varphi\,(x) & \text{iff for all } x\text{-variants } v' \text{ of } v,\ M, w \models_{v'} \varphi\,(x)
\end{array}$$

Comments on the semantics are postponed to the ensuing section.

Logic. A sound and complete two-sorted logic for the presented semantics can be found in [13]. The logic is here denoted $\mathsf{QS5}_{(\sigma_{LEX}, \sigma_{OBJ})}$.

4 Model Validation

As mentioned above, the modules of the functional architecture of the SLC is represented by model-theoretic structures, over which the agent's capabilities can then be expressed using the formal logical syntax. So far, the structures introduced bear little resemblance to the SLC, and it will be a first task to extract this hidden structure. Secondly, it is shown that the logical model can express the three competence types and that the dissociation properties are preserved in the logic.

4.1 Ontologies

The two sets of external objects and external words are easy to identify in the model-theoretic structure. The external objects constitute the sub-domain Obj, and are denoted in the syntax by the terms TER_{OBJ}, when these occur outside the scope of an operator. External words (proper names) constitute the sub-domain Nam denoted by the terms TER_{LEX}, when occurring outside the scope of an operator.

The word lexicon and the semantic system are harder to identify. The strategy is to extract a suitable notion from the already defined semantic structure. To bite the bullet, we commence with the more complicated semantic system.

Semantic System. In order to include a befitting, albeit very simple notion, define an *object indistinguishability relation* $\sim^a_w \subseteq Obj \times Obj$:

$$d \sim^a_w d' \text{ iff } \exists w' \sim w : \mathcal{I}(a, w) = d \text{ and } \mathcal{I}(a, w') = d'.$$

and from this define the agent's *individual concept class for a at w* by $C^a_w(d) = \{d' : d \sim_i^{a,w} d'\}$. The semantic system of agent i may then be defined as the collection of non-empty concept classes: $\mathsf{SS}_i = \{C^a_w(d) : C^a_w(d) \neq \emptyset\}$.

The set $C^a_w(d)$ consists of the objects indistinguishable to the agent by $a \in OBJ$ from object $d \in Obj$ in the part of the given model connected to w by $\sim$. As an example, consider a scenario with two cups (d and d' from Obj) upside down on a table, where one cup conceals a ball. Let a denote *the cup containing the ball*, say d, so $\mathcal{I}(a, w) = d$. If the agent is not informed of which of the two cups contains the ball, i.e. which is a, there will be an alternative w' to w such that $\mathcal{I}(a, w') = d'$. Hence, $d \sim^a_w d'$ so $d' \in C^a_w(d)$. The interpretation is that the agent cannot tell cups d and d' apart *with respect to which conceals the ball.*[8]

Important properties of individual concepts can be expressed in $\mathcal{L}$ (see [13]). For present purposes, most importantly we have that

$$|C^a_w(d)| = 1 \text{ iff } M, w \models_v \exists x K_i(x = a), \tag{1}$$

i.e. the agents has a singleton concept of a in w iff it is the case that the agent *knows which object a is*, in the readings of [9,5]. The intuition behind this reading is that the satisfaction of $\exists x K_i(x = a)$ requires that the interpretation of a is constant across i's epistemic alternatives. Hence, there is no uncertainty for i with respect to which object possesses feature a – i knows which object a is. Using a contingent identity system for objects, i.e. giving these a non-rigid interpretation as done in the semantics above, results in the invalidity of both $(a = b) \to K_i(a = b)$ and $(a = b) \to \exists x K_i(x = b)$. Hence, agent i does *not* by default know whether objects are identical when identified by different features, and neither is the agent able to *identify* objects by default – as in the example above. This is a good example of how the present is a weak, generic model: subject specific abilities regarding identificatory abilities needs to be made as further assumptions on a per subject basis.

Word Lexicon. A suitable representation of the word lexicon is simpler to extract than for the SS. This is due to the non-world relative interpretation of

[8] Though the agent may be able to tell them apart with respect to other features, like their color or position.

name constants $n \in LEX$, which so far has gone without comment. The interpretation function $\mathcal{I}$ of the name constants is defined constant in order ensure that the agent is *syntactically competent*. From the definition of $\mathcal{I}$, it follows that $(n_1 = n_2) \to K_i(n_1 = n_2)$ is valid on this class of models. This corresponds formally to the incontingent identity system used in [11]. The interpretation is that whenever the agent is presented by two name tokens of the same type of name, the agent knows that these are tokens of the same name type. The assumptions is adopted as the patients reviewed in [12] are able to recognize the words utilized.

Notice that identity statements such as $(n_1 = n_2)$ *do not* convey any information regarding the *meaning* of the names. Rather, they express identity of the two *signs*. Hence, the identity '*London* = *London*' is true, where as the identity '*London* = *Londres*' is false – as the two first occurrences of '*London*' are two tokens (e.g. $n_1, n_2 \in LEX$) of the same type (the type being $\dot{n} \in Nam$), whereas the '*London*' and '*Londres*' are occurrences of two different name types, albeit with the same meaning.

Due to the simpler definition of $\mathcal{I}$ for name constants, we can define i's *name class for* n directly. Where $\dot{n} \in Nam$ and $n \in LEX$ this is the set $C_i^n(\dot{n}) = \{\dot{n}' : \mathcal{I}(n) = \dot{n}'\}$. The word lexicon of i is then the collection of such sets: $\mathsf{WL}_i = \{C_i^n(\dot{n}) : n \in LEX\}$. Each name class is a singleton equivalence class, and WL_i is a partition of Nam. Further, (1) holds for name classes if suitably modified, and the construction of WL_i therefore fits nicely fit the assumption of syntactic competence.

4.2 Interlude: Word Meanings

In order to model knowledge of the meanings of word tokens $n \in LEX$, these must first be assigned a meaning. In the clinical trials reviewed in [12], applying a name to it's meaning is done by extension identification. Therefore, a simple purely extensional theory of meaning have been embedded in the framework: the function symbol μ of arity $TER_{LEX} \longrightarrow TER_{OBJ}$. A meaning function rather than a relation is used as only proper names are included in the agent's language, and for these to have unambiguous meanings, the function requirement is natural. Given it's defined arity, μ assigns an element of TER_{OBJ} to each name in TER_{LEX}. From the viewpoints of the agents, μ hence assigns an object (meaning) to each name.

On the semantic level, $M, w \models_v (\mu(n) = a)$ is taken to state that the meaning of name n is the object a in the actual world w. The reference map μ is defined *world relatively*, i.e. the value $\mu(n)$ for $n \in LEX$, can change from world to world. This is the result of the world relative interpretation of μ given in the semantics above. Hence, *names are assigned values relative to epistemic alternatives*.

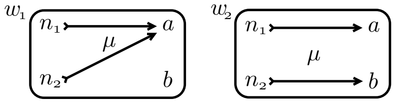


Fig. 2. The meaning function μ is defined world relatively, so meaning of a name may shift across epistemic alternatives.

The motivation for a world relative meaning function is the generic nature of the model. No agents will *by default* have knowledge of the meaning of words from their language, but further assumptions to that effect can be assumed in specific cases.

The simplifying assumption that the WL should include only proper names was technically motivated by the inclusion of μ. Had verbs been introduced in the agent's language, then μ should have assigned them relation symbols as meanings, and a second-order logic would be required.

4.3 Competence Types

Inferential Competence. Due to the restriction to proper names, the model is extremely limited in the features expressible regarding inferential competence. The expressible instances of inferential competence are limited to knowing relations between referring names and not inferential knowledge regarding names and verbs. As an example, one cannot express that the agent knows the true sentence 'name is verb' as the word lexicon does not contain a 'verb' entry. We are however able to express *knowledge of co-reference*:

$$K_i(\mu(n) = \mu(n')) \tag{2}$$

(2) states that i knows that n and n' co-refer, i.e. knows the two names to be synonyms. Based on (2), we may define that agent i is *generally inferentially competent with respect to n* by

$$M, w \models_v \forall \dot{x}((\mu(n) = \mu(\dot{x})) \rightarrow K_i(\mu(n) = \mu(\dot{x}))) \tag{3}$$

where $\dot{x} \in VAR_{LEX}$. If (3) is satisfied for all n, agent i will have full 'encyclopedic' knowledge of the singular terms of her language. This may however be 'Chinese Room style' knowledge, as it does not imply that any names can be applied nor that any objects can be named.

Referential Competence. Referential competence compromises two distinct relations between names and objects, relating these through the semantic system, namely *application* and *naming*. An agent can *apply a name* if when presented with a name, the agent can identify the appropriate referent. This ability can be expressed of agent i with respect to name n in w by

$$M, w \models_v \exists x K_i(\mu(n) = x) \tag{4}$$

i.e. there is an object which the agent can identify as being the referent of n. Given the assumption of syntactical competence, there is no uncertainty regarding which name is presented. Since the existential quantifier has scope over the knowledge operator, the interpretation of $\mu(n)$ is fixed across epistemic alternatives, and i thus knows which object n refers to.

To be able to *name an object*, the agent is required to be able to produce a correct name when presented with an object, say a. For this purpose, the *de re* formula $\exists \dot{x} K_i(\mu(\dot{x}) = a)$ is insufficient as $\mu(\dot{x})$ and a may simply co-vary across states. This means that i will be guessing *about which object is to be named*, and may therefore answer incorrectly. Since there may in this way be uncertainty regarding presented objects, naming must include a requirement that i can identify a, as well as know a name for a. This is captured by

$$M, w \models_v \exists x \exists \dot{x} K_i((x = a) \wedge (\mu(\dot{x}) = a)). \tag{5}$$

Here, the quantification and first conjunction ensures that i can identify the presented object a and the second conjunct ensures that the name refers to a in all epistemic alternatives.

Dissociations. As mentioned, inferential competence and naming are dissociated. This is preserved in the model in that neither (2) nor (3) alone imply (5). Nor does (5) alone imply either of the two. The dissociation of application from naming is also preserved, as (4) does not alone entail (5). That application does not imply naming is illustrated in Figure 3.

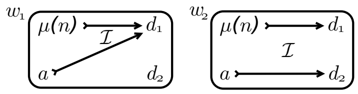


Fig. 3. Application and naming are not correlated. In actual world w_1, n refers to a and i can correctly apply n, but cannot name a using n: $w_1 \models_v (\mu(n) = a) \wedge \exists x K_i(\mu(n) = x)$, but $w_1 \models_v \neg\exists x \exists \dot{x} K_i((x = a) \wedge (\mu(\dot{x}) = a))$. Here, i cannot name a due to an ambiguous concept. a may be either of d_1 or d_2, and can therefore not be identified precisely enough to ensure a correct answer.

Whether application entails inferential competence, and whether naming entails application is not discussed in [12]. In the present model, however, these are modeled as dissociated in the sense that (4) does not entail, nor is entailed by, either (3) or (5). However, the modeled dissociations are *single instances* of the various abilities. Once more instances are regarded simultaneously, implicational relationships arise, as will be discussed below.

5 Hypotheses and Explanations

A motivation for constructing formal logical models is that precision and logical entailment can provide explanatory force and working hypotheses. One testable hypothesis of the present model predicts *lack of dissociation* between multiple application instantiations and inferential competence. Specifically, the model entails that subjects capable of applying two co-referring names will be knowledgeable of their co-reference:

$$\begin{gathered}\text{If } M, w \models_v (\mu(n) = \mu(n')) \text{ then} \\ M, w \models_v \exists x K_i(\mu(n) = x) \wedge \exists y K_i(\mu(n') = y) \rightarrow K_i(\mu(n) = \mu(n'))\end{gathered} \tag{6}$$

From this, an explanation why none of the studies reviewed in [12] show dissociation between application and inferential competence can be conjectured: simple inferential competence can come about as a bi-product of application, memory and deductive skill and may therefore require much damage before being severely impaired.

The formalizations of application and naming also suggests a reason why no cases where naming was intact, but application broken, was reported in [12]: the ability to name is very close to entailing the ability to apply a name. In fact, once (5) is instantiated with a specific name, it implies (4) for the same name. For application, the chosen object is identified by the subject *via* the mental representation 'the referent of n', whereas for naming, the presented object must first be identified by some other feature, e.g., a visual trademark. In case the mental representation in the SS of this feature is then identical to that of 'the referent of n', then the subject will be able to name a. Hence, one of the necessary conditions for naming almost implies the necessary and sufficient condition for application, why the latter will be observed accompanying the former.

Implicational relationships as the mentioned should allow for the refutation of the model. If subjects are found who possess abilities represented by the antecedents, but lack those of the consequents, the model can be regarded as refuted, though exactly what the problem is may not be obvious, qua the previously mentioned issue with logical omniscience.

As mentioned above, both the difference between orthographic and and phonological lexica as well as the difference between input and output lexica was ignored in this presentation. Since the logical model therefore is based on an arguably wrong functional architecture, it should be possible to find inconsistencies between model and observed subject behavior. This is indeed the case. For, if the presented model was correct, then the word lexicon entries should play both orthographic and phonological roles for words the agent knows in both speech and writing. Given such a word, an agent able to name with the word should always be able to do so both orally and in writing. That is, the hypothesis that agent i is able to name a, i.e. $\exists x \exists \dot{x} K_i((x = a) \wedge (\mu(\dot{x}) = a))$ ((5) from above), requires that the subject can produce name(s) $\dot{x}$ both orally and in writing. This, however, is not the case, as is illustrated, e.g., by the case of RCM, an 82-year old woman, reported on in [8, p. 191]. When prompted with a picture of a turtle, RCM was able to correctly name it orally using 'turtle', but named it incorrectly in writing, using 'snake'. As RCM repeatedly made similar mistakes with respect to written word tasks but not with oral naming tasks, this case can be taken to show that damage to the orthographic (output) lexicon does not imply damage to the phonological (output) lexicon. This is not possible in the presented model, why the model is refuted.[9]

6 Conclusions and Further Perspectives

In the present paper, we have looked at the possibilities of using epistemic logic as a modeling tool for cognitive neuropsychological theories by a toy model construction. It has been shown how the functional architecture can be represented using a combination of model theory and formal syntax. It was further shown that the constructed model respected important dissociations, but also how the model could be refuted by suitable empirical evidence contradicting an hypothesis of the model. In conclusion, though the model is incorrect and simplistic, a serious epistemic logical approach to modeling functional architecture theories from cognitive science could possibly be of value. An attempt at making a proper model of a full CN theory would be an obvious next step.

A clear limitation of the presented model is that the formal semantic system lacks content. The limitation to objects only should be lifted as to include various properties and relations as well, and moreover, the representation of objects are black boxed behind constants without a precise interpretation. Before a clear picture such concepts' role can be given, an explicit theory of *object recognition* must be incorporated. It would be interesting to see the effects of formalizing, e.g., *geon theory* of [2] and 'plug it in' in the place of the object constants.

[9] This can easily be remedied by distinguishing between phonological and orthographic word terms, i.e. by the addition of a further word sort.

More structure could also be provided by attempting to incorporate elements from *conceptual spaces theory* [6]. Finally, knowledge operators are too strong in some cases – RCM from above being a case in point. In situations where subjects answer consistently but wrongly, belief operators would be better suited. A range of competence levels could be captured using the various operators from [1].

The style of modeling semantic competence presented here differs from the way the conceptual theory of [12] and other cited literature tend to regard these matters. Here, competence types was defined relative to specific words, and competence judged on a case-by-case basis. Many studies from cognitive neuropsychology base their conclusions on percent-wise correct performance over one or more test batteries and therefore focus on impaired connections between modules. In order to facilitate comparison of formal models and empirical research, the case-by-case methodology must be reconciled with this more general approach.

References

1. Alexandru Baltag and Sonja Smets. A Qualitative Theory of Dynamic Interactive Belief Revision. In *LOFT7*, pages 11–58. Amsterdam University Press, 2008.
2. Irving Biederman. Recent Psychophysical and Neural Research in Shape Recognition. In I. Rentschler, I. Biederman, and N. Osaka, editors, *Object Recognition, Attention, and Action*, chapter 5, pages 71–88. Springer, 2007.
3. Max Coltheart. Cognitive Neuropsychology. In John Wixted, editor, *Methodology in Experimental Psychology*, chapter Cognitive Neuropsychology, pages 139–174. John Wiley & Sons, 2002.
4. Max Coltheart. Cognitive Neuropsychology. *Scholarpedia*, 3(2):3644, 2008.
5. Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge*. The MIT Press, 1995.
6. Peter Gärdenfors. *Conceptual Spaces*. The MIT Press, 2000.
7. Vincent F. Hendricks. *Mainstream and Formal Epistemology*. Cambridge University Press, 2006.
8. Argye E. Hillis. The Organization of the Lexical System. In Brenda Rapp, editor, *The Handbook of Cognitive Neuropsychology*, pages 185–210. Psychology Press, 2001.
9. Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press. Reprinted in 2005, prepared by V.F. Hendricks and J. Symons, King's College Publications, 1962.
10. Jaakko Hintikka. Different Constructions in Terms of 'Knows'. In Jonathan Dancy and Ernest Sosa, editors, *A Companion to Epistemology*. Wiley-Blackwell, 1994.
11. G.E. Hughes and M.J. Cresswell. *A New Introduction to Modal Logic*. Routledge, 1996.
12. Diego Marconi. *Lexical Competence*. The MIT Press, Cambridge, MA, 1997.
13. Rasmus Kræmmer Rendsvig. Towards a Theory of Semantic Competence. Master's thesis, Roskilde University, http://rudar.ruc.dk/handle/1800/6874, 2011.
14. Rasmus Kræmmer Rendsvig. Modeling Semantic Competence: a Critical Review of Frege's Puzzle about Identity. In Dan Lassiter and Marija Slavkovic, editors, *New Directions in Logic, Language, and Computation*. Springer, 2012.
15. Marie-Josèphe Tainturier and Brenda Rapp. The Spelling Process. In Brenda Rapp, editor, *The Handbook of Cognitive Neuropsychology*, pages 263–290. Psychology Press, 2001.

Easy Solutions For a Hard Problem? The Computational Complexity of Reciprocals with Quantificational Antecedents[⋆]

Fabian Schlotterbeck and Oliver Bott

Collaborative Research Center 833, University of Tübingen
{fabian.schlotterbeck,oliver.bott}@uni-tuebingen.de

Abstract. The PTIME-Cognition Hypothesis states that cognitive capacities are limited to those functions that are computationally tractable (Frixione 2001). We applied this hypothesis to semantic processing and investigated whether computational complexity affects interpretation preferences of reciprocal sentences with quantificational antecedents, that is sentences of the form *Q dots are (directly) connected to each other*. Depending on the quantifier, some of their interpretations are computationally tractable whereas others are not. We conducted a picture completion experiment and two picture verification experiments and tested whether comprehenders shift from an intractable meaning to a tractable interpretation which would have been dispreferred otherwise. The results suggest that intractable readings are possible, but their verification rapidly exceeds cognitive capacities if it cannot be solved using simple heuristics.

1 Introduction

In natural language semantics sentence meaning is commonly modeled by means of formal logics. Recently, semanticists have started to ask whether their models are cognitively plausible (e.g. Pietroski, Lidz, Hunter & Halberda (2009)). Obviously, for a semantic theory to be cognitively realistic the assumed meanings have to be bounded in computational complexity since they have to be computed in real time by processors with limited resources. In line with this consideration, Frixione (2001) has proposed the *PTIME-Cognition Hypothesis* (PCH) as a general research heuristic for cognitive science: cognitive functions have to be limited to those functions that are computationally tractable (see also van Rooij 2008).

[⋆] This research was funded by the German Research Foundation in the Project B1 of SFB 833. The authors would like to thank Fritz Hamm, Janina Radó, Wolfgang Sternefeld, Jakub Szymanik and three anonymous reviewers for valuable comments and discussions on earlier drafts of this paper. We would also like to thank our reviewers and audience at the *International Workshop on Computational Semantics 9* and the *International Conference on Linguistic Evidence 2012* where we have presented parts of the present paper. For their assistance in conducting the reported experiments we would like to give special thanks to Anna Pryslopska and Aysenur Sarcan.

However, computational complexity has hardly received any attention in formal semantics. In this paper we apply the PCH to semantic processing by looking at a particularly interesting test case that has been discussed by Szymanik (2010).

We investigated the interpretation of reciprocal sentences with quantificational antecedents of the form *Q (some N, all N, most N, ...) stand in relation R to each other*. What makes these sentences particularly interesting for our purposes is that, depending on the quantificational antecedent, the evaluation of one of their readings is NP-hard whereas other antecedents stay within the realm of PTIME verifiable meanings. We tested whether interpretations are limited to those for which verification is in PTIME (i.e. computationally tractable) as proposed in Szymanik (2010). In particular, we investigated whether comprehenders shift from the preferred interpretation to other – under normal circumstances dispreferred, but computationally tractable – readings in order to avoid having to deal with an intractable verification problem[1]. Consider (1).

(1) a. All of the students know each other.
b. All of the students followed each other into the room.

Reciprocals like (1) are notoriously ambiguous (Dalrymple, Kanazawa, Kim, McHombo & Peters 1998). The intuitively preferred reading of (1-a) is that any two members of the restriction (= *the students*) have to participate in the reciprocal relation (= *knowing one another*). By contrast, (1-b) states that the students entered the room one after the other, i.e. they can be ordered on a path. Quantified reciprocals like (1) may also exhibit a third reading, i.e. for any member of the restriction there is some other member and these two participate in the reciprocal relation. We dub the first interpretation a complete graph reading, the second a path reading and the third a pair reading. Note the logical dependencies. In fact, the complete graph reading is the logically strongest interpretation entailing the others.

To account for interpretation preferences of sentences like (1) Dalrymple et al. (1998) have put forward the *Strongest Meaning Hypothesis* which proposes that sentences like (1) receive the strongest interpretation consistent with the reciprocal relation. This hypothesis is not undisputed, however, and Kerem, Friedmann & Winter (2010) have presented empirical evidence that it does not hold in general, but that interpreters will choose the most typical interpretation, instead. They refined the original hypothesis accordingly and formulated their *Maximal Typicality Hypothesis*.

The strongest meaning of (1-a), which is presumably also the most typical one, is PTIME verifiable. Interestingly, once we replace the aristotelian quantifier *all* with a proportional quantifier like *most* or a cardinal quantifier like *exactly k*, verification of the strong meaning becomes NP-hard (Szymanik 2010) since it involves solving the CLIQUE problem. In contrast to the *Strongest Meaning Hypothesis*, the PCH therefore predicts that complete graph readings should not be possible for reciprocals with proportional or counting quantifiers as their

[1] A clarification may be in order: When we speak of intractable problems we always refer to NP-hard problems and assume silently that $\mathsf{P} \neq \mathsf{NP}$.

antecedents. As illustrated in this example, combining the *Strongest Meaning Hypothesis* and the PCH yields specific predictions (and similar predictions can be derived by combining the *Maximal Typicality Hypothesis* with the PCH). We can think of computational complexity as a filter acting on the possible meanings of reciprocal sentences. The effect of this filter should be that the logically strongest meanings is preferred, as long as it is computationally tractable.

The structure of the paper is as follows. In the next section we present a picture completion experiment which was conducted to elicit the preferred interpretation of reciprocals with three different quantificational antecedents. The results show that, in line with our predictions, the interpretation preferences clearly differed between aristotelian and proportional or cardinal quantifier antecedents: we observed only a small proportion of complete graph interpretations in the latter two quantifier types, whereas reciprocals with an aristotelian antecedent were ambiguous. We then present two picture verification experiments that tested whether an intractable complete graph reading was a viable interpretation for proportional and cardinal quantifiers, at all. The results show that potentially intractable complete graph readings are possible as long as they are tested in graphs of limited size. We conclude with a discussion of whether the obtained results require a parameterized version of the PCH (cf. van Rooij & Wareham (2008)) or whether the findings could also be explained if we assume that intractable meanings were approximated using simple heuristics that fail under certain conditions.

2 Picture Completion Experiment

We measured interpretation preferences in a picture completion experiment in which participants had to complete dot pictures corresponding to their preferred interpretation of sentences like (2). As outlined above, the *Strongest Meaning Hypothesis* predicts sentences like (2) to be preferably interpreted with their complete graph meaning (3).

(2) All/Most/Four of the dots are connected to each other.

(3) $\exists \mathrm{X} \subseteq \mathrm{DOTS}\,[\mathrm{Q}\,(\mathrm{DOTS}, \mathrm{X}) \wedge \forall x, y \in \mathrm{X}\,(x \neq y \leftarrow \mathrm{CONNECT}\,(x, y))]$, where Q is ALL, MOST or FOUR.

In combination, the PCH and the SMH predict interpretation differences between the three quantifiers. While the complete graph meaning of *reciprocal all* can be verified in polynomial time, verifying the complete graph interpretation of *reciprocal most* and *reciprocal k* (here: $k = 4$) is NP-hard. By contrast, the weaker readings are computable in polynomial time for all three types of quantifiers. It is thus expected that the choice of the quantifier should affect the preference for complete graph vs. path/pair interpretations: *reciprocal all* should preferably receive a complete graph reading, but *reciprocal most/k* should receive a path or pair reading. The *Maximal Typicality Hypothesis* is hard to apply to these cases because it is unclear what the most typical scenario for *to be connected* should look like. Most probably, the different readings shouldn't differ in typicality and should hence show a balanced distribution.

2.1 Method

23 German native speakers (mean age 24.3 years; 10 female) received a series of sentences, each paired with a picture of (yet) unconnected dots. Their task was to connect the dots in a way that the resulting picture matched their interpretation of the sentence. We tested German sentences in the following three conditions (*all* vs. *most* vs. *four*).

(4) Alle (die meisten/vier) Punkte sind miteinander verbunden.
All (most/four) dots are with-one-other connected.
All (most/four) dots are connected with each other.

All-sentences were always paired with a picture consisting of four dots, whereas *most* and *four* had pictures with seven dots. In addition to the fifteen experimental trials (five per condition), we included 48 filler sentences. These were of two types. The first half clearly required a complete graph. The second type was only consistent with a path. We constructed four pseudorandomized lists, making sure that two adjacent items were separated by at least two fillers and that each condition was as often preceded by a complete graph filler as it was by a path filler. This was done to prevent participants from getting biased towards either complete graph or path interpretations in any of the conditions. The completed pictures were labeled with respect to the chosen interpretation. We considered a picture to show a complete graph meaning if it satisfied the truth conditions in (3). A picture was taken to be a path reading if a sufficiently large subgraph was connected by a continuous path, but there was no complete graph connecting the nodes. A picture was taken to be a pair reading if the required number of nodes were interconnected, but there was no path connecting them all. Since we didn't find any pair readings, we will just consider the complete graph and path readings in the analysis. Cases that fulfilled neither interpretation were counted as mistakes.

2.2 Results and Discussion

The proportions of complete graph meanings were analyzed using a logit mixed effects model (cf. Jäger (2008)) with *quantifier* as fixed effect and random intercepts of participants and items. Furthermore, we computed three pairwise comparisons: one between *all* and *most*, one between *all* and *four* and one between *most* and *four*.

In the *all*-condition, participants chose complete graph meanings 47.0% of the time. By contrast, in the *most*-condition there were only 22.9% complete graph interpretations among the correct pictures. The number of complete graphs was even lower in the *four*-condition with only 17.4%. The statistical analysis revealed a significant difference between *all* and the other two quantifiers (*estimate* $= -1.87$; $z = 4.14$; $p < .01$). The pairwise comparisons revealed a significant difference between *all* and *most* (*estimate* $= -1.82$; $z = -3.99$; $p < .01$), a significant difference between *all* and *four* (*estimate* $= -3.16$; $z = -5.51$; $p < .01$), but only a marginally significant difference between *four* and *most* (*estimate* $= .80$; $z = 1.65$; $p = .10$).

The error rates differed between conditions. Participants were 100% correct in the *all*-condition. They made slightly more errors in the *four*-condition which had a mean of 94.8% correct drawings. In the *most*-condition the proportion of correct pictures dropped down to 83.5%. To statistically analyze error rates we computed two pairwise comparisons using Fisher's exact test. The analysis revealed a significant difference between *all* and *four* ($p < .05$) and a significant difference between *four* and *most* ($p < .01$).

The observed preference for path interpretations, and in particular the very low proportions of complete graph readings for *most* and *four*, matched the predictions of the PCH . Both, *most* and *four* reciprocals, constitute intractable problems and their strong interpretation shouldn't hence be possible. The error rates provide further support for the PCH. *Most* and *four* led to more errors than *all.* This can be accounted for if we assume that participants were sometimes trying to compute a complete graph interpretation, but due to the complexity of the task did not succeed.

An open question is whether the strong readings of *reciprocal most* and *reciprocal four* are just dispreferred or completely unavailable. This was addressed in a picture verification experiment.

3 Picture Verification Experiment 1

The second experiment tested reciprocals which were presented together with pictures that disambiguated graph from path readings. To achieve clear disambiguation, we had to use different quantifiers than in the previous experiment. This is because the quantifiers were all upward entailing and therefore complete graphs are also consistent with a path reading. In the present experiment, we used reciprocals with *all but one*, *most* and *exactly k*, as in (5). *All but one* and *exactly k* are clearly non-monotone and hence none of the readings entails the others. For *most* it is possible to construct complete graph diagrams in a way that the other readings are ruled out pragmatically if we take its scalar implicature (= *most, but not all*) into account. Crucially, although intuitively more complex than simple *all*, the complete graph reading of *all but one* is PTIME computable. A brute force algorithm to verify reciprocals with *all but one* requires approximately n-times as many steps as an algorithm to verify *all* reciprocals. In order to verify a model of size n, at most the n subsets of size $n-1$ have to be considered. By contrast, verifying the strong meaning of (5-b,c) is intractable. In particular, *exactly k* is intractable because k is not a constant, but a variable. In the experiment we kept *all but one* constant and presented *exactly k* with the values *exactly three* and *exactly five.*

(5) (a)Alle bis auf einen / (b)Die meisten / (c)Genau drei (/fünf)
(a)All but one / (b)The most / (c)Exactly three (/five)
Punkte sind miteinander verbunden.
dots are with-one-other connected.
(a)All but one/ (b)Most / (c)Exactly three (/five) dots are connected with each other.

We paired these sentences with diagrams disambiguating towards the complete graph or the path reading. Sample diagrams are depicted in the appendix A in Figure 4(a)/(e) and 4(b)/(f), respectively. As for complete graph pictures, the PCH lets us expect lower acceptance rates for (5-b,c) than for (5-a). In order to be able to find out whether the complete graph readings of (5-b/c) are possible at all we paired them with false diagrams which served as baseline controls (see Figure 4(c)/(g)). The controls differed from the complete graph pictures in that a single line was removed from the completely connected subset. If the complete graph reading is possible for (5-b/c), we should observe more "yes, true" judgments in the complete graph condition than in the false controls. As an upper bound we included ambiguous conditions compatible both with complete graph and path conditions (cf. Figure 4 (d)/(i)).

It is possible that people can verify the strong meaning of (5-b,c) given small graphs, but fail to do so for larger ones. Therefore, besides having the three types of quantifiers, another crucial manipulation consisted in the size of the graphs, i.e. the number of vertices. Small graphs always contained four dots (see Fig. 4(a)-(d)) and large graphs consisted of six dots (see Fig. 4(e)-(i)). This way, we also were able to keep the quantifier *all but one* constant and compare it to *exactly k* with variable *k*. *Exactly k* was instantiated as *exactly three* and *exactly five*, respectively. In total, this yielded 24 conditions according to a 3 (*quantifier*) × 4 (*picture type*) × 2 (*graph size*) factorial design.

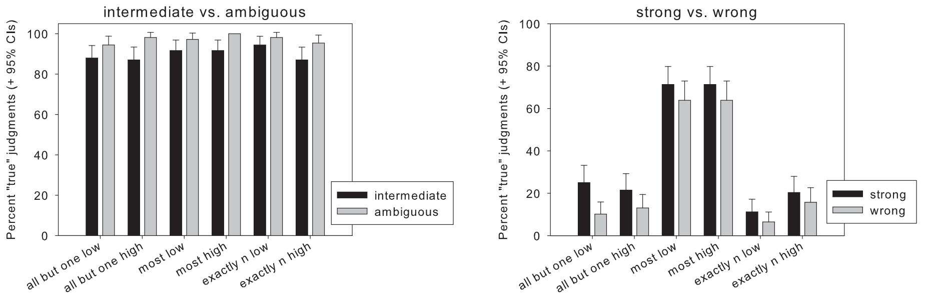
3.1 Method

We constructed nine items in the 24 conditions and used a latin square design with three lists to make sure that each picture only appeared once for each subject, but that the same picture appeared with all three quantifier types. Each participant provided three judgments per condition resulting in a total of 72 experimental trials. We added 66 filler trials.

36 German native speakers (mean age 26.9 years; 23 female) read reciprocal quantified sentences on a computer screen. After reading the sentence, they had to press a button, the sentence disappeared and a dot picture was presented for which they had to provide a truth value judgment.

3.2 Results

The mean judgments are presented in Figure 1. The proportions of "yes, true" judgments were subjected to two analyses. The lower bound analysis compared



(a) upper bound analysis

(b) lower bound analysis

Fig. 1: Mean judgments in Picture Verification Experiment 1 (*low*: pictures with 4 dots; *high*: pictures with 6 dots)

the complete graph conditions with the false baseline controls in order to determine whether complete graph pictures were more often accepted than the latter. The fixed effects of *quantifier* (three levels), *graph size* (two levels) and *picture type* (two levels) were analyzed in a logit mixed effects model with random intercepts of participants and items. Accordingly, upper bound analyses compared the path conditions with the ambiguous conditions.

Upper bound analysis: There was an across the board preference (7.3% on average) of ambiguous pictures over pictures disambiguating towards a path interpretation (*estimate* $= -2.37$; $z = -2.88$; $p < .01$). No other effects were reliable.

Lower bound analyses: Quantifiers differed with respect to how many positive judgments they received. *Most* received reliably more positive judgments than *exactly k* and *all but one* which led to a significant effect of *quantifier* (*estimate* $= 3.31$; $z = 8.10$; $p < .01$). There was also a marginally significant effect of *truth* (*estimate* $= 0.72$; $z = 1.77$; $p = .07$) which was due to slightly higher (7.9%) acceptance of the complete graph pictures as compared to the false baseline controls. No other effects were reliable.

3.3 Discussion

Most behaved rather unexpectedly. Surprisingly, it was rather often accepted in the false baseline controls. The high acceptance rates in these two conditions indicate that participants were canceling its scalar implicature (= *most, but not all*) and interpreted it equivalently to the upward monotone *more than half*. This also explains the high acceptance rates of *most* in the complete graph conditions which were, without the implicature, compatible with a path interpretation. Not being able to tell path interpretations apart from complete graph interpretations, we exclude *most* from the following discussion.

Overall, the path reading was strongly preferred and the complete graph reading was hardly available for any quantifier type. However, both the upper bound and the lower bound analyses provide evidence that the complete graph reading, even though strongly dispreferred, is available to some degree. Both analyses revealed an effect of picture type. Path diagrams were accepted somewhat less often than the ambiguous diagrams and complete graph diagrams were somewhat more often accepted than false diagrams. To our surprise, we didn't find any differences between quantifiers and graph sizes. However, we have to be careful in interpreting these effects because judgments were very close to the floor in the complete graph conditions.

Why did we observe only so few complete graph interpretations even for tractable *all but one* reciprocals? One possible explanation is that readers shifted to path interpretations in all three quantifier conditions because the complete graph readings may have been too complex. In the following we will show that this is not the case, but that the low acceptability of complete graph interpretations was for another reason. In particular, it is due to the reciprocal relation *be connected to each other*. When reconsidering this relation we were not sure whether it had the desired logical properties. It is possible that *to be connected* is preferably interpreted transitively. This could have led to the low number of positive judgments for the complete graph pictures, because, strictly speaking, the complete graph reading would then be false in the complete graph pictures. Note that there was a path connecting all the dots. Thus, assuming transitivity, any dot was pairwise connected to all the others, violating the complete graph reading of *all but one* and *exactly k*.

4 The Reciprocal Relation

To find out whether the reciprocal relation *be connected with each other* allows for a transitive interpretation we conducted a picture verification experiment with non-quantificational reciprocal sentences. We presented 80 participants with the picture in Figure 2 and presented one of the following sentences to each of four subgroups of 20 participants. Each participant provided only one true/false-judgment. This was done to make sure that the data are unbiased.

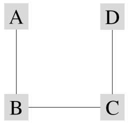


Fig. 2: Diagram used to test for (in)transitivity of the reciprocal relations.

(6) a. A und D sind miteinander verbunden.
A and D are connected to each other.

b. A und D sind nicht miteinander verbunden.
   A and D are not connected to each other.
c. A und D sind direkt miteinander verbunden.
   A and D are directly connected with each other.
d. A und D sind nicht direkt miteinander verbunden.
   A and D are not directly connected with each other.

The proportion of *yes, true* judgments of sentence (6-a) provides us with an estimate of how easily *be connected with each other* can be interpreted transitively. By contrast, judgments of the negated sentence in (6-b) inform us about how easily the relation *be connected with each other* can be understood intransitively. As a pretest for the next picture verification experiment, we included another relation, *be **directly** connected with each other*, which, according to our intuitions, should only be interpretable intransitively.

4.1 Results and Discussion

The distribution of judgments was as follows. (6-a) was accepted in 80% of the cases, (6-b) was accepted in 42%, (6-c) in 10% and (6-d) in 80% of the cases. Fisher's exact test revealed that the proportions of "yes, true" judgments in the conditions (6-a), (6-b) and (6-d) were significantly different from 0% ($p <$ .01), whereas condition (6-c) didn't differ significantly from 0% ($p = .49$). With respect to *to be connected* the results show that the relation is in fact ambiguous although there was a preference for the transitive reading as indicated by the higher proportion of "yes, true" judgments of sentence (6-a) than (6-b). The transitive preference might have confined the acceptability of complete graphs in the previous picture verification experiment. Regarding *to be directly connected* there is a strong preference to interpret the reciprocal relation intransitively as revealed by almost no acceptance of (6-c).

5 Picture Verification Experiment 2

In the first picture verification experiment the complete graph interpretation seemed to be possible across quantifiers. If this was really the case it would provide evidence against the PCH. In the present experiment we thus investigated whether this tentative finding can be replicated with a reciprocal relation like *be **directly** connected* which is fully consistent with complete graphs. If the PCH, in its original form, were correct, it would predict complete graphs to be acceptable in tractable *all but one* reciprocals. By contrast, *exactly k* should lead to an interpretive shift and reveal path interpretations, only. If, on the other hand, the findings of the first picture verification experiment could in fact be generalized to intransitive relations, we would expect to find complete graph interpretations of *exactly k*, contra the PCH. Extending this line of reasoning, we might expect the availability of a complete graph reading of *exactly k* reciprocals to be influenced by graph size. In small sized graphs this reading may be fully acceptable, whereas with increasing size it may well exceed processing capacity. As for *all but one*, both theoretical alternatives predict complete graphs to constitute possible interpretations irrespective of the size of the model.

(7) (a) Alle bis auf einen / (b) Genau drei (/fünf) Punkte sind
(a) All but one / (b) Exactly three (/five) dots are
miteinander verbunden.
with-one-other connected.
(a) All but one/ (b) Exactly three (/five) dots are connected with each other.

To test these predictions we combined the sentences in (7) with the pictures in Figure 5. The picture conditions were mostly identical to those of the first picture verification experiment. The only difference was that the ambiguous pictures were replaced by two other conditions (see Figure 5(d)/(h)). These were false baseline controls included to find out whether path readings are available with a clearly intransitive relation. The false controls depicted paths that were one edge too short. Altogether this yielded 16 conditions in a 2(*quantifier*) × 2(*reading*) × 2(*truth*) × 2(*graph size*) within subjects design.

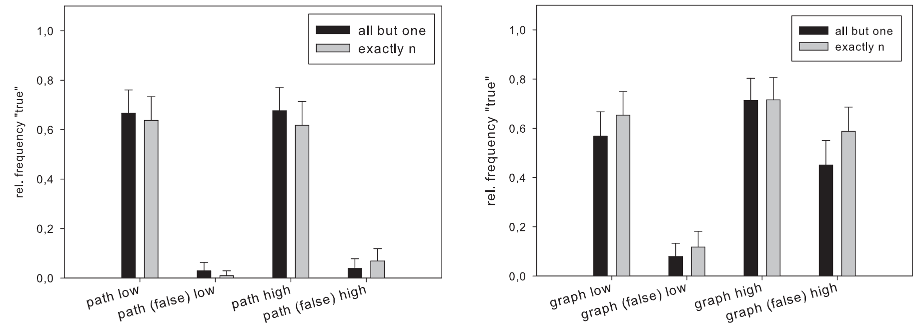
Thirty-four native German speakers (mean age: 27.5y, 20 female) took part in the study. Except for the mentioned changes everything was kept identical to the first picture verification experiment.

5.1 Results and Discussion

Mean acceptance rates are depicted in Figure 3. The path and complete graph conditions were analyzed separately using logit mixed effect model analyses. The path reading was generally accepted (true conditions: mean acceptance of 67.7%) and led to almost no errors (false conditions: mean acceptance 4.2%) with both quantifiers and graph sizes. The statistical analysis revealed that only the main effect of *truth* was significant ($estimate = 5.70$; $z = 8.70$; $p < .01$).

The true complete graph diagrams were also accepted across the board (66.3%). Further, true complete graph diagrams were accepted significantly more often than false ones (31%: $estimate = 1.94$; $z = 4.08$; $p < .01$). In the false complete graph conditions there were, however, clear differences between diagrams with four and six dots. In the former conditions participants made relatively few errors (9.9%), whereas error rates increased drastically in the latter conditions (52%). This led to a reliable effect of *graph size* ($estimate = -6.46$; $z = -3.99$; $p < .01$) and a significant interaction of *truth* and *graph size* ($estimate = 5.23$; $z = 3.18$; $p < .01$). Furthermore, the increase in error rates was greater for *exactly k* than for *all but one*. For *exactly k* we observed an increase of 47.0%, whereas there was only a 37.2% increase for *all but one*. Analyzing the false conditions separately we found a significant interaction of *graph size* and *quantifier* ($estimate = 18.6$; $z = 2.04$; $p < .05$).

The relation *to be directly connected* clearly allowed for complete graph readings. As opposed to the predictions of the PCH, this was the case for both *all*



(a) continuous paths

(b) complete graphs

Fig. 3: Mean judgments in Picture Verification Experiment 2 (*low*: pictures with 4 dots; *high*: pictures with 6 dots)

but one and *exactly k*. In the false conditions we did, however, find clear effects of semantic complexity. In these conditions errors drastically increased with the number of vertices. Especially, for the intractable *exactly k* reciprocals error rates increased with the number of dots.

Why did semantic complexity only affect the false conditions? We think that the reason for this may lie in the specific verification strategies participants were able to use in the relevant conditions. In the true complete graph conditions the complete subgraph was visually salient and could be immediately identified. In combination with the fact that the CLIQUE problem is in NP, the true complete graph conditions may, thus, in fact have posed a tractable problem to the participants. As a consequence, only in the false conditions with one edge removed from the graph participants really had to solve an intractable problem. Apparently, this was still possible when the relevant subgraph consisted of only three dots but already started to exceed cognitive capacities when it consisted of five dots.

5.2 Conclusions

We started off with hypotheses that made too strong predictions. Firstly, the linguistic work on reciprocal sentences by Dalrymple et al. (1998) let us expect that reciprocals should be interpreted in their logically strongest meaning. When the antecedent is upward entailing a complete graph interpretation should thus be chosen. Secondly, Szymanik (2010), building upon the PCH, predicted interpretation shifts if reciprocals are intractable due to their quantificational antecedents. Surprisingly, neither of these predictions was fully confirmed. In contrast to the first prediction, complete graph readings were not the default even for tractable reciprocals with an upward entailing antecedent. In the picture completion study path readings were equally often chosen for reciprocals with *all* as the stronger complete graph reading. At first sight the predictions of the PCH were borne out by the picture completion study. In this experiment the quantificational antecedent affected interpretation preferences according to the predictions of the PCH. When the complete graph reading was intractable it was only rarely chosen. However, in picture verification intractable readings were clearly available to our participants. This provides evidence against the PCH in its most general form.

Still, we did find effects of semantic complexity. Participants had problems to correctly reject pictures not satisfying the complete graph reading with one missing connection. This difficulty increased with the number of dots, especially for intractable *exactly k*. How can these effects be explained? Firstly, it is possible that participants approximated the intractable meaning of *exactly k*, thereby effectively computing tractable functions. These tractable approximations may have worked well in the true complete graph conditions but were inappropriate for the false complete graph controls. We think of specific graphical properties present in the true conditions, e.g. salience of the relevant subgraph which may have simplified the task. Secondly, it is possible that participants were able to compute intractable functions, but only within certain limits, e.g. up to a certain number of elements or in pictures where the relevant subgraph was graphically salient. No matter which of these explanations is correct our data provide an interesting challenge for the PCH. A promising perspective in this respect may be a parameterized version of the PTIME Cognition Hypothesis (cf., for instance, van Rooij & Wareham 2008) which allows us to take into consideration the exact instantiation of the problem. The presence or absence of a perceptually salient group of objects may be a crucial factor for identifying a clique and should, therefore, influence error rates. We plan to explore this in future research.

References

Dalrymple, Mary, Makoto Kanazawa, Yookyung Kim, Sam McHombo, & Stanley Peters (1998), Reciprocal expressions and the concept of reciprocity, *Linguistics and Philosophy*, 21(2):159–210.

Frixione, Marcello (2001), Tractable competence, *Minds and Machines*, 11:379–397.

Jäger, Florian T. (2008), Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models, *Journal of Memory and Language.*, 59:434–446.

Kerem, Nir, Naama Friedmann, & Yoad Winter (2010), Typicality effects and the logic of reciprocity, in *Proceedings of SALT XIX.*

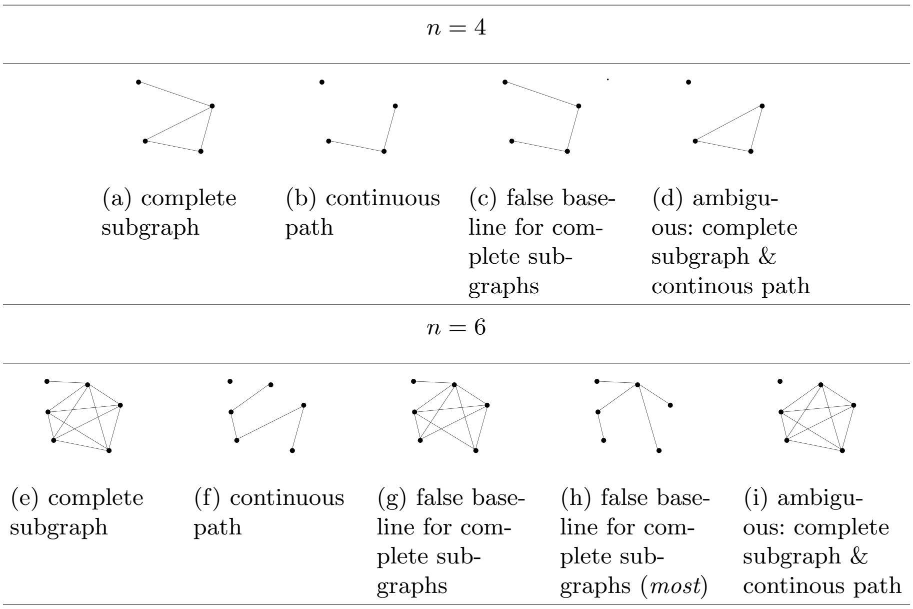
Pietroski, Paul, Jeffrey Lidz, Tim Hunter, & Justin Halberda (2009), The meaning of 'most': semantics, numerosity, and psychology, *Mind & Language*, 24(5):554–585.

Szymanik, J. (2010), Computational complexity of polyadic lifts of generalized quantifiers in natural language, *Linguistics and Philosophy*, forthcoming.

van Rooij, Iris (2008), The tractable cognition hypothesis, *Cognitive Science*, 32:939–984.

van Rooij, Iris & Todd Wareham (2008), Parametized complexity in cognitive modeling: foundations, applications and opportunities, *The Computer Journal*, 51(3):385–404.

A Sample diagrams



(a) complete subgraph

(b) continuous path

(c) false baseline for complete subgraphs

(d) ambiguous: complete subgraph & continous path

(e) complete subgraph

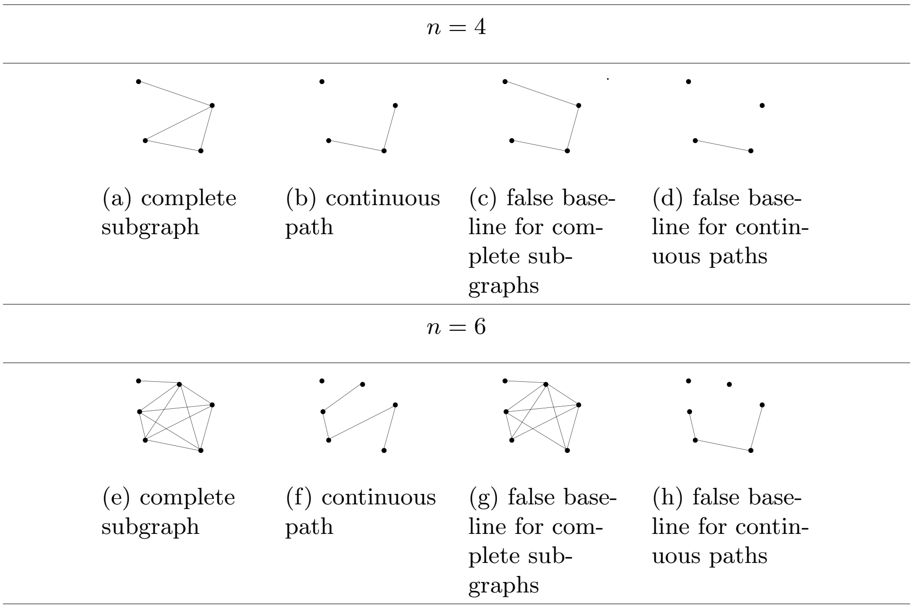
(f) continuous path

(g) false baseline for complete subgraphs

(h) false baseline for complete subgraphs (*most*)

(i) ambiguous: complete subgraph & continous path

Fig. 4: Sample diagrams presented in Picture Verification Experiment 1. The upper row represents graphs with four dots. Graphs with six dots are represented in the bottom row. The false baseline controls for complete graphs with six dots were slightly different for *all but one* and *exactly five* (g) than for *most* (h). In diagrams like (h), all dots were connected by a path, but in contrast to diagrams like (g) there was no subset containing four or more elements forming a complete graph. This way, for all three quantifiers falsity was due to a small number of missing connections.



(a) complete subgraph

(b) continuous path

(c) false baseline for complete subgraphs

(d) false baseline for continuous paths

(e) complete subgraph

(f) continuous path

(g) false baseline for complete subgraphs

(h) false baseline for continuous paths

Fig. 5: Sample diagrams presented in the Picture Verification Experiment 2. The upper row represents graphs with four dots. Graphs with six dots are represented in the bottom row.

Superlative quantifiers and epistemic interpretation of disjunction

Maria Spychalska

Ruhr-University Bochum, Department of Philosophy

Abstract. We discuss semantics of superlative quantifiers *at most n* and *at least n*. We argue that the meaning of a quantifier is a pair specifying a verification and a falsification condition for sentences with this quantifier. We further propose that the verification condition of superlative quantifiers should be interpreted in an epistemic way, that is as a conjunctive list of possibilities. We also present results of a reasoning experiment in which we analyze the acceptance rate of inferences with superlative and comparative quantifiers in German. We discuss the results in the light of our proposal.

1 Introduction

There is an ongoing debate (Look inter alia: (Geurts & Nouwen, 2007), (Koster-Moeller et al, 2008), (Geurts et al., 2010), (Cummins & Katsos, 2010), (Nouwen, 2010), (Cohen & Krifka, 2011)) concerning the right semantical interpretation of so-called superlative quantifiers, such as *at most n* and *at least n*, where *n* represents a bare numeral, e.g *two*. Generalized Quantifier Theory (referred here as a "standard account") defines superlative quantifiers as equivalent to respective comparative quantifiers: *fewer than n* and *more than n*, that is:

$$at\,most\,n\,(A,B) \iff fewer\,than\,n+1(A,B)^1 \tag{1}$$

$$at\,least\,n\,(A,B) \iff more\,than\,n-1(A,B) \tag{2}$$

It has been observed that in natural languages those equivalences (1) and (2) might not hold, or at least they might not be accepted by language users based on pragmatical grounds. There are numerous differences between comparative and superlative quantifiers involving their linguistic use, acquisition, processing and the inference patterns in which they occur. First of all, it seems that superlative and comparative quantifiers are not freely exchangeable in same linguistic contexts. Geurts & Nouwen (2007) provide, among many others, such examples:

(a) I will invite at most two people, namely Jack and Jill.
(b) I will invite fewer than three people, namely Jack and Jill.

[1] Q(A,B) means Q A's are B, where Q is a $\langle 1,1 \rangle$ generalized quantifier

where (a) is considered a good sentence, while (b) is less felicitous. The contrast between (a) and (b) suggest that while embedding an indefinite expression (*two*) in a superlative quantifier licenses a specific construal (*namely Jack and Jill*), the same is not licensed in the case of a comparative modifier.

Secondly, it has been demonstrated that superlative quantifiers are mastered later than the comparative ones during language development (Musolino, 2004), (Geurts et al., 2010). Furthermore, there is ample data concerning processing of those quantifiers. It has been for instance shown that verification of sentences with superlative quantifiers requires a longer time than verification of sentences with respective comparative quantifiers (Koster-Moeller et al, 2008),(Geurts et al., 2010). Moreover, the processing of quantifiers is influenced by their monotonicity. A quantifier $Q(A, B)$ is upward monotone in its first argument A if it licences inferences from subsets to supersets, that is if $Q(A, B)$ and $A \subseteq A'$, then $Q(A', B)$. A quantifier $Q(A, B)$ is downward monotone in its first argument A if it licences inferences from supersets to subsets, that is if $Q(A, B)$ and $A' \subseteq A$, then $Q(A', B)$. Understood as $\langle 1, 1 \rangle$ generalized quantifiers, *at least n* and *more than n* are upward monotone in both their arguments, while *at most n* and *fewer than n* are downward monotone. It has been shown that although the downward monotone *at most n* and *fewer than n* take a longer time to be verified than the upward monotone *at least n* and *more than n*, they are actually falsified faster (Koster-Moeller et al, 2008).

Finally, important arguments against the semantical equivalence between the comparative and superlative quantifiers come from the analysis of people's acceptance of inferences with those quantifiers. Empirical data show that a majority of responders usually reject inferences from *at most n* to *at most n+*, (where $n+$ denotes any natural number greater than n), although they accept presumably equivalent inferences with comparative quantifiers (Geurts et al., 2010), (Cummins & Katsos, 2010). To illustrate it with an example: while people are unlikely to accept that *if at most 5 kids are playing in this room, then at most 6 kids are playing in this room* (2-14% in Cummins' and Geurts' experiments for this inference scheme) they are more likely to accept that *if fewer than 5 kids are playing in this room, then fewer than 6 kids are playing in this room* (between 60-70% in Cummins'and Geurts'). Such data seem to directly contradict the standard account.

2 Modal semantics or clausal implicature?

Several theories have been developed to explain why seemingly logically equivalent quantifiers show such big differences in the way people use them in the language. Geurts (2007), (2010) proposes modal semantics for superlative quantifiers and rejects the assumption that equivalences (1) and (2) hold in natural languages. According to this proposal (referred here as a "modal account"), while *more than n* and *less than n* have a conventional meaning defined in terms of inequality relation, *at least n* and *at most n* have a modal component, namely:

> *at least n A's are B* means that: a speaker is certain that there are *n* elements which are both A and B, and considers it possible that there are more than *n*.
>
> *at most n A's are B* means that: a speaker is certain that there is no more than *n* elements that are both A and B, and considers it possible that there are *n* elements.

According to this proposal, as semantically richer, superlative quantifiers are expected to be harder to process than the respective comparative quantifiers. Finally, defined as above, *at most n A are B* does not imply *at most n+ A are B*: The latter implies that it is possible that there are *(exactly) n+ A that are B*, which is contradicted by the semantics of *at most n* in the premise.

There are strong arguments against the modal account. For instance this account seems unsatisfactory with regard to superlative quantifiers embedded in conditional and various other contexts. Authors (Geurts & Nouwen, 2007),(Geurts et al., 2010) realize themselves this problem and illustrate it with the following example:

> *If Berta had at most three drinks, she is fit to drive. Berta had at most two drinks. Conclusion: Berta is fit to drive.*

Such inferences, which are indeed licensed by the inference from *at most n (2)* to *at most n+ (3)*, are commonly accepted by speakers (over 96% in Geurts's experiment).

Furthermore, while inferences from *at most n* to *at most n+(1)* are rejected by majority of people (ca 84% in Geurts' experiment) there are subjects who do accept them (14% in Geurts' and even more in our experiment — ca. 23%). If *at most n* logically implies *possible that n and not possible that more than n*, then the inference from *at most n* to *at most n+* should be inaccessible (except for cases of random mistakes) for any language users, due to the apparent contradiction between the premise and the conclusion. Last but not least, to say that *possible that n* is a part of the semantics of *at most n*, implies that *at most n* cannot be paraphrased by *not more than n*. However such a paraphrase seems totally eligible.

A slightly different account was proposed by Cummins & Katsos (2010), who observe that the considered linguistical phenomena could be better explained on pragmatical grounds. The authors show that people do not evaluate *at most n* and *exactly n-1* as equally semantically incoherent as cases of obvious logical incoherence, e.g. *at most n* and *exactly n+1* or *more than n* and *exactly n-1*. While sentence pairs, such as:

> *Jean has at most n houses. Specifically she has exactly n+1 houses.*

get average coherence judgments very low, i.e. -4, in the scale from -5 (incoherent) to $+5$ (coherent), sentence pairs:

> *Jean has at most n houses. Specifically she has n-1 houses*

get already +1.9. This result speaks against the "modal account", whose direct consequence is semantical incompatibility of *at most n* and *exactly n-*.

Consequently, Cummins et al. agree with Geurts that *at most n* and *at least n* both imply *possible that n*, but they claim that this is a pragmatical rather than a logical inference, namely a co-called *clausal implicature* (Levinson, 1983). Clausal implicature is a quantity implicature inferred due to use of epistemically weak statement. Since the expressed statement with a superlative quantifier e.g. *at most n A are B*, as equivalent to a disjunctive statement *there are exactly n or fewer than n elements that are both A and B*, does not imply the truth of its subordinate proposition $p =$ *there are exactly n elements that are A and B*, the possibility that p might or might not be true is inferred.

Although we agree with the intuitions concerning a modal component in the reasoning with superlative quantifiers, we reject the assumption by Geurts et al. that this component is a part of their meaning. Furthermore, although we agree with Cummins et al. that the mechanism that results with the observed inference patterns is more of a pragmatical nature, we are not satisfied with the "causal implicature" account. What we lack is a deeper insight into the source of this kind of a pragmatical inference and how it interacts with the logical meaning of those quantifiers in different reasoning contexts.

Our motivation to further experimentally investigate reasoning with superlative quantifiers is based inter alia on: (i) the lack of data concerning whether people accept inferences: *at least n* → *at least n* -, (ii) the lack of satisfactory data about how people accept inference with logically equivalent forms of superlative quantifiers: such as *not more than n/not fewer than n* or *n or fewer than n/n or more than n*, as well as how they accept mutual equivalences between these forms, (iii) finally, the lack of data concerning people's acceptance of logically incorrect inferences with the quantifiers considered here.

3 Two semantic conditions for *at most n*

Krifka (1999) points out that semantic interpretation of a sentence is usually a pair that specifies when the sentence is true and when it is false. Following Krifka, we propose to define meaning of a quantifier as a pair $\langle C_F, C_V \rangle$, where C_V is a verification condition (specifies how to verify sentences with this quantifier) and C_F is a falsification condition (specifies how to falsify sentences with this quantifier). Furthermore, we propose that the interpretation of a quantifier depends on a semantic context in which this quantifier is used, namely whether the context requires the use of the verification condition or the falsification condition. Verification and falsification conditions are to be understood algorithmically, with the "else" part of the conditional instruction being empty - thus, they verify (or falsify) the formulas only if their conditional test is satisfied. From a perspective of classical logic, these conditions should be dual, namely if C is a C_V condition for sentence ϕ, then C is a C_F condition for sentence $\neg\phi$, and vice versa. We further, however, observe that in the case of superlative quantifiers, there is a split between these two conditions. We suggest, that this split is a result of a pragmatic focus on the expressed borderline n.

One can think of the meaning of logical operators, thus also quantifiers, in terms of algorithms, that have to be performed in order to verify (or falsify) sentences with those operators. (See also (Szymanik, 2009), (Szymanik & Zajenowski, 2010), (Szymanik & Zajenowski, 2009)) Krifka (1999) observes, that a sentence *at most n x:* $\phi(x)$[2] says only that *more than n x:* $\phi(x)$ is false, and leaves a truth condition underspecified. In other words, the meaning of *at most n* provides an algorithm for falsifying sentences with this quantifier, but not (immediately) for verifying them. This corresponds with the experimental data showing that it is easier to falsify sentences with *at most* than to verify them (Koster-Moeller et al, 2008). Consequently, the primal semantical condition of *at most n x:* $\phi(x)$ could be understood as an algorithm: "falsify when the number of x that are ϕ exceeds n", and would constitute what we understand by the falsification condition.

Definition 1 *(falsification condition for at most)*

$$C_F(at\,most\,x : \phi(x)) := If\ \exists^{>n}x(\phi(x)),\ then\ falsify$$

But how can we know when a sentence with *at most n* is true? From the point of view of an algorithm it is a so-called "otherwise" condition that defines in this case the truth-condition. However a negation of a falsification condition is in sense *informationally empty*: it does not describe any concrete situation in which the given sentence can be verified. As a result, in those contexts that require to directly verify a sentence, we refer to a verification condition, which is specified independently. As expressing a positive condition, *at most n* may be understood as a disjunction *n or fewer than n* ("disjunctive *at most*").

Definition 2 *(verification condition for at most)*

$$C_V(at\,most\,n : x\phi(x)) := If\ (\exists^{=n}x\phi(x) \vee \exists^{<n}x\phi(x)),\ then\ verify$$

The disjunction in 2 could be further broken down to: $\bigvee_{i=1}^{n} \exists^{=i}x\phi(x) \vee \neg\exists x\phi(x)$, in short: $\bigvee_{i=0}^{n} \exists^{=i}x\phi(x)$, where the disjunct $\exists^{=o}x\phi(x)$ means that $\neg\exists x\phi(x)$.[3]

And $\exists^{=n}x\phi(x)$ means *precisely n x are* ϕ, that is:

$$\exists^{=n}x\phi(x) \iff \exists x_1 ... \exists x_n[\bigwedge_{i=1}^{n} \phi(x_i) \wedge \bigwedge_{1 \leq i < j \leq n} (x_i \neq x_j) \wedge \forall y(\bigwedge_{i=1}^{n} y \neq x_i \rightarrow \neg\phi(y))] \quad (3)$$

[2] at most n x are ϕ

[3] Let us observe that $\exists^{<n}x\phi(x)$ is a short notation that can be misleading, since it is not an existential sentence. As existential, *fewer than n* would imply that there has to be at least one (though less than n) such x that is ϕ. However we would like a sentence *less than n x:* $\phi(x)$ to be also true if no x's are ϕ. Therefore, in fact, such a downward entailing sentence is a disguised universal sentence: $\forall x_1..x_n(\bigwedge_{i=1}^{n} \phi(x_i) \rightarrow \bigvee_{1 \leq i < j \leq n}(x_i = x_j))$

3.1 Epistemic interpretation of disjunction

Following Zimmermann (2000) we adopt the view that disjunctive sentences in natural language are likely to get so-called epistemic reading, that is they are interpreted as *conjunctive lists of epistemic possibilities.* According to the proposed solution a disjunction P_1 *or...or* P_n is interpreted as an answer to a question: *Q: What might be the case?* and, thus, is paraphrased as a (closed) list L:

> L: P_1 (might be the case) [and]... P_n (might be the case) [and (*closure*) nothing else might be the case].

This results in the following reading of a disjunctive sentence:

Definition 3 *(Zimmermann, 2000)*

$$P_1 \vee ... \vee P_n \iff \diamond P_1 \wedge ... \wedge \diamond P_n$$

and (closure):

$$\forall P[\diamond P \rightarrow [P \cap P_1 = \emptyset \vee ... \vee P \cap P_n = \emptyset]]$$

The character of the closure requires a bit of our attention. Zimmermann (2000) observes, that disjunctive sentences in natural languages could be understood as *closed* (exhaustive lists of possibilities) or open (when other possibilities are not excluded), which in the spoken language is usually marked by intonation. Closure in Definition 3 indicates that the list is exhaustive. There are good reasons to treat NL disjunctions as generally closed, and it would make sense obviously also in the case of superlative quantifiers. In classical logic the truth of $\exists^{=n}x\phi(x) \vee \exists^{<n}x\phi(x)$ semantically excludes the option that $\exists^{>n}x\phi(x)$, as contradictory. In our analysis, however, we want to treat closure as a merely optional condition. This results from regarding the verification and falsification conditions as independent from each other.

3.2 Verification condition for *at most*

If we assume that disjunctions in natural language are likely to be interpreted as conjunctions of epistemic possibilities, then we get the following verification condition for *at most*:

Definition 4 *(epistemic interpretation of the verification condition for at most)*[4]

[4] A detailed description of the modal predicate logic needed for providing semantics of this kind of sentences is beyond the scope of this paper. For our present purposes it is just enough to assume that, for each possible world, we have a different domain of objects over which we quantify. We assume also standard semantics for modal operators, and we restrict to reflexive and transitive Kripke models.

$$C_W(at\,most\,n\,x : \phi(x)) := \ If\ (\diamond\exists^{=n}x\phi(x) \wedge \diamond\exists^{<n}x\phi(x)),\ then\ verify$$

and (closure)

$$If\ \diamond\exists^{>n}x\phi(x),\ then\ falsify$$

Where:

$$(\diamond\exists^{=n}x\phi(x) \wedge \diamond\exists^{<n}x\phi(x)) \iff \bigwedge_{i=o}^{n} \diamond\exists^{=i}x\phi(x) \tag{4}$$

Now we can see the asymmetry between the falsification and verification condition of *at most n*. While the falsification condition specifies precisely when the sentence has to be rejected as false, the verification condition (in the epistemic reading) provides a conjunctive list of epistemic possibilities that should all be the case in order to verify it.

The epistemic reading of the verification condition is also what differentiates superlative quantifiers from the comparative ones. We propose that the disjunctive form of the verification condition in the case of superlative quantifiers is a result of the focus on the borderline n. The n mentioned in the superlative quantifier constitutes the borderline of the truth-conditions, however the numeral n that occurs in a comparative quantifier is not a part of the truth-conditions. The borderline of truth-conditions (that is *n-1* for *fewer than* n, and *n+1* for *more than n*) remains silent. Consequently, the disjunctive form of comparative quantifiers (and hence also the epistemic reading), though logically possible, is not pragmatically justified. In principle, a comparative quantifier can be as well interpreted in the epistemic way (as a conjunction of epistemic possibilities). Such a reading is, however, not equally likely to occur as in the case of a superlative quantifier, since the borderline is not explicitly expressed.

Let us observe that closure of the verification condition is stronger than the falsification condition. Intuitively, if $\neg\diamond\psi$ (equivalently $\Box\neg\psi$), then as well $\neg\psi$ (here: $\psi = \exists^{>n}x\phi(x)$), i.e. $\Box\xi \Vdash \xi$, however the theorem holds only in reflexive Kripke frames. Furthermore, the optional character of the closure bases on our assumption that the falsification and verification conditions are in a sense independent and only as a pair constitute the full semantic interpretation. Since the falsification condition, as defined in 2, is sufficient to account for the right semantical criterion of when the sentence with *at most n* is false, the closure of the verification condition is redundant and might or might not be considered in the reasoning process. The optional character of closure turns out crucial in evaluating validity of inferences with *at most n*.

When *at most n* is interpreted as in Definition 4, then the inference from *at most n* to *at most n+1* is not valid. Namely, from $\diamond\exists^{=n}\phi(x) \wedge \diamond\exists^{<n}x\phi(x)$ one cannot infer $\diamond\exists^{=n+1}x\phi(x) \wedge \diamond\exists^{<n+1}x\phi(x)$. It is easy to observe that the conjunct $\diamond\exists^{=n+1}x\phi(x)$ cannot be proven based on the premise, though it can be excluded only if the closure of the premise is applied.

On the other hand, the inference: *at most n* → *at most n-1*, which is invalid in the standard account, in the epistemic interpretation is blocked only due to closure of the conclusion. That is the $\diamond\exists^{=n}x\phi(x)$ implied by the premise is contradicted by the closure of the conclusion, i.e. $\neg\diamond\exists^{>n-1}x\phi(x)$. It is important to notice that without the closure the implication holds (if the epistemic reading of the verification condition is applied).

In some aspects our proposal might seem similar to Geurts' "modal" approach, namely we define the verification condition of *at most n* and *at least n* (see below) in modal terms. The main difference is that, in our account, it is *only* the verification condition that is defined modally, while the falsification condition remains standard. This results in a specific split or ambiguity in the meaning of superlative quantifiers.

4 *At least* and bare numerals

The quantifier *at least n* might seem perhaps less interesting than *at most n*, as it seems, as is also evident from our results (see Section (5)), less problematic from the point of view of reasoning. As an upward monotone quantifier, *at least n* appears to provide a clear verification algorithm: "verify when *n x* (that are ϕ) are found". Such a semantical interpretation would not, however, account for the linguistical differences between *at least n* and *more than n-1*.

Let us start with defining a falsification condition for *at least n* as follows:

Definition 5 *(*falsification condition for *at least)*

$$C_F(at\ least\ n : x\phi(x)) :=\ If\ \exists^{<n}x\phi(x),\ then\ falsify$$

What we expect from the verification condition is that it express the whole range of epistemic possibilities in which the sentence with *at least n* can be true. Understood as in formula (2), thus as equivalent to *more than n-1*, a sentence with *at least n* can be expressed as an existential sentence that merely says that *there are n* (*x* that are ϕ), but does not exclude the possibility that there are more:

$$\exists^{\geq n}x\phi(x) \iff \exists^{>n-1}x\phi(x) \iff \exists x_1...\exists x_n[\bigwedge_{i=1}^{n}\phi(x_i) \wedge \bigwedge_{1\leq i<j\leq n}(x_i \neq x_j)] \quad (5)$$

This formula however, which would be a right verification condition for the comparative *more than n-1*, does not outline the borderline *n*, which is emphasized in the superlative *at least n*. Therefore we further break down formula (5) into a disjunctive formula: *(exactly) n or more than n x are* ϕ, which constitutes our verification condition for *at least n*.

Definition 6 *(*verification condition for *at least n)*

$$C_V(at\ least\ n : x\phi(x)) := If\ (\exists^{=n}x\phi(x) \vee \exists^{>n}x\phi(x)),\ then\ verify$$

The latter can be handled as a conjunctive list of possibilities.

Definition 7 *(*epistemic interpretation of the verification condition for *at least n)*

$$C_F(at\ least\ n : x\phi(x)) := If\ (\diamond\exists^{=n}x\phi(x) \wedge \diamond\exists^{>n}x\phi(x)),\ then\ verify$$

and (closure)

$$If\ (\bigvee_{i=0}^{n-1} \diamond\exists^{=i}x\phi(x)),\ then\ falsify$$

Having introduced the semantical conditions for *at least n*, we further analyze how they affect inferences with this quantifier, in particular we mean here the inference (*at least n*→ *at least n-1*), as well as its presumably equivalent disjunctive form: (*n or more than n* → *n-1 or more than n-1*). We propose that the way people handle these inferences depends on how they interpret bare numerals, such as n.

From a logical perspective, a bare numeral n (e.g. "two") can be interpreted as denoting any set of *at least n* elements, or a set of *exactly n* elements.[5] Thus $\psi = nx : \phi(x)$ can simply mean that *there are n x that are* ϕ, without any further constraints on whether there are more. Then, ψ gets the reading as in formula (5). On the other hand, ψ could be understood with a kind of closure, that is that there are *exactly n x's are* ϕ, thus as in formula (3).

It has been a matter of a wide debate in formal semantics and pragmatics what is the right approach to interpreting bare numerals in natural languages. It has been proposed that: (i) the literal meaning of n is *at least n*, while the condition *exactly* comes as a scalar implicture (Horn, 1972), (ii) the basic meaning of n is *exactly n*, while both *at least* and *at most* readings would be context-based (Breheny, 2008) (iii) n is ambiguous between *at least n* and *exactly n* (Geurts, 2006), (iv) n is underspecified and can receive *at least*, *at most* or *exactly* readings depending on the context (Carston, 1998).

Let us now show how the interpretation of n interacts with the validity of inferences (*n or more than n*)→ (*n-1 or more than n-1*), given the epistemic interpretation of disjunction. Suppose now that n is interpreted with a closure: *exactly n*. It is easy to observe that, in such a case, *possible that n and possible that more than n* does not imply *possible that n-1 or possible that more than n-1.* The premise which is interpreted as in Definition 7 does not imply $\diamond\exists^{=n-1} \wedge \diamond\exists^{>n-1}$ (with closure $\bigwedge_{i=0}^{n-2} \neg\diamond\exists^{=i}x\phi(x)$) While $\diamond\exists^{>n-1}$ follows from both $\diamond\exists^{>n}$ and $\diamond\exists^{=n}$, the problematic element is $\diamond\exists^{=n-1}$, which is directly contradicted by the closure of the premise. But suppose that n does not get the "exact" reading, but it is interpreted barely as *there are n*, so as (5). Then from *possible that n* we can infer *possible that n-1*, since the latter does not exclude the possibility that there is a bigger set of elements.

[5] or a set of *at most n* elements, however this interpretation seems to be counterintuitive and allowed only in special contexts.

5 Experiment

We conducted a pilot reasoning experiment (in German) to check how people reason with superlative quantifiers: *at least n* (*mindestens n*)[6] and *at most n* (*höchstens n*), as well as with logically equivalent but linguistically different forms of those quantifiers: a comparative negative form and a disjunctive form. These were quantifiers: *not more than n* (*nicht mehr als n*) and *n or fewer than n* (*n oder weniger als n*) (as logically equivalent to *at most n*), and *not fewer than n* (*nicht weniger als n*) and *n or more than n* (*n oder mehr als n*) (as equivalent to *at least n*).

We were particularly interested in comparing subjects' acceptance of inferences from *at most n* to *at most n+1* (type B: see Table 1) with their acceptance of logically equivalent forms of this inference: type D and type F. We were also interested in people's willingness to infer *at most n* from *not more than n*, and vice versa (type G). Furthermore, we checked the respective inferences with *at least n*, that is: *at least n* → *at least n-1* (type A), and its equivalent forms: type C and type E. Finally, we checked the inferences between *not fewer than n* and *at least n* (type H). Last but not least we tested the respective *incorrect* inferences with the considered quantifiers, in all forms (see Table 2).

In the premises of our inferences we used four different quantifiers: *at least three*, *at most three*, *at least four*, *at most four*, and their equivalent forms. All the numbers were throughout spelled out in words according to the requirements of German grammar. There was only one example for each distinct inference relation (i.e. two per inference type). Every sentence content was different. Additionally, to introduce more variation, sentences which had *at most 4/at least 4*(or their equivalent forms) in the premise had a quantifier in the subject position (e.g. *At least four computers are broken in the lab*), sentences which had *at least 3 /at most 3* in the premise had a quantifier in the object position (e.g. *Arthur has at least three cars.*)

As fillers we used inferences with so-called bare numerals (e.g. *four*): those whose correctness depends on the "at least" reading of bare numerals, i.e. $n \rightarrow n-$ (e.g. $4 \rightarrow 3$); and those that are logically incorrect independently of the presumed reading, such as $n \rightarrow n+$ (e.g. $4 \rightarrow 5$).

5.1 Procedure

The experiment was conducted on German native speakers, mainly students of philosophy, psychology, neuroscience and computer sciences. There were 17 subjects (7 male). Subjects were asked to respond "yes" or "no" to the question whether the second sentence (below the line) has to be true, assumed that they know that the first sentence (above the line) is true. For a better understanding of the task two examples were given: one of a valid inference, that should be given a "yes" response:

[6] We give in brackets the German translation used in the experiment

$$\frac{\text{Inga has done three exercises.}}{\text{Inga has done more than two exercises.}}$$

and one of an invalid inference, that requires a “no” response:

$$\frac{\text{Eva has done three exercises.}}{\text{Eva has done fewer than two exercises.}}$$

Note that the examples were selected in such a way that their validity did not depend on the understanding of any of the tested inference relations, that is the examples served as an instruction of what is an inference in general, but not how to evaluate the inferences, that were tested in the experiment.

After reading the instruction subjects saw 40 randomly ordered reasoning tasks: one task at a time, displayed on a computer screen. There was no time limit in the test.

At the end of the experiment two additional control questions were asked, in which the inference from *at most n* to *at most n+1* was embedded in a deontic context. Note that the logically correct response to first question is “yes”, while to the second is “no”.

(1) Erika promised to drink at most six beers. She drank at most four. Did she keep her promise?

(2) Thomas is allowed to eat at most three cookies. He ate at most two. Did he break the rule?

Tables: 1 and 2 summarize the inferences as well as the results.

5.2 Results

Our first observation is that people accepted the logically correct inferences much more frequently than the logically incorrect ones. The incorrect inferences (apart from disjunctive inferences: *E'* and *F'* which turned out specially problematic) were mostly rejected and their acceptance rate was low enough ($1 - 9\%$) to be considered as a result of random mistakes (Table 2). On the other hand all correct inferences were accepted on the level of at least 20%[7], with high variance depending on the form of an inference, in this: inferences of type B and F seemed the most problematic.

The important result is that nearly 100% of responders did accept inference of type G and H, that is (*at most n* → *not more than n*) as well as (*not more than n* → *at most n*), and respective inferences between *at least n* and *not fewer than n*, which suggests that they do see those expressions as equivalent. Furthermore, while inferences from type B, namely the problematic (*at most n* → *at most n+1*) were accepted only by 23% of responders, the inferences of type D (*not more than n* → *not more than n+1*), were already accepted by 44%. The difference was statistically significant: $z = -2,333, p = .02, r = -.4$ [8] Thus, it seems that paraphrasing *at most n* to the negative form *not more than n* facilitates the inference.

Table 1. Logically correct inferences

	Premise	*Conclusion*	*Correct Responses*	*Percentage*
A *At least*	*at least 4*	*at least 3*	13	79%
	at least 3	*at least 2*	14	
B *At most*	*at most 4*	*at most 5*	3	23%
	at most 3	*at most 4*	5	
C *Not fewer than*	*not fewer than 4*	*not fewer than 3*	8	59%
	not fewer than 3	*not fewer than 2*	12	
D *Not more than*	*not more than 4*	*not more than 5*	7	44%
	not more than 3	*not more than 4*	8	
E *N or more than n*	*4 or more than 4*	*3 or more than 3*	10	67%
	3 or more than 3	*2 or more than 2*	13	
F *N or fewer than n*	*4 or fewer than 4*	*5 or fewer than 5*	6	23%
	3 or fewer than 3	*4 or fewer than 4*	2	
G *"At most" Equivalence*	*not more than 3*	*at most 3*	16	98%
	at most 3	*not more than 3*	17	
	not more than 4	*at most 4*	17	
	at most 4	*not more than 4*	17	
H *"At least" Equivalence*	*not fewer than 3*	*at least 3*	17	93%
	at least 3	*not fewer than 3*	17	
	not fewer than 4	*at least 4*	16	
	at least 4	*not fewer than 4*	13	
K *Numerical*	*5*	*4*	9	65%
	6	*2*	11	
	7	*6*	13	
	8	*5*	11	
embedded	*at most 4*	*at most 6*	17	100%
L	*at most 2*	*at most 3*	16	94%

Table 2. Logically incorrect inferences

	Premise	*Conclusion*	*Correct Responses*	*Percentage*
A' *At least*	*at least 4*	*at least 5*	1	6%
	at least 3	*at least 4*	1	
B' *At most*	*at most 4*	*at most 3*	0	6%
	at most 3	*at most 2*	2	
C' *Not fewer than*	*not fewer than 4*	*not fewer than 5*	1	9%
	not fewer than 3	*not fewer than 4*	2	
D' *Not more than*	*not more than 4*	*not more than 3*	1	9%
	not more than 3	*not more than 2*	2	
E' *N or more than n*	*4 or more than 4*	*5 or more than 5*	3	20%
	3 or more than 3	*4 or more than 4*	4	
F' *N or fewer than n*	*4 or fewer than 4*	*3 or fewer than 3*	8	50%
	3 or fewer than 3	*2 or fewer than 2*	9	
K' *Numerical*	*4*	*5*	0	1,5%
	2	*6*	1	
	6	*7*	0	
	3	*8*	0	

The inferences of type A, that is *at least n*→ *at least n-1*, turned out relatively unproblematic for subjects, who accepted them in ca. 80% of cases. Interestingly a paraphrase to the negative comparative form *not fewer than n*, made the task more difficult (59% accepted; means comparison: $z = -2.07, p = .038, r = -.355$). However, it is worth to note that inferences of type A were still rejected

[7] We give an overall result for a given type of an inference

[8] In all the cases we used Wilcoxon Signed Ranks test to compare means.

by ca. 20%, which suggests that there is some, at least pragmatic, mechanism suppressing this inference.

The results for the disjunctive inferences (E and F) are especially interesting. First of all the response pattern for disjunctive counterpart of *at least* corresponds with the predictions of classical logic: While logically valid inferences (E) were accepted on a relatively high level of 67% (which is lower, though not significantly lower, compared to the acceptance of the basic form (type A)), the invalid inferences (E') were mostly rejected (only 20% accept). The opposite effect, however, we got for the disjunctive form of *at most*. The logically valid inferences (F) were mostly rejected (only 23% accept), whilst invalid inferences (F') were accepted in exactly 50% of cases. Interestingly the acceptance rate of (F) inferences was similar to the acceptance rate of the basic form of inferences with *at most* (B). In both cases the differences between acceptance rate of correct and incorrect forms were statistically significant, and were as follows: The differences between correct and incorrect inferences with "disjunctive *at most*" (F and F'): $z = -2.491, p = .013, r = -.43$ and correct and incorrect "disjunctive *at least*" (E and E'): $z = -2.165, p = .030, r = -.37$. The differences between disjunctive *at most* and *at least*: incorrect (E' and F') $z = -2.057, p = .040, r = -.36$: and correct (E and F) :$z = -2.697, p = .007, r = -.46$.

The "correct" inferences with bare numerals were accepted in ca. 65% of cases. There was only one mistake in the incorrect inferences with bare numerals. Finally, both embedded *at most* inferences got nearly 100% correctness rate (one mistake only for question 2).

6 Discussion

Although our results cannot be treated as a final evidence of our theory, our experiment certainly provides various important observations that support the plausibility of our proposal.

First of all, all the implications between the negative comparative and superlative forms of considered quantifiers were almost without exceptions accepted by our subjects. This result supports the assumption that those are semantically equivalent forms in natural language.

Secondly, the inferences *(at least n → at least n-)*, although accepted by a majority of responders, were not as obvious as the standard theory would predict, and 20% of subjects rejected them (A, Table 1). This suggests existence of some, at least pragmatic, mechanism interfering in subject's reasoning with *at least*. What is also worth reminding, valid inferences with "disjunctive *at least*" were rejected even more often (F, Table 1). We consider that this effect can be explained in terms of the epistemic interpretation of *at least n* and its interaction with the reading of bare numerals. Let us notice that our results provide a weak evidence for the interplay between the reading of bare numerals and the treatment of "disjunctive *at least*" inferences. In our experiment inferences of type K: $n \rightarrow n-$, which base on the "at least" reading of numerals were accepted in ca. 65% of cases, thus our responders in 35% cases integrated the "exact" reading of bare numerals, which resulted in their rejection of considered inferences. However, "disjunctive *at least*" inferences (type E) ware rejected also in ca. 33%. We suggest that rejection of type E inferences was as well a result of an exact reading of a bare numeral n, as we have explained above. Although a correlation between subject's acceptance of type E and type K inferences failed to reach significance, it was close to significant (Spearman's rho= .426, $p = .08$) and we expect that with a bigger sample it could reach the significance level.

A similar effect is presumably the reason why 20% of subjects rejected type A inferences: (*at least n*→ *at least n-1*). Namely, the application of the epistemic verification condition of *at least n* together with an exact reading of bare numerals results in rejection of such inferences. This effect might be however weaker and less likely to occur than in the case when the disjunctive form is given explicitly.

Thirdly, the surprising result that subjects accepted the invalid inferences with "disjunctive *at most*" more frequently than the valid ones can be explained by our proposal. As we have proposed above, closure in the verification condition is optional, since the falsification condition is sufficient to account for the right semantics. However, if context enforces applying one of the semantical conditions (verification or falsification), then the other one tends to be ignored. While, from the perspective of classical logic it should be enough to use only one of the conditions (since the other can be defined via the first one), in the case of superlative quantifiers the epistemic reading of the verification condition creates the bifurcation in the meaning. This results in different inferential patterns in which those quantifiers occur, depending on what the context primarily enforced: the verification or falsification condition.

In what follows, and as we have explained above, when the verification condition is used, then *n or fewer then n* does not imply *n+ or fewer then n+* due to the epistemic interpretation. Though, it also does not exclude it if no closure is applied. However, *n or fewer then n* does imply *n- or fewer then n-* if the verification condition is used but no closure is applied. Now we can explain why inferences (F', Table 2): (*n or fewer than n*) → (*n-1 to fewer than n-1*) got a 50% rate of acceptance, although they are invalid both in the standard account and in the epistemic account. Based on Definition 4, $\exists^{=n}x\phi(x) \vee \exists^{<n}x\phi(x)$ is interpreted as $\diamond\exists^{=n}x\phi(x) \wedge \diamond\exists^{<n}x\phi(x)$ (with closure: $\neg\diamond\exists^{>n}x\phi(x)$). But $\diamond\exists^{<n}x\phi(x)$ can be broken down to: $\gamma = \diamond(\diamond\exists^{n-1}x\phi(x) \wedge \diamond\exists^{<n-1}x\phi(x))$ Now γ implies $\diamond\exists^{=n-1}x\phi(x) \wedge \diamond\exists^{<n-1}x\phi(x)$ (here we use the assumption that the world accessibility relation is transitive). In such a case it might happen that the closure of the conclusion, that is: $\neg\diamond\exists^{>n-1}x\phi(x)$ which contradicts the assumption that $\diamond\exists^{=n}x\phi(x)$ is ignored by subjects, which results in the high logical mistake ratio.

7 Conclusions

We have argued that the meaning of a quantifier can be defined as a pair $\langle C_F, C_V\rangle$, in which the verification (C_V) and the falsification (C_F) condition for sentences with this quantifier are specified separately. Though from the logical point of view those conditions should be dual, in the case of superlative quantifiers they are not. Namely, pragmatic focus on the borderline n in both *at least* n and *at most* n enforces a disjunctive verification condition, which is further interpreted as a conjunctive list of epistemic possibilities.

Finally, we would like to say few words about why we want to understand the verification and falsification conditions in terms of algorithms. Let us make an observation that semantical equivalence and procedural identity of algorithms are different things. Let us consider algorithms A_1 and A_2:

A_1 : Count all x that are ϕ. If the number m of x that are ϕ is smaller than $n-1$, then $verify$.

A_2 : Count all x that are ϕ. If the number m of x that are ϕ equals n or is smaller than n, then $verify$.

A_2 and A_3 are semantically equivalent, namely they verify logically equivalent formulas, e.g. ψ_1, ψ_2 and ψ_3, however in a sense of procedures that are executed they are not identical.

$$\begin{aligned}
&\psi_1 \iff \neg\exists^{>n} x\phi(x)\\
&\psi_2 \iff \exists^{<n+1}\phi(x)\\
&\psi_3 \iff \exists^{=n} x\phi(x) \vee \exists^{<n} x\phi(x)\\
&\psi_1 \iff \psi_2 \iff \psi_3
\end{aligned}$$

Furthermore, A_3

A_3 : When the number m of x that are ϕ is bigger then n, then $falsify$.

is dual to both A_1 and A_2, namely adding an *otherwise verify* condition to A_3 and *otherwise falsify* condition to A_1 and A_2 would make A_3 semantically equivalent to both A_1 and A_2. However, without the "otherwise" condition, A_3 does not allow to verify any of the given sentences, while A_2 or A_1 do not allow to falsify them. Then, $\langle A_1, A_3\rangle$, or $\langle A_2, A_3\rangle$ could be considered pairs of partial algorithms. Each pair could constitute a full semantical interpretation of each of the sentences ψ_1, ψ_2, ψ_3.

We consider that logically equivalent, but linguistically different, natural language sentences may trigger different kinds of such partial algorithms, or pairs of partial algorithms. First of all two equivalent sentences that differ in the linguistical form can trigger as primary only one of the algorithmic conditions: verification or falsification, while the complement condition is ignored. Secondly, they might trigger non-identical verification/falsification procedures. Consequently, it might happen that the executed procedures differ in complexity. Additionally, if we take into account that some extra mechanisms, e.g. the above-discussed epistemic interpretation of disjunctive conditions, might play a role, we not only obtain different algorithms in the procedural sense but also *semantically non-equivalent* algorithms for logically equivalent or even same sentences. For instance, A_2 would be replaced by:

> A_2': if [it is possible that there are exactly n x that are ϕ AND it is possible that there are fewer than n x that are ϕ], then $verify$.

While A_2 is dual to A_3, A_2' is not anymore. However a pair $\langle A_3, A_2' \rangle$ could constitute a semantical interpretation of a natural language sentence *at most* $n : x\phi(x)$, which would explain the non-semantically coherent (from the point of view of classical semantics) inference patterns in which this quantifier occurs.

Appendix: The complete list of pairs of sentences used in the experiment (premise, conclusion), translated from German.

At least four Anna's dress are red. At least three Anna's dresses are red.

Arthur has at least three cars. Arthur has at least two cars.

At most four books were stolen from the library. At most five books were stolen from the library.

Markus ate at most three pieces of cake. Markus ate at most four pieces of cake.

Not fewer than four cards are missing in the deck. Not fewer than three cards are missing in the deck.

A child has painted not fewer than three pictures. A chid has painted not fewer than two pictures.

Not more than four students came today to the philosophy seminar. Not more than five students came today to the philosophy seminar.

Sabine got not more than three presents. Sabine got not more than four presents.

Four or more than four students were sick this week. Three or more than three students were sick this week.

Christopher speaks three or more than three languages. Christopher speaks two or more than two languages.

Four or fewer than four students have passed the course. Five or fewer than five students have passed the course.

Beate has three or fewer than three children. Beate has four or fewer than four children.

Five people came to the party. Four person came to the party.

Monika invited six guests to her birthday party. Monika invited two guests to her birthday party.

Seven fruits in the basket have spoilt. Six fruits in the basket have spoilt.

Alicia bought eight bottles of beer. Alicia bought five bottles of beer.

At least four of Carolina's scarfs are blue. At least five of Carolina's scarfs are blue.

Thomas has read at least three books. Thomas has read at least four books.

At most four computers in the lab are broken. At most three computers in the lab are broken.

Andrea baked at most three pizzas. Andrea baked at most two pizzas.

Not fewer than four professors attended the meeting. Not fewer than five professors attended the meeting.

Hans had not fewer than three glasses of wine. Hans had not fewer than four glasses of wine.

Not more than four people have applied for this job. Not more than three people have applied for this job.

Natalie wrote not more than three exercises. Natalie wrote not more than two exercises.

Four or more than four girls in the class are good in arts. Five or more than five girls in the class are good in arts.

Christina's cat gave birth to three or more than three kittens. Christina's cat gave birth to four or more than four kittens.

Four or fewer than four students failed in the exam. Three or more than three students failed in the exam.

Tanja trains three or fewer than three times a week. Tanja trains two or fewer than two times a week.

Four new students joined the chess club this week. Five new students joined the chess club this week.

Stephanie baked two cakes for her birthday. Stephanie baked six cakes for her birthday.

Six members of the library club came to the meeting. Seven member of the library club came to the meeting.

Frank gave his mother three roses. Frank gave his mother eight roses.

Not more than three children have done their homework for today. At most three children have done their homework for today.

At most three girls took part in the maths competition. Not more than three girls took part in the maths competition.

Erika has not more than four necklaces. Erika has at most four necklaces.

Lena takes at most four courses at the university. Lena takes not more than four courses at the university.

Not fewer than three new animals were born in the city zoo. At least three new animals were born in the city zoo.

At least three exotic three have died in our botanic garden. Not fewer than three exotic threes have died in our botanic garden.

Daniel plays not fewer than four times a week football. Daniel plays at least four times a week football.

Albert has at least four exams this semester. Albert has not fewer than four exams this semester.

Bibliography

Breheny, Richard. 2008. A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics* 25.93.

Carston, Robin. 1998. Informativeness, relevance and scalar implicature. In *Relevance theory: Applications and implications*, ed. by R. Carston & S. Uchida, 179-236. Amsterdam: John Benjamins.

Cohen, Ariel & Krifka, Manfred. 2011. Superlative quantifiers as modifiers of meta-speech acts. In. Partee, B.H., Glanzberg, M., & Skilters, J. (Eds.) (2011). *Formal semantics and pragmatics. Discourse, context and models. The Baltic International Yearbook of Cognition, Logic and Communication,* Vol. 6 (2010). Manhattan, KS: New Prairie Press.

Cummins, Chris & Katsos, Napoleon. 2010. Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*, 27: 271-305.

Geurts, Bart. 2006. Take 'five': The meaning and use of a number word. In *Non-definiteness and Plurality*, ed. by S. Vogeleer & L. Tasmowski, 311-329. Amsterdam/Philadelphia: Benjamins.

Geurts, Bart and Nouwen, Rick. 2007. "At least" et al.: The semantics of scalar modifiers. *Language*, 83: 533-559.

Geurts, Bart, Katsos, Napoleon, Cummins, Chris, Moons, Jonas, and Noordman, Leo. 2010. Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, 25: 130-148.

Horn, Laurence R. 1972. *On the Semantics of Logical Operators in English.* Ph. D. thesis, Yale University dissertation.

Koster-Moeller, Jorie, Varvoutis, Jason & Hackl, Martin. 2008. Verification procedures for modified numeral quantifiers, *Proceedings of the West Coast Conference on Formal Linguistics*, 27.

Krifka, Manfred. 1999. At least some determiners aren't determiners. In K. Turner (ed.), *The semantics/pragmatics interface from different points of view.* (= *Current Research in the Semantics/Pragmatics Interface* Vol. 1). Elsevier Science B.V., 257-291.

Levinson, Stephen C. 1983. *Pragmatics.* Cambridge, England: Cambridge University.

Musolino, Julien. 2004. The semantics and acquisition of number words: Integrating linguistic and developmental perspectives, *Cognition*, 93-1, 1-41.

Nouwen, Rick 2010. Two kinds of modified numerals. *Semantics and Pragmatics* 3 (3).

Szymanik, Jakub. 2009.*Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*, DS-2009-01, Amsterdam.

Szymanik, Jakub & Zajenkowski, Marcin. 2010. Comprehension of simple quantifiers. Empirical evaluation of a computational model. *Cognitive Science*, 34(3), pp. 521-532.

Szymanik, Jakub & Zajenkowski, Marcin. Improving methodology of quantifier comprehension experiments. 2009. *Neuropsychologia*, Vol. 47, No. 12, pp. 2682-2683.

Zimmermann, Thomas Ede. 2000. Free choice disjunction and epistemic possibility, *Natural Language Semantics*, 8 (2000), 255 - 290.

Quantifiers and visual cognition: the processing of proportional and superlative *most* in Bulgarian and Polish

Barbara Tomaszewicz

University of Southern California

btomasze@usc.edu

Abstract. I provide experimental evidence that quantifier semantics guides visual verification processes (Lidz et al. 2011). I tested the processing of two majority quantifiers in Bulgarian and Polish: the proportional Most1, the counterpart of English *most,* and the superlative/relative Most2. Three obtained notable results have been obtained: (i) Most1 is verified by a Subtraction strategy, directly replicating the findings of Lidz et al. for Slavic; (ii) Most2 is verified by a Selection strategy in accordance with its lexical semantics; (iii) each verification strategy is consistently used even in cases where either strategy would yield the correct truth value.

1 Introduction

Lidz et al. (2011) argue that the verification of truth/falsity a declarative sentence is biased towards those procedures that are transparently associated with the semantic representation of that sentence. They show that sentences containing the quantifier *most* such as (1) are uniquely associated with truth conditions and a verification procedure involving subtraction (2), despite the availability of other semantically equivalent specifications (e.g. 3).

1. Most of the dots are yellow.
2. |Dot(x) & Yellow(x)| > |Dot(x)| – |Dot(x) & Yellow(x)|
3. |Dot(x) & Yellow(x)| > |Dot(x) & Red(x)| + |Dot(x) & Blue(x)| + |…|

I provide further experimental evidence that quantifier semantics guides the verification process. My evidence is based on the comparison of the verification patterns of two minimally distinct quantifiers and suggests that the properties of the linguistic input directly influence the unconscious visual processes.

The meaning of *most* intuitively refers to a comparison of quantities, where one of the quantities is greater than others. For countable objects what is compared are cardinalities. Visual perception of numerical information has been studied extensively and it is known in psychology that the visual selection of a target "can be influenced by expectations and strategies" (Trick, 2008). Manipulating a linguistic stimulus affects the patterns of the visual search for obtaining a cardinality (or its estimation), however, at the same time the choice of the visual verification strategy is also constrained by the psychophysics of visual cognition. We can both formulate hypotheses and interpret the visual response pattern on the basis of the findings about human perception in visual numerical judgment tasks. Under time pressure, when precise counting becomes impossible, people switch to the Approximate Number System (ANS) that generates a representation of magnitude and is governed by Weber's law, i.e. the greater the distance between two numbers the better discriminability (Dehaene, 1997). Numbers can be represented as 'noisy magnitudes' even for the purposes of basic arithmetic operations like addition and subtraction. Quantifiers can be verified against a visual display even when counting is blocked.[1] Psychophysical constraints, however, affect the accuracy of judgment.

Lidz et al. (2011) hypothesize that the procedure in (3) (selection of each individual color set in order to obtain the cardinality of the non-yellow set) should be computationally costly if the verification involves more than one non-yellow set, because of the evidence from Halberda et al. (2006) that on a 500ms display, multiple color sets can be enumerated in parallel, but only for the total set of dots and two color subsets. The procedure in (3) involves selection of each individual color set in order to obtain the cardinality of the non-yellow set. The subtraction procedure in (2), on the other hand, is independent of the number of color sets and is thus more suitable as a general verification strategy for the quantifier *most*.

I argue, however, that the choice of the procedure (2) over (3) for the verification of the English quantifier *most* is not forced by psychophysics. My evidence suggests that (3) is psychologically available as a procedure for visual verification, because a computationally similar procedure is employed by the speakers of Bulgarian (Bg) and Polish (Pl) when verifying the superlative majority quantifiers *naj-mnogo* (Bg) and *najwięcej* (Pl), cf. (4).

4. |Dot(x) & Yellow(x)| > |Dot(x) & Red(x)|,
   & |Dot(x) & Yellow (x)| > |Dot(x) & Blue(x)|,
   & |Dot(x) & Yellow (x)| > |Dot(x) & Green(x)|, & …

I tested the processing of two majority quantifiers in Bulgarian and Polish: the counterpart of English proportional majority quantifier *most* (*povečeto* in Bg, większość in Pl, henceforth Most1) and a superlative/relative majority quantifier (*naj-mnogo* in Bg, *najwięcej* in Pl, henceforth Most2). I obtained three notable results:

[1] Also Halberda, Taing and Lidz (2008) have shown that children who have not yet learned to count are perfectly able to understand sentences containing *most.*

- Most1 is verified by a Subtraction strategy as in (2) and not (3), directly replicating the findings of Lidz et al. for Slavic;
- Most2 is verified by a Selection strategy as in (4) in accordance with its lexical semantics;
- Each verification strategy remains to be used even in cases where either strategy would yield the correct truth-value.

The results have some immediate implications for the semantics of quantifiers and the interface of semantics with visual cognition. We can argue for the contribution of the individual morphemes not only to the meaning of Most1 vs Most2 but also to the interface with the visual cognition. The combined Bulgarian and Polish results further strengthen the conclusions I presented in Tomaszewicz (2011) that quantifier semantics provides a set of instructions to visual verification processes, since each of the two Polish Most1 and Most2 biases a distinct verification strategy.

.

2 Previous research

Pietroski et al. (2008), Lidz et al. (2011) devised experimental paradigms to look "beyond" the truth conditions of (1) to see how the meaning of a sentence containing *most* constrains the way people verify it against a visual scene. The two semantic specifications in (5) are truth conditionally equivalent, but they differ in how the cardinality of the non-yellow set of dots is arrived at. (5a) expresses a comparative relation between the cardinalities of two sets, while (5b) is a one-to-one correspondence function that maps an ordered pair of sets (X, Y) to a truth value.

5. (a) $|\mathrm{Dot}(x)\&\mathrm{Yellow}(x)|>|\mathrm{Dot}(x)\&{\sim}\mathrm{Yellow}(x)|$
   (b) $\mathit{OneToOnePlus}(\{\mathrm{Dot}(x)\&\mathrm{Yellow}(x)\},\{\mathrm{Dot}(x)\&{\sim}\mathrm{Yellow}(x)\})$

Pietroski et al. (2008) obtained evidence that even when the arrangement of dots favors the verification by strategy in (5b) (i.e. paired vs. unpaired arrangements of dots in two colors), this strategy is not used. Using the same experimental paradigm requiring visual verification under time pressure (screens displayed for 150 ms), Lidz et al. (2011) investigated how the cardinality of the non-yellow set in (5a) is estimated when this set contains dots in 1-4 different colors. They tested which of the two specifications in (6-7) *most* is verified with.

6. Selection strategy
   $|\{\mathrm{Dot}(x)\&\mathrm{Red}(x)\} \cup \{\mathrm{Dot}(x)\&\mathrm{Blue}(x)\} \cup \{\mathrm{Dot}(x)\&\mathrm{Green}(x)\} \cup \ldots|$
7. Subtraction strategy
   $|\mathrm{Dot}(x)| - |\mathrm{Dot}(x)\&\mathrm{Yellow}(x)|$

Their proposed Subtraction strategy is based on the psychological evidence that a heterogeneous set is not automatically selectable (i.e. red, green, blue dots are not automatically processed as one set unless it is the total set of dots), as well as on the findings of Halberda et al. (2006) that humans can automatically (i.e. without a prompt) compute the total number of dots and two color subsets but no more. Thus, Subtraction in (7) does not depend on the number of colors of dots, while Selection in (6) should.

In the experiment of Lidz et al. (2011) screens with dots in up to 5 colors in varying ratios (yellow to non-yellow dots) were flashed for 150ms. Twelve participants evaluated whether the sentence in (1) was true on each screen and the patterns of accuracy of their responses were analyzed. No difference in accuracy was found as the function of the number of colors of dots, but only as the function of the ratio (in adherence to Weber's law). This indicates that Subtraction was always used for the judgment of (1). Crucially, on the screens with just 2 colors, Selection would be computationally less costly (it would involve less steps than Selection as shown in Table 1) and thus more accurate.

Table 1. Steps involved in Selection as opposed to Subtraction on two-color screens.

Subtraction *(irrespective of no. of colors)*	*Selection* *two colors*
1. Estimate the total.	1. Estimate the target set.
2. Estimate the target set.	2. Estimate the distractor set.
3. Subtract the target set from the total.	3. Compare with the target set.
4. Compare the difference with the target set.	

Yet, even on the two-color condition Subtraction appeared to be used, since the accuracy was not higher than on the multi-color screens, i.e. the verification procedure failed to make use of the automatically obtained information, the cardinality of the two subsets that could be compared directly (Halberda et al. 2006). Thus, Lidz et al. conclude that Subtraction is the default procedure for verifying *most* under time pressure. On the basis of this finding they formulate the Interface Transparency Thesis: "the verification procedures that speakers employ, when evaluating sentences they understand, are biased towards algorithms that directly compute the relations and operations encoded by the relevant logical forms" (Pietroski et al. 2011).[2]

[2] What is crucial when comparing different strategies is evidence that a more advantageous strategy is failed to be used in favor of one that can be directly linked to semantics. Pietroski et al. (2011) "take it as given that speakers use various strategies in various situations. For us, the question is whether available procedures are *neglected*—in circumstances where they could be used to good effect—in favor of a strategy that reflects a candidate logical form for the sentence being evaluated." A strategy may be abandoned in favor of one that is unrelated to a semantic representation, but that cannot be taken as evidence against a particular semantic specification.

3 Most1 and Most2 in Bulgarian and Polish

Bulgarian and Polish have “two” versions of the English majority quantifier *most*. Most1 in both languages has the same proportional reading as the English *most* has, so that (8a) and (9a) are equivalent to (1).

8. (a) Povečeto točki sa žəlti. *Bulgarian*
Most1 dots are yellow
'**Most** dots are yellow.'
(b) Naj-mnogo točki sa žəlti.
Most2 dots are yellow
'Yellow dots form the largest subset.’

9. (a) Większość kropekjestżółta. *Polish*
Most1 dots is yellow
'**Most** dots are yellow.'
(b) Najwięcej jestkropekżółtych.
Most2 is dots yellow
'Yellow dots form the largest subset.'

Most2 in Bulgarian (8b) and Polish (9b) contains superlative morphology in contrast to Most1 as illustrated in (11). In accordance with the standard meaning of the superlative morpheme (the relative reading), Most2 modifying a noun says that what the noun denotes is the most numerous thing among other things of the same type, in our case, the set of yellow dots is more numerous than any other color set.

Table 2. The morphological make-up of Bulgarian and Polish Most1 and Most2.

Most1	Most2
“more than half”	*“largest subset”*
Bulgarian	
po-veče-to *er*+*many*+*the*	naj-mnogo **est**+many
Polish	
więk-sz-ość *many*+*er*+nominalizer	naj-więc-ej **est**+many+er

I predicted that Most1, being equivalent to the English most, should be compatible with the Subtraction strategy. Thus, the number of color sets was expected to not affect the accuracy of judgments with Most1 (it should only be affected by the ratio of yellow to non-yellow dots). Since the semantics of Most2 can be specified as in (10), which I call Stepwise Selection, I expected to find both an effect of ratio and of number of colors in contrast to Most1.

10. Stepwise Selection strategy
|Dot(x) & Yellow(x)| > |Dot(x) & Red(x)|, &
|Dot(x) & Yellow (x)| > |Dot(x) & Blue(x)|, &
|Dot(x) & Yellow (x)| > |Dot(x) & Green(x)|, & …

Both of the predictions were met. The results of the Experiment 1 (on Bulgarian) and of the Experiment 2 (on Polish) contain exactly the same main effects in the two conditions.

3.1 Experiments 1 and 2: Materials and methods

I conducted two on-line visual-display verification studies designed along the lines of the experiment of Lidz et al. (2011). A group of 39 native speakers of Bulgarian participated in Experiment 1 and 20 native speakers of Polish participated in Experiment 2. The Polish experiment is reported in Tomaszewicz (2011).

The procedure was identical in both experiments. The participants evaluated the truth of the sentences in (8-9) by pressing Yes or No buttons while viewing displays of arrays of colored dots on a black background, flashed on a computer screen for 200ms. I manipulated the ratio of the yellow target to the rest (1:2, 2:3, 5:6, i.e. 3 levels of the ratio variable) and the number of color sets (1, 2 or 3 other distractor colors, i.e. 3 levels of the distractor variable). The numbers of colors in each bin are presented in Table 5 in the Appendix.

As the schema in Fig. 8 in the Appendix shows, 360 displays were presented in 2 blocks (180 for Most1 and 180 Most2, half of each requiring a yes response and half a no response). Participants had 380ms to indicate their response by a button press. The experiment was performed using Presentation® software (Version 14.2, www.neurobs.com).

3.2 Experiments 1 and 2: Results

For Most1 accuracy rates were significantly affected only by ratio, and not by number of color sets (Table 3, rows (a),(c)). For Most2 accuracy rates were significantly affected both by ratio and by number of color sets (Table 3, rows (b),(d)). I analyzed each quantifier with a 3x3x2 Repeated Measures ANOVA crossing the 3 levels of the ratio variable, 3 levels of distractor, and the truth/falsity of screens:

Table 3. Accuracy rates.

			ratio				color sets			
			1:2	2:3	5:6		2	3	4	
(a)	Bg	Most1 (*povečeto*)	**.858**	**.778**	**.643**	p<.001	.764	.748	.767	p.=.321
(b)		Most2 (*najmnogo*)	**.827**	**.742**	**.617**	p<.001	**.807**	**.731**	**.648**	p<.001

(c)	Pl	Most1 (*większość*)	**.871**	**.785**	**.673**	p<.001	.797	.763	.769	p.=.215
(d)		Most2 (*najwięcej*)	**.866**	**.76**	**.63**	p<.001	**.801**	**.767**	**.688**	p<.001

The accuracy rates with Most1 in Bulgarian and Polish are significantly affected only by ratio[3] and not by number of color sets. These results are the same as for the English *most* in Lidz et al. (2011) and are entirely consistent with the prediction that Most1 is verified by Subtraction. The graphs in Fig. 2 and Fig. 3 clearly show the lack of a main effect of number (Bulgarian: F(2, 76) = 1.153, p = .321; Polish: F(1.47, 27.98) = 1.637, p = .215[4]).

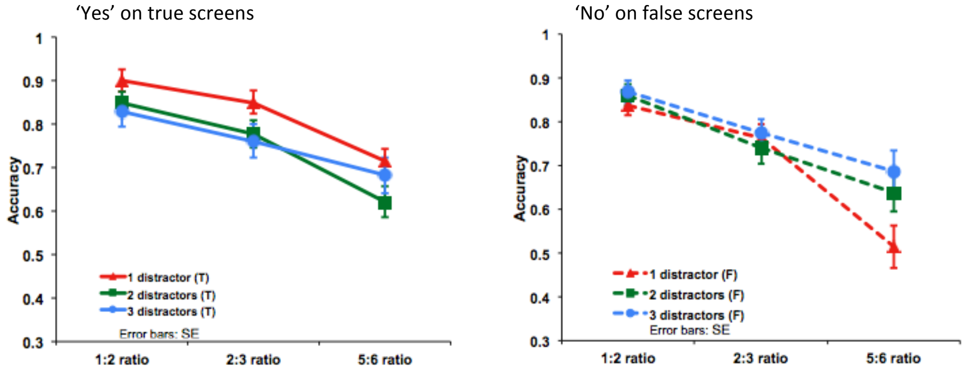


Fig. 2. Most1 in Bulgarian.

'Yes' on true screens

'No' on false screens

[3] Bulgarian: F(2, 76) = 171.791, p < .001, Polish: F(2, 38) = 76.072, p < .001. Post hoc tests using the Bonferroni correction revealed significant differences (p < .001) between all levels of the ratio variable.

[4] Because of the violations of sphericity (p = .019), we are reading the Greenhouse-Geisser corrected value. Whether or not we use this correction, there is still no significance: F(2, 38) = 1.64, p = .208.

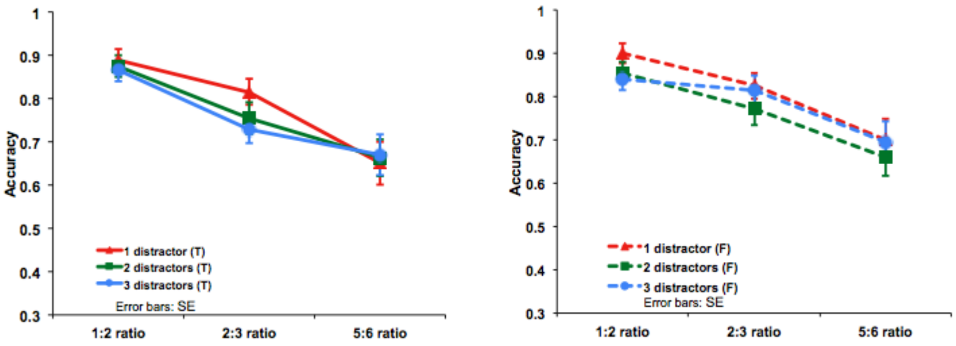


Fig. 3. Most1 in Polish.

The results for Most2 are entirely compatible with the view that it is verified by Selection. In both Bulgarian (Fig. 4) and Polish (Fig. 5) the accuracy rates are significantly affected both by ratio (Bulgarian: F(2, 76) = 182.449, $p < .001$, Polish: F(2, 38) = 124.77, $p < .001$) and number of color sets (Bulgarian: F(2, 76) = 72.612, $p < .001$, Polish: F(2, 38) = 17.34, $p < .001$).[5]

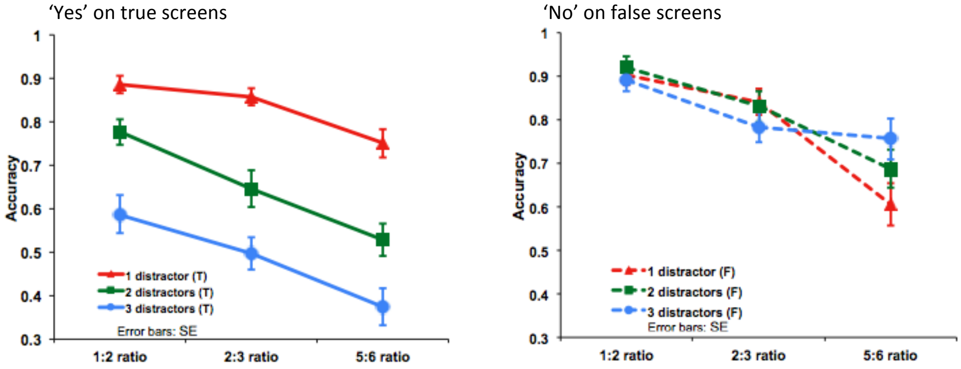


Fig. 4. Most2 in Bulgarian.

'Yes' on true screens 'No' on false screens

[5] Pair-wise comparisons for the main effect of ratio and the main in effect of distractor in Bulgarian (using a Bonferroni correction) revealed significant differences ($p < .001$) between all levels. For Polish the differences between all levels of the ratio variable were significant ($p < .001$). The differences between 1-3 and 2-3 distractors were significant ($p < .001$ and $p = .001$ respectively), while the difference between 1-2 distractors was not ($p = .316$). Note that the Polish sample (N=20) is much smaller than the Bulgarian sample (N=39).

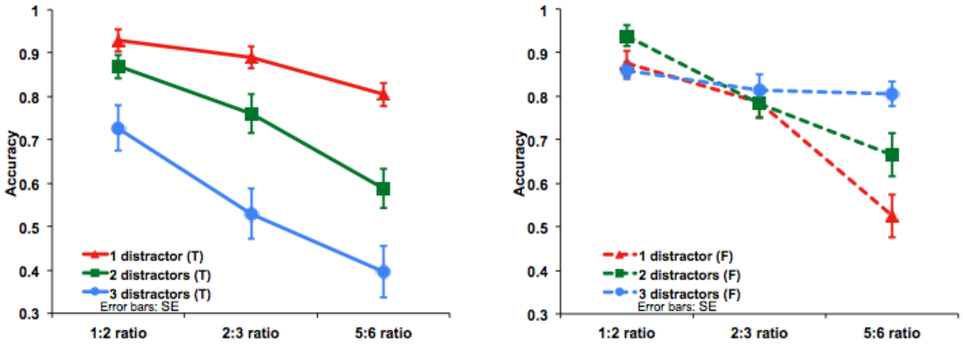


Fig. 5. Most2 in Polish.

It is also evident in the graphs in Fig. 5 and Fig. 6 that the accuracy with Most2 is affected by the truth/falsity of screens. The present design does not allow us to determine the reason for this, however, with Selection correct estimation of both the target set and each color set is expected to be affected by a higher number of factors than Subtraction.

Crucially, the significant effect of number of colors in addition to the effect of ratio indicate that both the yellow set and the other color sets are selected for the verification of Most2 in conformity with its semantics.[6]

Importantly, on screens with 2 color sets (identical for both quantifiers) both Bulgarian and Polish participants were significantly less accurate and slower confirming the truth of Most1 than of Most2. This indicates that Subtraction continues to be used with Most1 and Selection with Most2 even on the condition, where switching between the two procedures would provide more accurate results.

Participants could have used whichever strategy is computationally less costly/more accurate under time pressure, since both strategies are otherwise used by the speakers of Bulgarian and Polish. If the semantic representation guides verification, then with Most2 the non-yellow set should be selected directly – the accuracy should be greater than with Most1 where the non-yellow set is computed (cf. Lidz et al. 2011), which is exactly what we find on true screens.

[6] Note that successful selection and comparison of 3-4 color sets in 200ms is not inconsistent with the findings of Halberda et al. (2006). The three set limit is on the automatically obtained information without a stimulus that creates expectations and directs attention to some specific aspect of the display. The superlative morphology clearly contributes an expectation that multiple sets should be compared.

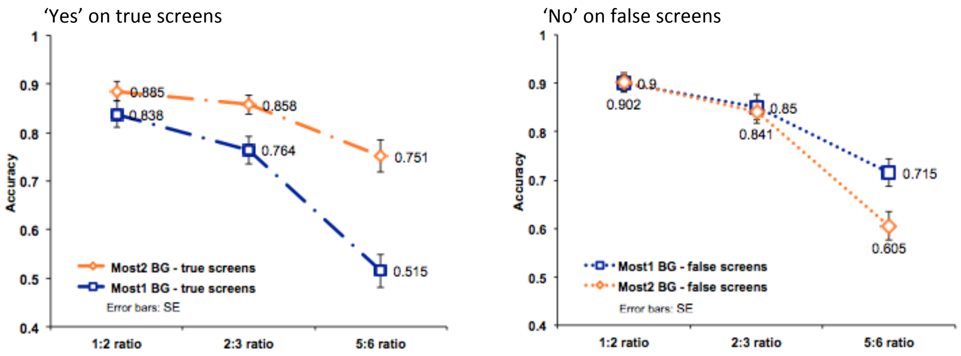


Fig. 6. Two color condition: Most1 vs. Most2 in Bulgarian.

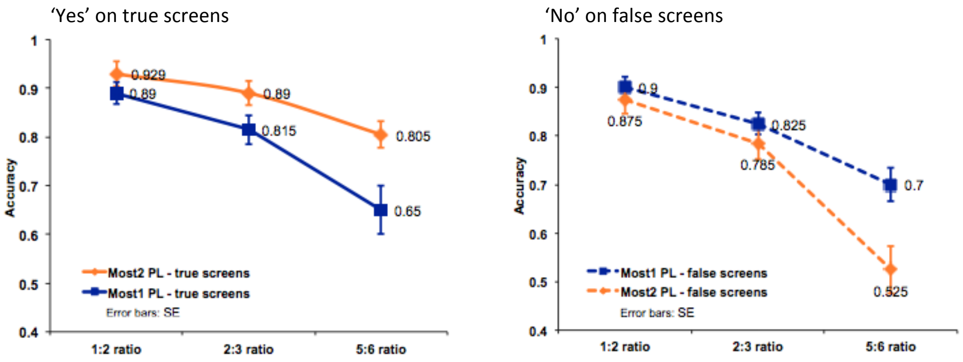


Fig. 7. Two color condition: Most1 vs. Most2 in Polish.

Both Bulgarian and Polish participants were significantly better with Most2 than Most1 on true screens (Bulgarian: (F(1, 38) = 32.970, p < .001, Polish: F(1, 19) = 10.49, p = .004). On false screens Most1 is significantly better than Most2 (Bulgarian: (F(1, 38) = 4.892, p = .033, Polish: F(1, 19) = 11.122, p =.003).

Notably, the two languages also behave exactly the same with respect to the reaction times. The accuracy is higher despite faster RTs and lower despite slower RTs (Table 4). On true screens Most2 is faster (Bulgarian: F(1, 38) = .587, p = .448, Polish: F(1, 19) = 5.173, p = .035). On false screens Most1 is faster (Bulgarian: F(1, 38) = 9.884, p = .003, Polish: F(1, 19) = .351, p = .561). See Table 6 in the Appendix for mean RTs. The RT data shows that it is not the case that people are more prone to errors as they make judgments faster. Instead, we can see that the procedure with Most2 on true screens is easier (faster, more accurate judgments) which is expected if the two color sets are selected directly. On false screens Most1 is judged faster and more accurately, which does not seem to follow from Subtraction vs. Selection difference. However, the correct disconfirmation probably involves more factors that cannot be identified on the present design.

Crucially, the accuracy patterns together with RTs consistent in both languages indicate that participants do not switch to the more advantageous strategy, e.g. they don't use Selection to more accurately confirm the truth of Most1. This is the more interesting given the findings of Halberda et al. (2006) that the cardinality of two color sets is automatically computed. Yet the semantics of Most1 apparently precludes the use of this automatically available information.

Different behavior with each quantifier on the very same screens indicates that participants do not switch between the procedures and that the way those procedures differ is specified by the semantics. Computation for both Most1 and Most2 involves the comparison between the yellow and the non-yellow set. The components provided by the visual system are exactly the same: yellow set, non-yellow set, superset. However, the algorithms must be different. To verify Most2 one has to (i) estimate target, (ii) estimate competitor, (iii) compare. To verify Most1 one needs to (i) estimate target, (ii) estimate total, (iii) subtract target from total. The lexical meaning of the functional morphemes that build up Most1 and Most2 and their logical syntax are interfacing with the visual system during the verification process.

4 Conclusions

In conclusion, our experiments indicate that semantics provides a direct set of instructions to the visual cognition processes, and that these instructions are followed even when computationally more advantageous strategies are available.

We have met the prediction that Bulgarian and Polish proportional majority quantifier Most1, just like English *most*, is verified using Subtraction strategy (we found a main effect of ratio and no effect of number of colors). The superlative/relative majority quantifier Most2 requires the Stepwise Selection strategy (as evidenced by the effect of ratio together with the effect of number of colors)[7]. Importantly, in a within-subject design the same group of participants behaves differently depending on the quantifier. The overall patters of accuracy are exactly the same in Bulgarian and Polish.

[7] As one of the reviewers observes, my evidence for the different verification processes for Most1 and Most2 is based on the use of the ANS representation of magnitude for the comparisons required by the semantics. If the superlative Most2 incurs a larger processing cost, it would be interesting to see if we find evidence for it also in experiments where counting is not precluded. Note, however, we cannot just "switch off" ANS, e.g. the effects of ratio-dependency characteristic of ANS are present also with judgments involving Arabic numerals,s although the quantities evoked by Arabic numerals may be more precise than those evoked by sets of dots (Dehaene 1997).

On two color screens (where Most1 and Most2 are either both true or both false) the verification procedure depends on the lexical item used. The patterns of accuracy for Most1 and Most2 were conspicuously different (but had the same direction in both Bulgarian and Polish) indicating that computationally Most1 and Most2 are different.

My results confirm and extend the findings of Pietroski et al. (2008) and Lidz et al. (2011) and indicate that semantics provides inviolable instructions to visual cognition processes.

References:

1. Dehaene, S. (1997) The number sense: How the mind creates mathematics. New York: Oxford University Press.
2. Halberda, J., Sires, S.F. & Feigenson, L. (2006) Multiple spatially overlapping sets can be enumerated in parallel. Psychological Science, 17.
3. Halberda, J., Taing, L. & Lidz, J. (2008) The development of "most" comprehension and its potential dependence on counting-ability in preschoolers. Language Learning and Development, 4(2).
4. Lidz, J., P. Pietroski, T. Hunter & J. Halberda (2011) Interface Transparency and the Psychosemantics of *most*. Natural Language Semantics 19.
5. Pietroski, P., J. Lidz, T. Hunter & J. Halberda. (2008) The meaning of *most*: Semantics, numerosity and psychology. Mind and Language, 24(5).
6. Pietroski, P., Lidz, J., Hunter, T.,Odic, D. and Halberda, J. (2011) Seeing what you mean, mostly. In J. Runner (ed.), *Experiments at the Interfaces*, Syntax & Semantics 37. Bingley, UK: Emerald Publications.
7. Tomaszewicz, B. M., (2011) "Verification Strategies for Two Majority Quantifiers in Polish", In Reich, Ingo *et al*. (eds.), *Proceedings of Sinn & Bedeutung 15*, Saarland Unversity Press: Saarbrücken, Germany.
8. Trick, L. M. (2008) More than superstition: Differential effects of featural heterogeneity and change on subitizing and counting. Perception and Psychophysics, 70.

Acknowledgements:

I would like to thank Roumi Pancheva as well as Elsi Kaiser, Barry Schein and Toby Mintz, for their invaluable comments and/or help with the experiments. Thanks are also due to Hagit Borer, Jeff Lidz, Martin Hackl, Victor Ferreira, Jon Gajewski, Robin Clark, Tom Buscher, Katy McKinney-Bock, Mary Byram, Ed Holsinger, Krzysztof Migdalski, Dorota Klimek-Jankowska, Grzegorz Jakubiszyn, Ewa Tomaszewicz, Marcin Suszczyński, Patrycja Jabłońska, Petya Osenova, Halina Krystewa, Petya Bambova, Ewa Panewa, Bilian Marinov, Marieta and the students of IFA at Wroclaw University and WSF Wroclaw and Polish Institute in Sofia.

My work was partially supported by an *Advancing Scholarship in the Humanities and Social Sciences Grant* from USC awarded to R. Pancheva.

Appendix:

Table 5. The numbers of dots in each bin.

Most1 - total number of screens: 180

no. of screens			ratio	dis-trac-tors	no. of yellow dots	no. of non-yellow dots	total no. of dots
30	10	true	1:2	**1**	**8 - 12**	**4 - 6**	**12 - 18**
	10			2	8 - 12	4 - 6	12 - 18
	10			3	8 - 12	4 - 6	12 - 18
30	10	true	2:3	**1**	**8 - 12**	**5 - 8**	**13 - 20**
	10			2	8 - 12	5 - 8	13 - 20
	10			3	8 - 12	5 - 8	13 - 20
30	10	true	5:6	**1**	**8 - 12**	**7 - 10**	**15 - 22**
	10			2	8 - 12	7 - 10	15 - 22
	10			3	8 - 12	7 - 10	15 - 22
30	10	false	1:2	**1**	**5 - 9**	**10 - 18**	**15 - 27**
	10			2	5 - 9	10 - 18	15 - 27
	10			3	5 - 9	10 - 18	15 - 27
30	10	false	2:3	**1**	**5 - 9**	**8 - 14**	**13 - 23**
	10			2	5 - 9	8 - 14	13 - 23
	10			3	5 - 9	8 - 14	13 - 23
30	10	false	5:6	**1**	**5 - 9**	**6 - 11**	**11 - 20**
	10			2	5 - 9	6 - 11	11 - 20
	10			3	5 - 9	6 - 11	11 - 20

Most2 - total number of screens: 180

no. of screens			ratio	dis-trac-tors	no. of yellow dots	no. of dots in closest competitor	total no. of dots
30	10	true	1:2	**1**	**8 - 12**	**4 - 6**	**12 - 18**
	10			2	8 - 12	4 - 6	13 - 23
	10			3	8 - 12	4 - 6	14 - 27
30	10	true	2:3	**1**	**8 - 12**	**5 - 8**	**13 - 20**
	10			2	8 - 12	5 - 8	15 - 27
	10			3	8 - 12	5 - 8	16 - 33
30	10	true	5:6	**1**	**8 - 12**	**7 - 10**	**15 - 22**
	10			2	8 - 12	7 - 10	19 - 31
	10			3	8 - 12	7 - 10	22 - 29
30	10	false	1:2	**1**	**5 - 9**	**10 - 18**	**15 - 27**
	10			2	5 - 9	10 - 18	18 - 35
	10			3	5 - 9	10 - 18	19 - 42
30	10	false	2:3	**1**	**5 - 9**	**8 - 14**	**13 - 23**
	10			2	5 - 9	8 - 14	16 - 31
	10			3	5 - 9	8 - 14	17 - 38
30	10	false	5:6	**1**	**5 - 9**	**6 - 11**	**11 - 20**
	10			2	5 - 9	6 - 11	14 - 28
	10			3	5 - 9	6 - 11	15 - 32

Table 6. Reaction Times (RTs, in tenth of millisecond).

<table>
<tr><td colspan="2"></td><td colspan="2">Most2</td><td colspan="2">Most1</td></tr>
<tr><td colspan="2"></td><td>true screens</td><td>false screens</td><td>true screens</td><td>false screens</td></tr>
<tr><td rowspan="2">1:2 ratio</td><td>BG</td><td>8784</td><td>9475</td><td>9066</td><td>9114</td></tr>
<tr><td>PL</td><td>8434</td><td>9922</td><td>8978</td><td>9706</td></tr>
<tr><td rowspan="2">2:3 ratio</td><td>BG</td><td>`9724</td><td>10316</td><td>9411</td><td>9423</td></tr>
<tr><td>PL</td><td>8797</td><td>10498</td><td>9892</td><td>9789</td></tr>
<tr><td rowspan="2">5:6 ratio</td><td>BG</td><td>9884</td><td>10793</td><td>10461</td><td>9730</td></tr>
<tr><td>PL</td><td>9569</td><td>10204</td><td>10217</td><td>10660</td></tr>
<tr><td></td><td></td><td colspan="2">Most2 slower on false screens,
BG: p = .002, PL: p <.001</td><td colspan="2">Most1: no sig. difference between true and false screens</td></tr>
<tr><td></td><td></td><td colspan="4">Most2 faster on true screens:
BG: p = .45 (ratio*quantifier p = .027), PL: p <.001
Most2 slower on false screens:
BG: p = .003, PL: p = .56 (ratio*quantifier p = .003, ratio p = .065)</td></tr>
</table>

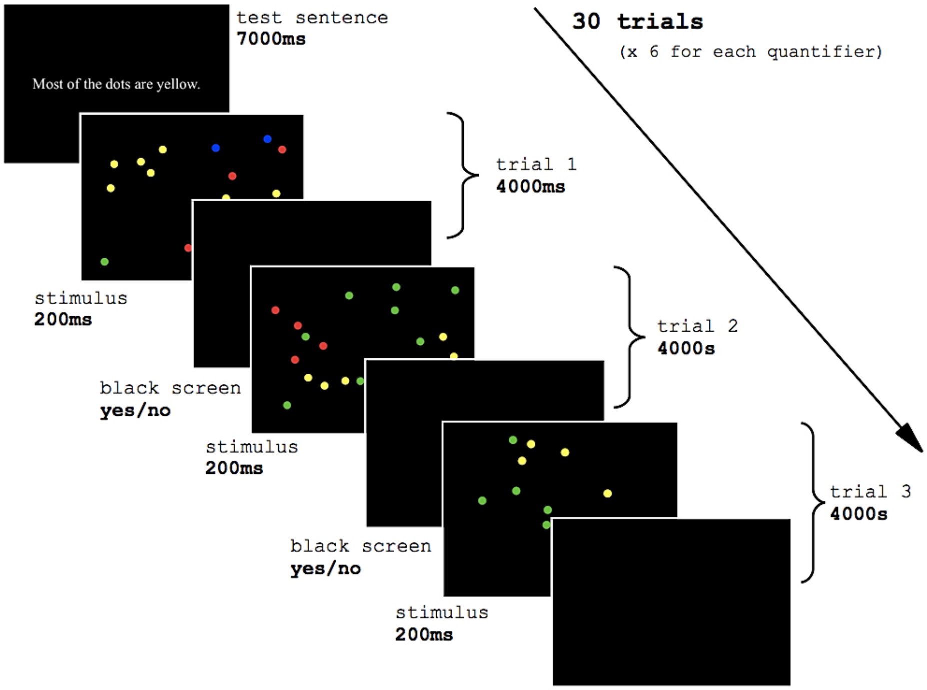


Fig. 8. A schema of the experimental procedure.

The experimental investigation of the updated traditional interpretation of the conditional statement

András Veszelka

Pellea Human Research & Development Ltd., Budapest, Hungary
andras.veszelka@gmail.com

Abstract. It is well known that the interpretation of the conditional statement in "everyday life" deviates from the official logical approach. It is conceivable, however, that the ancient logicians who first demonstrated the official approach erroneously characterised the "if P or R then Q" relationship in place of the "if P then Q" statement. When fixing this error, it turns out that the equivalent interpretation of the conditional statement, which is traditionally seen as one of the most common everyday fallacies, is in fact exactly the correct interpretation. Since classical logic has not been built on mathematical grounds but rather on philosophical argumentations and insights, its findings can be tested with the tools of today's human sciences, among others, with empirical experiments. The main experimental tools support this updated logical approach, and show that everyday thinking can be made compatible with logic. These results are summarized in this study.

Keywords. scholastic logic, psychology of reasoning, human experimentation, Wason selection task

1 Introduction

The conditional statement is a glaring example of how the abstractions of logic and the everyday use of the logical connectives deviate from each other. Many interpret this as the difference between formal and natural languages (e.g. [1]). This differentiation can be traced back to the beginning of the 20th century, where, for example, Frege [2] argued that the difference between the interpretations of the conditional statement as prescribed in logic and as used in "everyday life" reveals linguistic or psychological components. This is where the search for the linguistic or psychological components deemed different from logic began, first in the philosophy of language, then in linguistics and finally in psychology. However, the so-called everyday interpretation of the conditional statement does not merely deviate from such formal languages as propositional logic, which was born in Frege's era, but also from the classical interpretation of the conditional statement, which is basically the same as that of propositional logic, but which has been clearly created not in a mathematical, but in a linguistic, philosophical and psychological environment. Thus, the discrepancy between the everyday interpretation and the classical abstraction is within the same system.

1.1 The abstraction error in classical logic

That said, when looking back to the classical interpretation, it can be seen that it is erroneous. Instead of the "if P then Q" statement, classical logicians have erroneously abstracted the "if P or R then Q" statement. Let's take an example from Jevons [3] (p. 70), a late scholastic logician:

If the snow is mixed with salt, it melts

As is well known, in this if P then Q statement from the snow mixed with salt (P) antecedent, it is correct to infer the snow melts (Q) consequent (modus ponens, MP), and from the snow not melting (not-Q) it is correct to infer that it has not been mixed with salt (not-P) (modus tollens, MT). However, the denial of the antecedent (DA, if the snow is not mixed with salt (not-P), it does not melt (not-Q)) and the affirmation of the consequent (AC, if the snow melts (Q), it was mixed with salt (P)) are incorrect. Jevons has argued that, for example, from the snow melting (Q), it does not follow that it has been mixed with salt (P) because it can melt by other means as well. It is impossible to find any other explanation, even going back several hundred years, as to why these two latter inferences are incorrect. On the contrary, this interpretation can be traced back even to Aristotle, who wrote that:

The refutation, which depends upon the consequent, arises because people suppose that the relation of consequence is convertible. For whenever, suppose A is, B necessarily is, they then suppose also that if B is, A necessarily is. This is also the source of the deceptions that attend opinions based on sense perception. For people often suppose bile to be honey because honey is attended by a yellow colour: also, since after rain the ground is wet in consequence, we suppose that if the ground is wet, it has been raining; whereas that does not necessarily follow ([4], 167b1ff.).

To complete this argument, the inference that it has been raining does not necessarily follow because there are other possible means to make the ground wet. However, if we refer to additional possible causes, that is, to additional possible antecedents during the abstractions, these have to be denoted. In logic, it is fundamental that "we restricted ourselves to explicitly stated premises" ([5], p 6.). With their denotation, however, it can be seen that with the above explanations we characterized the “if P or R then Q” statement. Jevons' example was therefore the "If the snow is mixed with salt or, for example, the sun is shining, the snow melts" statement that he erroneously characterised in terms of P and Q only. These non-abstracted alternative antecedents can be found in every example provided for the interpretation of the conditional statement in the history of logic.

1.2 The correct abstraction of the conditional statement

The question arises therefore, of what the correct abstraction of the conditional statement can then be, that is, the abstraction of the relationship in which there is no wedging "or R" component. I believe the correct inference pattern is the equivalence, in which all the MP, MT, AC, DA inferences are valid. For example, we endorse the equivalent AC and DA inferences for the "if-then" connective, even in the case of the "if P or R then Q" statement. For instance from Q, we endorse the affirmation of the consequent (AC) inference, and we deduce to "P or R". As the traditional interpretation goes, *within this* we do not infer exclusively to P because it can be R as well. This can be demonstrated the same way in the case of all three other classical inferences as well. On the other hand, classical equivalent statements such as, for instance, "if the sun is in the sky then it is day" are equivalent because the context of these statements does not allow one to wedge any alternative antecedents. There can be day only if the sun is in the sky. Even propositional logic refers to the alternative antecedents, as when it differentiates the equivalence (to use another term, the biconditional) with the artificial expression "If P *then and only then* Q" from the "if P then Q" conditional, in the latter case of which, by parity of argument, several antecedents can lead to Q. All of this is illustrated in further detail by Veszelka [6]. When explaining the interpretation of propositional logic within the if-then statement, Geis and Zwicky [7] reinvented and employed the aforementioned scholastic interpretation, and by mentioning alternative antecedents they managed to block one of the most common fallacies that people commit in the case of the conditional statement, the equivalent/biconditional inferences. This approach has subsequently been implemented in psychology, and its mechanism has to an extent been experimentally verified. Byrne [8] has demonstrated that if the second antecedent is connected to the initial antecedent with an "and" connective, then in terms of P and Q only, the MP and MT inferences would be invalid and the DA and AC inferences would remain valid. This is the case, for instance, in the example of the "If the snow is mixed with salt and it is not extremely cold, it melts" (If P and R then Q) statement. These are very interesting relationships, however, and as a consequence of the historical reasons illustrated in the beginning of this study, this phenomenon is interpreted in linguistics and in psychology as a linguistic, pragmatic effect, which is contrary to logic. It was nevertheless demonstrated above that this phenomenon is actually the update of the classical logical interpretation of the conditional statement. It is the exact definition of what differentiates between the two well-known inference patterns on the conditional statement, the traditionally accepted conditional inference pattern, which allows only MP and MT, and the equivalence. In antiquity, the rule of thumb used was that since the conditional statement can evoke both conditional and equivalent inferences, one should label only those inferences that are prescribed by both of them as valid, that is, the MP and the MT [9]. Obviously, the new definition is more accurate. However, many important psychological experiments are in conflict with this approach.

2 The experimental investigation of the conditional statement

2.1 The demonstration of the biconditional inferences

The most important task of this type, the “single most investigated problem in the literature on deductive reasoning” ([10], p. 224) is Wason's abstract selection task. Consequently, Byrne, who introduced the study of the alternative antecedents into psychology, has rejected the basic biconditional interpretation of the conditional statement [8]. In this task, participants are shown four schematic cards having a letter on one side and a number on the other. Participants are then asked what card or cards they would turn over in order to decide whether, for example, the “If there is a letter E on one side there is a number 4 on the other side” conditional statement is true. On the cards, the “E” (P), “K” (not-P), “4” (Q) and “7” (not-Q) can be seen. In this task, abstract letters and numbers are used in order to assure that the context and the content have no influence on the results and so they accurately display how people interpret the if-then statement itself. The traditional conditional interpretation would be selecting the cards P and not-Q, since these could have falsifying instances on their other side, while the biconditional interpretation would be the selection of all four cards. The customary response is, however, merely the P and Q value. In the psychological field on logical reasoning, the logical negation is expressed in three different ways. It can be implicit (e.g. “A”, and its negation = “K”), explicit (“A” and its negation “not-A”) and dichotom, which is the same as the implicit, but in which the task instruction states that only two possible values can be found (e.g. “A” and “K”). The result is P and Q with all three negatives [11]. This result constitutes an important basis for many theories in the field. There are three additional main tasks:

- Truth-table evaluation task, in which the given co-occurrences of the truth table of propositional logic, for instance the co-occurrence of P and Q, must be evaluated in terms of whether it verifies or falsifies the conditional statement, or is irrelevant to it.
- Inference task, in which on the basis of the provided conditional statement, people must decide if the given conclusions follow from the minor premises or not, for example whether or not from not-P, not-Q follows.
- Inference production task, in which participants themselves write down what follows from the minor premises.

The available results from the combination of the four tasks and the three types of negatives are shown in Table 1.

Table 1. Results of the main task types with the tree types of negatives

	Implicit neg.	Explicit neg.	Dichotom neg.
Selection task	P&Q	P&Q	P&Q
Truth table task	Def. table [12]^1	Def. table [12]^1	≡ (83%) [13]^2
Inference task	≡ (48%) [14]^2	?	≡ (60%) [14]^2
Inf product.task	?	?	≡ (92%) [13]^2

^{1} Defective truth table
^{2} Biconditional

As can be seen in Table 1, although there are biconditional solutions, the results are generally inconsistent and there are missing data. For this reason, I have retested all of the tasks [15] except for the abstract selection task, which has robust results for all three types of negatives. Consequently, for the selection task, I tested two thematic tasks that have an evidently biconditional context in order to check if the results of these tasks deviate from the results of the abstract selection task, or if they also evoke the preference of the P and Q values, as was already observed by some researchers. My results are shown in Table 2.

Table 2. My results of the main task types with the tree types of negatives [15]

	Implicit neg.	Explicit neg.	Dichotom neg.
Selection task	P&Q	P&Q	P&Q
Truth table task	Def. truth table	Def. truth table	Bicon (50%)
Inference task	Def. truth table	Def. truth table	Bicon (42%)
Inf product.task	Bicon (73.3%)	Bicon (52%)	Bicon (67%)

The reasoning contained in the defective truth tables[1] require further analysis, although there are several explanations for this phenomenon that are compatible with the updated scholastic approach. It is still apparent that the predominant response is the biconditional. With regard to the selection task, instead of the biconditional responses, both tested biconditional problems evoked the P and Q preference, the characteristic response of the abstract selection tasks. According to my hypothesis [15], which has been also formulated and partially tested by Wagner-Egger [14] one month prior to my study, people avoid the selection of all the cards in the selection task. They believe that selecting all four cards would be contrary to the task instruction, which in

[1] In the defective truth table the co-occurrence of “P and Q” verifies the conditional statement, the co-occurrence of “P and not-Q” falsifies it, and the “not-P and Q” and the “not-P and not-Q” co-occurrences are irrelevant to it.

fact requires them to select *from among* the cards. This is fairly apparent in the case of the following task, which was one of the tasks involving biconditional context that I have tested:

On one side of each card, there is the name of a city and on the other side there is a mode of transportation. Let us suppose that when someone goes to Budapest, he always goes by car, and when he goes to Szeged, he always goes by train. Likewise, when he travels by car, he always goes to Budapest, and when he travels by train, he always goes to Szeged. Mark the card or cards that must be turned over in order to decide whether this is true.

The following statements were printed on the cards: “going to Budapest”, “going to Szeged”, “going by train” and “going by car” ([15], Experiment 3).

In this task, which was tested on 2x20 participants, everyone produced the biconditional answer in the inference production task, but only 10% did so in the selection task. However, as it can be seen, the task was in fact a pseudo-problem, because it contained a clear description of what follows from what, or what value has to figure on the other side of the cards. I obtained the same result on another clearly biconditional problem, the so-called “ball-light” problem [16]. This problem is commonly accepted in the literature as a biconditional task which, being tested on 2x30 participants, has produced biconditional answers in 96% of inference production tasks, but only in 23% of selection tasks [15]. Since participants do not find a better solution than the avoided biconditional response, they finally select those instances that are named in the conditional statement, the P and Q values. Thus, the main experimental tasks altogether support the biconditional approach.

2.2 The demonstration of the response traditionally deemed correct

One half of the updated classical interpretation of the conditional statement, the basic equivalent interpretation, can be therefore experimentally demonstrated. Another empirical obstacle to this approach is to trigger the P and not-Q answer, the traditionally expected response in logic. The elicitation of the “correct” answers has so far been studied almost exclusively with selection tasks, and in so-called thematic selection tasks researchers obtained the allegedly correct P and not-Q response several decades ago. The most cited task of this type is the drinking-age task [17], in which participants have to imagine that they are on-duty police officers who must check if everyone observes the rule that “If someone is drinking beer, he must be older than 18 years”. “Drinking beer”, "Drinking soft drink”, “21 years old,” and “17 years old” appear on the cards. A large proportion of participants select “Drinking beer” and the “17 years old” cards in this task—that is, the P and not-Q cards. This result is interpreted to arise from various effects, such as from pragmatic reasoning schemas [18], from relevance [19], from deontic context [20], from precautions [21], from cheater detection [22], from altruist context [23], from perspective switching [24], or from benefits or costs [5]. However, these are not normatively valid explanations, because in classical logic or propositional logic, where the abstraction itself has been defined, such components were clearly not present. This can be easily seen in the examples mentioned at the beginning of this study as well. It can be observed, however, that there is a wedging of information in the easy-to-resolve selection tasks as well, which mainly correspond with the effect of the alternative antecedents in the updated interpretation of classical logic: In the above task everyone knows that people above 18 years can drink both alcohol and other beverages, although this is not explicitly communicated in the task instruction. One of the experiments of Hoch and Tschirgi [25] can be seen as a means to test this additional information, in which they used in an abstract task, with the appropriate substitutions, the statement that "Cards with a P on the front may only have Q on the back, but cards with not-P on the front may have either Q or not-Q on the back" ([25], p 203). Although this cue facilitation produced 56% correct results in the experiment of Hoch and Tschirgi [25], in the replication of the experimental condition [26] the rate was only 36% in the usual experimental population, and participants with knowledge of logic were not filtered out; this could evidently improve the result. With the usual experimental population, only a modest improvement was received with this type of facilitation [27]. This task has so far been tested only in selection tasks. In an unpublished experiment (with 21 participants), I also received correct answers only in 14% of selection tasks, but the rate was 76% when the very same task was presented in the inference-production task Pearson Chi-Square (1, 42) = 16.243, p < .0001, Cramer's V = .622. Perhaps in this case once again, people in the selection task would test the complete relationship, and test, for instance, that both P and not-P can figure on the card with a Q on its other side. This would again require turning over all four cards, and as such the distorting effect mentioned above could reappear. Similarly, it can be observed that, contrary to this facilitation attempt, in the above easy-to-resolve drinking-age task the relationship that people above 18 years can also drink soft drinks is from outside of the task, it is not included in the investigated conditional statement. As a result, it must not be part of the examination. To test this assumption in an abstract selection task, in an unpublished experiment I used the following task:

Imagine that four cards are lying in front of you on the table. On one side of the card there is either the number 4 or the number 6; on the other side, there is either „divisible by two" or „divisible by three". Your task is to check whether the four cards on the table each conforms with the reality, namely, with the rule that

If the number is 4, then it is divisible by only two

Which card or cards would you turn over to check this?

In the control task I replaced 6 with 3 in the instruction and on the second card, and in order to assure better text comprehension, I removed the word "only" from the if-then statement. According to my interpretation, therefore, the two tasks evoke two different relationships as shown in Figure 1.

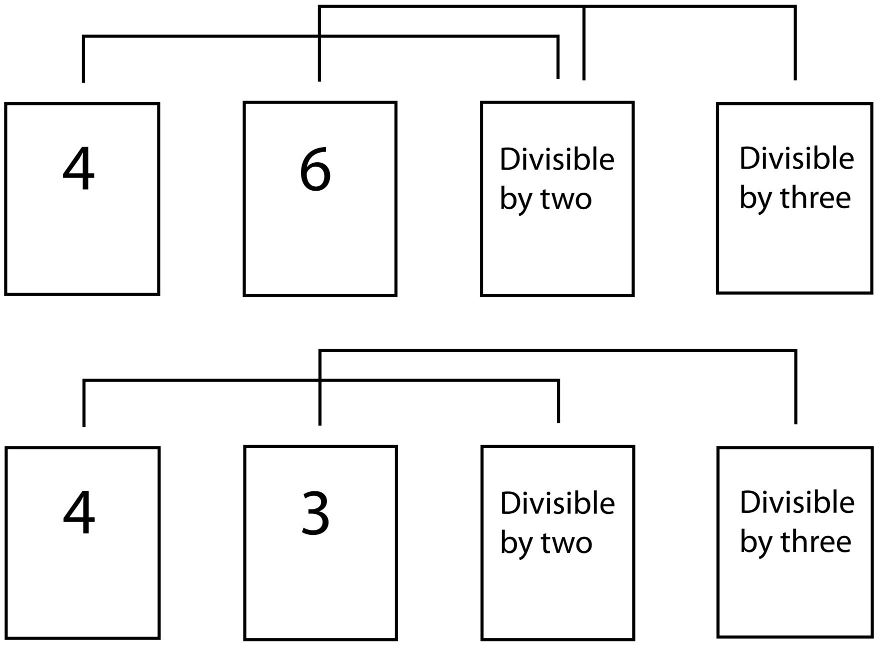


Fig. 1. Cards and evoked relations in the experiment on the "easy to resolve" abstract selection task

With number 6, the task produces the conditional inference pattern, and with number 3 the biconditional pattern. Indeed, with 21x22 participants I obtained P and not-Q responses on the conditional task in 41% of cases. The rate of the "P and Q" and "all" responses being in conformity with the biconditional was only 13.5%, whereas in the control task the rate of the "P and not-Q" answers was merely 4%, and the "P and Q" and "all" comprised 86% of the results. The difference is obvious with Pearson's Chi-Square (3, 43) = 20.157, p < .0001, with Cramer's V = .685, and because of the fact that the minimal differences between the tasks explain themselves. However, at a different university, where participants were given twice as much time to resolve the task, I failed to reproduce these results. It was then raised by a colleague that IQ scores have a similar difference between the two universities, and that IQ could possibly also play a role in the way the tasks are resolved. For this reason, with Anikó Kecse Nagy, we tested the task in the summer camp of Mensa HungarIQa. This organization collects Hungarians older than 17 years of age and who obtained a score on the Raven Advanced Matrices IQ test higher than 98% of the general Hungarian population. In this experiment, performed with 20x16 participants, people with and without knowledge of logic participated equally. The results altogether were 69% vs. 30% "P and not-Q" answers, Pearson Chi-Square (1, 36) = 5.355, p < .021, Cramer's V = .0386 for the conditional task, versus the biconditional task. The difference among the "all" biconditional answers was also significant in the opposite direction

(25% vs. 0%), Pearson Chi-Square (1, 36) = 4.654, p < .031, Cramer's V = .359. Deconstructing the results further, 55% of the participants with no knowledge of logic (9 subjects) gave "P and not-Q" answers to the conditional task, while 25% of them gave the same answers to the biconditional task. Even participants with knowledge of logic produced significantly more P and not-Q answers to the conditional task (86% vs. 33%), Pearson Chi-Square (1, 19) = 4.866, p < .027, Cramer's V = .506, and more "all" responses to the biconditional task (0 vs. 33%), Pearson Chi-Square (1, 19) = 2.956, p < .086, Cramer's V = .394). Although the 55% rate of correct responses of the Mensa members not familiar with logic is still below the 70-75% rate of easy-to-resolve thematic tasks, Fiddick and Erlich ([28], Exp 1) have received P and not-Q selections in only 54% of cases even when the participants were explicitly instructed to search for the falsifying co-occurrence of P and not-Q in the abstract selection task. It is therefore conceivable that this is the maximum one could obtain from this task.

3 Conclusion

In general, the functioning of the updated classical logical interpretation of the conditional statement can be demonstrated by the main experimental tasks used in experimental psychology. People basically interpret the conditional statement as an equivalent relation, and with the effect of the alternative antecedents this modifies into the relationship known as the conditional. This approach can be defended from the point of view of the history of logic [6], and is normatively valid. Many researchers assume that the description of human inferences necessitates the introduction of non-monotonic logics, or that the everyday interpretation of the conditional statement is not truth-functional [30]. Still, the results presented here could be well described with a merely slightly updated classical logic. In addition, this approach can also describe the everyday interpretation of syllogisms [29]. Non-monotonic logics (e.g. default logic [31], defeasible logic [32]) are introduced with reference to the effect of a certain type of context, without, however, denoting this context. To reiterate, this seems to be a mistake, as in logic "we restricted ourselves to explicitly stated premises" ([5], p 6). Leaving the context undenoted, or for example the traditional interpretation of logical necessity and logical truth is probably the heritage of a classical logic that, in consequence of the erroneous interpretation of the conditional statement, was rigid and unable to develop, and did not allow to describe the effect of the context. When fixing this error, however, the basic effect of the context can be seen even when the equivalent relationship transforms into a conditional relationship—and this can be quite precisely described. The purpose of non-monotonic logics is also to describe such belief revisions. A similar example of the basic context is the otherwise mathematical content, which can be observed at the end of this study in the easy-to-resolve abstract selection task. The conditional statement itself is basically the same in the two experimental conditions "If the number is 4, then it is divisible by (only) two", the underlining relationships (3 or 6) are, however, different. Still, these underlying relationships can be precisely described, they do not require to introduce a specific apparatus just because the conditional statement in one of the cases evokes equivalent, and in the other case a conditional relationship, and with the addition of further contextual components, could behave again in quite a different way. I believe that more complex contexts and even the concepts themselves behave in accordance with the same principle. Naturally, in a more complex case, we cannot predict the exact context or conceptual network in someone's mind, but without precise information we cannot predict which numbers someone is adding in his mind either. We could make only vague or probabilistic predictions, just as happens in the case of the vague or probabilistic approaches of the conditional statement. However, addition and subtraction written down on paper are still very useful tools.

In another logical approach of the field, Stenning and van Lambalgen [33], [34] worked precisely on defining the components behind the differences of such individual inferences. According to them, participants in the abstract selection task have first to define the parameters, and the differently chosen parameters produce the many different answers, all of which are correct within the given parameters. The authors themselves note that the parameters discussed by them are difficult to demonstrate experimentally, and they assume that further parameters could be discovered. In this respect, this study also defines such parameters, with markedly significant results, such as, for example, the basic equivalent inference, the avoidance of the selection of all cards and the effect of the alternative antecedent. And, of course, the whole literature investigating the relationship between logic and everyday inferences can be interpreted as the search for and testing of such parameters—components that influence how people resolve the tasks. According to this study, however, the many different answers appearing in the abstract selection task are merely artefacts resulting from the avoidance of the equivalent inferences. The same many different answers making altogether the preference of P and Q cards also appear in the evidently biconditional ball-light selection task already mentioned in this study [15]. It is, however, evident that the equivalent inference is the only correct solution in this thematic task. So, in the selection task, the search for the parameters that follows the rejection of the correct equivalent responses does not necessary reveal much about the basics of the inferential processes. Still, they can provide important information on how people try to resolve a situation that was made logically ambiguous. It is true that in the verbal reports presented by Stenning and van Lambalgen participants do not speak about avoiding the equivalent response. However, if logic has been unsure about the interpretation of the conditional statement for 2,400 years, layman participants cannot be expected to formulate a clear picture about this in the 5-10 minutes that they are given to resolve the tasks. They particularly cannot be expected to be so sure about their interpretation that, on the basis of this, they question the hidden instruction in the task, going against the equivalent responses. As a matter of fact, even the good performance on the easy-to-resolve drinking-age thematic task already mentioned in this study drops back to half (75% to 35%) by presenting only two P and two not-Q cards to the subjects, hence requiring the turning over of each of them [35].

In this study, instead of analysing the individual responses I intended to define the overall correct responses and to demonstrate empirically that people generally adhere to them. According to this approach, the greater the extent to which a task can be resolved in the same way, the more it appears easy and evident to the experimental participants. As the rate of characteristic response drops from 100% to just 20-30%, so the task becomes more and more obscure to the participants. The more the task become obscure, the more contextual effects activate in their mind in a great variation—giving a wider variety of parameters. The most characteristic solution for a task is a sort of vote on what people believe is the correct solution in that task. This study demonstrates that this voting/belief can be equated with some logical rules, which are very simple and hence can probably also be easily programmed into a machine.

Acknowledgements. I am grateful to Professor László Bernáth for all his support during this research, for Péter Bencsik for creating Figure 1 and for the comments of three anonymous reviewers.

References

1. Copi, I. M. (1986). Introduction to logic. (7th ed.) NY: Macmillan.
2. Frege, G. (1923). Compound thoughts. In E. D. Klemke (ed.) Essays on Frege. (pp. 537-558) Urbana, Ill: University of Illinois Press.
3. Jevons, W. S. (1906). Logic. London: Macmillan Philip.
4. Aristotle. (1928). De sophisticis elenchis: The works of Aristotle, edited and translated under the supervision of W.D. Ross. Oxford: Oxford University Press.
5. Evans, J. St. B. T., & Over D. E. (2004). If. Oxford University Press.
6. Veszelka, A. (2008). A feltételes állítás ekvivalens értelmezése: hibás a feltételes állítás tradícionális absztrakciója? (The equivalent interpretation of the conditional statement: Error in the traditional abstraction?), Világosság, 48(1), 15-22. English version available.
7. Geis, M., & Zwicky, A. M. (1971). On invited inferences. Linguistic Inquiry, 2, 561-566.
8. Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. Cognition, 31, 61-83.
9. Maróth, M. (1983). Arisztotelésztől Avicennáig. [From Aristotle to Avicenna] Budapest, Hungary: Akadémiai.
10. Evans, J. St. B. T. (1996). Deciding before you think: relevance and reasoning in the selection task. British Journal of Psychology, 87(2), 223-240.
11. Johnson-Laird, P. N., & Wason, P. C. (1977). A theoretical analysis of insight into a reasoning task In P. N. Johnson-Laird & P. C. Wason (Eds.), Thinking: Readings in cognitive science. Cambridge, UK: Cambridge University Press
12. Evans, J. St. B. T., Clibbens, J., & Rood, B. (1995). Bias in conditional Inference: Implications for mental models and mental logic. Quarterly Journal of Experimental Psychology, 48A, 644-670.
13. George, C. (1992). Rules of inference in the interpretation of the conditional connective. Cahiers de Psychologie Cognitive/ Current Psychology of Cognition. 12, 115-139.
14. Wagner-Egger, P. (2007). Conditional reasoning and the Wason selection task: Biconditional interpretation instead of a reasoning bias. Thinking & Reasoning, 13(4), 484-505.
15. Veszelka, A. (2007). A feltételes állítás kísérleti vizsgálata: mégis van benne logika? (The experimental investigation of the conditional statement: Does it have logic?), Magyar Pszichológiai Szemle, 62(4), 475-488. English version available.
16. Legrenzi, P. (1970). Relations between language and reasoning about deductive rules. In Giovanni B. Flores D'Arcais & Willem J. M. Levelt (Eds.), Advances in Psycholinguistics. Amsterdam: North-Holland.

17. Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. British Journal of Psychology, 73, 407-420.
18. Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. Cognitive Psychology, 17, 391-416.
19. Sperber, D., Cara F., & Girotto, V. (1995). Relevance theory explains the selection task. Cognition 57, 31-95.
20. Oaksford, M., & Chater, N. (1998). Rationality in an uncertain world: Essay on the cognitive science of human reasoning. Hove, UK: Psychology Press.
21. Manktelow, K. J., & Over, D. E. (1990). Deontic thought and the selection task. In K. J. Gilhooly, M., Keane, R. Logie, and G. Erdos (Eds). In Lines of thought: Reflections on the referpsychology of thinking. Chichester: Wiley
22. Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with Wason selection task. Cognition, 31, 187-316.
23. Brown, W., & Moore, C. (2000). Is prospective altruist-detection an evolved solution to the adaptive problem of subtle cheating in cooperative ventures? Supportive evidence using the Wason selection task. Evolution and Human Behavior, 21, 25-37.
24. Staller, A. Sloman, S., & Ben-Zeev, T. (2000). Perspective effects in nondeontic versions of the Wason selection task. Memory and Cognition, 28, 396-405.
25. Hoch, S. J., & Tschirgi, J. E. (1983). Cue redundancy and extra logical inferences in a deductive reasoning task. Memory & Cognition, 11, 200-209.
26. Hoch, S. J., and Tschirgi, J. E. (1985). Logical knowledge and cue redundancy in deductive reasoning. Memory and Cognition,13, 435-462.
27. Platt, R. D., & Griggs, R. A. (1993). Facilitation in the abstract selection task: The effect of attentional and instructional factors. Quarterly Journal of Experimental Psychology, 46A, 591-613.
28. Fiddick, L., & Erlich, N. (2010). Giving it all away: altruism and answers to the Wason selection task. Evolution and Human Behavior, 31, 131-140
29. Veszelka, A. (in press). Köznapi következtetések és kategórikus szillogizmusok. (Everyday inferences and categorical syllogisms). Világosság. English version available.
30. Edgington, D. (2001). Conditionals. In Lou Goble (ed.). Philosophical Logic. Blackwell.
31. Reiter, R. (1980). A logic for Default reasoning. Artificial Intelligence 13, 81-132.
32. Nute, D. (2003). Defeasible logic. Lecture Notes in Computer Science, vol. 2543/2003, 151-169.
33. Stenning, K., van Lambalgen, M. (2000). Semantics as a foundation for psychology: a case study of Wason's selection task. Journal of Logic, Language and Information, 10, 273-317.
34. Stenning, K., van Lambalgen, M. (2008). Human reasoning and cognitive science. MIT Press.
35. Veszelka, A. (1999). Feltételesen plasztikus (Conditionally pliable). Master's thesis, ELTE University, Hungary