Monotonicity and the Complexity of Reasoning with Quantifiers

Jonathan Sippel (jonathan.frederik.sippel@gmail.com)

Jakub Szymanik (jakub.szymanik@gmail.com)
Institute for Logic, Language, and Computation
University of Amsterdam

Abstract

We present a natural logic for reasoning with quantifiers that can predict human performance in appropriate reasoning tasks. The model is an extension of that in (Geurts, 2003) but allows for better fit with data on syllogistic reasoning and is extended to account for reasoning with iterated quantifiers. We assign weights to inference rules and operationalize the complexity of a reasoning pattern as weighted length of proof in our logic – this results in a measure of complexity that outperforms other models in their predictive capacity and allows for the derivation of empirically testable hypotheses.

Keywords: Reasoning; Quantifiers; Natural Logic; Psychology; Syllogisms

Of Logic and Reasoning

Logic and psychology can look back on a shared history that is full of twists and turns, with the two moving back and forth on their commitment to one another. Recently, some researchers proclaimed the time of logic in psychology to be over (Evans, 2002) while others argue that most of such criticism can be traced back to an uncalled-for equalization of logic with "standard logic", an umbrella-term for both predicate calculus and propositional logic.[1].A prominent example is Wason's infamous selection task, in which, if propositional logic is taken to be the adequate normative standard for human reasoning, only 5% of all participants manage to solve their task properly (Wason, 1983). But "the unargued adoption of classical logic as criterion of correct performance is thoroughly antilogical" (Stenning & van Lambalgen, 2012, 45). Fortunately, there are other alternatives. Braine (1978) already brought forward a *natural logic* with a directional entailment relation that accounts for Wason's results. Natural logics, a collection of various logical formalisms, emphasize the fact that some important and recurring natural language expressions are not only carriers of information but allow for reasoning, see, e.g., Icard III and Moss (2014).

The phenomenon of reasoning that caught our eye is that of inferences with quantifiers, e.g. ALL, NO, and MOST and their iterations, e.g. in "MOST pigeons annoyed AT LEAST THREE tourists". If $Q(A,B)$ is a quantifier, we can usually define it by only referring to the two sets A and B:

$\text{ALL}(A,B) \Leftrightarrow A \subseteq B$
$\text{SOME}(A,B) \Leftrightarrow A \cap B \neq \emptyset$
$\text{MOST}(A,B) \Leftrightarrow |A \cap B| > |A - B|$
$\text{MORE THAN } 2(A,B) \Leftrightarrow |A \cap B| > 2$

We will refine and extend a natural logic proposed by Geurts (2003) that captures the essential inferential properties of single and iterated quantifiers. At the semantic center of our logic is the notion of *monotonicity*. Consider the following example:

(i) All flowers are vermilion. ($\text{ALL}(F,V)$)
(ii) No flowers are red. ($\text{NO}(F,R)$)

Sentence (i) entails "All flowers are red" ($\text{ALL}(F,R)$) while (ii) entails "No flowers are vermilion" ($\text{NO}(F,V)$) because the set of all vermilion things is a subset of all red things ($\text{ALL}(V,R)$). We will usually say that the quantifiers ALL and NO are right-side upward monotone and right-side downward monotone, respectively, or just speak of their *directionality*. This kind of inference can be generalized to *iterated quantifiers*: we can infer that "MOST pigeons annoyed AT LEAST THREE humans" from "MOST pigeons annoyed AT LEAST THREE tourists" because we know that the set of all tourists is contained in the set of all humans.

[1]Researchers are often in effect equating the both: "...standard logic, which mental logic and mental models assume to be normative..." (Oaksford & Chater, 2001, 349).

As logics tend to do, this has some normative import: a logic often defines some inferences as *good* (see the examples above) and others implicitly as *bad*. We claim that it is possible to find a measure of *complexity* in a natural logic that aligns with the variation in *cognitive difficulty* that is observed in experiments, operationalized as mean success rate, thereby carrying our logic beyond notions of good and bad.

Reasoning with Quantifiers

For one who is not familiar with their centuries-old notation of Syllogisms, it must seem extremely cumbersome. We will thus stick to examples and keep our treatment of syllogisms as short as possible, though a proper introduction can be found in Khemlani and Johnson-Laird (2012). The syllogistic fragment is a natural language fragment that builds around inferences using the quantifiers EVERY, SOME, NO, and NOT ALL. Syllogisms consist of three quantified sentences (two premises and one conclusion) and three variables A, B, and C. Somewhat surprisingly, psychological studies that are concerned with human reasoning using quantifiers are more often than not restricted to syllogistic reasoning (results are assembled, for instance, in a meta-study by Chater and Oaksford (1999)). For $i \in \{1,2,3\}$, let Q_i be any of the four quantifiers above. We consider four variable configurations:

$$\frac{Q_1(B,C) \quad Q_2(A,B)}{Q_3(A,C)} \qquad \frac{Q_1(C,B) \quad Q_2(A,B)}{Q_3(A,C)} \qquad \frac{Q_1(B,C) \quad Q_2(B,A)}{Q_3(A,C)} \qquad \frac{Q_1(C,B) \quad Q_2(B,A)}{Q_3(A,C)}$$

All combinations of quantifiers and variables considered, we end up with 256 possible syllogisms, most of which are *not* good inferences in any sense of the word, where the definition of *good* varies across different cognitive models. Khemlani and Johnson-Laird (2012) however note, that there are 512 syllogisms, if one allows conclusions of the form $Q_3(C,A)$ (as it was done in scholastic logic). Case in point are the two syllogisms below:

$$\frac{\text{ALL}(C,B) \quad \text{ALL}(B,A)}{\text{SOME}(A,C)} \qquad \frac{\text{ALL}(C,B) \quad \text{ALL}(B,A)}{\text{ALL}(C,A)}$$

The "Aristotelian" one on the left-hand side with its restriction on the form of the conclusion is endorsed much less by participants in experiments than its counterpart on the right (Khemlani & Johnson-Laird, 2012) – a fact that our model will later offer an explanatory account for. Syllogistic reasoning patterns are readily extended to reasoning with other quantifiers, such as MOST, that are beyond first order logic. The results of Chater & Oaksford's meta-study unsurprisingly show that some *good* inferences are easier than others. We will henceforth refer to this as the *cognitive difficulty* of an inference (mean success rate in experimental settings). But syllogisms is not all there is to reasoning with quantifiers: Geurts and van der Silk (2005) did an experiment on reasoning with iterated quantifiers investigating how their combined monotonicity properties interact with the cognitive difficulty of inferences. Participants in their study had to determine whether reasoning patterns of the form

$$\frac{\text{Q}_A \text{ A played against Q}_B \text{ B.} \quad \text{All B were C. / All C were B.}}{\text{Q}_A \text{ A played against Q}_B \text{ C.}}$$

were valid or not with $Q_A \in \{$EVERY, MOST, AT LEAST, SOME, AT MOST, NO$\}$ and $Q_B \in \{$MORE THAN, FEWER THAN$\}$ and only one of the two possibilities of the minor premise (second line) present. These inferences are exclusively concerned with the monotonicity-properties of the second argument.

Monotonicity and Symmetry

We have seen examples of monotonicity above. While quantifiers can be increasing in one of the arguments and being decreasing in the other, some, like MOST and TWO, do not show monotonicity properties on either side. To capture this variation, we will henceforth talk about a quantifier's *monotonicity profile*. As an example, instead of stating that Q is left-side downward monotone and right-side upward monotone, we will say that its monotonicity profile is $\downarrow\uparrow$ or write $\downarrow Q \uparrow$. The lack of monotonicity-properties on either side will be indicated by a dot. Examples are $\downarrow$ ALL $\uparrow$, $\uparrow$ SOME $\uparrow$, $\downarrow$ NO $\downarrow$, $\uparrow$ NOT ALL $\downarrow$, $\cdot$MOST $\uparrow$.

Another property of quantifiers that allows for inferences is *symmetry*. A quantifier Q is called *sym-*

metric if and only if, for all A and B, $Q(A,B)$ implies $Q(B,A)$. The inference associated with this property is called *conversion*: if SOME pigeons are birds, then SOME birds are pigeons - whereas the same inference is clearly not good for the quantifier ALL. Let us now quickly look at iterated quantifiers. We can extend the notion above to combinatorial monotonicity profiles (CMP). So, for example, $\downarrow Q_1, Q_2 \uparrow$ means that the iteration of Q_1 and Q_2 puts their first argument in a downward entailing position and their second argument in an upward entailing position. The interaction between the monotonicity properties of single quantifiers is thus reminiscent of how subtraction and addition interact in arithmetics[2] Thus, if Q_1 is right-side downward entailing, this reverses the direction of entailment of the second quantifier - a downward entailing first quantifier switches the directionality of the second.

Reasoning with Quantifiers

It is now time to present our natural logic for reasoning with quantifiers. The following inference rules allow for proving all syllogisms that are valid in predicate calculus and / or Aristotelian logic, see Sippel (2017) for details.

$$\text{Mon}\uparrow \quad \frac{\mathrm{Q}\uparrow(A,B) \quad \textsc{all}(B,C)}{\mathrm{Q}\uparrow(A,C)} \qquad \text{Mon}\downarrow \quad \frac{\mathrm{Q}\downarrow(A,B) \quad \textsc{all}(C,B)}{\mathrm{Q}\downarrow(A,C)}$$

$$\uparrow\text{Mon} \quad \frac{\uparrow \mathrm{Q}(A,B) \quad \textsc{all}(A,C)}{\uparrow \mathrm{Q}(C,B)} \qquad \downarrow\text{Mon} \quad \frac{\downarrow \mathrm{Q}(A,B) \quad \textsc{all}(C,A)}{\downarrow \mathrm{Q}(C,B)}$$

$$\text{Conv} \quad \frac{\mathrm{Q}_s(A,B)}{\mathrm{Q}_s(B,A)} \qquad \text{pConv} \quad \frac{\textsc{no}(A,B)}{\textsc{all not}(A,B)}$$

$$\text{exImp} \quad \frac{\textsc{all}(A,B)}{\textsc{some}(A,B)}$$

Where Q_s denotes any symmetric quantifier (NO, SOME, all cardinal quantifiers, etc.) and all quantifiers have the indicated monotonicity properties. With the generality in quantifier assignment and the rules ↓Mon and ↑Mon, this already extends beyond Geurts' model. The following inference rules account for the reasoning task with iterated quantifiers in Geurts and van der Silk (2005).

$$\text{Mon}\uparrow\uparrow \quad \frac{Q\uparrow Q_M \uparrow \phi(A,B) \quad \textsc{all}(B,C)}{Q\uparrow Q_M \uparrow \phi(A,C)}$$

$$\text{Mon}\uparrow\downarrow \quad \frac{Q\uparrow Q_F \downarrow \phi(A,B) \quad \textsc{all}(C,B)}{Q\uparrow Q_F \downarrow \phi(A,C)}$$

$$\text{Mon}\downarrow\uparrow \quad \frac{Q\downarrow Q_M \uparrow \phi(A,B) \quad \textsc{all}(C,B)}{Q\downarrow Q_M \uparrow \phi(A,C)}$$

$$\text{Mon}\downarrow\downarrow \quad \frac{Q\downarrow Q_F \downarrow \phi(A,B) \quad \textsc{all}(B,C)}{Q\downarrow Q_F \downarrow \phi(A,C)}$$

Where Q is any binary quantifier with the indicated monotonicity properties, and Q_M and Q_F are MORE THAN and FEWER THAN, respectively. We can naturally extend this to account for left-side monotonicity inferences, whose directionality only depends on the first quantifier (Sippel, 2017).

Complexity

As for our complexity measure, there are three crucial ideas we borrow from Geurts (2003): Firstly, the number of reasoning steps from premises to a conclusion is the length of its minimal proof in a natural logic (i.e. how much solving a problem as efficiently as possible costs in our natural logic). Secondly, some reasoning steps are harder than others. Thirdly, we can account for this variation in difficulty by assigning a cost to inference rules, summing up to different overall costs for different proofs. As far as possible, we will motivate weights on semantic grounds.

The four combinations of right-side monotonicity properties for iterated quantifiers are ↑↑, ↑↓, ↓↑ and ↓↓. We immediately see how many of the quantifiers involved "go up" and whether both have the same directionality. Similar as Geurts (2003), we propose a cost-based system in which less favorable inferential

[2] In Sippel (2017), we show that this scheme is restricted to the quantifiers used in Geurts and van der Silk (2005). While all iterated quantifiers have clear monotonicity properties, they do just not generally follow this simple interaction system.

properties add to the cognitive cost of a reasoning task: firstly, upward is easier than downward (Clark, 1974). Secondly, inferences are harder when the they do not have the same directionality - call this *harmony*, or rather its absence (Geurts & van der Silk, 2005, 104). Thirdly, if the first quantifier is downward entailing, it turns the entailment direction of the second quantifier upside down, hence requiring additional processing. Furthermore, if an iterated quantifier is negative (e.g. *not all*), the associated monotonicity-inference is harder. This is embedded in a rich research-tradition showing that negation is harder to process (e.g. Wason (1961)).

From a semantic point of view, ALL and NO are the most *informative*. They are somewhat on top of a monotonicity-based semantic food chain and allow for inferences that less informative ones do not:

$$\text{ALL}(A,B) \Rightarrow Q\uparrow(A,B)$$
$$\text{NO}(A,B) \Rightarrow Q\downarrow(A,B)$$

For any quantifier with the right monotonicity properties. Let us now turn to symmetry and its associated inference of *conversion*. Geurts (2003) claims that while pConv has small cognitive cost, Conv itself has none. This was criticized by Newstead (2003) who is especially reluctant to accept the low cognitive cost of pConv. In absence of any empirical evidence on this (Newstead, 2003, 195), we will settle somewhat on the middle ground: Conv is *not* without any, but with very small cognitive cost, so is pConv.

The assumption that statements using ALL refer to non-empty sets allows for the inference exImp and is usually taken for granted in experimental design but unpopular with some researchers (Chater & Oaksford, 1999). The work of Katsos, Cummins, et al. (2016), while concerned with quantifier acquisition, offers some important insight on why that might be. As part of their study, they investigate how adults deal with underinformative quantifiers. In 84% of all cases were the statement was true but underinformative, the statement was rejected by the participants (Katsos et al., 2016, 9246). The weights for inferences on single quantifiers that we propose according to our considerations on cognitive difficulty above are as in table 1.[3] The weights in table 2 are directly derived from those in table 1 using the considerations on monotonicity interactions that make inferences easy or difficult above. Where there are two values, the first one holds when the first quantifier is NO or ALL and the second one if not (this reflects above fact that inferences on informative quantifiers are easier). The weights are interpreted as cost that is subtracted from an initial "cognitive reservoir" of 100 units (as done in Geurts (2003)) - this move does in fact not impact the results of our statistical analysis but allows for better readability and stronger hypotheses: where we before could only observe correlations between weights and cognitive difficulties, we can now make *predictions*. Tables 3 and 4 provide with two examples of proofs that show how this works.

[3]Zhai, Szymanik, and Titov (2015) had the weights for a similar logic learned from data.

Table 1: Weights for the inference rules used on the syllogistic fragment. Mon stands for all monotonicity inferences and Mon_N for monotonicity inferences involving NOT ALL or ALL NOT.

exImp	Mon_N	Mon	pConv	Conv
60	30	10	5	5

Table 2: Weights for the inference rules on iterated quantifiers. All numbers are rounded up.

Mon↑↑	Mon↑↓	Mon↓↑	Mon↓↓
0/15	23/38	38/53	30/45

Evaluation of the Model

We prove all syllogisms that are valid in Aristotelian logic or predicate calculus (or both) and compute the model's predictions for all of them, see details in Sippel (2017). The results, i.e. the model's predictions for valid syllogisms, can be seen in table 5 – we

Table 3: Complexity = $Conv + Mon = 15$, thus predicted mean success Success = $100 - 15 = 85$ (actual mean success rate in experiments: 89%).

[1]	ALL(M,P)	*premiss*
[2]	SOME(M,S)	*premiss*
[3]	SOME(S,M)	*Conv on* [2]
[4]	SOME(S,P)	↑*Mon on* [1] *and* [3]

Table 4: Complexity = $pConv + Mon_N + exImp = 95$, this predicted mean success Success = $100 - 95 = 5$ (actual mean success rate in experiments: 1%).

[1]	ALL(P,M)	*premiss*
[2]	NO(S,M)	*premiss*
[3]	ALL NOT(S,M)	*pConv on* [2]
[4]	ALL NOT(S,P)	*Mon↓ on* [4] *and* [1]
[5]	SOME NOT(S,P)	*exImp on* [4]

Table 5: Comparison of the cognitive difficulty of valid syllogisms (Chater & Oaksford, 1999) in brackets, our model, #2, and Geurts' model (Geurts, 2003), #3. The codes of three letters and one number denote syllogisms as it is usually done in the literature, see e.g. Khemlani and Johnson-Laird (2012). Predictions that are more than 10% off from mean success rates in experiments are marked as gray.

Syll	#1	#2	#3	Syll	#1	#2	#3
AI1I	(92)	90	80	EI2O	(52)	60	60
IA4I	(91)	85	80	EI3O	(48)	60	60
AA1A	(90)	90	80	AA3I	(29)	30	60
AI3I	(89)	85	80	EI4O	(27)	55	60
EA2E	(89)	85	80	EA3O	(22)	5	40
AE2E	(88)	90	80	AA4I	(16)	25	60
EA1E	(87)	90	80	EA4O	(8)	0	40
AE4E	(87)	85	80	AA1I	(5)	30	60
IA3I	(85)	90	80	EA1O	(3)	5	40
OA3O	(69)	70	70	EA2O	(3)	0	40
AO2O	(67)	70	70	AE4O	(2)	0	40
EI1O	(66)	65	60	AE2O	(1)	5	40

obtain $r^2 = 0.93$ and Pearson r= 0.96. The model thus already outperforms that of Geurts (2003), the increased performance is however best visible by the fact that predictions that are more than 10% off on those 24 syllogisms went down from 13 to 4. An interesting direction for further research would be to investigate whether this increased performance is due to more adequate weights or rather to the inclusion of left-side monotonicity rules. The evaluation of the model on inferences with iterated quantifiers can be seen in table 6: we obtain $r^2 = 0.88$ and Pearson r= 0.94. The proposed natural logic is thus well capable of capturing the general trends and predicts much of the variance in the empirical data.[4]

Table 6: Comparison of the cognitive difficulty of reasoning with iterated quantifiers (Geurts and van der Silk (2005), in brackets) and our model's predictions.

Det_A	Det_B	Minor	%	Model
AT LEAST↑	MORE THAN↑	ALL(B,C)	(96)	85
EVERY↑	MORE THAN↑	ALL(B,C)	(91)	100
MOST↑	MORE THAN↑	ALL(B,C)	(91)	85
SOME↑	MORE THAN↑	ALL(B,C)	(87)	85
NO↓	FEWER THAN↓	ALL(B,C)	(73)	70
EVERY↑	FEWER THAN↓	ALL(C,B)	(71)	77
MOST↑	FEWER THAN↓	ALL(C,B)	(62)	62
SOME↑	FEWER THAN↓	ALL(C,B)	(60)	62
NO↓	MORE THAN↑	ALL(C,B)	(53)	62
AT LEAST↑	FEWER THAN↓	ALL(C,B)	(53)	62
AT MOST↓	MORE THAN↑	ALL(C,B)	(38)	47
AT MOST↓	FEWER THAN↓	ALL(B,C)	(36)	55

Conclusions & Predictions

We have successfully refined the model in Geurts (2003), especially its complexity measure, and extended it to reasoning with iterated quantifiers. Performance on the data collected by Geurts and van der Silk (2005) also indicates good predictive capacities for reasoning with iterated quantifiers – but while the logic and its complexity measure is grounded in semantic relationships and psychological evidence, it might also seem somewhat post hoc – there is barely *direct* evidence accounting for the weight-assignments but mostly *related* evidence. Luckily, the model allows for empirically testable predictions, e.g. that reasoners accept exImp-inferences while being reluctant to draw them themselves (a hypothesis, which would, in fact, explain reasoner's preference for the right-hand side syllogism over the left-hand side one above), that left-side monotonicity inferences are not harder that their right-side counterparts (in fact, on iterated quantifiers, they should be easier as they involve no change of directionality), and that the model can be extended to account for further experiments on reasoning with quanti-

[4]As for the limited space, we have not yet talked about competing cognitive models for syllogistic reasoning. In Sippel (2017), we show that the main competing models by Geurts (2003), Chater and Oaksford (1999) and Johnson-Laird and Bara (1984) all give raise to identical categories of cognitively difficult syllogisms. A decision for one model over its competitors can thus not be made on grounds of good fit with empirical data. Geurts' model – as we have shown throughout this work – however gives us a strong, flexible measure of complexity and is the only one that can be easily extended to reasoning with iterated quantifiers.

fiers. With the pluralistic view of logic in mind, we can thus conclude that it is worthwhile to reevaluate logic's possible contributions to the science of reasoning. Our model indicates that a natural logic can in fact predict human performance for more complex symbolic reasoning patterns. It confirms that, as proposed, e.g., by Isaac, Szymanik, and Verbrugge (2014); Szymanik (2016), logic can contribute to our understanding why some cognitive tasks are easier than others.

Acknowledgments

Jakub Szymanik has received funding from the European Research Council under the European Unions Seventh Framework Programme (FP/20072013)/ERC Grant Agreement n. STG 716230 CoSaQ. The paper is based on the master thesis of the first author Sippel (2017).

References

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *22*, 97–117.

Clark, H. (1974). Semantics and comprehension. *Current Trends in Linguistics*, *12*, 1291–1428.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*, 223–251.

Geurts, B., & van der Silk, F. (2005). Monotonicity and processing load. *Journal of Semantics*, *38*, 191–258.

Icard III, T., & Moss, L. (2014). Recent progress in monotonicity. *LiLT (Linguistic Issues in Language Technology)*, *9*.

Isaac, A. M. C., Szymanik, J., & Verbrugge, R. (2014). Logic and complexity in cognitive science. In A. Baltag & S. Smets (Eds.), *Johan van Benthem on Logic and Information Dynamics, Outstanding Contributions to Logic 5.* Switzerland: Springer International Publishing.

Johnson-Laird, P., & Bara, B. (1984). Syllogistic inference. *Cognition*, *16*, 1–61.

Katsos, N., Cummins, C., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*, 9244–9249.

Khemlani, S., & Johnson-Laird, P. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*.

Newstead, S. (2003). Can natural language semantics explain syllogistic reasoning? *Cognition*, *90*, 193–199.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *TRENDS in Cognitive Science*, *5*.

Sippel, J. (2017). The good, the bad, and the difficult: Complexity in a monotonicity-grounded natural logic for reasoning with generalized quantifiers. `https://www.illc.uva.nl/Research/Publications/Reports/MoL-2017-25.text.pdf`.

Stenning, K., & van Lambalgen, M. (2012). *Human reasoning and cognitive science*. Cambridge: MIT Press.

Szymanik, J. (2016). *Quantifiers and cognition. logical and computational perspectives*. Springer.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, *52*.

Wason, P. C. (1983). Realism and rationality in the selection task. In S. B. T. Evans (Ed.), *Thinking and reasoning.* Routledge.

Zhai, F., Szymanik, J., & Titov, I. (2015). Toward probabilistic natural logic for syllogistic reasoning. In T. Brochhagen, F. Roelofsen, & N. Theiler (Eds.), *Proceedings of the 20th Amsterdam Colloquium* (pp. 468–477). Amsterdam.