university of groningen / faculty of science and engineering

The effect of quantifier entailment on sentence verification

Bachelor's Project Thesis

Jordy de Lange, s3171205, j.r.de.lange@student.rug.nl,
Supervisor: dr. J.K. Spenader

There are many theories that try to explain how syllogistic reasoning works in our brain. These are usually based on predicate logic, mental-model theories or a heuristic theory. Bart Geurts (2003) points out several problems with each method and proposes his own theory based on concepts from natural language semantics. In this theory he claims that sentences with only upward entailing quantifiers are easier than those that also include downward entailing quantifiers. In this study we aim to test if this has an effect on the verification times of sentences. We tested this through an experiment where participants need to verify if a sentence is true or false for a picture. The verification time and accuracy was then compared between the two types of quantifiers. There was no significant difference observed, which suggests that quantifier entailment does not influence the way we verify sentences.

1 Introduction

Deductive reasoning is an essential skill for us humans. A lot of research has therefore been done to try and explain the internal processes used when reasoning. A type of reasoning that has been studied extensively is syllogistic reasoning. This type of reasoning involves a conclusion being drawn from two given premisses. An easy example would be the following premisses and their conclusion:

Premiss 1:	All surgeons are golfers
Premiss 2:	All golfers are club members
Conclusion:	All surgeons are club members

We as humans are actually rather good at this type of reasoning. Correct lines of reasoning are often recognized as such, and most mistakes that are made are on syllogisms that are very similar to valid syllogisms (Geurts 2003).

1.1 Theories of reasoning

A lot of theories have been developed over the years to try and explain how we arrive at this conclusion. Some theories are based on predicate logic (Geurts 2003). In these theories, it is usually assumed that we as humans have some sort of internal natural deduction, allowing us to apply inference rules to arrive at a conclusion. This makes sense for the example given above, which can be represented in predicate logic quite easily:

Premiss 1:	$\forall x[surgeon(x) \wedge golfer(x)]$
Premiss 2:	$\forall x[golfer(x) \wedge member(x)]$
Conclusion:	$\forall x[surgeon(x) \wedge member(x)]$

One of the problems with this theory is that there is no good way to represent the quantifier "most" and similar quantifiers, like "less than half", in predicate logic. This in and of itself is not a big problem, were it not that we know from experimental evidence that we can reason with the word "most" as easily as we can with "all" (Oaksford and Chater 2001). This is best illustrated by changing one of the premisses:

Premiss 1:	Most surgeons are golfers
Premiss 2:	All golfers are club members
Conclusion:	Most surgeons are club members

This line of reasoning would be impossible to express in predicate logic, but we are just as fast in coming up with the conclusion here as we are in the example where all the quantifiers were 'all'.

Additional problems arise when reasoning with quantifiers like "at least one". To represent this in predicate logic one would need variables. Translated into predicate logic, it would look like this:

While one is a manageable number to keep track

Premiss 1:	At least one surgeon is a golfer
Premiss 2:	All golfers are club members
Conclusion:	At least one surgeon is a club member

Premiss 1:	$\exists x[surgeon(x) \wedge golfer(x)]$
Premiss 2:	$\forall x[golfer(x) \wedge member(x)]$
Conclusion:	$\exists x[surgeon(x) \wedge member(x)]$

of in our head, we can also reason perfectly fine with "more than a thousand". If we were utilizing some predicate logic for syllogistic reasoning, we would need to keep track of a thousand separate individual variables to represent the expression, which is obviously not feasible.

Another theory that tries to tackle the problem of how we reason is based on mental models (Bucciarelli and Johnson-Laird 1999). These models predict that we keep track of the information in a premiss by representing it in terms of individuals. This theory quickly runs into one of the same problems logic based theories run into. It can not efficiently represent quantifiers like "at least three". If we were creating a mental model of the situation described by the premiss, a premiss like "at least a thousand surgeons are golfers" would have us imagine a thousand individual surgeons that are also golfers. As said before, this is simply not feasible.

A different approach is taken by models based on probability (Oaksford and Chater 2001). The reasoning behind this is that we as humans are adapted to reason with a certain degree of uncertainty. This leads to the premiss "all surgeons are golfers" having the representation that each surgeon has a 100% probability of being a golfer. This theory is well suited for handling "most". "Most surgeons are golfers" will simply be represented as each surgeon has a probability of more than 50% to be a golfer. It does run into problems with quantifiers that call for a specific amount. In order to represent "exactly 6 surgeons are golfers" one would need to know the cardinality of the set of all surgeons. Say there are 10 surgeons; this would lead to the premiss having the representation that there is a 60% chance of a surgeon being a golfer. Now, if there were 20 surgeons, suddenly the representation would need to change to each surgeon having a 30% chance of being a golfer. Further complicating the case for probabilistic reasoning is that we as humans do not need to know the cardinality of a set in order to reason about it in terms of exact numbers.

1.2 Entailment and reasoning

So far none of the discussed theories have given a satisfying answer to the question of how we, as humans, reason. In order to try and circumvent the problems other theories are having, Bart Geurts (2003) proposed a theory based on entailment properties of the quantifiers used in syllogistic reasoning.

A quantifier can be upward or downward entailing. An example of an upward entailing quantifier is "some". If it is given that "some surgeons are golfers", one can conclude that "some doctors are golfers" as well. This ability to generalize "surgeons" to its superset "doctors" makes "some" upward entailing. We can not generalize to a subset of "surgeons", as is demonstrated by the fact that the sentence "some brain surgeons are golfers" does not follow from the given premiss. If we take a look at a downward entailing quantifier like "all" however, the situation is reversed. From a given sentence "all surgeons are golfers", it can be concluded that "all brain surgeons are golfers". It however does not lead to the conclusion that "all doctors are golfers". The ability to generalize to a subset of "surgeons" makes "all" downward entailing. Entailment is something that we use without even realizing it. There are studies that show that even very young children are aware of certain entailment properties (O'Leary and Crain 1994).

In their paper from 2005 Bart Geurts and Frans van der Slik predict reasoning with upward entailing quantifiers to be easier than reasoning with downward entailing quantifiers. The theory behind this is as follows. In many languages there exist a lot of word pairs describing the same property of something. Two examples from English would be "big" and "small", which both describe size, and "old" and "young", which both describe age. Since they both describe the same property, it would be reasonable to assume that each word in a pair could be used fairly interchangeably. This however is not the case, as there is a fundamental asymmetry between the words in a pair; one of them carries am implicit judgment about the property it describes. If you were to ask someone how old she is or how tall she is, that would not carry any extra meaning. In contrast, if you ask how young or how short

someone is, you are implying that that person is quite young or quite short. These two words are marked, in the sense that they convey implicit extra meaning. This asymmetry between marked and unmarked words exists across languages and in all cases the unmarked word is used more often and generally preferred.

The fundamental property that differs between marked and unmarked words is that of boundedness. From the earlier examples, "big" and "old" are unbounded. There is no limit to how big or how old something can be. This is not the case for "small" and "young" however; these both do have a limit. Geurts (2003) then compares this to quantifier entailment. Upward entailing quantifiers can keep on generalizing for ever; there is no limit to the size of a set. Downward entailing quantifiers on the other hand run into the hard limit of not being able to generalize further than the empty set. In this way, upward entailment is similar to the unmarked words and downward entailment is similar to the marked words. So if unmarked words are preferred and easier for us, maybe upward entailment is also preferred and easier for us.

Geurts and van der Slik (2005) then go on to show with an experiment that people indeed make fewer mistakes when reasoning with upward entailing quantifiers only than when downward entailing quantifiers are also involved. In this experiment participants were presented with syllogistic lines of reasoning in the form seen in here:

Premiss 1:	X A played against Y B
Premiss 2:	"All B were C" or "All C were B"
Conclusion	X A played against Y C

X and Y would be one of a number of quantifiers and A, B and C would be names describing arbitrary groups of people. The participants were then asked to determine whether the conclusion given is valid when taking the premisses as being true. The results confirmed their suspicions of upward entailing quantifiers being easier, with 79,75% correct when only upward entailing quantifiers are involved and 61,40% correct when there are also downward quantifiers involved.

1.3 Beyond syllogistic reasoning

Given that we prefer upward entailing quantifiers when reasoning, maybe we also prefer it for other tasks. One of these other tasks potentially impacted by quantifier entailment is sentence verification. If reasoning with upward entailing quantifiers is easier, it sounds plausible that the verification of sentences with upward entailing quantifiers is also easier. To determine whether this is the case, we pose the following research question: Are people better at picture-sentence verification tasks that involve the upward entailing quantifiers "Some" and "Only" than those that involve the downward entailing quantifiers "No" and "All"?

1.4 The experiment

In order to answer this question an experiment was performed. In this experiment participants were shown a sentence along with an accompanying picture and they had to determine if the sentence was true given the situation in the picture. Similar sentence verification tasks are sometimes paired with an additional working memory load (Neys and Schaeken 2007). This is done both to possibly discern an influence of working memory capacity on the amount of correct answers or the speed of answering and to ensure that the task is not too easy by limiting the mental resources the participant has access to. In the cited paper 3 by 3 grids that had a dot in 4 of the cells were used. The participant is then asked to remember the configuration of one of these grids while they perform tasks for the experiment. The same technique was also used in our experiment. An example of one of these grids is given in section 2, figure 2.3

Previous papers that also described picture-sentence verification tasks used very abstract stimuli (Zajenkowski, Szymanik, and Garraffa 2013). For example, simple white screens filled with several black or white circles. This has the disadvantage of making sentences like "all circles are black" trivially easy, since it can be seen almost instantly if one of the circles is white in the picture. To prevent this, more realistic pictures were used. Because these pictures have more details they require more of a verification strategy than simply seeing that there are white circles as well, which should make the effect the quantifier has more pronounced.

These more realistic pictures features figurines of pirates and policemen each of which may or may not be holding an object.

Four different quantifiers were used in the sentences, “some”, “only”, “no”, and “all”. The first two of those are upward entailing, the last two are downward entailing. To prevent the verification strategy from influencing the result, not all quantifiers can be applied to the same picture. From the four quantifiers, two pairs were made. In each pair the quantifiers have very similar conditions for when they are true or false. The first pair is “some” and “no”, and the second pair is “only” and “all”.

By using these pairs where the conditions are similar the only difference between the verification of the sentences is the entailment of the quantifier. To illustrate these similar conditions, take a look at this example for “some” and “no”. If you have the sentence “Some/No figurines have a bucket”, you get the truth conditions specified in table 1.1. This table makes it clear that the truth conditions, and therefore the verification strategies, of a true “some” sentence are the same as those of a false “no” sentence and vice versa. The “only” and “all” pair is very similar in this regard, as can be seen in table 1.2. The only complication comes from the fact that for the “All pirates have a bucket” sentence to be true, all pirates need to have a bucket, while the “Only pirates have a bucket” needs no policeman to have a bucket and at least one pirate to have a bucket. This issue is averted by making the picture-sentence combinations in such a way that the situation where the figurine featured in an “only” or “all” sentence does not have the object mentioned in the sentence never occurs.

2 Method

2.1 Participants

33 native English speaking participants completed the experiment. 15 of the participants were women and their mean age was 31. All participants gave informed consent to be a part of the experiment. The participants were recruited through a Prolific, a service where people earn money in reward for participating in experiments. The participants were therefore rewarded with £2.25.

2.2 Materials

The experiment was made using Psytoolkit (Stoet 2010, Stoet 2016), which is a web based experiment development environment that provides free hosting. The stimuli were made with Playmobil figurines and several Playmobil accessories. Perhaps the best way to illustrate the structure of the pictures is with some examples. Figure 2.1 shows a picture that was used in the experiment to explain to the participants what they were going to see.

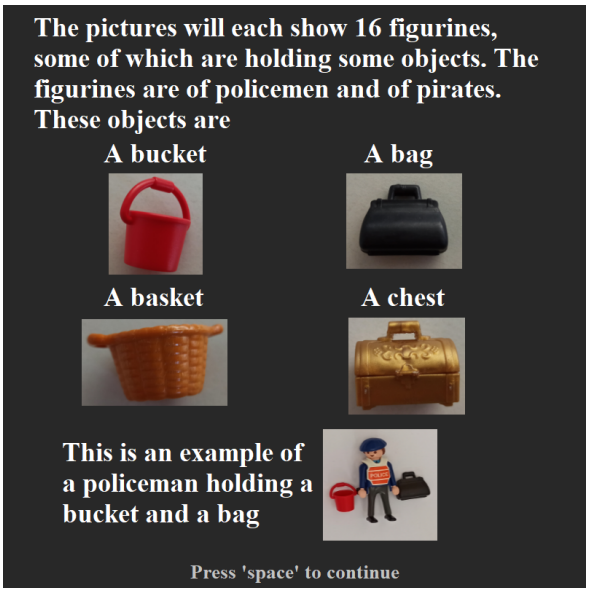

Figure 2.1: An explaining picture used in the experiment

Figure 2.2 shows a picture from the experiment. This specific picture could have appeared with either the sentence “Only policemen have a bag” or the sentence “All policemen have a bag”. In the first case, the correct response is “true”, since there is at least one policeman with a bag and there are no pirates that have a bag. In the second case, the correct response is “false”, since while some policemen have a bag, there are also policemen that do not have one.

2.3 Procedure

Each participant was recruited through the online service Prolific, and gave informed consent prior to starting the experiment. Then, after some

Table 1.1: Truth conditions for "Some" and "No"

Sentence	True when	False when
Some figurines have a bucket	At least one figurine has a bucket	No figurine has a bucket
No figurines have a bucket	No figurine has a bucket	At least one figurine has a bucket

Table 1.2: Truth conditions for "Only" and "All"

Sentence	True when	False when
Only pirates have a bucket	No policemen have a bucket	At least one policeman does not have a bucket
All pirates have a bucket	All pirates have a bucket	At least one pirate does not have a bucket

Figure 2.2: An example picture from the experiment

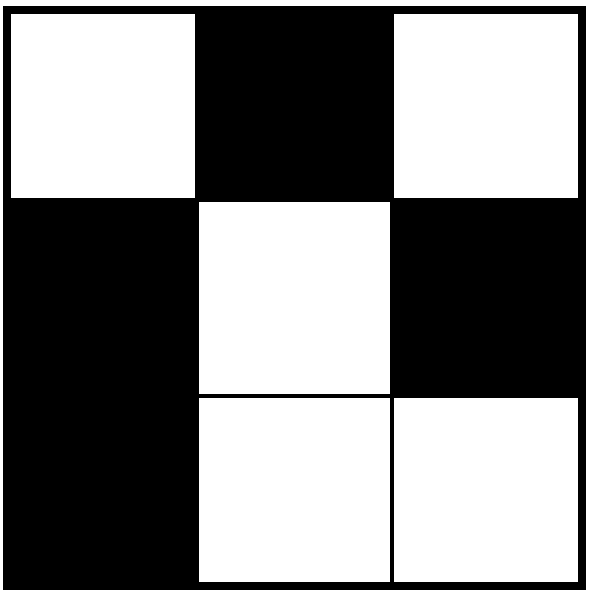

Figure 2.3: An example working memory grid from the experiment

explanation about what to expect, the practice block begins. This consists of four sentence verification tasks with sentences in the form "All pirates/policemen have a chest/basket/bucket/bag and the policemen/pirates do not". This is a sentence type that is not featured in the rest of the experiment.

For each sentence verification task a participant is first told to remember the grid that is going to flash on screen. The grid then flashes on screen for 850 milliseconds. One of these grids is shown in figure 2.3. Then the sentence appears, and the participant has two seconds to read it before the picture appears. At this point the participant needs to press a key depending on whether he/she believes the sentence to be true or false within 15 seconds. Lastly, nine grids are show on screen and the participant needs to select the one they were shown before the task. This is shown in figure 2.4. Everything then repeats itself for the next task.

When the practice block has been completed the two experiment blocks are done, each of which consists of 24 trials. Between these two blocks the participant is given an opportunity to take a break. After completing all items the participant is sent back to Prolific to receive payment.

2.4 Design

There are 48 experiment trials in total. This is made up of 24 pictures for the "some" and "no" pair and 24 pictures for the "only" and "all" pair. Within each pair each quantifier receives 12 pictures. These 48 trials are then randomly divided

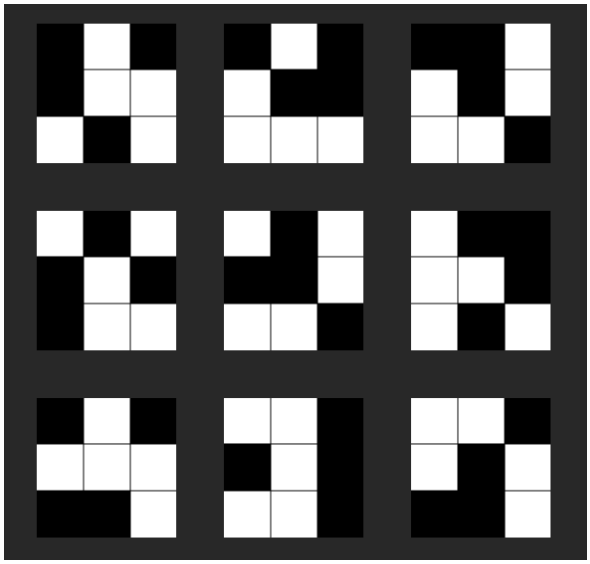

Figure 2.4: An example working memory grid selection screen from the experiment

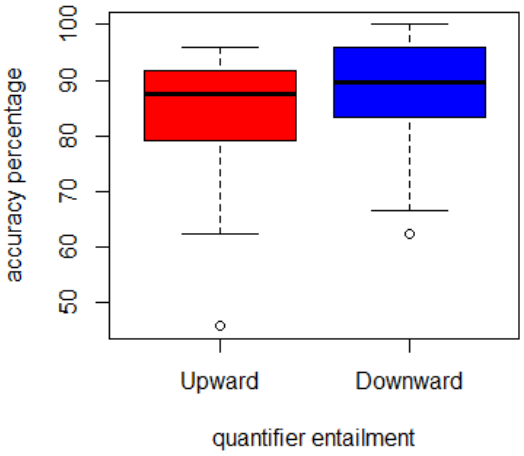

Figure 3.1: Boxplot of the accuracy per entailment ("some" and "only" are upward entailing, "no" and "all" are downward entailing)

into two groups of 24 trials; the first and second block. In between these two blocks is a break.

The measured variables were whether the answer to the verification task was correct, the speed of the verification, and whether the answer to the working memory grid was correct.

3 Results

Three of the participants were removed from the dataset because it was clear they were not performed correctly. Two participants rapidly clicking random answers, and the third participant did not do anything and simply timed out each trial. The whiskers boxplots used to illustrate the data are drawn to the datapoint furthest from the mean but not further than 1.5 times the interquartile range.

3.1 Accuracy of verification

The first result we are going to take a look at is that of accuracy per entailment. The performance of each participant was taken as a data point and plotted as a boxplot in figure 3.1. A generalized linear mixed effect model was used that predicted the correctness of the task using the fixed effect of entailment type and the random effects of participant ID and picture ID. This model was compared using ANOVA to a simpler model that did not use the fixed effect. This analysis showed that there is no significant difference between the two models (P = 0.292), meaning that the entailment type is not a significant factor in determining accuracy.

If we look at the accuracy for each quantifier separately, we get the boxplot in figure 3.2. An analysis using a general linear mixed effect model predicting the correctness of the task using the fixed effect of the quantifier and the random effects of participant ID and picture ID was compared using ANOVA to a similar model without the fixed effect. This analysis showed no significant difference between the two models (P = 0.087)

3.2 Verification time

For the analyses using the verification time the time was first log-tranformed. In the boxplot in figure 3.3 the verification time per entailment is shown. A linear mixed effect model that predicted the verification time using the fixed effect of entailment and the random effects of participant ID and picture ID was compared using ANOVA to a similar model, but without the fixed effect. The analysis

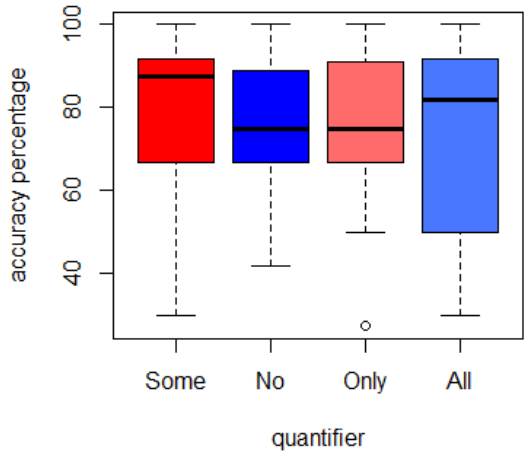

Figure 3.2: Boxplot of the accuracy per quantifier

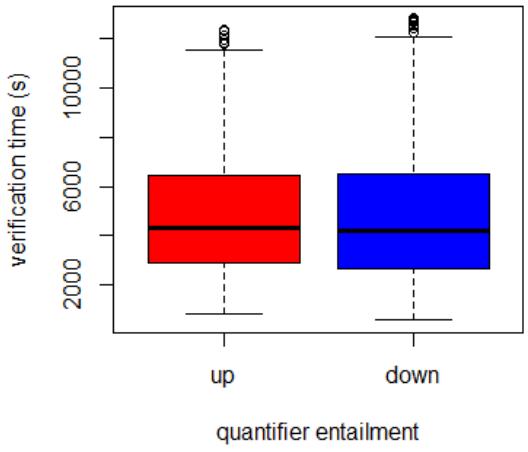

Figure 3.3: Boxplot of the verification times per entailment ("some" and "only" are upward entailing, "no" and "all" are downward entailing)

showed that the entailment type does not significantly affect the verification time (P = 0.683).

As can be seen in figure 3.4, when looking at the differences between individual quantifiers the

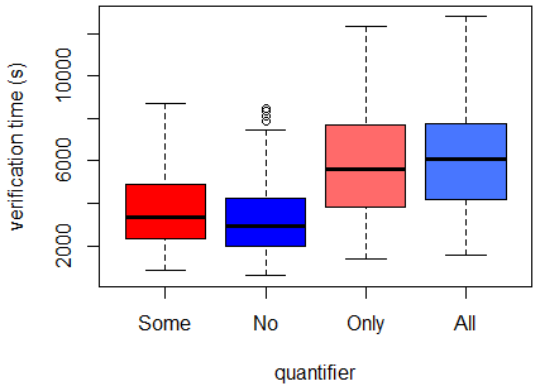

Figure 3.4: Boxplot of the verification times per quantifier

quantifier pairings are nicely reflected in the reaction times. "Some" and "no" are fairly equal, as are "only" and "all". Within each pairing there does not seem to be very big differences though. The models used in the following analysis are identical to the one described previously, except for their fixed effect.

First, the effect of the quantifier overall. A linear mixed effect model was used to predict the verification time with the fixed effect of the quantifier and the random effect of participant ID and picture ID. A comparison using ANOVA with another model with the same random effects but without the fixed effect suggests that the quantifier plays a role in verification time ($P < 0.001$).

3.3 Accuracy of grid recall

The last measurement is the accuracy of the grid recall after each task. This data is shown in the boxplot in figure 3.5. For the analysis a categorical linear mixed effect model was used. It predicted the correctness of the grid recall using the fixed effect of entailment and the random effects of participant ID and picture ID. A comparison with a simpler model that did not have the fixed effect showed that the difference between the models was significant ($P \leq 0.030$), with upward entailing quantifiers having a higher accuracy percentage. This model is shown in Appendix A, Table A.1.

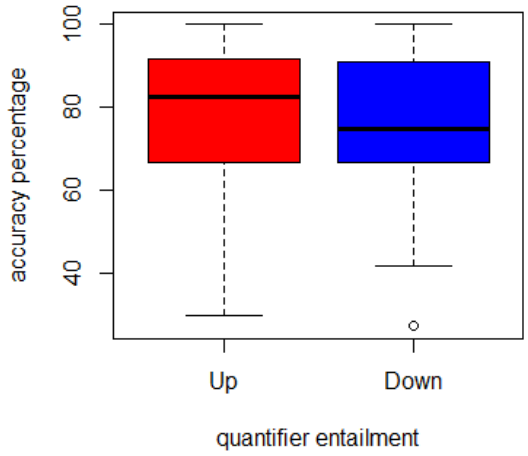

Figure 3.5: Boxplot of the grid accuracy per entailment ("some" and "only" are upward entailing, "no" and "all" are downward entailing)

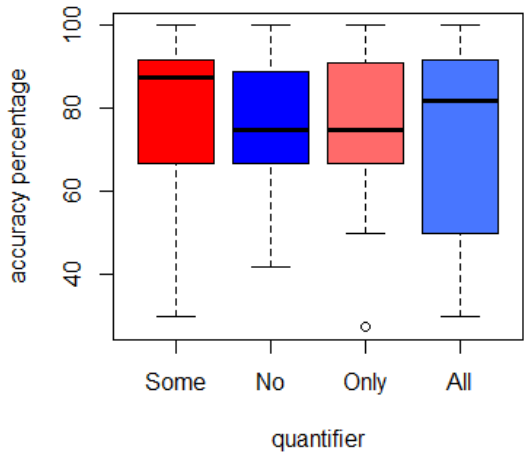

Figure 3.6: Boxplot of the grid accuracy per quantifier

Figure 3.6 shows the grid recall percentages per quantifier, which were analysed using an ANOVA comparison between a categorical linear mixed effect model predicting the correctness of the grid recall using the fixed effect of entailment and the random effects of participant ID and picture ID and a similar model without the fixed effect. This analysis showed that the quantifier had a significant effect on the correctness of the grid recall ($P \leq 0.036$). The quantifier "some" was taken as a reference level. The estimate of the intercept was significantly different for "all" ($\beta = 0.549$, z value = 2.812, $p \leq 0.005$) and for "no" ($\beta = 0.423$, z value = 2.146, $p \leq 0.032$). This model is presented in Appendix A, Table A.2. When choosing "no" as a reference level, the only significantly different interval is "some" (β = -0.423, z value = -2.146, $p \leq 0.032$).

4 Discussion

The results in section 3.3 show that it is easier to remember the working memory grid when then upward entailing quantifiers are used. This could mean that while we don't verify upward or downward entailing quantifiers at different speeds or with different accuracies, they take less resources to verify. This however is not a definite proof, since it could very well be that this difference between the entailment types is solely caused by the fact that "some" has a very high accuracy percentage, as can be seen in figure 3.6.

One of the things that might have given rise to some issues is the difficulty of our task. The aim was to make it more realistic and more difficult than comparable experiments in the past. This was meant to enlarge the effect that the quantifier would have. Maybe we went a bit overboard and made the pictures too complex. This would mean that the participant spend a large part of the reaction time simply looking at the picture. The amount of time that an easy/difficult quantifier may save/cost is then very small compared to the search time. This would obscure the effect of entailment.

As mentioned in the conclusion, it could also be the case that picture-sentence verification simply uses different internal processes than syllogistic reasoning. If that is true, then it makes sense that Geurts' theory does not apply in picture-sentence verification tasks.

Finally, due to the Covid-19 epidemic the experiment had to be performed online. This took all

control of the environment out of our hands. It also made it difficult to ensure that every participant understood everything correctly and completed the tasks seriously, since we could not be present when the participant took the experiment.

In the future it would be useful to look into the processes used for picture-sentence verification, and if they are the same as those used in syllogistic reasoning. If this is not the case, then a study similar to this one but with a task more closely related to syllogistic reasoning may be useful in determining if entailing have an effect on verification. One possibility is to not rely on pictures to provide the situation the sentence is applied to, but to instead use the participants knowledge. Sentences like "All pidgeons are birds" and "Some insects can fly" could be used here. This however can be difficult, since it is hard to control the knowledge of the participants.

5 Conclusion

None of the analyses that say anything about the research question show a significant result that agrees with our prediction that upward entailing quantifiers make for easier sentence verification than downward entailing quantifiers. This suggests that entailment is not a significant factor in the difficulty of sentence verification.

The only statistic that had participants perform better with the upward entailing quantifiers was the grid recall accuracy percentage. It however is a stretch to take this single statistic and use it to conclude that upward quantifiers are easier to verify.

This does not also mean that Geurts is wrong. It could be that the processes involved with sentence verification are simply different from those used in syllogistic reasoning, or that there were problems with our approach to measuring the impact of entailment on sentence verification difficulty.

Another possibility is that the mechanisms behind Geurts' results involves the interplay between upward en downward entailing quantifiers. In our experiment there was only ever one quantifier at a time, so there was no opportunity for interplay between two quantifiers of the same or different entailment.

Finally, it could be the case that for some reason the quantifiers that were chosen were not good choices. Maybe, with different quantifiers, a clear difference between the upward and downward entailing quantifiers would have arisen.

References

Monica Bucciarelli and P.N. Johnson-Laird. Strategies in syllogistic reasoning. *Cognitive Science*, 23(3):247–303, July 1999. doi: 10.1207/s15516709cog23031.

Bart Geurts. Reasoning with quantifiers. *Cognition*, 86(3):223–251, January 2003. doi: 10.1016/s0010-0277(02)00180-4.

Wim De Neys and Walter Schaeken. When people are more logical under cognitive load. *Experimental Psychology*, 54(2):128–133, January 2007. doi: 10.1027/1618-3169.54.2.128.

Mike Oaksford and Nick Chater. The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, 5(8):349–357, August 2001. doi: 10.1016/s1364-6613(00)01699-5.

C. O'Leary and S. Crain. Negative polarity items (a positive result), positive polarity items (a negative result). *Paper presented at the 19th Boston University Conference on Language Development*, 1994.

Gijsbert Stoet. PsyToolkit: A software package for programming psychological experiments using linux. *Behavior Research Methods*, 42(4):1096–1104, November 2010. doi: 10.3758/brm.42.4.1096.

Gijsbert Stoet. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31, November 2016. doi: 10.1177/0098628316677643. URL `https://doi.org/10.1177/0098628316677643`.

Marcin Zajenkowski, Jakub Szymanik, and Maria Garraffa. Working memory mechanism in proportional quantifier verification. *Journal of Psycholinguistic Research*, 43(6):839–853, December 2013. doi: 10.1007/s10936-013-9281-3. URL `https://doi.org/10.1007/s10936-013-9281-3`.

A Appendix

Table A.1: Model = GridCorrect ~ Entailment + (1|ParticipantID) + (1|Picnum)

Predictor	Estimate	Standard Error	z-value	p-value
(Intercept)	-1.3746	0.2520	-5.455	<0.001 ***
EntailmentDown	0.2977	0.1357	2.193	0.0283 *

Table A.2: Model = GridCorrect ~ Quantifier + (1|ParticipantID) + (1|Picnum)

Predictor	Estimate	Standard Error	z-value	p-value
(Intercept)	-1.5643	0.2743	-5.703	1.18e-08 ***
QuantifierAll	0.5489	0.1952	2.812	0.00493 **
QuantifierNo	0.4232	0.1972	2.146	0.03187 *
QuantifierOnly	0.3652	0.1975	1.849	0.06440 .