# Training away over-exhaustive errors with distributive quantifiers in children with a single training session

Bachelor's Project Thesis

Christian Roest, s2682192, c.roest@student.rug.nl,
Supervisors: dr. J.K. Spenader

**Abstract:** Children between the ages of 5 to 9 years often make language errors called "overexhaustive errors". If we show them an image with 3 boys each playing with a cat, they reject a statement like "Every boy plays with a cat", if there is an extra cat in the image which no boy is playing with. There is a clear polarization between the responses that children at this age give; they will either always make the mistake, or never make the mistake. This suggests that a full understanding of distributive quantifiers is not learned gradually, rather that there is a trigger effect after which they learn the correct understanding quickly. We wanted to find out whether exposing children to informative examples, in a single training session, will trigger children to learn the correct understanding of distributive quantifiers. We designed a study to test this using the Dutch quantifier "elke". Results of picture verification tasks before training, after training, and 5 weeks after training, show that 10 out of 24 children improved to correctly use the quantifier most of the time. This suggests that a single session of informative examples is not enough for most children to trigger a correct understanding.

## 1 Introduction

Over-exhaustive errors are a type of grammatical error that is common in children between the ages of 5 to 9. They occur when children reject a universally quantified sentences with an additional object. (Drozd, 2001) Among 5-to-9-year-old children around 50% of the children make them. An adult would accept a statement like Example (1), with the situation depicted in 1.1. A child that makes over-exhaustive errors would reject it. When asked why, they often explain that one sandcastle has no girl. Over-exhaustive errors are often called 'spreading errors', and the people that make them are called 'spreaders'. In this paper these terms will also be used.

(1) Every girl is building a sandcastle.

The reason why children make spreading errors has been a topic of discussion for a long time in language acquisition research, yet there is still no definitive theory. In order to contribute to this topic, we decided to look at what it takes for children to attain adult-like competence.

Previous research showed extreme response patterns where children will make the errors all the time, or not make them at all. We think the extreme response patterns may indicate a knowledge leap, as opposed to a gradual learning process.

We designed a training study with children that make spreading errors, to see if this is the case. Children were tested using picture verification tasks, and we compared their performance before and after a single training session. If there is a knowledge leap in the acquisition process, then it should be possible to trigger it by presenting the children with some highly informative examples.

The results from the experiment support the existence of a knowledge leap. Performance increased after the training session, indicating a lack of knowledge causing the errors rather than cognitive factors.

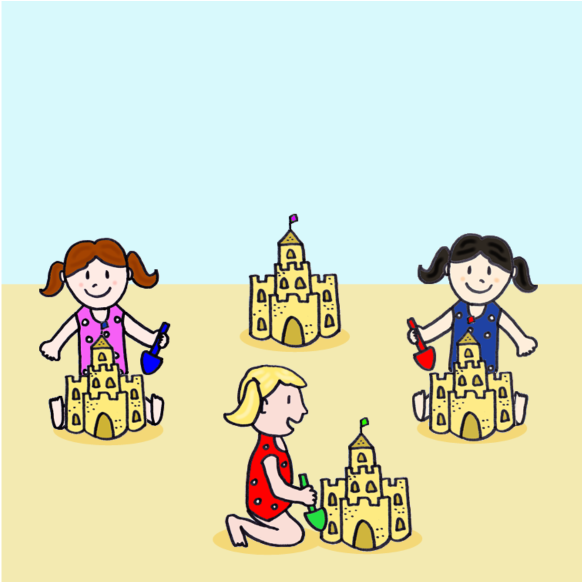


**Figure 1.1: An example image depicting an spreading scenario.**

## 1.1 Background

Acquisition of universal quantifiers has been the subject of research for decades, which has given rise to different theories about the nature of the process. The topic first gained interest after Inhelder and Piaget (1959) first wrote about the difficulty that children have with universally quantified sentences. They were interested in children's semantics pertaining to sets and did a study with children where they presented them with colored objects of different shapes. They found that children would often reject sentences that an adult would say is true, when there is an extra object in the context.

Multiple explanations have since been proposed by researchers, with most of these explanations either focusing on the cognitive or the linguistic aspect of the acquisition process.

Linguistic approaches give less credit to cognitive factors, and instead focus on how language processing could be different in children and adults.

Roeper et al. (2005) discuss several different possible explanations for the difference in interpretation of universal quantifiers in children.

One account from Philip (1994) highlighted in the paper by Roeper et al. (2005), describes the semantic representation of children to be focused around events. This translates itself into the idea that children that spread, think of the sentence "Every boy plays with a cat" as though it means that every event should be an event where a boy plays with a cat. If there is an extra object in the presented image, the extra object becomes an extra event and they will reject the statement.

An account by Sauerland (2003), also discussed in Roeper et al. (2005), revolves around the idea that grammars have a silent 'always', which causes children that don't understand the meaning of 'every' to ignore the word, and the meaning of the sentence becomes "It is always the case that a boy plays with a cat." Because the word 'every' is not common in a child's language production besides its use in compounds, it seems plausible that children don't understand its meaning yet.

Roeper et al. (2005) propose their view of how the quantifier 'every' is learned. The paper also investigates spreading errors with pictures that include objects which are not in the question. They call this unmentioned-object spreading, or 'bunny-spreading', which occurs with very young children. An example of bunny-spreading is shown in Example (2) with Figure 1.2.

(2) Every girl is washing a dog.

In this example we see that there are two events involving girls washing a dog, and one event of a boy petting a snail. When a young child hears the sentence in Example (2), they will often reject it. (Roeper et al., 2005)

According to Roeper et al., acquisition involves three stages between which children reevaluate their grammar to go to the next stage:

1. In the first stage, children will reject every situation that includes any event that is not an event described in the question. This can be captured by the event-based theory, as discussed by (Philip, 1994). At this stage the previously described 'bunny-spreading' occurs.

2. In the second stage, they argue that children's semantic representation allows the quantifier to influence both noun phrases, by placing the quantifier in the wrong location in the syntactic tree. Children in this stage no longer make spreading errors over objects and subjects which are not mentioned in the question,

**Figure 1.2: An example of a situation which can cause young children to show bunny-spreading.**

but still make 'classic' spreading errors where are the objects in the scenario are mentioned in the sentence.

3. Then in the third stage, children reach adult-like interpretation, by moving its location in the syntactic tree to where it does not influence the predicate.

By contrast, cognitive approaches rely on the idea that children have an adult-like understanding of universal quantifiers, but still make errors because of other factors. Crain et al. (1996) concludes that children have a full grammatical competence regarding universal quantifiers, and that the mistakes that they make arise from the context of the experiment. With age, children become less susceptible to indicative contexts of the experiments causing them to make less errors.

By studying bilingual adults in an eye-tracking experiment, Brooks and Sekerina (2005) conclude that because of the fact that non-native speakers make the same kind of errors that children do regarding universal quantifiers, that the errors are not caused by a lack of grammatical competence. Rather, they suggest that spreading errors are the result of a cognitive overload caused by the extra object in the situation. Errors are caused by mapping the syntax to an incorrect semantic representation, which would lead to shortcuts in processing. These shortcuts are enough to understand most contexts, but in the case of spreading questions it leads to oversimplification.

For cognitive theories such as those of Crain et al., and Brooks and Sekerina, attaining adult level competence is dependent on an increase in the cognitive capacities.

An interesting characteristic of spreading errors is that mostly the children that make them will make them all the time, and the children that don't make them never make them (Philip, 1994) (Aravind, 2017). Why do we rarely encounter children which make them only some of the time, if these are the ages where they learn it? This fact seems to suggest a process where the children change their grammar for the quantifier 'every' in a very short period of time. This could caused by a trigger, such as informative examples which do not match with their current understanding of the quantifier.

## 1.2 Training studies

Previously, Arslan et al. (2015) have been successful in letting children revise their reasoning strategies in a training study with children on second order theory of mind. In Arslan et al.'s training study only 'correct' or 'wrong' was given as feedback, together with the correct answer. They argue that it is in most cases possible for children to overcome their incorrect reasoning by repeated exposure to this type of feedback, despite previous studies in this field suggesting cognitive constraints are preventing the child from improving. This made us ask the question: can we apply this same strategy to let children successfully revise their reasoning on spreading questions?

Furthermore, Rohlfing (2006) conducted a training study on teaching children the meaning of the word "under" in Polish. They found a difference in performance between familiar situations and unfamiliar situations. Familiar situations are situations that are logical and relatable for the participating children, and the unfamiliar situations require the child to apply the same principle to a situation in which the word "under" does not seem logical, or with unfamiliar objects.

We wanted to test whether the obtained knowledge in our experiment will also show a difference between familiar and unfamiliar / abstract situations, or that the knowledge will transfer from the familiar situations to unfamiliar situations without a loss in performance. We included an additional test in our experiment design to see whether this is the case, which will be explained in section 2.10 of this paper.

This paper aims to contribute to the discussion about the language acquisition process for universal quantifiers, by testing whether it is possible to trigger children to rapidly learn the correct usage of the Dutch quantifier 'elke' ('every'), by exposing them to informative examples in a single training session.

We designed a three session training study to test performance on spreading questions before training, directly after training, and 4-5 weeks after training to find out whether children can be triggered to make a knowledge leap to adult-like use of the quantifier.

# 2 Experiment

The experiment consisted of four tests and a training session, divided over three visits. The pre-training test was taken on the first visit. The training and the post-training test were done during the second visit. The 5-weeks-after test and abstract test were both taken on the third visit, approximately five weeks after the second visit.

## 2.1 Participants

The analysis included the results of 65 Dutch 5- to 8-year-old children (24 female, $M_{age} = 6.25$, $SE = 0.09$, range=5-8). All children were recruited from two primary schools in the area of Groningen, Netherlands. They were in group 2 and 3 of a Dutch primary school, and come from mostly upper-middle-class families. The children all spoke Dutch as their first language. The children were tested in a quiet room in their school building.

For this study we are interested in the children who make the spreading errors in the first place. Therefore we conducted a pre-training test to see which children make the mistakes. In the end, 24 children were included in the training and post-test. (14 female, $M_{age} = 6.5$, $SE = 0.16$, range=5-8).

## 2.2 Materials

The tests consisted of Picture Verification Tasks (PVT). All of the questions were true-or-false questions. Images were shown one after another on a 15.6" Lenovo Y50 laptop screen. A quiz program was made specifically for this experiment, with simple interface showing only the picture, the question, and the option buttons for "yes" and "no". The program was designed to present the questions in a random order. Each combination of activity and subject would only appear once for each test. Balance among question types and anticipated answers was taken into account. The images were also balanced over the genders of protagonists and the different activities. The main question types for the experiment are as shown with examples in 2.1. Example sentences for each question type are:

(a) Every girl is building a sandcastle.
*(Correct answer: yes)*

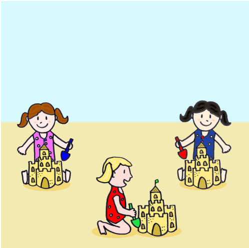
(b) Every girl is washing a dog.
*(Correct answer: no)*

(c) Every boy is building a snowman.
*(Correct answer: yes)*

(d) The monkey is pulling a horse.
*(Correct answer: yes)*

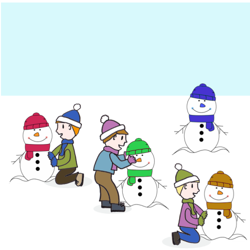
## 2.3 Pre-training test

The pictures for the pre-training test were colorful and contained cartoon boys, girls, and monkeys, performing different activities which are familiar for children from 5 to 9 years. Activities included actions like building a sandcastle, painting a fence, and pulling a toy car. For each combination of subjects and activities, the following three image types were made: one-to-one (OO), which shows an equal distribution between the objects and the subjects in the images (Figure 2.1a), under-exhaustive (UE), which shows more objects than subjects in the images (Figure 2.1b), and spreading (OE) which shows more subjects than objects in the images (Figure 2.1c). In the images it is clear that the subjects are interacting each with a separate object. In the UE and spreading images, the extra object or subject is clearly separated from

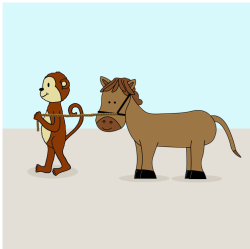


(a) one-to-one

(b) under-exhaustive



(c) over-exhaustive (spreading)



(d) control (true)

**Figure 2.1: Example images from the pre-training test**

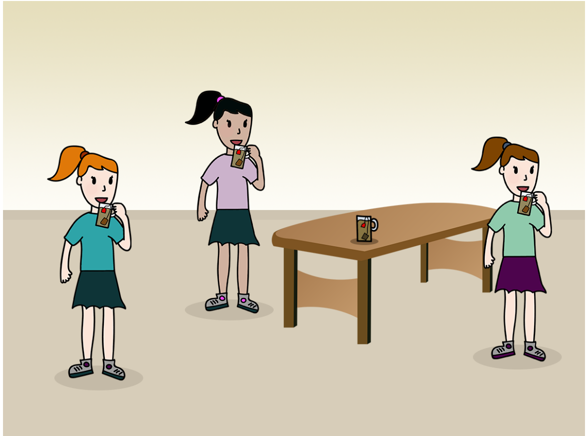


**Figure 2.2: An example of an incorporated question type. Note that the images for this question type are interchangeable with the spreading questions, since there is an additional object in the image.**

| Question type | $M_a$ | $SE_a$ | $M_s$ | $SE_s$ |
|---|---|---|---|---|
| Spreading (OE) | 0.64 | 0.02 | 0.15 | 0.03 |
| Under-exhaustive | 0.89 | 0.01 | 0.96 | 0.01 |
| One-to-one | 0.99 | 0.00 | 0.98 | 0.01 |
| Incorporated | 0.73 | 0.02 | 0.33 | 0.04 |
| Control true | 1.00 | 0.00 | 0.99 | 0.01 |
| Control false | 0.99 | 0.01 | 1.00 | 0.00 |

**Table 2.1: Mean correct response (1.00 = all correct, 0.00 = all incorrect) and standard error per question type. $M_a$ is the average over all pre-training test children, $M_s$ is the average over all children that had a score of 0.5 or less on the spreading questions.**

the other objects and subjects. The tests also contained control questions, where the image depicted a simple situation, to test basic knowledge and attention.

The children heard the sentences through the speakers. The audio was recorded in a recording room at the university, and the speakers spoke at a slow speed. The children reported no problems with understanding the audio, although occasionally the experimenter was asked to repeat the question because of other noises (e.g. they wanted to make a comment while the next recording was already playing, or somebody was passing by the room).

The question type 'incorporated' was included as an extra question type in the pre-training test. The images used for this type, like the spreading questions, have an additional object in the picture, but the question was formulated with the verb incorporated into the noun. Example (3) and Figure 2.2 show an example of this question type.

(3) Every girl is a tea-drinker.

As a result of including the activity in the object, the object becomes less salient in the sentence, and the children might be less inclined to make one-to-one pairings between the objects and subjects. We included six questions of this type in the pre-training test.

All in all, the pre-training test contained 8 spreading, 8 under-exhaustive, 8 one-to-one, 8 control, 6 incorporated, and 2 practice questions, for a total of 40 questions. Example images from the pre-training test can be seen in Figure 2.1

## 2.4 Results pre-training

Table 2.1 shows the mean correct answer per question type accross all children participating in the pre-training test. Table 2.2 shows the correct an-

| Question type | Correct answer |
|---|---|
| Spreading (OE) | yes |
| Under-exhaustive | no |
| One-to-one | yes |
| Incorporated | yes |
| Control true | yes |
| Control false | no |

**Table 2.2: An overview for the question types included in the pre-training test and the correct answers for each question type.**

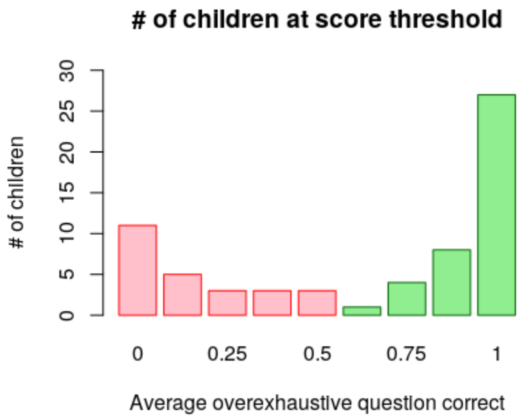



**Figure 2.3: Pre-training test results showing the number of children at each score threshold for the spreading questions. The red bars indicate the 24 children which were selected to take part in the rest of the experiment based on their scores.**

swer for each question type. The spreading questions were answered correct by the participants with a 0.64 accuracy. The incorporated questions were answered with an accuracy of 0.73. The children which had a score of 0.5 or less on the spreading questions in the pre-test were selected for the rest of the experiment. Of this group, the average correct answer on the spreading questions was 0.15, and the average correct answer on the incorporated questions was 0.33. Figure 2.3 shows how the scores of the children are distributed across score thresholds.

Using logistic mixed-effects models (Bates et al., 2015) the responses of the children were analyzed. A step-wise deletion procedure was used to find the best model for the data, with a complex model only being preferred over a simpler model only if its AIC value is two or more points lower than the simpler model. The fixed effects were Question type (spreading, under-exhaustive, one-to-one) and the random effects of Participant, Item, and Age (in months). We also tested the fixed effect as a random slope with all random factors (Jaeger, 2008). The resulting best model retained the Question type as a fixed effect, and Participant and Item as a random factor.

The model shows that there is a significant difference between the performance on the spreading question type, and the one-to-one, under-exhaustive, and incorporated question types: incorporated ($P < 0.05$), one-to-one ($P < 0.001$), under-exhaustive ($P < 0.001$). Appendix A.1 shows a summary of the best model for the pre-training test.

After the pre-training test, the 24 children (14 female, $M_{age} = 6.5$, $SE = 0.16$, range=5-8) which scored 0.5 or less on the spreading questions were selected to participate in the rest of the experiment

## 2.5 Training

For the training session, the children were introduced to "Tiger", a hand puppet that played the role of a non-Dutch native trying to learn Dutch. The children were told that the other experimenter was going to teach Tiger Dutch, and the children were recruited to help him as a "teacher". The purpose of this was to present the information in a way that they did not feel like they were being taught, rather that they were playing in a game. It also enables us to show children the correct way to do it without correcting them, by correcting the tiger instead.

Physical objects and subjects were used during the training session. The objects were 30cm tall laminated paper cutouts of three boys and three girls. The subjects were sets of small separate toys (cars, hamburgers, dice, etc.), as well as several partitionable objects (cake, lasagna, pie).

Different scenarios were presented in front of the children using the objects and subjects. After presenting a new situation, the experimenter always asked Tiger the question in the same form, as shown in Example (4).

(4) Heeft elke jongen een auto?
(English: Does every boy have a car?)

Tiger would then respond by either making a mistake, or by answering correctly. The child was always asked to verify Tiger's answer, and explain why it was correct or incorrect. After that, the experimenter would explain why Tiger was or was not correct. These corrections were aimed at Tiger, not at the children. The experimenter put the emphasis on the fact that Tiger should look at whether all the boys or girls had an object, and that any additional objects did not affect the answer.

The four different scenarios which we presented during the training sessions were one-to-one (each object has a subject), under-exhaustive (one subject with no object), spreading (one extra object, clearly separated from the objects), and many-left (multiple extra objects).

The first scenario was always an under-exhaustive situation with separate objects (Figure 2.4). This was to test whether they still make under-exhaustive errors. The second scenario was always an spreading situation with separate objects. (Figure 2.5). The rest of the training consisted of repeating the four different scenarios. Five minutes into the training passed, the objects were changed from boys to girls or from girls to boys. In each session at least four scenarios with divisible objects were presented. When after a while children showed good understanding, they were asked if they wanted to make a question for Tiger, and they would take some objects and make scenario themselves. After approximately ten minutes we stopped the training, and asked them to do another test to see if the training had an effect on their spreading errors.

After the training session, all the children were able to correctly identify spreading mistakes, and explain that only the subjects need to be considered in the question.

## 2.6 Post-training test

The post-training test was conducted directly after the training session. Similar pictures were used, but depicting situations different from the ones in the pre-training test. There were no audio recordings,


**Figure 2.4: Under-exhaustive scenario presented to child during the training session. Tiger guesses correctly, the child is asked if Tiger is correct, and to explain why.**

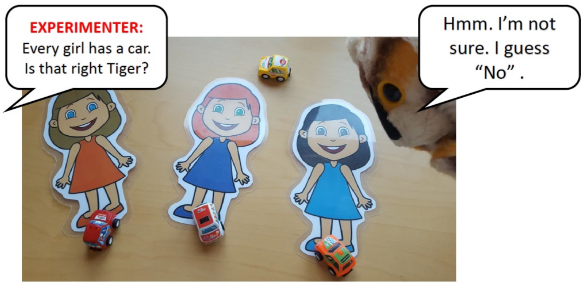



**Figure 2.5: spreading scenario presented to child during the training session. Tiger guesses incorrectly, and the child is asked to correct Tiger, and to explain why it is incorrect.**

| Question type | $M$ | $SE$ | Correct answer |
|---|---|---|---|
| spreading | 0.51 | 0.05 | yes |
| Under-exhaustive | 0.95 | 0.02 | no |
| One-to-one | 0.99 | 0.01 | yes |

**Table 2.3: Post-training test results showing the mean correct response per question type (1.00 = all correct, 0.00 = all incorrect).** $M$ **and** $SE$ **are the mean and standard error over children participating in the post-training test.** $N = 24$

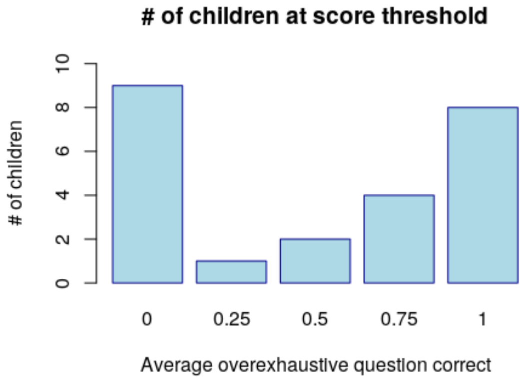



**Figure 2.6: Post-training test results showing the number of children at each score threshold for the spreading questions.** $N = 24$

instead one of the experimenters read the questions off the screen. The test was shorter because we could not take the children out of class for a longer time. The purpose was to give an indication of whether the training had affected their performance. The list of questions included 4 spreading questions, 4 under-exhaustive, and 4 one-to-one questions, for a total of 12 questions.

## 2.7 Results post-training

Table 2.3 shows the mean correct answer per question type across all children participating in the post-training test. These are the children that scored 0.5 or less on the spreading questions in the pre-training test. The spreading questions on the post-training test were answered correct by the participants with a 0.51 accuracy. Figure 2.6 shows how the average correct answers are distributed across the participants.

12 out of the 24 children have improved from 0.5 or lower on the pre-training test, to a score higher than 0.5 on the mid-test, of which 8 children have improved to a perfect score on the post-training test.

The same process from the pre-training test was used for the post-training test, using a logistic mixed-effects model to analyze the responses of the children on the different question types. The fixed effects was Question Type and the random effects were Participant, Item, and Age (in months). The fixed effect was also tested as a random slope with all random factors. The resulting best model retained the fixed effect Question Type, with a random effect for Participant.

The model shows that there is a significant difference between the performance on the spreading question type, and the one-to-one, and under-exhaustive question types: one-to-one ($P < 0.001$), under-exhaustive ($P < 0.001$).

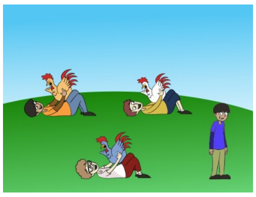
A.2 shows a summary of the best model for the post-training test.

## 2.8 5-weeks after test

Five weeks after the training session another Picture Verification Task was conducted. The participants are the same children that took part in the training session and the post-test (24 children, 14 female, $M_{age} = 6.5$, $SE = 0.16$, range=5-8) The images for this test were different from the pictures which were used in the previous tests. Example images for this test can be seen in Figure 2.7. The pictures featured different animals as objects, and activities such as petting, feeding, and playing with the animals. The subjects were boys and girls. The images were accompanied with audio recordings of the questions. When the children did not hear the question fully, the experimenter repeated the question. The test included 8 spreading, 8 one-to-one, 6 under-exhaustive, 8 control, and 2 practice questions, for a total of 32 questions. Example images from the 5-weeks-after test can be seen in figure 2.7

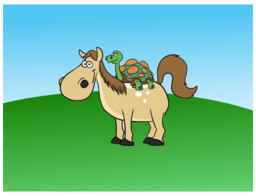
For determining the contents of the tests, it was taken into consideration how long the children are approximately able to focus on the task. For this reason we had a shorter task during the post-training test, since the mid-test was on the same

(a) one-to-one



(b) under-exhaustive

(c) spreading (OE)



(d) control

**Figure 2.7: Example images from the 5-weeks after test**

| Question type | $M$ | $SE$ | Correct answer |
|---|---|---|---|
| Spreading (OE) | 0.44 | 0.04 | yes |
| Under-exhaustive | 0.92 | 0.02 | no |
| One-to-one | 0.99 | 0.01 | yes |
| Control true | 1.00 | 0.00 | yes |
| Control false | 1.00 | 0.00 | no |

**Table 2.4: Mean correct answer and standard error per question type on the post-test (1.00 = all correct, 0.00 = all incorrect).** $N = 24$

day as the training. The post-test was also shorter than the pre-training test because we anticipated that it would start to bore the children to do the same task for the third time.

## 2.9 Results 5 weeks post-training

Table 2.4 shows the mean correct answer per question type across all children participating in the 5-weeks after training test. The spreading questions on the 5-weeks after training test were answered correct by the participants with a 0.51 accuracy. Figure 2.8 shows how the average correct answers are distributed across the participants.

Ten children retained a score above 0.5 on the spreading questions in the 5-weeks after training test, of which seven children had a perfect score on the spreading questions. The average score across all children on the spreading questions was 0.44,

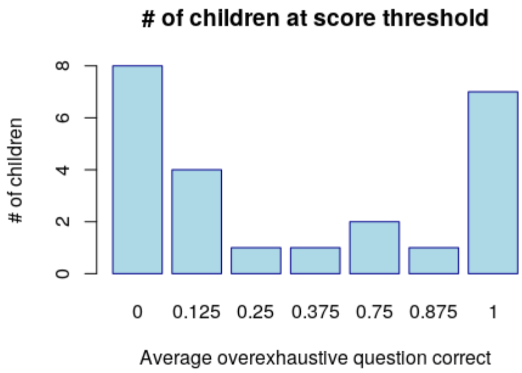



**Figure 2.8: 5-weeks after training test results showing the number of children at each score threshold for the spreading questions.** $N = 24$

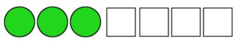
with a standard error of 0.04.

Compared to the post-training test, six of the eight children which had a perfect score on the post-training test (only looking at the spreading questions), also had a perfect score on the 5-weeks after training test. In addition, one child that scored a 0.50 on the post-training test, improved to a perfect score on the 5-weeks after training test. Two children that scored perfect on the post-training test got a lower score on the post-test (new scores 0.125 and 0.875).

Appendix Table A.3 shows a summary of the best model for the 5-weeks after training test, and Appendix Table A.5 shows the model comparing the accuracy on spreading questions on the pre-training test to the accuracy on the 5-weeks after training test. Appendix Figure A.1 shows a graph with bars for each of the children, showing the pre-training test scores, and the post-training test scores side by side.

## 2.10 Abstract test

After the 5-weeks-after test had concluded, we took another short test directly after, to test whether the obtained knowledge would transfer from familiar situations (i.e. the situations we tested with previously, and animate objects), to unfamiliar situations. This task used images with three circles and four squares on a white background. An example image for this task can be seen in Figure 2.9. The



(a) one-to-one

(b) under-exhaustive

(c) over-exhaustive (spreading)

**Figure 2.9: Example images from the abstract test**

question for this question type was always "Is every circle [color]?", where the color varied between blue, green, red, pink, and yellow. There were three different question types included in this test, which represented the under-exhaustive, one-to-one, and spreading cases from the other tests. The under-exhaustive case had an unfilled circle. The one-to-one case had all the circles filled, and no squares filled. The spreading case had all the circles filled, and one square filled, leaving the other squares unfilled. In the images the shapes were grouped together, and used only one color per image. The questions appeared in a random order and could also appear horizontally mirrored.

Again, the 24 children which were selected during the pre-training test were included in the analysis (14 female, $M_{age} = 6.5$, $SE = 0.16$, range=5-8).

The abstract test included 4 spreading, 4 one-to-one, and 4 under-exhaustive questions, for a total of 12 questions. Example images from the abstract test can be seen in Figure 2.9

## 2.11 Results abstract test

The abstract test was made very well by the children, only 2 of the 24 children had a score of 0. For the rest, all the children scored 0.75 or higher on the abstract test. Figure 2.10 shows the distribution of scores on the OE questions.

Appendix Table A.4 shows a summary of the best model for the abstract test.

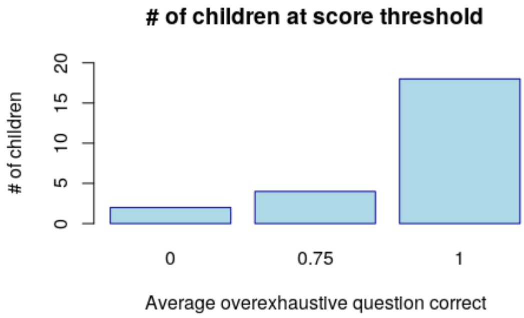



**Figure 2.10: Abstract test results showing the number of children at each score threshold for the spreading questions.** $N = 24$

# 3 Discussion

It is clear that the training had a positive effect on the performance of the selected children. Out of the 24 children, 7 children did not make any mistakes on the final test. This shows that it is possible for children to very quickly improve their performance on over-exhaustive questions when they pick up on examples which don't match their current grammar. This supports the idea that over-exhaustive errors are caused by missing knowledge. By exposing the children to examples where the quantifier works different than how they use it, they learned how the quantifier is used correctly. Because every child was able to identify and correct the mistakes during the training session, we conclude that children actually are cognitively capable of doing this task, making the theory of Brooks and Sekerina unlikely, since their theory suggests that children are cognitively not on a level where they can process the task correctly.

However, more than half of the children in the experiment scored 50% or less on the 5-weeks-after test, from which we conclude that a single training session will not be enough to overcome over-exhaustive errors in most cases. It would be interesting to see how multiple training sessions would affect the performance on these tasks. A ten minute training session is quite short, and although all the children were able to identify errors that the puppet made, they may not have extrapolated what they learned to situations outside of the training.

Recall that around 50% of the children was expected to make over-exhaustive errors, but out of the 65 children that participated in the experiment only 24 made over-exhaustive errors. We suspect that this has to do with the fact that the ages of the participating children from one of the two schools was skewed more to the higher end. About the children from this school it was also said that they had a strong class in general this year. In this school, we recorded considerably less over-exhaustive errors.

Because we only did a single training session, this outcome of this experiment is particularly vulnerable for outside effects. It could be that some children were not able to focus on the task and reevaluate their grammar, because of lack of concentration on that particular day, or other possible influences.

About the ability to transfer the knowledge which was obtained during the training session to unfamiliar situations we unfortunately cannot say very much. The results of the abstract test show that even the children which struggled with the regular 5-weeks after test, have good scores on the abstract test. The results show that the test was too easy for most children, possibly because of the visual spacing (all the same shapes were next to each other), or too little diversity among images. It would be interesting to see the abstract test redone, with a difficulty more on par with the difficulty of the other tests.

Another possible explanation for the high scores on the abstract test is that the children are less emotionally involved with pictures of shapes and colors, as opposed to pictures of animals and children. This could cause them to base their reasoning less on emotions and making less mistakes. Further research will have to determine what caused the children to perform so well on this test compared to the regular tests.

# 4 Conclusion

We can conclude from this study that it is possible to train away spreading errors by only exposing children to highly informative examples. This indicates a knowledge leap in the acquisition process for universal quantifiers. Further research should focus on the effect of multiple training sessions, and also on finding out if knowledge obtained from training sessions transfers from familiar situations to unfamiliar situations.

# References


Aravind, A. e. a. (2017). Childrens quantification with every over time. *Glossa: a journal of general linguistics*, 43(1(2)).

Arslan, B., Verbrugge, R., Taatgen, N., and Hollebrandse, B. (2015). Teaching children to attribute second-order false beliefs: A training study with feedback.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Brooks, P. J. and Sekerina, I. (2005). Shortcuts to quantifier interpretation in children and adults. *Language Acquisition*, 13(3):177–206.

Crain, S., Thornton, R., Boster, C., Conway, L., Lillo-Martin, D., and Woodams, E. (1996). Quantification without qualification. 5:83–153.

Drozd, K. F. (2001). *Children's weak interpretations of universally quantified questions*, page 340376. Language Culture and Cognition. Cambridge University Press.

Inhelder, B. and Piaget, J. (1959). La gense des structures logiques lmentaires: Classifications et sriations. english translation (1964): The early growth of logic in the child: Classification and seriation. routledge and kegan paul, london. *Routledge and Kegan Paul, London.*

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

Philip, W. (1994). Event quantification in the acquisition of universal quantification. *Doctoral dissertation, University of Massachusetts, Amherst.*

Roeper, T., Strauss, B., and Pearson, B. Z. (2005). The acquisition path of the determiner quantifier every: Two kinds of spreading. *Ms., University of Massachusetts, Amherst.*

Rohlfing, K. J. (2006). Facilitating the acquisition of under by means of in and on - a training study in polish. *Journal of Child Language.*

Sauerland, U. (2003). Universal quantification in language acquisition and disorder,.

# A Appendix

**Table A.1: Pre-training test: Model = Correct responses $\sim$ Picture type + (1 | id) + (1 | question)**

| Predictor | Estimate | Standard Error | z-value | p-value |
|---|---|---|---|---|
| intercept (spreading) | 1.1735 | 0.3430 | 3.421 | 0.000623 *** |
| incorporated | 0.7839 | 0.3408 | 2.300 | 0.021464 * |
| one-to-one | 5.7716 | 0.6191 | 9.323 | <2e-16 *** |
| under-exhaustive | 2.3430 | 0.2235 | 10.483 | <2e-16 *** |

**Table A.2: Post-training test: Model = Correct responses $\sim$ Picture type + (1 | id)**

| Predictor | Estimate | Standard Error | z-value | p-value |
|---|---|---|---|---|
| intercept (spreading) | 0.03728 | 0.51449 | 0.072 | 0.000623 |
| one-to-one | 5.91222 | 1.23169 | 4.800 | 1.59e-06 *** |
| under-exhaustive | 3.97124 | 0.66920 | 5.934 | 2.95e-09 *** |

**Table A.3: 5-weeks after training test: Model = Correct responses $\sim$ Picture type + (1 | id)**

| Predictor | Estimate | Standard Error | z-value | p-value |
|---|---|---|---|---|
| intercept (spreading) | -0.2687 | 0.4010 | -0.670 | 0.503 |
| one-to-one | 6.5016 | 1.0467 | 6.212 | 5.24e-10 *** |
| under-exhaustive | 3.6510 | 0.4286 | 8.518 | <2e-16 *** |

**Table A.4: Abstract test: Model = Correct responses ∼ Picture type + (1 | id) + (1 | question)**

| Predictor | Estimate | Standard Error | z-value | p-value |
|---|---|---|---|---|
| intercept (spreading) | 3.0476 | 0.8388 | 3.633 | 0.00028 *** |
| one-to-one | 2.2849 | 0.8557 | 2.670 | 0.00758 *** |
| under-exhaustive | 2.2849 | 0.8557 | 2.670 | 0.00758 *** |

**Table A.5: Pre-training test compared to 5-weeks after training test accuracy on spreading questions**
**Model = Correct responses $\sim$ Experiment + (1 | id)**

| Predictor | Estimate | Standard Error | z-value | p-value |
|---|---|---|---|---|
| intercept (pre-training) | -3.2370 | 0.6148 | -5.265 | 1.40e-07 *** |
| 5-weeks after test | 2.6835 | 0.3948 | 6.797 | 1.07e-11 *** |

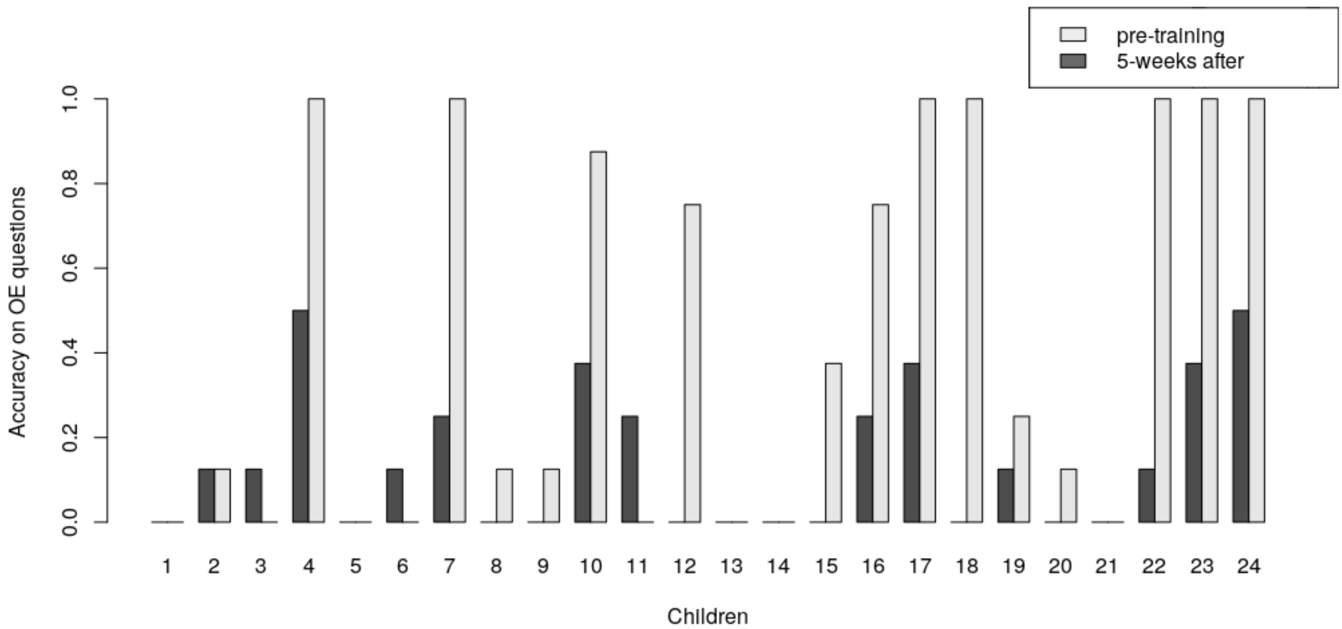



**Figure A.1: Comparison scores pre-training to 5-weeks after training**