Does every child produce "every" correctly?

UNIVERSITY OF MASSACHUSETTS OCCASIONAL PAPERS 34

CURRENT ISSUES IN FIRST LANGUAGE ACQUISITION

2006

Edited by

Tanja Heizmann

Published by
GLSA
(Graduate Linguistic Student Association)
Department of Linguistics
South College
University of Massachusetts
Amherst, MA 01003-7130
U.S.A.

glsa@linguist.umass.edu
http://www.umass.edu/linguist/GLSA

Cover design: Jan Anderssen

ISBN: 1-4196-5076-9

TABLE OF CONTENTS

Does every child produce "every" correctly?*

Emily Altreuter and Jill de Villiers

Smith College

1 Introduction

In this paper we will examine the classic phenomena associated with children's interpretation of "every," and present an experimental study that looks at production and comprehension in the same children. First, consider the sentence "Every cat has an apple," and the different simple stimulus arrays that have been used to test children's comprehension of the sentence.

> Type A: three cats each have an apple, and there is an extra apple. Adult answer: "yes."
> A common child error on Type A is to answer "no," due to the extra apple with no cat. This has also become known as the "spreading error," also "exhaustive pairing."
>
> Type B: three cats each have an apple and one cat has a banana. Adult answer: "no."
> An error on Type B would be to say "yes," misunderstanding "every." This is sometimes called an "underexhaustive search" error.
>
> Type C: three cats each have an apple and a dog has a banana. Adult answer "yes."
> An error on Type C, less studied, is to answer "no" because of the extra dog and banana. This has been called a "perfectionist" error.

* We would like to acknowledge the help of the children at the Smith College Campus school, their teachers and parents, and Cathy Yarnell for her patient help with our requests. Thanks are also due to Gabrielle Merchant, Clara Feldmanstern and Alison O'Connor for help testing and transcribing. We also thank the audience at the Fall 2005 UUSLAW for their helpful comments on an earlier version of this paper.

Tanja Heizmann (ed.): Current Issues in Acquisition. University of Massachusetts Occasional Papers in Linguistics 34, 1-19.
GLSA Amherst.

Type D: Four cats each have an apple. Adult answer "yes."
An error on Type D would suggest a failure to understand "every."

There are several distinct hypotheses that have been forwarded for these basic phenomena, particularly errors on Type A (from now on, we simply will call these "Type A errors"). Philip (2004) gives a very explicit comparison of two such theories, called the Event Quantificational Account (EQA, a revision of Philip 1995) and the Relevance Account (RA) a revision of the Presuppositionality Account of Drozd and van Loosbroek (2004). In addition, there is the Weak Quantifier Account (WQA) of Guerts (2003), and the Developmental Account (DA) of Roeper, Strauss and Pearson (this volume). There is no room to consider all the fine points of comparison but the major ones are reviewed to set up our experiment.

1.1 The Event Quantification Account

The EQA describes the child's semantic representation informally as follows:

> "Every minimal event which is a temporal subevent of a minimal cat-holding-apple event and in which a cat or an apple or both is a participant is a minimal event of a cat holding an apple."

In other words, if there is anything in the picture that is either cat or an apple, the child asks whether it is true that a cat is holding an apple. The child then sums across all such events to ask if is true that in every case, a cat is holding an apple. The EQA attributes to the child a semantic representation for "every" roughly equivalent to an adverbial quantifier such as "always" (see also Roeper & de Villiers, 1993). Since there are languages in the world that only have adverbial-type quantifiers, Philip asserts that the child is still following constraints on UG in positing such an interpretation. However, it is an immature representation for English. The consequence of such an interpretation is that a child would object to scenes in which, for example there was an additional apple, as in Type A. However, the EQA does not predict a mistake on Type C, where there is an extra dog holding a banana, since neither a dog nor a banana is a subevent of cat-holding-apple. As a result, errors on Type C are regarded by the EQA as a cognitive mistake of "perfectionism." In such a case a child has a processing failure of a much more basic kind, and cannot retain the lexical items called for in the representation. Similarly, underexhaustive responses (saying yes to Type B) are not given any linguistic account by the EQA, and are considered a separate cognitive error.

1.2 The Relevance Account

On this account, the child has full linguistic representation of the adult form. The particular problem manifest in exhaustive pairing is that the child imagines a fourth unseen object, say a cat missing from Type A. This happens because of symmetry requirements in the child's mental model. The child's verification process is non-adult like because s/he assumes that the missing cat is relevant to the verification of the sentence. Philip (2004) has demonstrated that this effect can be ameliorated in children by environments that discourage visual symmetry, and can be stimulated in adults by situations in which the missing agent is assumed relevant. Like with the EQA, Type B and Type C errors are considered different in kind from Type A and from each other. The RA then differs from the EQA in assuming full adult competence except for a pragmatic difficulty in determining relevance.

1.3 The Weak Quantifier Account

Geurts (2003) proposed that children construe the strong determiner as if it were *weak*. Children are said to adopt an interpretation of the quantifier as weak rather than strong because weak quantifiers are easier. He argues that the problem lies in the mapping between syntactic and semantic representations, namely a parsing problem, which is then "patched" by pragmatics. The adult interpretation (in Guerts' framework) of "every cat holds an apple" is:

(1) <every> (x,y: cat (x) apple (y):x holds y]

However the child's semantic interpretation begins as this:

(2) [...:....] <every> [x,y: cat(x), apple (y), x holds y.]

The front brackets contain an open variable for the domain of quantification. So, depending on the salience in the context, the child might interpret this as being about cats or about apples. In the case where the apples are in focus, the reading will be:

(3) [y:apple(y)] <every> [x: cat(x), x holds y]

Thus the WQA accounts for Type A errors, because if there are apples without cats, the child will say "No", because every apple is not being held by a cat. Unlike the other theories, Geurts explains Type B or underexhaustive errors by the same mechanism: If there are cats without apples, the child might say "yes", because every apple *is* being held by a cat.

With some modification, WQA can also account for the Type C error, by arguing that the child quantifies over all animate objects, as if to say roughly "everything that is an animal is an animal holding an apple." As in the other accounts, in the WQA the Type C error is regarded as more serious than exhaustive pairing of Type A, and it resonates with the account in the EQA that the child making such an error fails to pay attention to the lexical items. However, Geurts emphasizes that more work is needed on Type C errors.

1.4 Developmental Account

Roeper, Strauss and Pearson (this volume) provide an account that considers closely the comparison of Type A and Type C errors, as well as Type B or underexhaustive responders who they class as "perseverators" or "yes-men." In a very large sample of children aged 4 through 9, they find a consistent age difference between the types of responders, with the Type B underexhaustive group being the youngest at 4;7, followed by Type C, then Type A, then target or adult-like children by age 8. However, responses of all types were found at all ages. They propose that children change in their semantic representations that are tied to the syntax of their quantifiers. At the start, children treat quantifiers adverbially, or via event quantification. At this stage, Type A and C errors will occur. At the next stage, the child assimilates "every" primarily to "each," which can "float" in the syntax in the adult language:

(4) Each cat has an apple → Cats each have an apple

The claim is that at this stage a child can misconstrue this as quantifying over both subject and object:

(5) Each cat has each apple

At this stage the quantifier is an NP-quantifier moved to the Focus phrase, where it can c-command elements in the VP. The child requires exhaustive pairing so commits Type A errors, but Type C errors should decline. What happens next on the DA story is the acquisition of the particular properties of quantifiers, e.g. whether they are in a DP, and if so, where. When the child discovers that *every* is not inherently distributive, i.e. it can also take a collective reading. Roeper et al argue that the child no longer raises it to the Focus position, and "every" instead gets fixed to the DP and no longer floats. As a result, the children will have adult like interpretations.

The advantage of the DA account is that it links semantic changes to syntactic developments and predicts a step-wise learning path that accommodates within the *grammar* three possible types of responses. However, it is at the cost of proposing non-adult-English grammatical options en route to the target form. Furthermore, it does not have an account of Type B underexhaustive readings. Children who give underexhaustive readings are said to perseverate on "yes" responses, which says nothing about their grammars.

In sum, the EQA and the RA consider only Type A exhaustive pairing of interest linguistically. In a careful comparison within studies designed to tease them apart, Philip (2004) favors the RA. Geurts' WQA predicts that Type A and Type B errors should co-occur as they are due to the same process. The DA predicts an ordered development, but considers Type B errors to be cognitive in origin.

The following chart attempts to capture which types of responses each theory captures or excludes:

Type of response:	EQA	RA	WQA	DA
Type B	Cognitive error	Cognitive error	Weak reading + salience	Cognitive error
Type C	Lexical error	Lexical error	Animate reading	Event quantification
Type A	Minimal event quantification	Symmetry + relevance failure	Weak quantifier + salience	Event quantification OR Floated NP quantification
Target	Event→ NP quantification	Adult pragmatics	Strong quantifier	DP quantification

How would the different accounts predict production of "every" by children? The DA most clearly would predict Type A and C errors in production, as it is the developing grammar that is being described. The EQA and the RA might predict Type B errors in production too, as those errors are attributed to cognitive mistakes. It is unclear whether the RA, EQA or WQA would expect Type C errors, as the lexical errors are meant to be processing problems that would be less likely in production. The EQA might predict Type A errors if the child's grammar allows them. However, since Philip attributes the error under the EQA to insufficient processing resources to maintain an adult LF, that could be taken to imply that the child would not have the same difficulty in production. The RA might also predict that Type A production errors should occur, since the same issues of symmetry and relevance would operate on production too. On both the RA and the EQA, the different error types have different causes, and so could presumably co-occur in the same child. The DA does not predict that, since the errors represent different stages of grammar. The exception is Type C errors, which should never appear without Type A, but Type A could occur without Type C.

The WQA is described as a processing account, so Geurts presumably would not predict parallel failures in production if the grammar is adult-like. Or would the child stumble over fixing the appropriate domain for the weakly construed quantifier even in producing such sentences? Guerts would at least need to predict Type A and Type B errors should co-occur, though none of the other theories predict this.

We are in a position to compare the various hypotheses against data from our own experiment that looks at both comprehension and production across the four types of scenario. In sum:

DA: Errors of different types should occur in different children, not in the same child. The different grammars should be consistent across comprehension and production. Underexhaustive or Type B errors are given no account, and might not be expected in production if they are simply "yes" perseveration. Type C errors should not occur alone.

EQA: No relationship is expected between Type A errors and other types. Type A errors should be consistent across production and comprehension. Different errors can co-occur in the same child but should not correlate as their sources are distinct.

RA: No relationship is expected between Type A errors and other types. Type A errors may not be found in production. Different errors can co-occur in the same child but should not correlate as their sources are distinct.

WQA: Type A and Type B errors should co-occur in comprehension and if they occur at all, in production too. Other errors should occur independently

There has to our knowledge been no work done to elicit production of "every" in an experimental setting of this sort. We designed a study to address the question of whether the error types A-D are restricted to comprehension. We also wanted to see if children knew that "every" had to be exhaustive, rather than just "plural," given the ambiguity of the ways in which children use "every" in spontaneous speech. Merchant (2005) searched the files of 18 CHILDES corpora and showed that children used "every" very rarely, with children under four or five almost never using it. When "every" was used, it was most often in the frozen forms such as "everyday" or "everybody," rather than "every toy" or "every apple." In fact, Merchant identified only 10 instances of "every N," all occurring late in the transcripts.

Of course there is a good reason why elicited production has not been tried, namely, how can a child be induced to use "every" if the context provided does not fit what the child's grammar demands? Following Chomsky (1965), we had to be cunning. We designed the stimuli to allow an alternative response that could be chosen by the child whose grammar did not allow the use of "every", but it required the child to choose something that had not been modeled. In this way, we could examine the child's resistance to the usual adult form.

2 Method

2.1 Subjects

Sixty-four children aged 5;0 to 7;11 were tested, with a mean age of 6;3. The subjects were in kindergarten, first or second grade at the Smith College Campus School. There were 26 boys and 38 girls.

2.2 Procedure

2.2.1 Comprehension

The children were taken one by one to the testing area, where a laptop computer presented the stimuli in a PowerPoint presentation that included the pre-recorded narration for each stimulus. The children all received the comprehension trials on one day, then the production trials on the second day. In this first study, we did not

counterbalance the order. Children received a pretest slide with all the animals pictured, and we first ensured that they could easily identify the animals. Then the children were told that they would see some pictures on the computer with these animals in them, and hear the computer present a sentence for each picture. The children were warned that some of the sentences would be true, but others would not match the picture. The children were asked to say "yes" if they thought the sentence matched the picture and "no" if it did not.

Four stimulus conditions were presented. Consider the sentence "every cat has an apple" to illustrate what might be pictured in each stimulus type:

a) Type A, the source of the "spreading" error, in which there were three cats holding apples and a fourth apple on a table.
b) Type B, the "not-every" case as the source of perseveration or underexhaustive error, in which one cat held a different thing, say, a book.
c) Type C, the source of the "perfectionist" error, in which there were three cats holding apples and say, a sheep holding a book.
d) and Type D, the uncontroversial case, in which four cats are each holding an apple, that could provide the source for a "no"-bias error.

(See Appendix A and B for examples of each type of stimulus). Each randomization included five of each type of stimuli. The pictures were set up in line-ups, not randomly arranged, in such a way that they maximized symmetry. According to Philip (2004), that should also maximize the potential for spreading errors (also Rahklin, 2005).

There were two randomizations of the 20 stimuli, A and B, that were each given to half the children at each age. When the children had completed the 20 trials, they were thanked for their participation and told that they would return the next day for some further examples.

2.2.1 Production

The day after testing comprehension, the same children were tested again in production. Three children did not provide production data; in two cases they were absent and in one case the child refused to be recorded. The subjects were reminded that the day before, some of the computer sentences had not matched the pictures. They were told that we thought they could do a better job, and we liked having kids' voices on the computer, so today we would record them saying the sentences to get them right. The children were then shown a Powerpoint presentation of new pictures similar to the ones used in comprehension and told to make "a true sentence that starts with "Every…" The narrative that the child created was recorded on the same Powerpoint, a procedure used before with success and good recording fidelity (de Villiers, Cahillane & Altreuter, 2006). However, every session was also video-taped from behind the child so the stimulus was visible.

The production stimuli also had all stimulus types A-D represented, but a change was made so that the animals were all shown wearing hats or shoes. These items were not

particularly salient, but the child would have a way to provide alternative true sentences for the array presented. This was necessary for Type B, where "every cat has an apple" was simply false. However, a child could say "every cat has a hat" and therefore make a true sentence. This option existed also for a child who was a classic "spreader" and therefore did not want to say that every cat had an apple if there was an apple on the table. That child also had the option of saying instead that "Every… cat has a hat" for Type D. The perfectionist could say "Every *animal* has a hat" for Type C. (See Appendix B for examples of stimuli.)

3 Results

3.1 Coding

All data were transcribed into a FileMaker Pro database. In the case of comprehension, the child's answers were tallied for correctness by type of stimulus A-D. Only in the case of Type B was a "no" answer considered correct. For production, all the children's responses were transcribed verbatim and then coded as below.

We first removed those children who made more than one mistake on Type D questions (these tended to be the "nay-sayers": children who answered "no" to everything). This removed 13 children: nine 5 year olds, three young 6 year olds and one young seven year old. This left 51 children. 5 of these children failed to do Production: one refused and 4 were absent, leaving 46 for production.

3.2 Comprehension

There was no major difference between the two randomizations so we ignored it in the analyses. We also tested for the effects of gender and found no difference in comprehension, so it also removed from consideration.

Table 1: Mean Number Correct/5 by Age and Type

	AGEGRP	Mean	Std. Deviation	N
TYPEA	5.00	4.75	.910	20
	6.00	4.41	1.064	17
	7.00	4.79	.579	14
TYPEB	5.00	4.05	1.877	20
	6.00	4.59	1.004	17
	7.00	4.93	.267	14
TYPEC	5.00	3.65	2.033	20
	6.00	3.65	1.618	17
	7.00	3.93	1.592	14
TYPED	5.00	5.00	.000	20
	6.00	5.00	.000	17
	7.00	5.00	.000	14

Table 1 shows the mean number correct out of 5 for each type for each age group. A repeated-measures analysis of variance was run on the number correct in comprehension, with age as the group variable and type of stimulus (A,B,C,D) as the repeated measure. There was a significant difference across types ((F(3,48) =11.35, p< .001), but no significant effect of age nor interaction with age.

Table 2 shows the results of post-hoc paired samples t-test across type to locate which types were statistically different from each other. The asterisks indicate a significant difference between those question types.

Table 2: Paired Samples Test comparing performance across types of stimulus

Compare:	Paired Differences Mean	Std. Deviation	t	df	Sig. (2-tailed)
*Type D – Type B	.53	1.347	2.807	50	.007
*Type B – Type C	.75	2.448	2.173	50	.035
*Type D – Type A	.35	.890	2.831	50	.007
Type B – Type A	-.18	1.682	-.749	50	.457

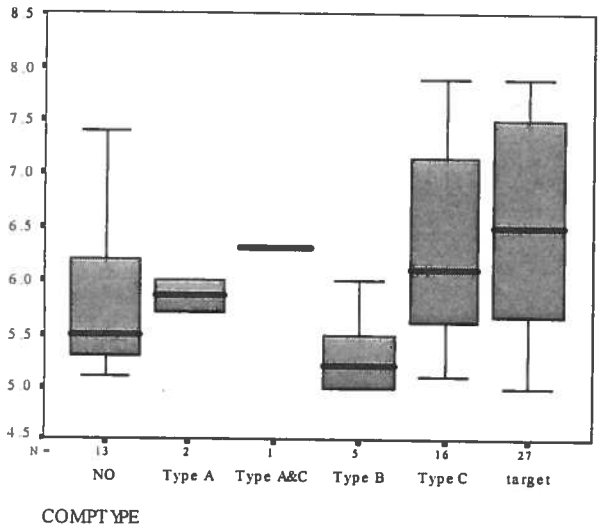
The most difficult across age is Type C, with Type D being errorless. Type B and A are equally difficult.

To investigate the individual patterns more closely, we divided the children into five types of responders. It should be noted that these divisions were very easy to do: children were for the most part remarkably consistent in their responses for each type of stimuli. The majority of the data consists of 0s and 5s.

- "Target" children gave adult readings (allowing one error on any type).
- "Type A" children were the classic spreaders, saying more than one "no" just to Type A stimuli.
- "Type B" children answered "yes" across the board, even to Type B questions where the answer was clearly "no". It is possible that they allow "every" to mean "a lot of", i.e. non-exhaustive. However, it should be noted that this response could be a simple yes-bias as Roeper et al claim. The production data should help distinguish the alternatives.
- "Type C" children were the perfectionists, saying "no" more than once to Type C.
- "Type A & C" children made both types of error.

The first question is, are the types of children of different ages? The box–plot graph in Figure 1 below shows the mean age of each type and the spread of ages around it. It includes the excluded group who said "no" to everything. Type B children tended to be the youngest. There were only two Type A responders, intermediate ages. Type C kids are slightly younger on average than the target children. The Target children tend to be the oldest, but they have the greatest range. If the two Type A and the single Type A&C responders are removed, a univariate ANOVA with age as the dependent variable is statistically significant, with Type B < Type C <Target ($F(2, 45)=3.79$, $p<.03$).

Figure 1: Box plot of ages of children by each comprehension type



By seven years old, the majority of children (11 out of 15) were target responders. Without enough Type A children, the results are still not conclusive that there is a specific path of development that the DA alone would predict, since all the theories would concur that Type B is the most primitive error, and that Type C should precede adult-like responses.

Recall that the different theories make different predictions about co-occurrence of different errors. On the WQA, Types A and B should co-occur, but not the others as they are from independent sources. The correlations are in fact all very weak, with no sign of a significant correlation. This is what would be predicted under EQA, RA and DA. The DA predicts that Type C errors should not occur without Type A, but not vice versa. This is contradicted by the large number of Type C only children.

3.3 Production

Recall that the stimuli in production were designed to allow an escape hatch for children, in that they could choose to say something true about every animal if they wanted symmetry, such as "every cat has a hat." We thought that Type B stimuli would require this option. Instead, many children simply used negation, e.g. "every cat doesn't have an apple," even though this was never modeled. It was rare to choose the "hat" or "shoes" option below age seven, but the seven year olds took advantage of it, and sometimes used it for Type A and C stimuli as well as B stimuli. Only five responses mentioned the clothing for Type D, suggesting it was not salient. In contrast, 25 children constructed such sentences as "every cat has shoes" in the case of Type C. However, only 9 such responses used a general word such as "every animal has shoes", all from two children. There were 16 of the evasive kind for Type A.

If the child produced a response to either Type A or Type C that suggested that they were avoiding the standard answer, such as "Every animal has a hat," this was not counted as an error, but we took note of it in the coding. A statement counted as an *error* only if the child made a factually incorrect statement.

Evidence of Type A "spreading" in production was generally seen in negation, such as:

- "Every dog doesn't have an apple" (counted as an error)

but also may be evident in the following productions:

- "Every dog has an apple and one apple's on the table" (not counted as an error)
- "Wait a minute there's a balloon in midair! Every cat has shoes" (not counted as an error)

Production *errors* on Type B scenarios included "Not every elephant has a ball or a milkshake," and "Every dog has an apple." A type C *error* generally consisted of a negated statement. The following are some examples of what might be equivalent to Type C "spreading" in production:

- "Every sheep doesn't have a balloon"
- "Every cat has a watermelon and a hat and the elephant has a letter and a hat" (counted as correct)
- "Every dog has a watermelon but not the sheep. The sheep has a letter." (counted as correct)
- "Every elephant has a book except for the mouse" (counted as an error in quantifier scope)
- "Every animal has a hat" (counted as correct)

How do the errors correlate in production? To the extent that the WQA predicts production errors at all, it would presumably predict correlation of Types A and B, but this is not seen. And Type C errors do occur alone in three subjects, against the DA.

3.3 How do the two tasks relate?

We next tried to determine the relationship between response type in comprehension and production. Although some of the qualified production sentences above might be evidence that the child would reject *un*qualified sentences in comprehension, we cannot be sure. In the following analysis we only considered clear production errors for each stimulus type.

A series of one-way ANOVAs was conducted in which the grouping variable was not age but type of responder in comprehension, and the dependent measure was accuracy of production across each scene type A, B, C and D. The results in each case showed a significant match between the comprehension style of the child (Comp type) and the kinds of production errors made (See Table 3 for summary). The breakdown of the mean responses by type are shown by the following tables 4-6.

Table 3: Effect of comprehension error type on production errors of each type:

Production of:	variable	df	Mean Square	F	Sig.	Partial Eta Squared	Observed Power
Type A	Comp type	4	6.128	176.551	.000	.945	1.000
Type B	Comp type	4	12.906	3.554	.014	.262	.825
Type C	Comp type	4	2.194	3.094	.026	.232	.764

Table 4 reveals that although there were only three Type A error children in comprehension (counting the one Type A&C child), their errors carry over into their production.

Table 4: Type A errors in production by comprehension error type

Comp type	Mean	Std. Deviation	N
Type A &C	**5.0**	.	1
Type B	0.0	0.0	5
Target	0.0	0.0	25
Type C	0.1	0.3	13
Type A	**0.5**	0.7	2

Table 5 reveals that the children who made Type B errors in comprehension (possibly taking "every" to mean "a lot of") did significantly worse at Type B questions in production than other types of responders. In other words, their errors in comprehension carried over into production. For these children, the fact that not every cat was holding an apple in the picture was not a problem: it was enough that most of them were, so they freely said, "every cat is holding an apple."

Table 5: Type B errors in production by comprehension error type

Comp type	Mean	Std. Deviation	N
Type A&C	0	.	1
Type B	**4.5**	1.0	4
Target	0.8	1.9	25
Type C	1.4	2.2	13
Type A	0.0	0.0	2

In Table 6, we see children who made Type C errors in comprehension (including the one Type A & C child) making significantly more Type C errors in production than other types of responders.

Table 6: Type C errors in production by comprehension error type

Comp type	Mean	Std. Deviation	N
Type A&C	**2.5**	.	1
Type B	0.0	0.0	5
Target	0.0	0.0	25
Type C	**0.6**	1.6	13
Type A	0.0	0.0	2

4 Discussion

Children in the age range five through seven years, the major population studied in quantifier comprehension testing, apparently do not have full adult competence in producing sentences containing "every." Their mistakes in comprehension carry over to their production.

Here we will review the results in light of the different models proposed in the introduction. Consider first the DA. It predicts that errors of different types should occur in different children, not in the same child. The fact that we had little trouble classifying the children suggest this is so, with one overlapping case of Type A and C, as the DA would predict. The different grammars should be consistent across comprehension and production, and they are. That is, the same kinds of errors get produced by the same children in production as in comprehension, though with lower frequency. Furthermore, the developmental ordering seems roughly correct. On the DA, Type C errors should not

occur alone, but this is where the DA fails to predict the data as there are 13 such children.

As for the EQA, it predicts no relationship between exhaustive errors and other types, and that is true in the present data. Exhaustive errors should be consistent across production and comprehension, which they are. However, because they have different sources, different errors should be able to co-occur in the same child, but they rarely do in our data. They do not correlate as their sources are distinct in the EQA, which is confirmed. It is not clear that the EQA would predict the finding that the same errors occur in production as in comprehension for all types of error.

The RA also predicts no relationship between exhaustive errors and other types. It is not clear that exhaustive errors should be found in production, but they are found here. Again, the possibility that different errors can co-occur in the same child is not borne out, but they do not correlate, as predicted.

The WQA predicts that Type A and Type B errors should co-occur in comprehension and if they occur at all, in production too. Do these exhaustive and nonexhaustive errors pattern together? The answer is "no" - there is no correlation in comprehension and the errors are prevalent in different children at different ages. Nevertheless, it would be more clear if we had more Type A answers. Other errors should occur independently, and they do.

Roeper et al (this volume) argue for a comprehensive grammatical explanation (the DA) that would expect the errors to appear in production as well as comprehension. However, they propose a developmental progression beginning with both Type C and Type A errors, followed by only Type A. The very small number of classic Type A errors does not allow us to verify this claim. In addition, we found a significant number of children who made only Type C errors, which most of the literature would suggest is a somewhat unusual result. The DA predicts no stage at which Type C errors should occur in the absence of Type A errors, but we find 13 such responders. Although the DA cannot explain this fact, we look to a reason for why Type A and Type C did not *co*-occur more often.

We examine two potential explanations for this problem. First, there were 13 children who *did* make errors both on Type A and Type C, but because they also made errors on Type D, we excluded them as nay-sayers. Are we right in doing this, or should they count as representing the first stage in the DA?

One answer is that they are willful nay-sayers who do not have a logical reading for "every." If the children are really just *mindless* "nay-sayers," then they should be weak in *production* of Type B, where the adult answer really is "no," as well as the rest. In fact, they score very poorly in production of Type B. A univariate ANOVA shows this is a significant failure rate compared to the rest of the subjects ($F(1,54)=6.99$, $p<.01$)

A second answer about Type D errors is that it is a failure of accommodation or domain selection. One child gave us a possible explanation for the strategy when it came to production: he protested that you couldn't say "every cat" has an apple because in real life they don't. In other words, he judged the truth generically, and not with respect to the pictures.[1] Two children made consistent errors on Type D in production as well as saying no to them in comprehension. This phenomenon deserves more attention, as there is not only a logical requirement but also a pragmatic requirement in such experiments that the subject make an accommodation to the pictured context. However, this does not seem to be a consistent belief on the part of all the excluded children. The other eleven children excluded above because of Type D errors in comprehension nevertheless produced Type D correctly. It could be that they made an effort at compliance by the second day, when they came to realize that if they were to make the sentences, we meant them to focus only within the pictures. If this accommodation problem is *independent* of the other problems with quantifiers, then the 11 children who complied in production should show the same range of errors as the remainder of the group. However as we have seen, they do not: they make significantly more Type B errors.

Third, suppose that the children are really the kind who make Type A+Type C errors, because they are at that stage of the DA. The computer sentences would already violate their grammars 75% of the time, which might push them to find reasons to reject Type D as well. They then reject Type D for a somewhat unlikely reason such as the generic reading: cats don't all have apples in the world. This predicts that in production, these same children should be very prone to Type A and Type C errors. However, they do not make any more than the non-excluded children, in fact they do quite well (4.45/5 correct on each).

These facts lead us to reject the idea that the Type D nay-sayers are either accommodators-in-recovery or really Type A+Type C in disguise, and to admit that we do not know what these children think 'every' means. So we cannot count the nay-sayers as "lost" children from the first stage of the DA who make both Type A and Type C errors.

A second consideration is the fact that we found very few children who made classic Type A exhaustive paired answers, despite our attempts to a) sample the age range in which this has been reported as most prevalent and b) design stimuli that maximized attention to symmetry and c) provide little other pragmatic support and d) have no foils to break potential set effects (Philip, 2004; Rakhlin, 2005). In other words, we did everything that has been claimed might minimize the child's success and increase Type A errors, but without success. Why? Just possibly our stimuli might have broken symmetry by showing a table on which to rest the remaining object. Alternatively, perhaps our objects were just too small and non-salient to encourage the kind of refocusing that

[1] The fact that it was the younger children who did this suggests the error is a developmental one, not a smart-aleck response. One of us had a real-life experience with a precocious four year old that illustrates this point. She was trying to persuade the child to try sesame noodles, by using the argument that "Most people like them." The four year old responded quite seriously, "You haven't met most people."

Guerts and Philip predict leads to exhaustive pairing. Many such studies have included large animals as objects ("every boy rides on a horse") or sizeable inanimate objects (bikes, cars, ladders). However, Roeper & de Villiers (1993) had stimuli of boys drinking milkshakes, so it never occurred to us that the Type A error required large objects. If the presence of the table, and the size of the objects, are yet further contextual variables that influence the likelihood of Type A errors, this would have to be chalked up as a victory for the pragmatic side, i.e. RA or WQA, not the DA.

A significant number of children gave perfectionist or Type C errors. Those children who made spreading errors of either type in comprehension were likely to also make errors on the same stimuli in production. It cannot be estimated how many other true but qualified productions such as "well, every dog has an apple but one apple is on the table" might be due to the same discomfort with the unqualified "every" statements for Types A and C.

An unexpected group judged Type B sentences as true. These children might be rejected in comprehension studies as being "yes-bias" children, but our production results suggest at most of these same children are prone to the identical error in production: they use "every" for a majority but not an exhaustive group as in Type B. This is not an occasional error but a major form of response for these children. Prior work had suggested that children may consider "every" to be a plural, perhaps even a majority, but not necessarily an exhaustive majority. We contend that this be taken seriously as a stage in the development of the meaning of "every." The production data make it apparent that we should look with greater scrutiny at the concept of a "perseverator" whose data should be discarded. Geurts (2003) wrote "the determiner's lexical meaning is transparent enough, it is just the mapping from form to meaning that goes awry," but this may be too quick a conclusion.

In summary, these data are insufficient to decide among the different accounts, primarily because of the low number of Type A errors. The WQA is found wanting in respect of the clearly different nature of Type A and Type B errors. The DA made strong predictions that incorporated most response types under the grammar, instead of invoking other types of explanation for different responses. It correctly predicts the progression of stages, and the consistency of answers across comprehension and production, but it fails to account for the children who produce only Type C errors. We have suggested some "excuses" for this based on our stimuli, but those very excuses lean on the pragmatic factors that the RA and WQA invoke. So we cannot yet say with certainty whether our data support their developmental hypothesis. Furthermore, none of the different theories as they stand allow for either the "non-accommodators" or the "Type B/a lot of" children.

Like every researcher who embarks on a study of how children use "every," we now know every thing we wish we had done. However, we hope to have inspired researchers to take note of the potential for including production data in their models of how children learn "every" and other quantifiers.

References

Chomsky, N. (1965) *Aspects of the theory of syntax*. Cambridge, Ma: MIT press.

de Villiers, J.G., Cahillane, J & Altreuter, E. (2006). What can production tell us about Principle B? In B. Schwarz (ed) *Proceedings of the first GALANA conference, Hawaii*. U.Conn/MIT Working Papers in Linguistics.

Drozd, K & van Loosbroek, E. (2004). The effect of context on children's interpretation of universally quantified sentences. Ms.,University of Copenhagen.

Geurts, B. (2003). Quantifying kids. *Language Acquisition* 11: 197-218.

Merchant, G. (2005). Children's use of the quantifier "every". Lab report, Smith College.

Philip,W. (2005). *Event quantification in the acquisition of universal quantification*. Ph.D. dissertation, University of Massachusetts, Amherst, Ma.

Philip, W. (2004). Two theories of exhaustive pairing. Ms., Utrecht University.

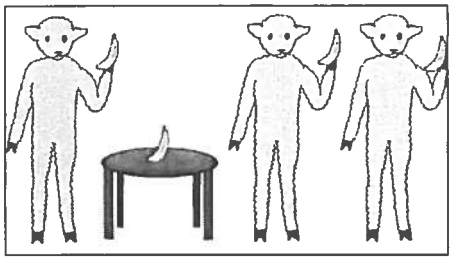
Rakhlin, N. (2005) New forms of Quantifier-Spreading in Acquisition. Acquisition Laboratory meeting presentation, University of Massachusetts, Amherst.

Roeper, T. & de Villiers, J.G. (1993). The emergence of bound variable structures. In E. Reuland & W. Abraham (eds) *Knowledge and Language: Orwell's Problem and Plato's Problem*. Dordrecht: Kluwer.

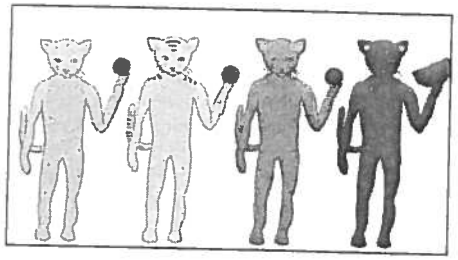
Roeper,T. Strauss, U. & Pearson, B. (this volume) The acquisition path of the determiner quantifier every: Two kinds of spreading.

Appendix A

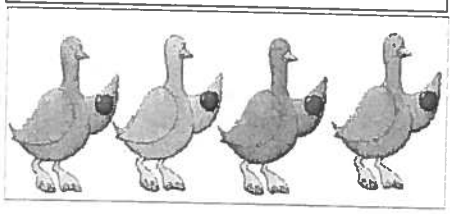
Type A: Every sheep has a banana.



Type B: Every cat has a cookie.

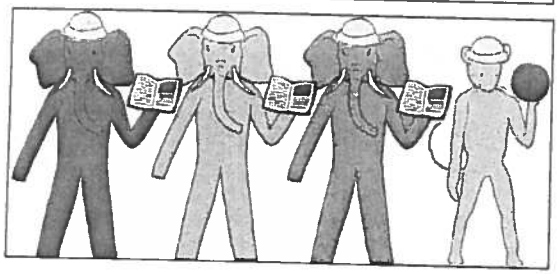


Type D: Every duck has an apple.

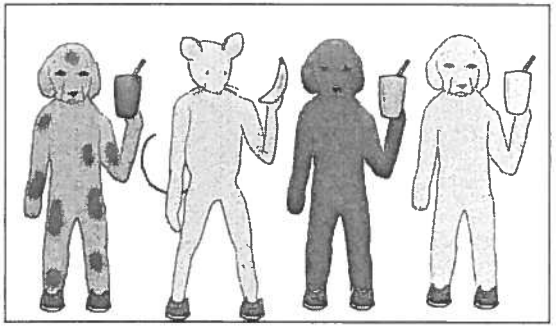


Appendix B

Type C in Production:
Every elephant has a book
OR
Every elephant has a hat
OR
Every animal has a hat



Type C in Production:
Every dog has a milkshake
OR
Every dog has a hat
OR
Every animal has a hat



Box 6067
Smith College
Northampton
Ma 01063

ealtreut@smith.edu

Department of Psychology
Smith College
Northampton
Ma 01063

jdevil@smith.eu