A large-scale investigation of scalar implicature

Petra Hendriks[1], John Hoeks[1], Helen de Hoop[2], Irene Krämer[2], Erik-Jan Smits[1], Jennifer Spenader[1] & Henriëtte de Swart[3]

[1]University of Groningen, [2]Radboud University Nijmegen, [3]Utrecht University

1. Introduction

Many utterances not only carry a literal meaning, but are often used with a related but stronger meaning. For example, the sentence *Some elephants have trunks* is literally true, but many people consider this sentence infelicitous because a more informative statement could have been used, namely *All elephants have trunks*. This meaning effect is caused by the quantifier *some*. If *some* is taken to literally mean 'at least one (and possibly all)' (Chierchia, 2004; Noveck, 2001), it should be possible to use *some* to describe a situation in which in fact all elephants have trunks. However, in everyday conversation language users often pragmatically strengthen the meaning of *some*, and use *some* with the meaning 'at least one but not all'. This stronger meaning is assumed to arise from a Gricean scalar implicature. Because listeners assume speakers to be co-operative, and because the speaker did not use a stronger form (*all*) on the same scale, the listener may draw the conclusion that apparently the speaker is not in a situation to use the stronger form, for example because the stronger form yields a false description of the situation. Thus the meaning of the weak scalar item *some* is pragmatically strengthened from 'at least one (and possibly all)' to 'at least one but not all'.

It has often been noted that children have difficulty accessing the pragmatically strengthened meaning of weak scalar items such as *some*. Instead, they interpret a sentence such as *Some elephants have trunks* logically, i.e., according to its literal meaning. For example, Noveck (2001) found that 7-8 and 10-11 year-old children were more likely to respond logically to this sentence than the adult controls did. On the other hand, Papafragou and Musolino (2003) and Guasti et al. (2005) showed that if an explicit training session is provided before the test session, children have less difficulty generating scalar implicatures. However, adding an explicit training session to the experiment does not tell us much about the ability to spontaneously generate scalar implicatures. Pouscoulous et al. (2007) found that children as young as 7 are able to spontaneously generate scalar implicatures at an adult rate if the task is made easier by using action-based judgments rather than verbal judgments, and by a particular choice of scalar expression. But what about children younger than 7? And what happens at a later age? According to Feeney et al. (2004), some adults develop the ability to inhibit a pragmatic response in favour of a logical response. If their explanation is on the right track, at what age does this ability to inhibit a pragmatic response arise? Or do some adults just not fully develop the ability to generate implicatures?

To answer these questions, an investigation of scalar implicatures is needed across a broad age range using the same test items for all age groups. For practical reasons, most studies of implicatures focus on only a few age groups and use a relatively small group of participants. In this study, however, we investigate the development of the ability to generate scalar implicatures across the entire age range. As we will discuss in more detail below, this is feasible because of the unique design of our experiment: as an unsupervised task on a personal computer in a museum, as part of a small temporary exhibition on communication. This design allowed us to gather data from participants across the entire life span. As we discovered later, an additional advantage was that it resulted in an unusually large number of participants (> 4,000). So our first research question is how the ability to generate scalar implicatures develops across age.

In our study, we focus on scalar implicatures in Dutch. Dutch allows us to study an important further issue, namely whether the referential properties of the scalar noun phrase (NP) are relevant for generating scalar implicatures. Some NPs are able to receive a referential (for example, a specific or partitive) reading as well as a non-referential (for example, an existential or predicative) reading. For example, the NP *two horses* can refer to two specific horses (the referential reading), but also to two arbitrary horses (the non-referential reading). Evidence that NP interpretation may be relevant for generating implicatures comes from a study on French by Pouscoulous et al. (2007). Pouscoulous et al. argue that children have more difficulty generating implicatures with French *certains* ('some') than with French *quelques* ('some'), because *certains* contains a notion of partitivity and hence is semantically more complex than *quelques*. In adult language use, this difference in processing difficulty has disappeared, they claim, but children are still sensitive to the difference in complexity. In Dutch, referential readings are dependent on syntactic position. Subjects in their canonical sentence-initial position preferably receive a referential reading, whereas subjects in sentence-internal position and objects in canonical position preferably receive a non-referential reading (de Hoop & Krämer, 2005/6). If the interpretation of the scalar NP is relevant in generating implicatures, we would expect to see differences with respect to the syntactic position of the scalar NP in Dutch. In particular, if referential NPs discourage scalar implicatures, we expect subjects in canonical position in Dutch to give rise to fewer implicatures than subjects in existential sentences and objects do, at least for children. So our second research question is whether the syntactic position of the scalar NP influences the number of implicatures being generated.

In section 2 we describe our three experiments, together with their results. Section 3 discusses the theoretical implications of our results. In section 3.1, we discuss the general pattern that was found, as well as the differences between the three lexical items studied. In section 3.2, we address the influence of syntactic position on scalar implicatures. Finally, in section 3.3 the developmental pattern we found is related to previous studies on the development of scalar implicatures.

2. The experiments

The experiments were part of an exhibition on communication. Because the exhibition was on display for almost a year, we were able to run three experiments as part of this exhibition. The first experiment was conducted at the University Museum Utrecht, and the second and third experiments at the University Museum Groningen. The experiments had exactly the same design and used the same picture stories, and only differed in the scalar item used. In Experiment 1 the Dutch numeral *twee* ('two') was investigated, in Experiment 2 the Dutch unstressed indefinite article *'n* ('a'), and in Experiment 3 the Dutch existential quantifier *enkele van de* ('some of the').

2.1 Materials and design

We used a modified version of the Truth Value Judgment Task (Crain & Thornton, 1998) with a 3 (NP position) x 2 (picture match/mismatch) design. The following three sentence types were used:

(1) *Subject-Initial*:

a. Twee konijnen kropen in de kast. (+ picture match scenario)

‘Two rabbits crawled into the cupboard’

b. Twee paarden sprongen over het hek. (+ picture mismatch scenario)

‘Two horses jumped over the fence’

(2) *Existential*:

a. Er sprongen twee kikkers in de vijver. (+ picture match scenario)

‘There jumped two frogs into the pond’

b. Er klommen twee poezen in de boom. (+ picture mismatch scenario)

‘There climbed two cats into the tree’

(3) *Object*:

a. De olifant gooide twee emmers om. (+ picture match scenario)

‘The elephant overturned two buckets’

b. De aap at twee appels. (+ picture mismatch scenario)

‘The monkey ate two apples’

These sentences all contain a scalar NP, in this case *twee N* (‘two N’), occurring in different positions in the sentence. Sentences (1a) and (1b) contain the subject in its canonical sentence-initial position. In the existential sentences (2a) and (2b), the expletive *er* (‘there’) occupies the sentence-initial position and therefore the subject follows the finite verb. In sentences (3a) and (3b), finally, the scalar item is the direct object of the verb.

Each of these sentence types was presented in a picture match scenario, where the sentence gave a fully informative description of the last picture in a picture story, as well as in a picture mismatch scenario, where the sentence gave an under-informative description of the last picture. The picture stories consist of three pictures each. Each story is introduced by a statement such as “Look, three rabbits” or “Look, an elephant” above the pictures. Below the first and second pictures of the story, a one-sentence description is given of the situation in the picture. Below the third picture Bennie the sheep’s utterance, describing the situation in the third picture, is shown in a text balloon. In Experiment 1, Bennie’s utterance consists of one of the 6 sentences in (1)-(3) and thus refers to two entities performing or undergoing an action. In a picture match scenario, the third picture shows *two* entities performing or undergoing an action (see Figure 1). The target answer to these items is “yes”. In a picture mismatch situation, the third picture shows *three* entities performing or undergoing an action (see Figure 2). Depending on whether participants generate an implicature or not, the expected answer is “no” (corresponding to a pragmatic response, i.e., an implicature) or “yes” (corresponding to a logical response, i.e., no implicature).

Figure 1: A test item in a picture match scenario. Text below picture 1: The rabbits are sitting next to the cupboard. Text below picture 2: They want to crawl into it. Sentence uttered by Bennie the sheep: Two rabbits crawled into the cupboard (= sentence (1a)).

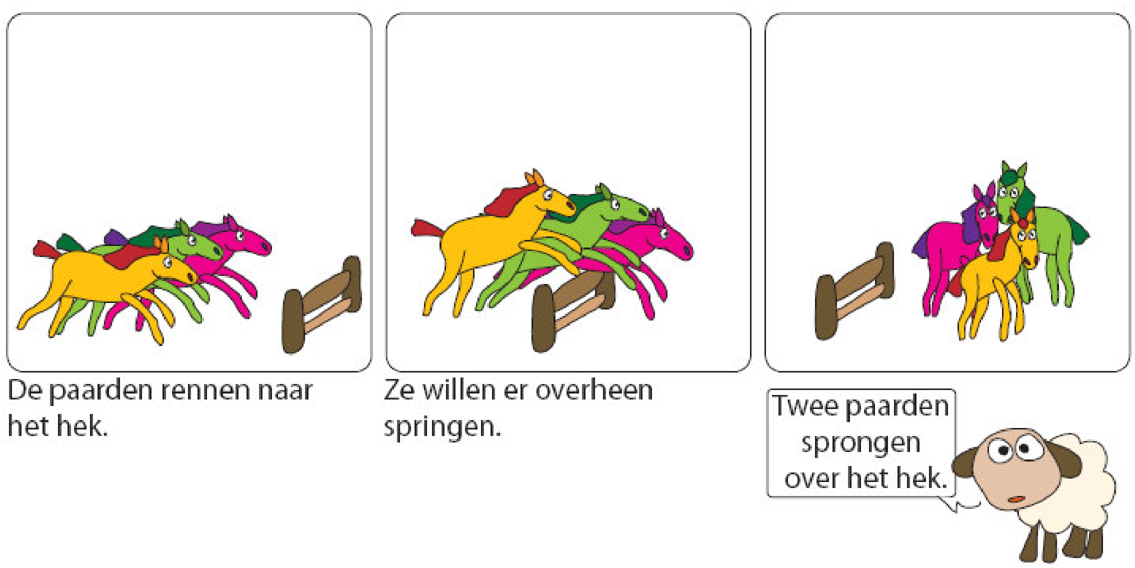

Figure 2: A test item in a picture mismatch scenario. Text below picture 1: The horses are running towards the fence. Text below picture 2: They want to jump

over it. Sentence uttered by Bennie the sheep: Two horses jumped over the fence (= sentence (1b)).

In addition to these 6 test items, we included 1 control item that gave a truth-conditionally false description of the series of pictures:

(4) Drie honden gingen in het hok liggen. (+ picture mismatch scenario)
'three dogs went into the doghouse'

The series of pictures for the control item show two dogs going into the doghouse and one dog remaining outside. The target answer therefore is "no".

We tested the sentence types in (1)-(3) using the Dutch numeral *twee* ('two') in Experiment 1, the Dutch unstressed indefinite article *'n* ('a') in Experiment 2, and the Dutch quantifier *enkele van de* ('some of the') in Experiment 3. Although the three experiments differed in the scalar items used, the same sentence types and picture stories were used across experiments. So the picture stories in Figure 1 and Figure 2 were also used in Experiments 2 and 3, with the items *'n* and *enkele van de* replacing the numeral *twee* in the test sentence. Besides the existential quantifier *enkele* ('some') that we used in Experiment 3, Dutch also has a partitive quantifier *sommige* ('some'). However, because *sommige* is infelicitous in existential sentences such as (2a) and (2b) and we chose to keep the experimental design constant across experiments, *sommige (van de)* could not be used in our experiments.

The experiments were conducted unsupervised on a personal computer with a touch screen. The brief instructions were presented as written text on the computer screen. If a child participant could not yet read, we expected (based on observations in the pilot phase of the experiment) that parents or older siblings would read the text aloud to the child. The personal computer for the experiment was located in the same room as the other components of the temporary exhibition on communication, which formed only a small subset of the entire collection of the museum.

The experimental session started with the introduction of a cartoon character named Bennie. Bennie was introduced as a sheep that could talk but was very stupid. We expected that our emphasis on the sheep's stupidity would make it easier (especially for children) to give "no" answers (corresponding to implicatures in the under-informative conditions). The session continued with 2 practice items not involving any scalar items, one with a target "yes" answer ("The cat is green") and the other with a target "no" answer ("The panda bear is eating an ice cream"). Participants were asked to judge whether the sentence uttered by Bennie matched a single picture or not. They had to press a button on the bottom of the screen to enter their response (yes/no) and then received feedback about the correct response on the practice item. Next, participants were instructed to enter their age and gender. When participants had done so, the test session started. The 6 test items and 1 control item were presented in random order. Each participant received the same 2 practice items, 6 test items and 1 control item. The entire session lasted only a few minutes, which was a very stringent restriction on the design of the experiment. If the session would have lasted longer, we may have lost many more participants than we did now.

2.2 Participants

Experiment 1 was conducted at the University Museum Utrecht, and Experiment 2 and Experiment 3 at the University Museum Groningen. The participants in the three experiments were regular paying visitors of the museum. In addition, several school classes from local elementary schools visited the museum and participated in the experiments.

In total 4,090 participants took part in the three experiments reported here (Experiment 1: N=2,549; Experiment 2: N=961; Experiment 3: N=580). To facilitate the analysis of data from so many participants, we grouped together people with ages that were not more than 4 years apart, such that the first group contained all participants from age 5 up to and including age 9, the second group contained all participants from age 10 up to and including age 14, and so on, with the final group 12 containing all participants aged 60 up to and including 64 (see Table 1).

Age group	1	2	3	4	5	6	7	8	9	10	11	12
Age range (in years)	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64

Table 1: Age groups in the experiment

Participants were included if they met a number of criteria: 1) they had given a response on all seven items; 2) they had given the correct response on the control item; and 3) the age group to which they belonged consisted of 25 or more participants. On the basis of the third criterion, we excluded from further analysis participants between the ages of 1-4 and 65-100. We did not use the two practice items to remove any participants, nor did we analyze the results of the practice items or the control item. The practice items were meant solely to familiarize the participants with the task. Because the experiment was conducted unsupervised, the control item was our only means to exclude participants who failed to understand the task or did not take the task seriously.

	Age group												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Exp. 1	405	451	241	233	157	116	97	164	154	141	74	60	2293
Exp. 2	99	140	138	111	69	26	29	37	70	48	38	26	831
Exp. 3	31	58	50	88	61	36	29	28	43	43	31	25	523
Total	535	649	429	432	287	178	155	229	267	232	143	111	3647

Table 2: Number of participants per experiment and per age group

Thus, the data from 3,647 people were included in the actual statistical analysis (Experiment 1: N=2,293; Experiment 2: N=831; Experiment 3: N=523), in which each experiment was analyzed separately.

2.3 Results

Statistical analysis comprised a number of steps. First, for each experiment separately, we determined whether there were significant differences between the three conditions *Subject-Initial*, *Existential*, and *Object* in any of the age groups. To this end we determined Cochran's Q (which is suited for the analysis of within-subject nominal data, Field (2005)) for each of the twelve age classes, controlling for a possible increase in Type-I error due to multiple testing by using Bonferroni correction. Surprisingly, there were no significant differences between these conditions in any of the three experiments. This suggested that the three conditions behaved rather similarly, and indeed, reliability analysis showed that the three items used to measure each of the three conditions respectively were highly interrelated in each of the three experiments (Cronbach's alpha was .91 for Experiment 1, .93 for Experiment 2, and .76 for Experiment 3). This led us to treat these items as a three-item measurement scale, measuring the degree to which participants are inclined to choose either a logical or a pragmatic response to an under-informative statement.

For each participant, then, we calculated the proportion of logical responses over the three under-informative items, and also over the three fully informative items. These proportions were then entered into two ANOVAs. First, to determine whether participants responded differently to the fully informative items than they did to the under-informative items, we conducted a Repeated Measures ANOVA with Informativity (informative vs under-informative) as a within-subjects factor. Secondly, as we are mainly interested in the pattern of results of the *under-informative* items, we conducted a univariate ANOVA, with Age Class as the between-subjects factor (in both of these analyses we actually used the arcsine transformation of the square root of each proportion to guard against possible deviations from the normal distribution).

Let us start with the first ANOVA, concerning the possible differences between proportion of logical responses in informative versus under-informative items. Please note that only in Experiments 1 and 3 we are actually able to distinguish between informative versus under-informative items. In Experiment 2, the sentence *'n paard sprong over het hek* ('a horse jumped over the fence') would be under-informative both in a situation in which three horses jumped over the fence and in a situation in which two horses jumped over the fence. In Experiment 1, we found a large main effect of informativity ($F(1,2292)=9416.1$; $p<.001$). On average, participants gave a pragmatic response by rejecting under-informative statements with the numeral *twee* ('two') in the majority of the cases. They gave a logical (yes) response on these under-informative statements in only 15% of the cases. In contrast, fully informative statements with *twee* were accepted in 91% of the cases. A similar pattern was found in Experiment 3 with the quantifier *enkele van de* ('some of the'): Participants rejected under-informative items with *enkele van de* in the majority of the cases. They gave a logical (yes) response on these under-informative statements in only 14% of the cases, whereas they accepted fully informative statements with *enkele van de* in 92% of the cases ($F(1,522)=1820.7$; $p<.001$). In contrast to Experiments 1 and 3, we found only a small difference between the two items in Experiment 2. This was as expected, because in fact both items were under-informative with respect to the scenario presented in the pictures. Looking at the items that were considered under-informative in the other two experiments (i.e., with pictures showing three entities performing or undergoing an action), statements with the indefinite unstressed article *'n* ('a') were accepted in 46% of the cases. With respect to the items that were considered informative in the other two experiments, but are in fact also under-informative in Experiment 2 (i.e., with pictures showing two entities performing or undergoing an action), statements with *'n* were accepted in 42% of the cases. The fact that this difference was nevertheless significant ($F(1,830)=21.2$; $p<.001$) is hard to interpret. Perhaps it has something to do with the perceived contrast between the weaker and the stronger form on a scale: The larger the contrast between the weaker form and the stronger form, the higher may be the chances that an implicature is generated with the weaker form. This explanation however is dependent on the assumption that a stronger contrast exists between a singular indefinite and the plural *two* than between a singular indefinite and the plural *three*. We leave this for further study.

In our main analysis we focused on the responses to the three items that were under-informative in each of the three experiments, and to which participants could respond either logically or pragmatically. We already established (see above) that there were no significant differences between the three conditions Subject-Initial, Existential, and Object, so we entered the proportion of logical responses (i.e., over the three items that were answered by each participant) as a dependent variable into a univariate ANOVA, with Age Class as the between-subjects factor. We found significant effects of Age Class in the first two experiments (Experiment 1: $F(11,2281)=5.88$; $p<.001$; Experiment 2: $F(11,819)=4.33$; $p<.001$), but not in Experiment 3: $F(11,511)=1.1$; $p=.369$). As for Experiment 1, visual inspection of the data suggested a rise in logical responses from age group 1 onwards, reaching a kind of plateau at age group 4, with a sudden decline after age group 10 (see Figure 3). This pattern was input into a statistical contrast analysis, which turned out to be highly significant ($F_{contrast}(1, 2281)=30.34$; $p<.0001$). This analysis was supported by the results of the posthoc pairwise comparisons (see Table 3).

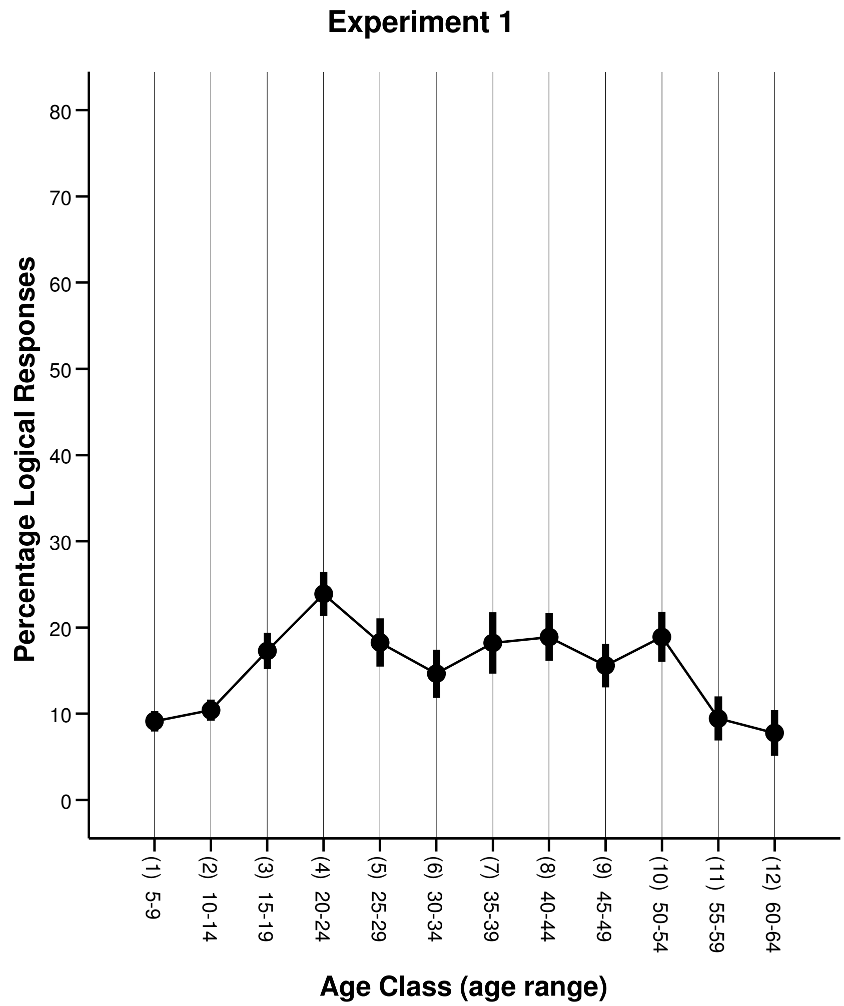

Figure 3: Percentage of logical responses in Experiment 1 (*twee*)

Experiment 1		Experiment 2	
1 vs	<u>3</u>, <u>4</u>, <u>5</u>, 6, 7, <u>8</u>, 9, <u>10</u>	1 vs	3, 4, 5, 6, <u>7</u>, 8, 9
2 vs	3, <u>4</u> ,5, 7, <u>8</u>, 9, 10	2 vs	<u>3</u>, <u>4</u>, <u>5</u>, 6, <u>7</u>, 8, <u>9</u>, 10, 11, 12
3 vs	<u>1</u>, 2, 4, 11, 12	3 vs	1, <u>2</u>, 7
4 vs	<u>1</u>, <u>2</u>, 3, 5, 6, 7, 8, 9, <u>11</u>, <u>12</u>	4 vs	1, <u>2</u>
5 vs	<u>1</u>, 2, 4, 11, 12	5 vs	1, <u>2</u>
6 vs	1, 4	6 vs	1, 2
7 vs	1, 2, 4, 11, 12	7 vs	<u>1</u>, <u>2</u>, 3, 9, 10, 11, 12
8 vs	<u>1</u>, <u>2</u>, 4, 11, 12	8 vs	1, 2
9 vs	1, 2, 4, 12	9 vs	1, <u>2</u>, 7

10 vs	1, 2, 11, 12	10 vs	2, 7
11 vs	3, 4, 5, 7, 8, 10	11 vs	2, 7
12 vs	3, 4, 5, 7, 8, 9, 10	12 vs	2, 7

Table 3. Significant results of the pairwise comparisons between age groups in Experiments 1 and 2. Only results that are significant at α = .10 are presented, both for tests with (underlined numbers) and without Bonferroni correction for multiple testing.

Experiment 2 showed an even more striking pattern of results: After a (non-significant) decline from age group 1 to age group 2, there was a steep increase in the percentage of logical responses ending up to and including age group 7, followed by a steep decrease in logical responses returning to the level of logical responding seen in age group 2 (see Figure 4). Again, this visual pattern was strongly supported by the matching contrast analysis ($F_{contrast}(1, 819)=52.27$; $p<.0001$). This analysis was supported by the results of the posthoc pairwise comparisons (see Table 3).

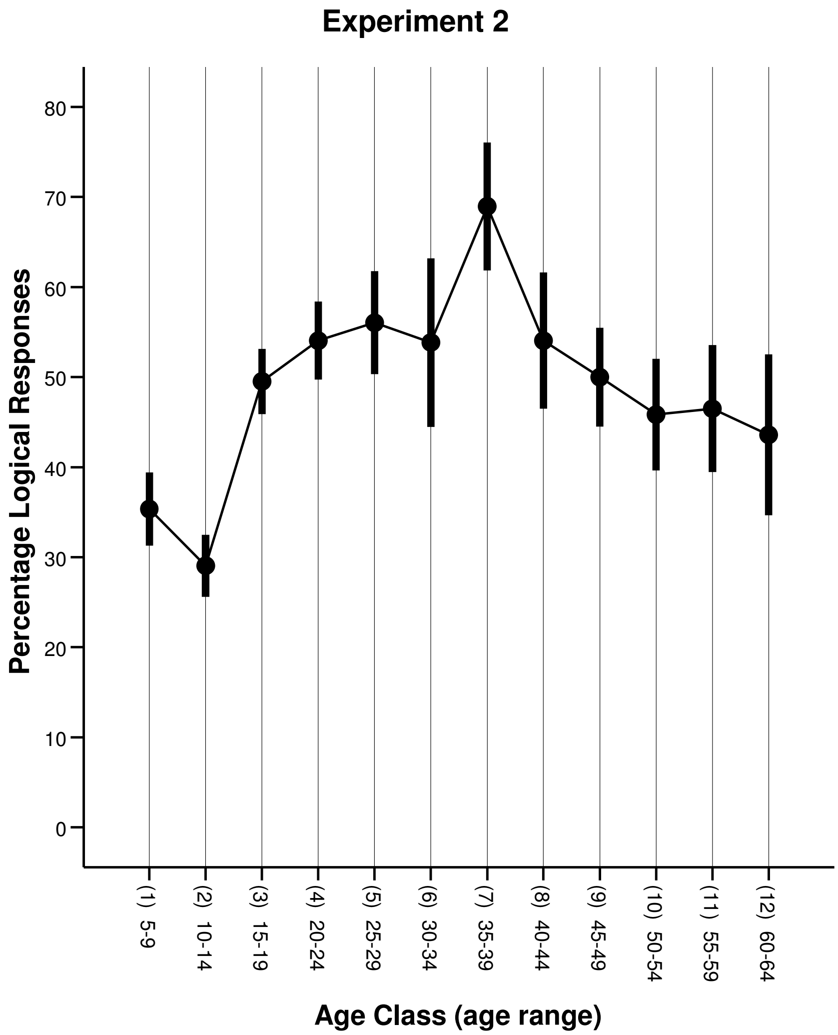

Figure 4: Percentage of logical responses in Experiment 2 (‘*n*)

Figure 5 gives the results of Experiment 3. As mentioned above, we did not find a significant effect of Age Class in this experiment.

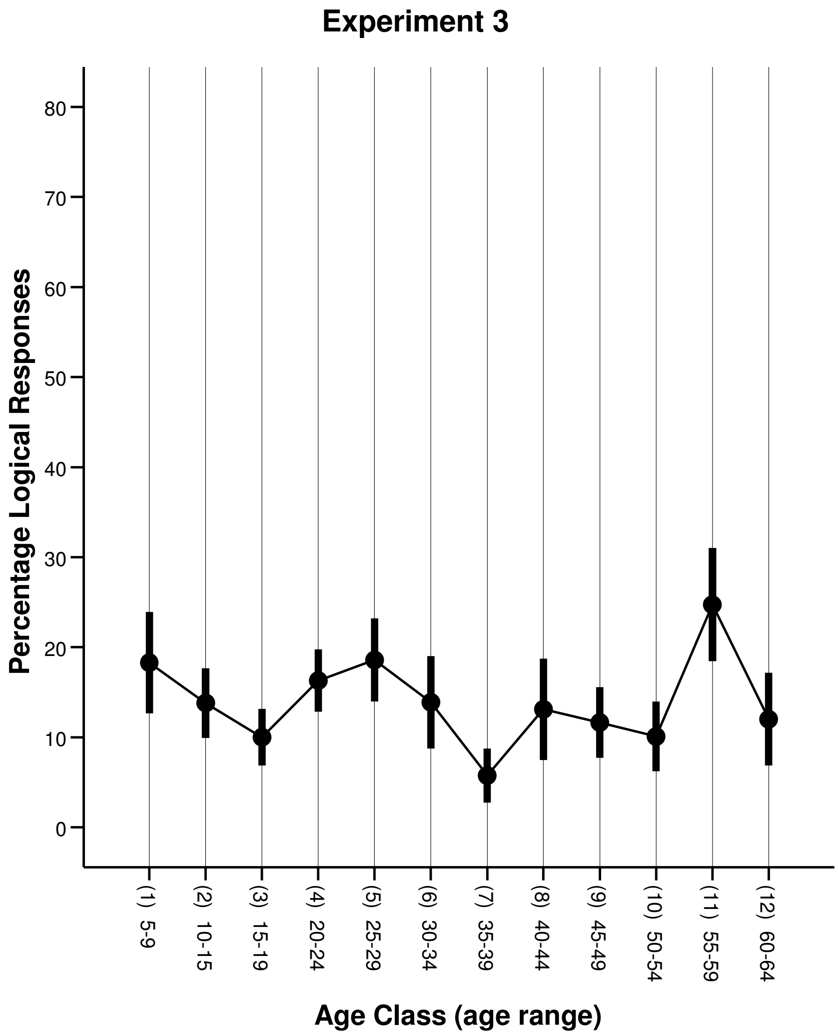

Figure 5: Percentage of logical responses in Experiment 3 (*enkele van de*)

In the next section, we will discuss the most important theoretical implications of the results presented in this section.

3. General discussion

Our study had two aims. First of all, we wanted to know how the ability to generate scalar implicatures develops across age. The second question we aimed to answer was whether the syntactic position of the scalar NP influences the number of implicatures being generated. We address the second question regarding the syntactic position of the scalar NP in section 3.2, and the first question regarding the development of scalar implicatures in section 3.3. But first, we turn to a more general issue and consider the question whether we have indeed found evidence of scalar implicatures in our experiments.

3.1 Evidence of scalar implicatures in Dutch

In Experiment 1 (with the numeral *twee* ('two')) and Experiment 3 (with the quantifier *enkele van de* ('some of the')), but not in Experiment 2 (with the indefinite article *'n* ('a')), we found a large main effect of informativity. Under-informative items were rejected in the majority of cases in these two experiments (85% in Experiment 1 and 86% in Experiment 3), suggesting that participants were generating scalar implicatures. Although the under-informative statements were truth-conditionally correct descriptions of the situation presented in the pictures, participants overwhelmingly rejected these statements. This may have been caused by the fact that the stronger form on the scale would have formed a more appropriate description of the situation: *three* on the scale <*two*, *three*> in Experiment 1, and *all* on the scale <*some*, *all*> in Experiment 3. In Experiment 2, in contrast, there appeared to be no clear consensus among participants about the availability of such a scale. Since participants' responses were equivocal in Experiment 2, some participants may have interpreted the indefinite as the weak expression on a scale with the plural as the stronger expression, thus generating an implicature, whereas others may have interpreted the indefinite as merely existential.

With respect to numerals such as *two*, a number of recent studies have suggested that they do not give rise to implicatures (see Breheny, 2005, for an overview). As Hurewitz et al. (2006) show, 3-year-olds are already able to assign a strong 'exact' interpretation to numerals, but the same children fail to assign a strong interpretation to the quantifier *some*. However, the pattern we find in Experiment 1 with *twee* is similar to that in Experiment 3 with *enkele van de*. Moreover, as will be discussed in more detail in section 3.3, our adult participants do not reject the under-informative use of *two* at a rate close to 100%, which is to be expected if *two N* yields a truth-conditionally false statement in a scenario with three entities. Rather, we find a very specific pattern of responses in adults. Therefore, we will continue to talk about the interpretation of numerals as involving implicatures.

Participants in our experiment generally generated implicatures at a relatively high level, namely around 85% in Experiments 1 and 3. This could be due to either the kind of participants in our study, or the design of the experiments. Our participants were regular paying visitors of the museum, who may have been more co-operative than the standard undergraduate students participating in many other experiments. In addition, our experiments differed from those of Noveck (2001) and Feeney et al. (2004) in that we did not use statements that required world knowledge for their evaluation (such as *Some elephants have trunks*). Rather, the statements in our experiments had to be evaluated on the basis of a picture story. The entire picture story was visible on the computer screen while the statement was being read or heard, so participants did not have to rely on their short or long term memory to evaluate the statement. This undoubtedly will have made the experiment easier for the child participants.

Crucially, the participants in our experiments did not receive any training in dealing with under-informative sentences, in contrast to the participants in the experiments of Papafragou and Musolino (2003) and Guasti et al. (2005). The only training our participants received consisted in two practice items targeting truth-conditional aspects of meaning. However, this did not prevent them from generating scalar implicatures with the majority of under-informative items. Interestingly, since the picture stories simply described a particular action and did not place the participants in a position where they had to detect whether the speaker was trying to intentionally deceive the listener, there was no clear relevance of the implicature for the listener. Nevertheless, the level of pragmatic responses was quite high.

3.2 The influence of syntactic position on scalar implicatures

In their third of a series of three experiments, an action-based judgment task, Pouscoulous et al. (2007) found that 9-year-old French children were more likely to generate implicatures with *quelques* (0% logical response) than with *certains* (42% logical response). Their adult controls were not affected by the choice of item (7% vs 21% logical response). Pouscoulous et al. attribute the different performance of children with respect to the two scalar items (both meaning 'some') to the fact that *quelques* is a simple existential, while *certains* is partitive. As a result, they claim, *certains* is more complex than *quelques*. Although children understand the meaning of both lexical items, according to Pouscoulous et al. the added processing cost of *certains* makes the task of generating an implicature harder, thus reducing children's rate of implicature production. With respect to adults, they contend: "If anything, French adult native speakers would tend to make more implicatures with *certains* than with *quelques* precisely because it is a partitive and therefore raises the salience of a larger set."

One of our research questions was whether the syntactic position of the scalar NP influences the number of scalar implicatures being generated. If referential (e.g., partitive) NPs discourage implicature production, we expect subjects in canonical position in Dutch to give rise to fewer implicatures than subjects in existential sentences and objects. This is because, in Dutch, subjects in canonical position preferably receive a referential interpretation, whereas subjects in existential sentences and objects in canonical position preferably receive a non-referential interpretation (de Hoop & Krämer, 2005/6). However, our results showed no significant differences between conditions in any of the three experiments and in any of the age groups. This means that, contrary to our expectations, the syntactic position of the scalar NP, and hence the referential status of the scalar NP, does not appear to influence the rate of implicatures. Even in the youngest age group in our study, children between 5 and 9 years old, syntactic position of the scalar NP did not seem to matter much: In a post-hoc analysis we only found a significant effect of syntactic position for the five-year olds (Cochran's $Q(2)=10.33$; $p<.05$, $N=35$); in the other groups there were no significant effects of condition ($p>.12$). The five-year olds gave significantly more logical responses in the Object condition (Mean=20%; SD=41%), than in the Subject-Initial (Mean=6%; SD=24%; $Q(1)=5.00$; $p<.05$)) or the Existential condition (Mean=3%; SD=17%; $Q(1)=6.00$; $p<.05$); the latter two conditions did not differ significantly ($p>.30$). These results strongly suggest that referential status of the scalar NP is not relevant for generating implicatures, even for the youngest age group in our study. Because the five-year olds did not give more logical responses in the Existential condition compared to the Subject-Initial condition, their larger amount of logical responses in the Object condition cannot be attributed to the referential status of the scalar NP. Why generating implicatures is nevertheless harder for young children if the scalar item occurs in object position remains to be seen.

If referential status of the scalar NP is not relevant for generating scalar implicatures, what could be the explanation for the difference Pouscoulous et al. observe between partitive *certains* and non-partitive *quelques*? One possible explanation is frequency. Apart from the notion of partitivity, *certains* also differs from *quelques* in that it is less frequent in children's written production as well as in children's books (see Pouscoulous et al. for a discussion of this observation). Perhaps the relative frequency of the two scalar items in the language output of the child or the language input to the child causes the two items to behave differently with respect to implicatures. It is conceivable that more frequent items are more likely to give rise to an implicature than less frequent items, for example because their literal meaning can be accessed faster, or because their pragmatically strengthened meaning has already been used more often and is therefore computed more easily.

Another possible explanation that may be worth looking into is the felicity of a partitive interpretation in the context of the task. Pouscoulous et al. employed an action-based judgment task, where participants saw a number of cardboard boxes, some of them containing a token and others empty. Consider the scenario where each box contained a token. A puppet would then say: "I would like some boxes to contain a token". The participant's response could be to leave the boxes unchanged (the logical response in this scenario), or to remove one or more tokens (the pragmatic response in this scenario). To use *certains* in a felicitous way with a partitive reading, the listener should be able to determine which subset of a contextually salient larger set the speaker is referring to by the expression *certains N*. This subset should be identifiable by means of a particular property that its elements share, for example the property of being large or the property of being green (see de Hoop, 1995, for a discussion of Dutch partitive *sommige*, which seems to be related to Dutch non-partitive *enkele* in the same way as French partitive *certains* to French non-partitive *quelques*). However, nowhere in their description of the experiment do the authors mention that it is possible to identify a subset of the set of four cardboard boxes by means of, e.g., their size or colour. Apparently, the four cardboard boxes are completely identical. This makes it impossible to determine the relevant subset required to assign a partitive interpretation to *certains N*. If adult interpretation is somewhat more robust than children's interpretation, the infelicitous use of *certains* in the test item will have had a smaller impact on their responses. Note that this problem does not arise in a verbal judgment task like ours, where the relevant subset is identifiable by simply looking at the pictures. For example, if the sentence *Some rabbits crawled into the cupboard* is uttered in a scenario where two of the three rabbits are inside the cupboard and one of them is outside the cupboard (see Figure 1), the relevant subset that the expression *some rabbits* refers to is the set of rabbits that share the property of being inside the cupboard in the third picture. In addition, the elements in the set referred to by the scalar NP were given distinct colours in the pictures. Thus the relevant subset may also be identified by means of the colour of its elements.

3.3 The development of scalar implicatures

Our second main research question was how the ability to generate scalar implicatures develops across age. The results of Experiment 1 (with the scalar item *twee*) and Experiment

3 (with the scalar item *enkele van de*) show that even our youngest group of children (5-9 years old) is already sensitive to scalar implicatures at an adult level. On average, participants responded with an implicature in 85% of the cases in Experiment 1, and in 86% of the cases in Experiment 3.

Focusing on the responses within the youngest age group in Experiment 1, where we had a sufficiently large set of participants of different ages to compare 1-year subgroups, we did not find a developmental effect between the 5-, 6-, 7-, 8-, and 9-year-olds: 5-year-olds did not produce more logical responses than 9-year-olds. Furthermore, the children in the youngest age group did not produce a lower rate of pragmatic responses than the adults in our experiments did. On the contrary, what we observe in the data of Experiment 1 is a gradual increase in *logical* responses, starting around age 9 and continuing until the age of 20-24 years old. After that age, the rate of logical responses to statements with *twee* remains more or less constant, until it drops again after the age of 50-54.

Although the development of the younger participants in our experiments is in line with earlier results, the observed development of participants from age 9 seems incompatible with a traditional two-stage model. According to the two-stage model, children start out with a preference for logical responses, and gradually develop the pragmatic ability to generate scalar implicatures until an adult level is reached. Previous studies have found that, although 4-year-olds still frequently give logical responses to under-informative sentences, children from the age of 7 are already able to produce implicatures at an adult level if the task is made easy (Feeney et al., 2004; Guasti et al., 2005; Pouscoulous et al., 2007). Moreover, 5-year-olds already produce implicatures with regularity. Assuming that the task in our experiments was sufficiently easy, our results with respect to the youngest age group are compatible with previous findings.

Our results with respect to the older age groups, however, seem incompatible with a traditional two-stage model of the development of scalar implicature. According to the two-stage model, when children have mastered the ability to generate scalar implicatures, they will do so. But Feeney et al. (2004) argue that some adults develop the ability to inhibit a pragmatic response in favour of a logical response. They base their claim on the observation in their Experiment 3 that adults' logical responses to infelicitous *some* take significantly longer to make than their logical responses to felicitous *some*. This suggests that the logical response to infelicitous *some* is accompanied by additional cognitive processing. Feeney et al. argue that this additional cognitive processing includes generating the implicature as well as subsequent inhibition of the implicature. Thus their model assumes that people start out with a logical interpretation, which can be strengthened into a pragmatic interpretation, which can then be inhibited to yield a logical interpretation again. Support for their three-stage model of the development of scalar implicature comes from the observation that the adults in their study (undergraduates) that tended to respond logically to infelicitous *some* generally also scored higher on a counting span task. If inhibition of a pragmatic response requires additional processing costs, it is predicted that the capacity to do so is positively correlated with memory span or other cognitive measures.

Since the participants in the youngest age group in our Experiment 1 were generally able to give pragmatic responses with the item *twee*, it seems plausible to assume that the participants in age group 4 (the 20-24-year-olds) are also able to do so. Why then do we find an increase in the number of logical responses from age 9 to age 24? This observed increase in the number of logical responses fits in with the three-stage model of Feeney et al.: Perhaps several of the adolescents and young adults in our study inhibited a pragmatic response in favor of a logical response. Since memory span decreases again with age, this would also explain the drop in logical responses again after age 50-54. Interestingly, we see a similar developmental pattern, but more pronounced and with a later peak at around age 35-39, in Experiment 2 with the indefinite article *'n*. The results of Experiment 3 with the quantifier *enkele van de* also support the general developmental pattern, although we cannot draw solid conclusions from the data in this experiment because of the relatively small number of participants.

So perhaps adults sometimes inhibit a pragmatic response in favour of a logical response. Alternatively, it may be that several of the adolescents and young adults in our study simply did not compute a scalar implicature. Because of their familiarity with tests and test situations in their daily (educational) life, it is conceivable that they viewed the experiment as a typical test situation where a logical response is required. However, this would not immediately explain why the drop in logical responses for the item *twee* only starts after age 50. A clear disadvantage of an off-line task measuring yes/no responses only, as the one we employed in our study, is that this issue cannot be settled. The only way to distinguish between a logical interpretation as the result of not computing a pragmatic interpretation and a logical interpretation as the result of inhibition of a pragmatic interpretation is by using on-line measures.

4. Conclusions

In our large-scale study of scalar items in Dutch as an unsupervised experiment in two university museums, which attracted over 4,000 participants, we found that participants accepted under-informative statements with the numeral *twee* ('two') in only 15% of the cases, and with the existential quantifier *enkele van de* ('some of the') in only 14% of the cases. This suggests that participants generally tended to generate scalar implicatures with these items. In contrast, participants' responses with the unstressed indefinite article *'n* ('a') were equivocal. In this study, we addressed two questions: (i) how does the ability to generate scalar implicatures develop across age?, and (ii) does the syntactic position of the scalar NP influences the rate of implicatures? Children's performance with the three lexical items was adult-like already from the youngest age group of 5-year-olds. We found no increase of pragmatic responses with age. We did find a *decrease* of pragmatic responses with age with the numeral *twee*, however, starting from the age of 9. This decrease of pragmatic responses continued until the age of 24, and may support a three-stage model of the development of implicatures. Scalar implicatures appeared not to be influenced by the syntactic position of the scalar NP in Dutch. As we argued, this implies that the referential status of the scalar NP, and whether the scalar NP receives a partitive interpretation or not, does not have any effect on the rate of implicatures.

Acknowledgments

We would like to thank the Netherlands Organisation for Scientific Research (Cognition programme), CLCG and BCN at the University of Groningen, CLS at the Radboud University Nijmegen, and UiL OTS at Utrecht University, for sponsoring the exhibition "Bent u een goed verstaander?". Furthermore, we thank the University Museum Utrecht and the University Museum Groningen for their cooperation, all participants in the experiments for their willingness to participate in our experiments, and the audience of the Experimental Pragmatics 2007 conference in Berlin and an anonymous reviewer for this volume for their helpful suggestions.

References

Breheny, R. (2005). Some scalar implicatures really aren't quantity implicatures – but *some*'s are. *Proceedings of Sinn und Bedeutung* 9, 57-71.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (ed.), *Structures and beyond*. Oxford University Press, 39-103.

Crain, S. & R. Thornton (1998). *Investigations in Universal Grammar. A Guide to Experiments on the Acquisition of Syntax and Semantics*. Cambridge, MA: MIT Press.

De Hoop, H. (1995). On the characterization of the weak-strong distinction. In E. Bach, E. Jelinek, A. Kratzer, and B.H. Partee (eds.), *Quantification in Natural Languages*. Vol. 2. Kluwer, Dordrecht, 421-450.

De Hoop, H. & I. Krämer (2005/6). Children's optimal interpretations of indefinite subjects and objects. *Language Acquisition* 13, 103-123.

Feeney, A., S. Scrafton, A. Duckworth, & S. Handley (2004). The story of *some*: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, 58:2, 121-132.

Field, A. P. (2005). *Discovering statistics using SPSS* (Second Edition). London: Sage.

Guasti, M.T., G. Chierchia, S. Crain, F. Foppolo, A. Gualmini, & A. Meroni (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes* 20, 667-696.

Hurewitz, F., A. Papafragou, L. Gleitman, & R. Gelman (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development* 2, 77-96.

Noveck, I.A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition* 73, 165-188.

Papafragou, A. & J. Musolino (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition* 86, 253-282.

Pouscoulous, N., I. Noveck, G. Politzer, & A. Bastide (2007). Processing costs and implicature development. *Language Acquisition* 14:4, 347-376.