The investigation of quantity implicatures during typical development: a systematic review

Anna Teresa Porrini & Luca Surian*

Abstract. The present work is a systematic review of the acquisition of quantity implicatures in typically-developing children. The references were selected through the PRISMA method. The criteria for eligibility were that the articles should be peer-reviewed, published articles written in English, containing empirical data on the comprehension of quantity implicatures in first language acquisition during typical development. The aim of this review is three-fold. First, to provide a picture of what empirical data tell us about the acquisition of quantity implicatures, based on both lexical and ad-hoc scales, potentially contributing to theoretical accounts of the phenomenon. Second, to analyze the methodologies that have been used to test children and their adequacy. Lastly, to evaluate whether systematic review is an accurate analysis method for this type of varied and often complicated data. The results suggest that children improve in implicature derivation with age, especially with lexical scales, and that action-based tasks not based on meta-linguistic evaluations might be better suited to test these inferences, especially as opposed to Truth Value Judgment tasks. The fact that the systematic analysis confirms previously individuated trends in the acquisition of implicatures confirms that this is in fact a useful methodology to analyze the data, despite some limitations.

Keywords. Developmental pragmatics; implicatures; quantity maxim; systematic review.

1. Introduction. Quantity implicatures are enrichments on the meaning of an utterance derived by appealing to the Gricean maxim of Quantity (Grice 1975), which states that, when in conversation, one should make their contribution as informative as is required for the current purposes of the exchange, and not more. Quantity implicatures arise by the use of linguistic items often called scalar items, which can be situated within lexical scales – such as the <some/all> scale but also the <or/and> and <might/must> scales, among others – or by the use of contextually dependent, ad-hoc expressions. The following sentences are examples of both types of implicature respectively:

(1) John ate some of the cookies.
→ John ate some but not all of the cookies.

(2) *In a context with two shirts, one with polka dots and the other with polka dots and stripes.*
Give me the shirt with polka dots.
→ Give me the shirt with polka dots and no stripes.

From this section on, for the sake of clarity, we will refer to the former as lexical quantity implicatures and to the latter as ad-hoc quantity implicatures.

* Anna Teresa Porrini, Università degli Studi di Trento (annateresa.porrini@unitn.it) & Luca Surian, Università degli Studi di Trento (luca.surian@unitn.it).

Data on acquisition of quantity implicatures suggest that children may have difficulties in deriving them correctly and often interpret <some> as meaning <some and possibly all> (e.g., Guasti et al. 2005, Huang & Snedeker 2009, Noveck 2001, Papafragou & Musolino 2003, Sullivan et al. 2019). This seems to be in contrast with the fact that children are very capable from younger ages when it comes to other pragmatic abilities (Condry & Spelke 2008, Matthews et al. 2012, Tomasello 2003), and begs the question of whether their difficulties with quantity implicatures may be due not to general pragmatic language delay, but to some other factors. For instance, the data underline a distinction between lexical and ad-hoc scales, with the former being significantly more difficult than the latter (Foppolo et al. 2020, Horowitz et al. 2017, Kampa & Papafragou 2019, Stiller et al. 2015, Wilson & Katsos 2021, Yoon & Frank 2019, Zhao et al. 2021). In a way, this seems to suggest that the difficulty children have with implicatures lays in their lexical knowledge. At the same time, however, the only corpus study available so far suggests that in production children are competent in their use of lexical scales from a very young age (Eiteljeorge et al. 2018).

As a matter of fact, several hypotheses have been put forward to explain children's delay in acquiring lexical quantity implicatures (Barner et al. 2011, Foppolo et al. 2012, Katsos & Bishop 2011, Pouscoulous et al. 2007, Reinhart 2004, Skordos & Papafragou 2016 among others). Research to disentangle this issue is still ongoing, and experimental data on the acquisition of quantity implicatures are abundant. There is, however, great variety within the available data: children have been tested in different languages, at different ages and with different tasks, and the phenomenon of quantity implicature has been presented using different scales and in co-occurrence with other linguistic or cognitive factors.

The various layers of complication present in the literature make it difficult to compare studies directly. Still, the considerable number of studies conducted allows for an analysis of available data in the form of a systematic review, which could give the possibility of a more objective viewpoint on the available data as a whole. The methodology is, however, not without its limitations, as it does not allow for analyses of the details of each study and of some of the relevant differences between different experiments, tasks and items. The aim of this review is therefore not only to shed light on the phenomenon of quantity implicatures during typical development, and how they have been tested so far, but also to validate systematic reviewing as a methodology for analyzing available data.

2. Methodology. Through a synthetic search and evaluation of multiple studies, we concentrated on quantity implicatures in an attempt to describe the data collectively. The references for this review were selected through the PRISMA method, and initially the search was extended to any type of implicature.[1] At first, a search through keywords was performed on three databases: Scopus, Web of Science and APA PsycInfo. Then, the data were screened in a four stage process, in order to only include articles that fit our eligibility criteria. After collecting the articles and extracting the data, we decided to concentrate only on quantity implicatures. The main reason for this was that the data were much richer for this type of implicature and allowed for more interesting comparisons.

There were various eligibility criteria selected for this search: first, the selected references needed to be peer-reviewed, published articles written in English after the year 2000. Second, they should contain empirical quantitative data on the comprehension of implicatures in first language acquisition during typical development. Moreover, there needed to be a clear classifica-

[1] Prisma flowcharts, as well as the final dataset and R script, are available on OSF: https://osf.io/g4edn/.

tion of what type of implicature was being tested and in what way, with examples. Finally, the authors needed to have performed a replicable statistical analysis on the data and there needed to be indication of the age range and mean age of the participants. In order to make the data more easily comparable, one last criterion was that the articles should all present their results in terms of percentage of success in implicature derivation (or a measure that could be converted to this).

3. Results.

3.1. COLLECTED DATABASE. In the end, 44 papers were deemed eligible for the analysis, all published between the years 2001 and 2021.[2] Within these references, a total of 158 different findings in terms of percentage of success was obtained across the different experiments, implicature types, tasks and groups tested within the 44 references. The minimum age tested was 2 years old and the maximum age tested was 13 years and 4 months old. Information on how many findings were found for each age group can be found in Table 1.

Mean age in years	Findings per age group
2	2
3	11
4	42
5	54
6	9
7	21
8	3
9	5
10	8
11	3

Table 1: Distribution of findings by age

The experiments were run in eight different languages: Dutch, English, French, Greek, Italian, Japanese, Mandarin Chinese and Spanish. The results are generalizable beyond the scope of just one language, as there is no detectable difference in percentage of success among the eight languages. In fact, while a Kruskal-Wallis rank sum test reports a chi-square of 17.102 and a p-value of 0.017, which shows a significant effect of language on performance, a subsequent Dunn test reveals that there is no statistically significant difference between any two languages.

Six different task types were used to test quantity implicatures within the dataset. A summary of the tasks used and how many findings were collected with each can be seen below in Table 2. The most frequently used tasks were the Truth Value Judgment Task (TVJT), the Felicity Judgment Task (FelJ) and the Referent Selection Task (RefS). In TVJT experiments, participants are asked to make a binary choice regarding the truthfulness (or correctness) of an uttered sentence, while in FelJ they are simply asked to evaluate whether the speaker had "said something well", and therefore to make a judgment on how felicitous the sentence was, rather than true. The difference between the two tasks is not always unequivocal, but experimenters tend to specify which sentences they used to ask for judgment, which made the categorization of the two during data collection for the present review less arbitrary. Referent Selection Task

[2] The data collection was performed in August 2021.

(RefS), on the other hand, is in this instance an umbrella term for any task that required participants to pick a referent for a determinate sentence or utterance, be it a picture, a person or an object. As for the less used tasks, Action Based Tasks (ActB) are those in which children are asked to perform an action after hearing a sentence, instead of being asked for judgment. In Communicative Context Assessment (CCA), children are asked to give a judgment on an event that took place after a sentence (an instruction) was uttered, instead of after the sentence itself, as is the case for TVJT and FelJ. The last task is the Speaker Selection Task (SpeS), in which participants need to select which of two speakers uttered a determinate sentence, and it can be for instance that one of the speakers has full knowledge of what happened, while the other does not.

Within these six task types, four possible types of output variable types were presented to participants: binary, ternary, quaternary or performative.

Task	Findings per task
Action based	10
Communicative context assessment	2
Felicity judgment	43
Referent selection	54
Speaker selection	9
Truth value judgment	40

Table 2: Distribution of findings by task

3.2. DATA ANALYSIS. The data analysis was performed using RStudio (R 4.1.0) in different ways: first, a Generalized Linear Model was fitted to analyze the data as a whole, inserting percentage of success as a dependent variable and mean age, task type and implicature type (lexical or ad-hoc) as independent variables. Variable type and language were not selected as factors for the GLM because their inclusion did not guarantee a better fit. The results of this analysis can be seen in Table 3. The results of the GLM suggest that age does have an effect on performance, as well as implicature type and task type. Then, non-parametric tests were performed at different stages of the analysis. We will discuss the effects of age, implicature type and task in the following section.

	Estimate Std.	Error	z value	p-value	
(Intercept)	1.085	1.074	1.010	0.313	
Mean age	0.017	0.008	2.066	0.039	*
Lexical scale	-1.115	0.475	-2.346	0.019	*
CCA	-0.255	1.688	-0.151	0.880	
FelJ	-1.047	0.839	-1.248	0.212	
RefS	-0.982	0.848	-1.158	0.247	
SpeS	-0.924	1.059	-0.872	0.383	
TVJT	-1.538	0.841	-1.828	0.068	.

Table 3: Results of the Generalised Linear Model
(with ad-hoc scale and ActB as default)

4. Discussion.

4.1. AGE. It appears from the analysis that age does, in fact, have an effect on children's performance with implicature derivation. This should not come as a surprise, since improvements with age were detected in virtually every experiment on the subject. As can be seen in Figure 1, the effect is confirmed to be a positive one.

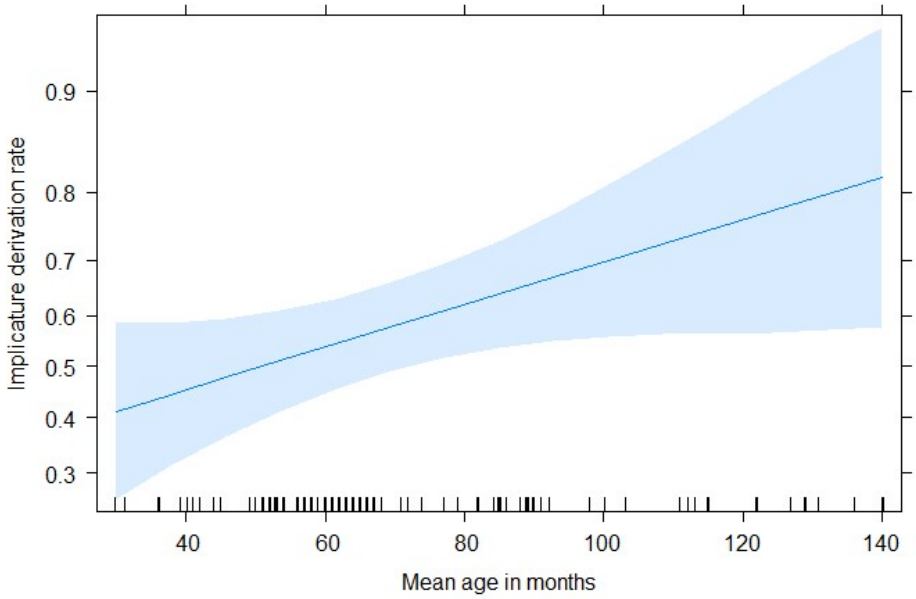

Figure 1: Effect of age

4.2. IMPLICATURE TYPE. The GLM seems to suggest that ad-hoc implicatures are easier to derive as compared to lexical ones, since the lexical implicature type has a significative negative effect in the model. A Wilcox rank test confirmed that the difference in percentage of success between the two implicature types is significant ($p < 0.001$).

If we look at the data more closely, however, this difference is less and less detectable as children get older, as seen in Figure 2. After dividing the dataset by age, roughly based on

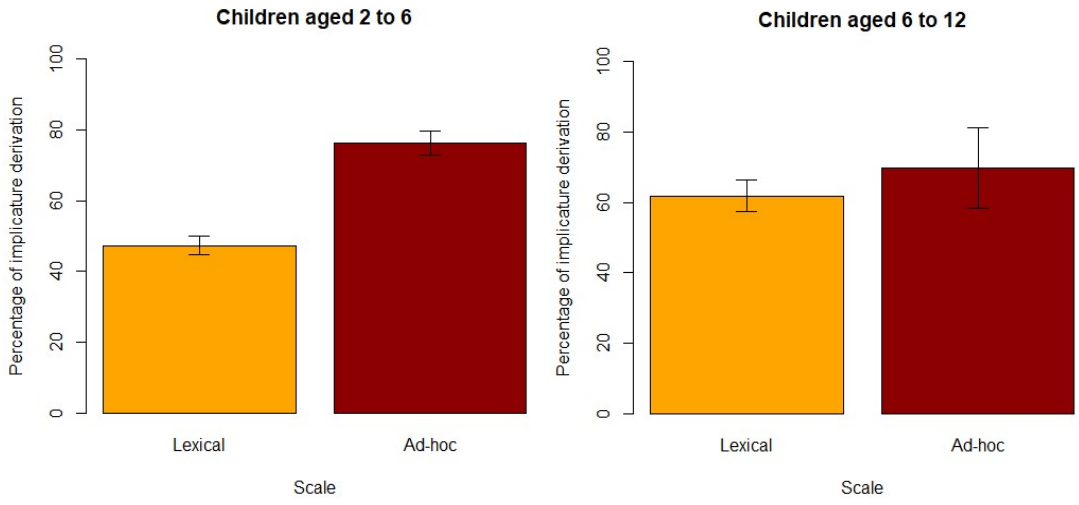

Figure 2: Difference between scales at different age ranges

whether children could already be in school or not (age 2 to 5;11 on one side, age 6 to 13;4 on the other), two new Wilcox rank tests show that in the case of preschool children the difference is indeed significant ($p < 0.001$), while it is not for older children ($p=0.5$).

Another thing worth noticing on the topic of implicature type is that there is a number of Horn scales that might be used to test lexical quantity implicatures. In the majority of cases, however, children are tested on the <some/all> scale. To be more precise, out of 39 papers that experimented using lexical scales, 26 tested only the <some/all> scale, thus only 13 of the remaining also included the <or/and> scales and modal or aspectual verbs. A Kruskal-Wallis rank sum test was performed on the results for lexical quantity implicatures, analyzing the difference in performance based on the scales used. This resulted in a chi-square of 14.872, and a p-value of 0.005, signaling that there is indeed a difference in how successful children are based on the lexical scale used. A subsequent Dunn test confirmed that the significance was driven by the difference between the <some/all> and the <or/and> and modal verb scales. The <some/all> scale appears significantly easier than the other two for children ($p=0.0123$ for modals, $p=0.0012$ for <or/and>). The lack of more data on scales other than <some/all>, however, makes more accurate analyses on the potential differences difficult to perform.

4.3. TASK. With regards to task type, the GLM shows a marginally significant negative effect of the TVJT. This task has been previously criticized in the literature for its inaccuracy when dealing with pragmatic phenomena (e.g. Katsos & Bishop 2011). One of the issues of the TVJT is that it requires participants to make judgments about truthfulness, and for this reason it might fail to detect fundamental aspects of pragmatic processing. Pragmatic meaning, in fact, is concerned with informativeness, felicity and optimality more than it is with truthfulness itself, which is more a domain of semantics. There is work that provides proof that the TVJT is indeed a good methodology to test implicatures (Guasti et al. 2005, Foppolo et al. 2020), but the analysis of systematically collected data does not support this view. In fact, it seems that the among the three most used tasks in the dataset, the TVJT is the only one for which age does not predict performance, and children show lower accuracy even at older ages (Figure 3 for reference).

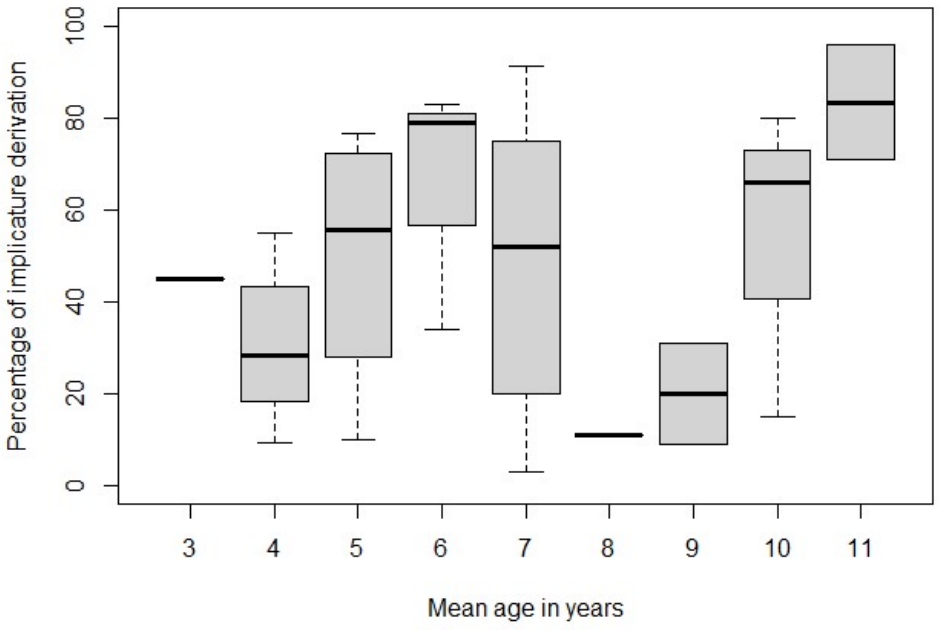

Figure 3: Mean success rates and boxplots at different ages for TVJT

Interestingly, an exploratory analysis of the data seems to suggest that some of the less used methodologies might grant children better performance on quantity implicatures. As Figure 4 shows, children seem to perform better with Action Based Tasks (ActB) and Communicative Context Assessment Tasks (CCA). These are both tasks that do not require children to make any meta-linguistic judgments, which might be a source of difficulties at younger ages. A Kruskal-Wallis rank sum test was performed on the data to verify whether task had an effect on success with implicature derivation. This test showed that the difference between tasks is indeed significant (chi-square = 13.595, df = 5, p-value = 0.0184). More in-depth analysis through a Dunn test revealed that the significant difference that drives this effect is between TVJT and ActB, in line with the result of the GLM (p= 0.0118).

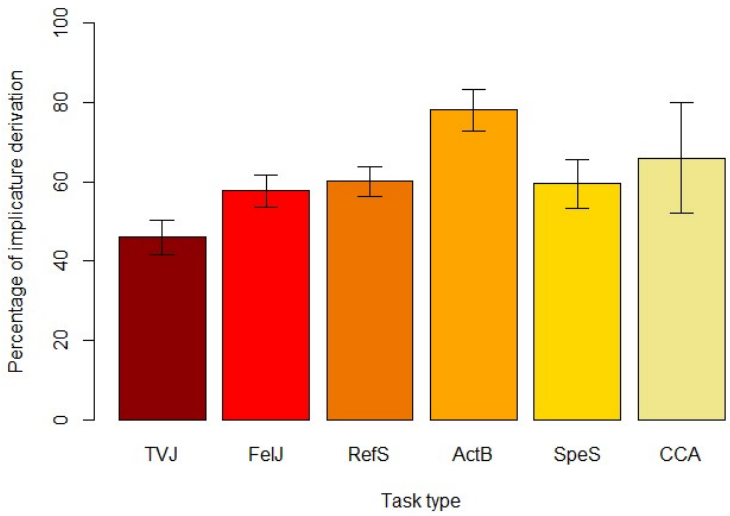

Figure 4: Difference in performance across tasks

In Speaker selection tasks children do not seem to be as good as they are with ActB and CCA, and this may be due to the fact that it requires more meta-representational abilities such as reasoning on other people's knowledge. This is a very speculative analysis, however, because the scarceness of data for all three task types hardly allows for any conclusions to be drawn from statistical analysis.

4.4. VARIABLE TYPE. Output variable type was not included as a factor in the GLM because it did not guarantee a better fit of the model with the data, nor did it seem to interact with other factors such as task type or scale. Previous literature, however, suggests that in Felicity Judgment Tasks (FelJ), when children are asked to judge someone's utterance by giving them a prize, they perform significantly better when they are given the opportunity to express judgment on a ternary scale (small prize, medium prize and big prize) as opposed to a binary choice (prize or no prize). When given a ternary option, children seem to be able to distinguish between an incorrect sentence, a correct sentence and a non-optimal but correct sentence, which is pragmatically infelicitous. An analysis of the available data confirms this, as can be seen from Figure 5. Children's performance in FelJ tasks when the outcome variable is ternary is better than when it is binary.

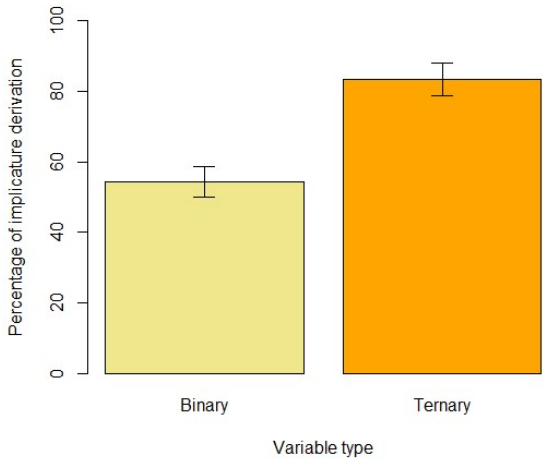

Figure 5: Success rates for binary and ternary outcome variables in FelJ

A Wilcoxon rank sum test with continuity correction was performed to confirm this effect, and it resulted in a p-value of 0.0124. This suggests that the difference is indeed significant, although it needs to be specified that within the dataset the instances of binary outcome variables were considerably more compared to the ternary ones.

5. Conclusion. What the systematic review as a whole seems to point to is that children improve in deriving quantity implicature with age, which is an expected result. It also confirms that there is a detectable difference between ad-hoc and lexical scales that makes the latter more difficult for children, especially in their preschool years.

Another important result of this systematic analysis of experiments on quantity implicatures during development regards experimental factors. It seems that the task used to test children may have a considerable impact on how they perform in implicature tasks. What the literature suggests, in this case, is that tasks that do not rely on children's meta-linguistic or meta-representative abilities may be better suited for these investigations. The data also points to an inadequacy of TVJT, a task that revolves around truthfulness instead of felicity. Although the result is relevant, as it reveals a trend across different experiments throughout the past 20 years, it needs to be pointed out that the methodology of systematic reviewing ignores the presence of potential experimental manipulations that might make the task more or less adequate to test implicatures on children. On another note implicit tasks would probably provide good results as well. These tasks were not included in this review due to their outcome measure being too different from the others, and thus not easily comparable to the rest of the dataset, but based on the reasoning that meta-linguistic and meta-representative judgments exert a significant toll on children, an implicit methodology such as eye-tracking might be an optimal way of testing them.

As a conclusion, we would argue that the systematic methodology is probably not sufficient by itself to analyze data in depth, as it does not afford attention to details, features and manipulations within each experiment which might be of considerable interest to researchers. Nevertheless, it is an extremely useful and apparently accurate methodology if the aim is to get a broader, comprehensive view of the available data, in which salient factors to be kept in mind can be individuated at a glance. A systematic review can be an excellent starting point for re-

search on a topic, as it also shows in what aspects the current literature could be enriched: with regards to quantity implicatures during development, we would suggest that this review recommends the following. It could be interesting to test a wider variety of lexical scales, for one, and to use more implicit tasks or tasks that do not require meta-linguistic abilities. Another interesting suggestion might be to try and deviate from binary outcome variables more often. These future directions might be a good step in the direction of getting a more comprehensive picture of what children are really capable of in terms of quantity implicature comprehension.

References

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition* 118. 87–96. https://doi.org/10.1016/j.cognition.2010.10.010.

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General* 137(1). 22–38. https://doi.org/10.1037/0096-3445.137.1.22.

Eiteljoerge, S. F. V., Pouscoulous, N., and Lieven, E. V. M. (2018). Some pieces are missing: implicature production in children. *Frontiers in Psychology* 9:1928. https://doi.org/10.3389/fpsyg.2018.01928.

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language Learning and Development* 8. 365–394. https://doi.org/10.1080/15475441.2011.626386.

Foppolo, F., Mazzaggio, G., Panzeri, F., & Surian, L. (2020). Scalar and ad-hoc pragmatic inferences in children: guess which one is easier. *Journal of child language* 48(2). 350-372. http://dx.doi.org/10.1017/S030500092000032X.

Grice, H. P. (1975). *Logic and conversation*. In P. Cole & J. L. Morgan (Eds.). *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). New York: Academic Press.

Guasti, T. M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes* 20(5). 667–696. https://doi.org/10.1080/01690960444000250.

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2017). The trouble with quantifiers: Exploring children's deficits in scalar implicature. *Child Development* 8(6). e572–e593. https://doi.org/10.1111/cdev.13014.

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology* 45(6). 1723–1739. https://doi.org/10.1037/a0016704.

Kampa, A., & Papafragou, A. (2019). Four-year-olds incorporate speaker knowledge into pragmatic inferences. *Developmental Science* 23(3). e12920. https://doi.org/10.1111/desc.12920.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition* 120. 67–81. https://doi.org/10.1016/j.cognition.2011.02.015.

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions to context: Effects of distracters and feedback on referential communication. *Topics in Cognitive Science* 4(2). 184–2010. https://doi.org/10.111/j.1756-8765.2012.01181.x.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigation of scalar implicature. *Cognition* 78. 165-188. https://doi.org/10.1016/s0010-0277(00)00114-1.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition* 86. 253–282. https://doi.org/10.1016/S0010-0277(02)00179-8.

Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). Processing costs and implicature development. *Language Acquisition* 14. 347–375. https://psycnet.apa.org/doi/10.1080/10489220701600457.

Reinhart, T. (2004). The processing cost of reference set computation: acquisition of stress shift and focus. *Language Acquisition* 12. 109–155. https://doi.org/10.1207/s15327817la1202_1.

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition* 153. 6–18. https://doi.org/10.1016/j.cognition.2016.04.006.

Stiller, A., Goodman, N. D., & Frank, M. C. (2015). Ad hoc implicature in preschool children. *Language, Learning and Development* 11(2). 176–190. https://doi.org/10.1080/15475441.2014.927328.

Sullivan, J., Davidson, K., Wade, S., & Barner, D. (2019). Differentiating scalar implicature from exclusion inferences in language acquisition. *Journal of Child Language* 46. 733–759. https://doi.org/10.1017/S0305000919000096.

Tomasello, M. (2003). *Constructing a Language: A Usage Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Wilson, E., & Katsos, N. (2021). Pragmatic, linguistic and cognitive factors in young children's development of quantity, relevance and word learning inferences. *Journal of Child Language*. 1-28. https://doi.org/10.1017/S0305000921000453.

Yoon, E. J., & Frank, M.C. (2019). The role of salience in young children's processing of ad hoc implicatures. *Journal of Experimental Child Psychology* 186. 99–116. https://doi.org/10.1016/j.jecp.2019.04.008.

Zhao, S., Ren, J., Frank, M. C., & Zhou, P. (2021). The Development of Quantity Implicatures in Mandarin-Speaking Children, *Language Learning and Development*. 343-365. https://doi.org/10.1080/15475441.2021.1886935.