Scalar Implicature: Gricean Reasoning and Local Enrichment

CHAO SUN

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Linguistics

University College London

2017

Declaration

I, Chao Sun, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

The first author was supported by a scholarship under the State Scholarship Fund from China Scholarship Council.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Richard Breheny, who has been an incredible supervisor over the years. I started out knowing little about linguistic research, Richard introduced me to the field of experimental pragmatics. He is always patient, supportive and extremely generous with his time and thoughts. I really appreciate the tremendous amount of time and effort that he spent for me. Without his insights and guidance this work would not have been possible.

I thank Andrea Santi and Nathan Klinedinst, who gave me different perspectives and feedback during my upgrade and cared about my research at every stage. I thank Ye Tian, who introduced me to corpus research and opened my mind about real-life application of linguistic research.

My thanks also go to Robyn Carston, Wing-Yee Chow, Judith Degen, Alison Hall, Napoleon Katsos, Stephen Politzer-Ahles, Nausicaa Pouscoulous, Tim Pritchard, Jesse Snedeker, Yasu Sudo, Bob van Tiel. The work presented in this thesis greatly benefited from your discussions.

I thank my wonderful PhD peers in Chandler House for friendship and support. Thank you Zoë Belk, Florian Breit, Caitlin Canonica, Giulio Dulcinati, Elizabeth Eden, Patrick Elliott, Diana Mazzarella, Emilia Molimpakis, Nick Neasom, Caterina Paolazzi, Lewis Pollock, Irene Symeonidou, Kevin Tang, Xiaobei Zheng, Ziren Zhou.

Last but not least, I want to thank my parents for always believing in me and supporting my choice. My greatest gratitude is reserved for Yiming, my dearest husband. Thank you for your love, understanding and support.

ABSTRACT

This thesis investigates the cognitive underpinnings of Scalar Implicature phenomenon. Here I present a series of experiments in three domains of research for scalars: (i) scalar diversity phenomenon, (ii) implicature priming and (iii) the time course of access to pragmatic enrichments. I adopt a broadly Gricean theoretical approach with local pragmatic enrichment to the design of the studies and argue that this approach can shed light on the phenomena. The results of the experiments also lend support to the theoretical perspective taken.

This thesis introduces a new perspective to interpret scalar diversity phenomenon. Given the observation that different scalar terms give rise to unembedded scalar implicatures at different rates, experiments presented in Chapter 2 and 3 suggest that one source of scalar diversity is the strength of association between a scalar term and its upper-bounding local enrichment. It provides indirect evidence that local enrichment impacts on the interpretation of unembedded scalars. More direct evidence of an effect of local enrichment in unembedded scalars is found in implicature priming. Experiments presented in Chapter 4 find unembedded and embedded scalar enrichments could prime each other, indicating local pragmatic enrichment as a shared mechanism involved in both. In addition, this thesis presents research on the time course of access to local pragmatic enrichment of 'some', which reveals no delay in pragmatic enrichment vis a vis semantic interpretation.

Overall, this thesis argues for an integrated Gricean system that allows for scalar phenomena to be explained by two mechanisms, a global inference mechanism and a local enrichment mechanism.

TABLE OF CONTENTS

Acknowledgements 5
Abstract 7
List of Figures 12
List of Tables 14
Chapter 1 Introduction 15
1.1 Scalar Implicature 16
1.2 The Gricean view of pragmatics 18
1.2.1 'Standard' Gricean account of Scalar Implicature 19
1.2.2 Relevance theory 21
1.2.3 RSA approach 25
1.3 The Grammatical Theory 30
1.4 Thesis overview 31
Chapter 2 Scalar Diversity and its causes. 33
2.1 Theoretical Background 33
2.1.1 The basic phenomenon 33
2.1.2 Neo-Gricean accounts 35
2.1.3 Alternative approaches 37
2.1.4 The uniformity assumption 39
2.2 Previous studies 42
2.2.1 Doran et al. (2009, 2012) 42
2.2.2 van Tiel et al. (2016) 43
2.3 Potential factors affect the status of alternatives 48
2.3.1 Scale homogeneity 48
2.3.2 Local enrichability 48
2.4 Experiment 1 50
2.4.1 Overview 50
2.4.2 Methods 51
2.4.3 Results 51
2.4.4 Discussion 54
2.5 Experiment 2 54
2.5.1 Overview and prediction 54

2.5.2 Methods 55
2.5.3 Results 56
2.6 Experiment 3 57
2.6.1 Overview and prediction 57
2.6.2 Methods 58
2.6.3 Results 58
2.7 Combined analysis 59
2.8 Discussion of Exp. 2 and 3 61
2.9 Conclusion 63
Chapter 3 Scalar Diversity – A corpus study 65
3.1 Introduction 65
3.1.1 The availability of scalar implicatures and the role of context 66
3.1.2 Re-examining the uniformity assumption 68
3.2 Creating a tweet corpus 70
3.2.1 Collection 70
3.2.2 Annotation 71
3.3 Paraphrase task 73
3.3.1 Overview and prediction 73
3.3.2 Methods 74
3.3.3 Results 75
3.3.4 Discussion 79
3.3.5 Combined analysis 81
3.4 General discussion 83
3.5 Conclusion 86
Chapter 4 Shared mechanism underlying unembedded and embedded enrichments 88
4.1 Introduction 88
4.1.1 One mechanism or two? 89
4.1.2 Enrichment priming 91
4.1.3 Rationale and predictions 94
4.2 Experiment 1 98
4.2.1 Overview and prediction 98
4.2.2 Method 100
4.2.3 Data treatment and analysis 102

4.2.4 Results and discussion 103
4.3 Experiment 2 104
4.3.1 Method 105
4.3.2 Data treatment and analysis 106
4.3.3 Results and discussion 106
4.4 Inverse Preference and Frequency of Local Enrichment 107
4.5 Conclusion 109
Chapter 5 What would a compositional listener do? – Another look at the time course of scalar implicatures. 111
5.1 Introduction 111
5.2 Experiment 1 116
5.2.1 Experiment 1(a) 116
5.2.2 Experiment 1(b) 119
5.3 Experiment 2 120
5.3.1 Experiment 2(a) 120
5.3.2 Experiment 2(b) 137
5.4 Summary of Experiments 1 and 2 145
5.5 Experiment 3 148
5.5.1 Method 148
5.5.2 Data analyses and Results 151
5.6 General Discussion 159
Chapter 6 Conclusions 162
References 164
Appendix A: List of experimental items 171
A.1 Items used in Experiment 1 (Chapter 2) 171
A.2 Items used in Experiment 2 (Chapter 2) 172
A.3 Items used in Experiment 3 (Chapter 2) 173
A.4 Scales used in Corpus and Paraphrase task (Chapter 3) 174
A.5 Filler items used in Experiment 1 (Chapter 4) 174
A.6 Filler items used in Experiment 2 (Chapter 4) 176

LIST OF FIGURES

Figure 1 Sample item in van Tiel et al. (2016) - Experiment 2 44

Figure 2 Mean inference ratings for Experiment 1 53

Figure 3 Sample item in Experiment 2 55

Figure 4 Negative correlation between the absence of homogeneity and inference rate 56

Figure 5 Sample item in Experiment 3 58

Figure 6 Positive correlation between the propensity of local enrichment and inference rate 59

Figure 7 Word sense disambiguation task example item 73

Figure 8 Paraphrase task example item 75

Figure 9 The hierarchical structure of the dataset 76

Figure 10 Negative correlation between the mean rating and the mean entropy 77

Figure 11 Mean inference ratings for the paraphrase task 78

Figure 12 Example items in Bott & Chemla (2016) 92

Figure 13 Proportion of strong responses for within-category and between-category priming in Bott & Chemla's Experiment 1 94

Figure 14 Critical items for Embedded Target trials in Experiment 1 and 2 95

Figure 15 Discarded displays 96

Figure 16 Example display where the global reading is true 96

Figure 17 Sample items of Experiment 1 98

Figure 18 The proportions of enriched responses across conditions in Experiment 1 103

Figure 19 Sample items of unembedded target condition in Experiment 2 105

Figure 20 The proportions of enriched responses across conditions in Experiment 2 106

Figure 21 Examples for experimental items used in Experiments 1(a) and 1(b) 117

Figure 22 Example displays in Experiment 2(a) 121

Figure 23 Example displays in Experiment 2(a) 122

Figure 24 Log ratios of percentage of looks to target over competitor by Determiner from the instruction onset to the determiner window offset in Experiment 2(a) 125

Figure 25 Log ratios of percentage of looks to target over competitor by Determiner and Target size from the instruction onset to the determiner window offset in Experiment 2(a) 127

Figure 26 Average ln((Target)/(Competitor) by Determiner from the modifier onset to the instruction offset (e.g. 'stripy squares') in Experiment 2(a) 128

Figure 27 Average ln(P(Target)/P(competitor) by Determiner and Target size from the modifier onset to the instruction offset in Experiment 2(a) 128

Figure 28 Bias to residue set (empirical logits) by Determiner from the instruction onset to the instruction offset in Experiment 2(a) 131

Figure 29 Log ratios of percentage of looks to target over competitor by Determiner from the instruction onset to the determiner window offset in Experiment2(b)..... 138

Figure 30 Log ratios of percentage of looks to target over competitor by Determiner and Target size from the instruction onset to the determiner window offset in Experiment 2(b) 139

Figure 31 Log ratios of percentage of looks to target over competitor by Determiner from the modifier onset to the instruction offset in Experiment 2(b) 140

Figure 32 Log ratios of percentage of looks to target over competitor by determiner and target sizes from the modifier onset to the instruction offset in Experiment 2(b) 141

Figure 33 Bias to residue set (empirical logits) by Determiner from the instruction onset to the instruction offset in Experiment 2(b) 143

Figure 34 Example displays in Experiment 3 150

Figure 35 Log ratios of percentage of looks to target over competitor by Determiner from the display onset to the instruction offset in Experiment 3 152

Figure 36 Log ratios of percentage of looks to target over competitor by Determiner and Target size from the display onset to the instruction offset in Experiment 3 152

Figure 37 Bias to residue set (empirical logits) by Determiner from the instruction onset to the instruction offset from the display onset to the instruction offset in Experiment 3 155

LIST OF TABLES

Table 1 Results of multiple linear regression for inference ratings of Experiment 1..... 52

Table 2 Results of combined analysis 60

Table 3 Overview of the role of context in Gricean derivations of scalar implicatures . 68

Table 4 Environments prohibit the scalar inference 72

Table 5 Results of combined analysis with the inference rating from the paraphrase task as dependent variable..... 82

Table 6 Results of combined analysis with the entropy value from the paraphrase task as dependent variable..... 83

Table 7 Design of experimental items in Experiment 1..... 101

Table 8 Experimental Design of Experiment 2(a) 122

Chapter 1 INTRODUCTION

In human communication, what speakers mean often goes beyond or departs from what they say. The question is how listeners infer what speakers mean. This thesis focuses on one particular phenomenon – Scalar Implicature, and presents a series of experiments that hopefully shed light on the phenomenon.

In the short history of experimental approaches to pragmatics, Scalar Implicature is by far the most exhaustively studied. Despite this there is still a great deal of experimental work to do. In this thesis, I focus mostly on 'Gricean' perspectives on scalars. I outline the mechanisms that have been posited to explain the full range of scalar phenomena and how different Gricean approaches incorporate these into their theoretical framework. In particular, I am interested in how post-Gricean theories, such as Relevance Theory and the Rational Speech Act (RSA) approach, deal with so-called embedded scalar implicatures using the idea of local pragmatic enrichment. In a series of experiments, I show how a Gricean theoretical approach with local pragmatic enrichment can shed light on various phenomena that have been investigated in the past few years. Particularly, this includes scalar diversity and implicature priming. The results of my studies can provide support for these Gricean accounts. In addition I present a paper that re-examines the issue of the time course of access to pragmatic enrichments – in fact local pragmatic enrichments. In addition to the contribution of the individual studies to the main theoretical thread of this thesis, the chapters themselves make other contributions to the general experimental pragmatics programme. These will be also highlighted along the way.

The outline of this chapter is as follows. I will first present some background on the phenomenon of scalar implicature. I will then summarise the standard Gricean approach and alternative approaches in the spirit of Grice, Relevance Theory and RSA. I will also briefly give some details on non-Gricean, Grammatical Theory and locate that theory in the context of this thesis. This theoretical summary sets the scene for subsequent chapters, which are summarised in an outline in the last section of this chapter.

1.1 SCALAR IMPLICATURE

In this thesis, the term 'Scalar Implicature' refers to a wide range of phenomena. As there are theoretical differences as to how to categorise and explain these phenomena, I will follow standard practice in identifying Scalar Implicature by prototypical example. Generally speaking, there are three widely discussed subclasses of Scalar Implicature which, following Breheny (2018) I will call 'Straight Scalar' (SS), 'Ignorance Inference' (II) and 'Embedded Enrichment' (EE).

Straight Scalars

Straight scalars are the most commonly discussed cases of Scalar Implicature. They are usually described as carrying implications that a relevant and more informative alternative proposition is not true. The alternative proposition could be available in virtue of lexical association (Horn, 1972, 1984) or in virtue of contextual salience (Carston, 1998; Hirschberg, 1985). Consider the following example taken from Bott & Noveck (2004):

(1) Some elephants are mammals.
    ~> Not all elephants are mammals.

The literal interpretation of (1) is that at least some of the elephants are mammals. But (1) is also likely to convey the negation of the more informative alternative that can be derived from replacing 'some' with 'all'. As the result of this, several studies have found around 60% of the participants judged (1) to be false (Bott & Noveck, 2004; Noveck, 2001; Pouscoulous et al., 2007; Zondervan, 2010).

Another example is given in (2). B's utterance implies that Ivan did not wash the car. In this case, the more informative alternative *Ivan cut the grass and washed the car* is made available by the context.

(2) A: Ivan was planning to wash the car and cut the grass on the weekend. I wonder how he got on.
    B: He cut the grass.
    ~> He did not wash the car.

As is common in the literature on this topic, I will refer to cases like (1) as *lexical scalars*, in contrast to *ad hoc scalars* like (2). Both are instances of Straight Scalar Implicature.

Ignorance Inferences

Consider the following scenario discussed in Grice (1975):

(3) A is planning with B an itinerary for a holiday in France. Both know that A wants to see his friend C, if to do so would not involve too great a prolongation of his journey.
A: Where does C live?
B: Somewhere in the south of France
~> The speaker does not know where in the south of France

B's utterance conveys an implication that the speaker does not know exactly where C lives. This implication arises in the following way: B's utterance is clearly less informative than what A expects; since it is plausible to assume that B is aware that more specific information is required, it follows that B lacks evidence for the more informative statement.

To illustrate the similarity between II and SS, consider the example in (4). As a case of SS, ((4)) could be interpreted as *Not all of the students got an A on the test*. However, if the speaker who uttered ((4)) is not in a position to know how many students got an A, then (4) would trigger an ignorance inference that the speaker does not know whether or not all of the students got an A on the test.

(4) Some of the students got an A on the test.

Embedded enrichments

A simple characterization of embedded scalar enrichments is that SS can be generated by sub-constituents of sentences as part of semantic interpretations. Consider the following examples:

(5) # Mary saw a dog or an animal.

(6) [Mary solved the first problem or the second problem] or both problems.

(7) a. Exactly one player hit some of his shots.
b. Exactly one player hit some but not all of his shots.

According to Hurford (1974), the infelicity in ((5)) is an instance of a general constraint that blocks disjunction where one disjunct entails the other. If this constraint is applied to (6), we should expect similar infelicity since the literal meaning of the disjunct in brackets is entailed by the other. It is argued in Chierchia, Fox, & Spector (2012) that the felicity of (6) is due to the fact that the bracketed disjunct is understood as, *Mary solved the first problem or the second problem and not both*. This would be a case where a typical enrichment of a segment when unembedded ('Mary solved the first problem or the second problem'), becomes an enrichment of that segment when embedded.

In ((7)), ((7)a) was used in Potts et al. (2015)'s experimental work and they reported that participants frequently judged the situations that match the reading in ((7)b) to be true. This result indicates that the embedded enrichment in ((7)b) is optional but indeed available.

Unlike SS and II, many cases of embedded enrichments cannot be explained by conjoining the literal meaning of the sentence with the negation of some other alternative proposition. For instance in ((7)), the EE interpretation in ((7)b) is logically independent from the literal meaning of ((7)a).

So far I have introduced three sub-subclasses of Scalar Implicature: Straight Scalar, Ignorance Inference, and Embedded (Scalar) Enrichment. In the next section I outline the standard Gricean account of scalar phenomenon, but set it in the context of a more general Gricean programme. I highlight some of the shortcomings of that account and then go on to summarise two other accounts that broadly adopt the Gricean view.

1.2 THE GRICEAN VIEW OF PRAGMATICS

Grice (1957, 1969) introduced the notion of utterer's occasion-meaning, an initial characterisation of which is given below:

"*U* meant something by uttering *x*" is true iff, for some audience *A*,

U uttered *x* intending:

i. A to produce a particular response r
ii. A to think (recognize) that U intends (i)

iii. A to fulfil (i) on the basis of his fulfilment of (ii)

Grice's framework for explaining speaker meaning is quite general. In priniciple, any communicative stimulus produced by an agent (an 'utterance' in Grice's terminology), if manifesting an intention to $Mean_{NN}$ something would be interpreted in context to mean anything, taking whatever is common ground between speaker and hearer as the basis of the inference. The idea is that what an utterance means depends ultimately on the speaker's intention.

1.2.1 'Standard' Gricean account of Scalar Implicature

In later work, Grice (1967) set out a framework for deriving certain subclasses of speaker meaning as conversational implicatures. This framework, unlike Grice's foundational work on speaker meaning, took for granted that a literal interpretation could be derived for an utterance based on the sentence structure and semantic rules. Implicatures are attributed to the speaker as implications that are intended in addition to (or instead of) the literal proposition expressed. The basis for deriving conversational implicatures are expectations about how speakers will conduct themselves. These expectations are captured in Grice's (1975) Cooperative Principle and more specific maxims:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (Grice, 1975, p. 45)
>
> *Quality*:
>
> 1. Do not say what you believe to be false.
> 2. Do not say that for which you lack adequate evidence.
>
> *Quantity*:
>
> 1. Make your contribution as informative as is required (for the current purposes of the exchange).
> 2. Do not make your contribution more informative than is required.
>
> *Relation*: Be relevant
>
> *Manner*: Be perspicuous.

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

What has become known as the standard Gricean view of Scalar Implicature was developed in Horn (1972), Gazdar (1979) and elsewhere using Grice's ideas. According to the standard view, Scalar Implicatures can be explained as conversational implicatures. That is to say, Scalar Implicature is not encoded in the linguistic meaning of the utterance, rather it results from reasoning about the speaker's intentions. Geurts (2009, 2010) outlined a 'standard Gricean formula' for deriving SS. For instance, given ((8)):

(8) Bonnie stole some of the pears.

   i. The speaker has said ((8)). This could be true whether or not she stole all of the pears, as long as she stole at least two.
   ii. It is clear that if the speaker had thought that ((8)*) *Bonnie stole all of the pears*, he would not have been observing the maxim of quantity.
   iii. It must be that it is not true that the speaker thinks that Bonnie stole all of the pears, i.e. $\neg BEL_C((8)^*)$.
   iv. It seems clear that the speaker would know or have an opinion about how many pears Bonnie stole, i.e. $BEL_C((8)^*) \vee BEL_C(\neg((8)^*))$.
   v. So it must be the case that the speaker thinks that Bonnie did not steal all of the pears, i.e. $BEL_C(\neg((8)^*))$.

Geurts (2010) argued that the assumption that the speaker is competent with respect to the truth of the stronger statement, e.g. whether or not ((8)*) is true, is not always satisfied. Therefore, the standard derivation predicts both an ignorance inference $\neg BEL_C((8)^*)$ when the competence assumption is not met, and a straight scalar $BEL_C(\neg((8)^*))$.

There are two problems for such a global derivation mechanism. The first one is often referred as the symmetry problem. The symmetry problem is about the choice of the more informative alternative in step (ii) above. The problem is to explain why *all* rather than *some but not all* is considered as the alternative. More specifically, in the

given context, *Bonnie stole some but not all of the pears* would be equally relevant and informative as *Bonnie stole all of the pears*.

The second problem has to do with the phenomenon of Embedded Enrichment. Consider the following example:

(9) a. Every player hit some of his shots.

b. Every player hit some but not all of his shots.

(10) Every player hit some of his shots and not every player hit all of his shots.

Applying the standard Gricean formula to ((9)a), the best one could derive is given in ((10)). This interpretation is weaker than the interpretation in ((9)b). It has been argued that the embedded enrichment in ((9)b) can be explained by an extension of the standard Gricean formula involving an extra assumption (Geurts, 2009, 2010; Sauerland, 2004). As mentioned, the standard formula could derive ((10)). To yield the embedded enrichment in ((9)b), an extra assumption like (vi) is required. As a result, the conjunction of ((9)b) and (vi) is equivalent to ((10)).

vi. Either every player hit all of the shots or every player did not hit all of his shots.

Although cases like ((9)) can be explained by an extension of the standard Gricean formula, consider ((7)a) again, it is impossible to derive the interpretation in ((7)b) from the standard Gricean formula, even with extra assumptions. The standard Gricean formula derives the enriched interpretation by conjunction of literal meaning and some other proposition. Since the embedded enrichment in ((7)b) is logically independent from the literal meaning in ((7)a), there is no way that the Gricean theory could explain such embedded enrichment.

Although the standard Gricean account cannot account for these problems, other theories of inferential pragmatics that are in the spirit of Grice explain embedded enrichments by allowing a separate pragmatic mechanism – local pragmatic enrichment.

1.2.2 Relevance theory

Under the Relevance Theory (RT) view, utterance interpretation is guided by the Communicative Principle of Relevance and the notion of optimal relevance. In Sperber & Wilson (2004) the Relevance-theoretic comprehension procedure is given as follow:

Relevance-theoretic comprehension procedure

a. Follow a path of least effort in computing cognitive effects: Test interpretive hypotheses (disambiguations, reference resolutions, implicatures, etc.) in order of accessibility.
b. Stop when your expectations of relevance are satisfied.

Unlike Grice's Cooperative Principle and maxims, the Relevance comprehension procedure is not assumed to operate only on a full, literal proposition. More in the spirit of Grice's original proposals about speaker meaning, RT allows that an utterance can achieve relevance via local adjustments to the encoded meaning. It is taken for granted in RT that an utterance activates encoded meanings for lexical items and these, together with the syntactic structures used in the utterance are the basis for computing a semantic representation (logical form). In this phase, adjustments to the literal proposition may take place. This aspect of RT's framework takes into account the apparent fact that embedded enrichments are widespread in language use (Carston, 1988; Cohen, 1971; Wilson, 1975, among others).

To illustrate using an example in Cohen (1971), consider that ((11)-(12)) carry different 'result' implicatures, as indicated:

(11) A republic has been declared and the old king has died of a heart attack.
~> The heart attack was a result of the declaration

(12) The old king has died of a heart attack and a republic has been declared.
~> The declaration was a result of the king's death.

The observation is that, when ((11)) and ((12)) are embedded in conditional sentences, the resulting sentences express different propositions, consistent with the putative implicature in ((11)-(12)) becoming part of the antecedent proposition:

(13) If a republic has been declared and the old king has died of a heart attack, the king's supporters will revolt.

(14) If the old king has died of a heart attack and a republic has been declared, the king's supporters will revolt.

These 'embedded' effects can apparently arise from any pragmatic phenomena, including strengthening/enrichment ((15)) - see Carston (2002), loose use ((16)), metaphor ((17)), and other figures of speech:

(15) a. John has money.

~> John has a significant amount of money.

b. Buying a house is easy if you have money.

(16) a. The train arrives at quarter past ten.

~> The train arrives at around quarter past ten.

b. Every morning, a night train from Paris arrives at quarter past ten.

(17) a. Her student residence was a dungeon.

~> Her student residence was ill-lit, prison-like etc.

b. At exam revision time, every student spends the evenings in their dungeon.

According to RT (see Carston, 2002; Sperber & Wilson, 1998), the input to deriving contextual implications is potentially an enriched proposition. What determines whether a sentence meaning is enriched is the drive for relevance. For example, ((15)a) is apparently trivially true while ((15)b) would clearly express a falsehood in the contemporary property market, given the literal meaning of 'money'. The idea is that easily accessible contextual implications ('implicatures') can be derived if 'money' is understood to mean *a significant amount of money*. Then, depending on the context these examples could lead to implicatures such as, *John can easily afford a house* for ((15)a); or *the speaker does not have a significant amount of money* for ((15)b). Thus, in RT, pragmatic effects can occur at two levels: at the level of enriching the proposition expressed (the 'explicature') and at the level of intended implications ('implicatures').

It is important to note that, pragmatic effects at the level of the explicit proposition in RT can also occur in apparently unembedded cases. To return to ((15)a), it was noted that it could be used to imply in context that John could easily afford a house. This would be an implicature. But the example also involves a pragmatic enrichment of the explicit proposition. In fact, in this example, the implicature could not be derived without the prior enrichment of the explicit proposition.

In this thesis, I will refer to cases where the explicit proposition is enriched via pragmatic reasoning as 'local pragmatic enrichment'. RT allows for this local pragmatic enrichment but also allows for conversational implicatures. Conversational implicatures in RT, somewhat like in Grice's theory, are contextual implications of the explicature that the speaker intends the hearer to draw on the assumption that the speaker follows conversational principles. The difference is that in RT, intended contextual implications may follow from an enriched explicit proposition.

Because there are two routes to obtaining pragmatically derived effects in RT, it follows that there might be two routes to explaining scalar phenomena in RT. In cases where embedded scalar enrichment occurs in the scope of non-monotonic (or downward monotonic) operators – as in ((7)a,b) – the only route to explaining this effect in RT is as a local pragmatic enrichment. However, unembedded examples, as in ((1)), could be explained in RT either as a conversational implicature or as a local pragmatic enrichment. Where examples such as ((1)) or the ad hoc case in ((2)) are derived as conversational implicatures in RT, the derivation is somewhat like the standard Gricean formula. In that case one reasons that the speaker who uttered a sentence with a scalar expression intends the audience to be aware that they are not in a position to provide a contextually salient, more relevant proposition. Alternatively, as argued in Noveck & Sperber(2007), Straight Scalars could result from a local pragmatic enrichment of the literal meaning of the scalar expression. Consider ((18)) which is an example of straight scalar:

(18) Most Americans are creationists and some even believe that the Earth is flat. (Noveck & Sperber, 2007)

The denotation of the linguistic encoded meaning of ‘some’ is a subset of at least two Americans and at most all American creationists. According to Noveck & Sperber(2007), to make the utterance relevant enough, the hearer should adjust the meaning of ‘some’ by narrowing at both ends. As a result, the locally enriched denotation of ‘some’ would result in implying that the set of flat-earth believers is much larger than a set of two Americans but smaller than the set of American creationists.

Noveck & Sperber(2007) do not spell out what factors might lead to deriving a scalar inference as a local enrichment or as a conversational implicature. In the example they discussed, their argument is that the literal meaning needs to be enriched to make the utterance relevant. This seems plausible on the lower bound. If just two or three Americans believed some strange things, that would be hardly surprising or relevant to the apparently intended point. However, at the other end, in this case, simple world knowledge rules out the upper-bound: that all Americans or even all American creationists also believe the earth is flat. So, it is not clear in this case that the upperbounding on the set of Americans who are flat-earthers is obtained by local pragmatic enrichment. In fact, it is possible that this example is understood so the second clause is understood, *some but not most Americans believe that the earth is flat.* This enrichment seems just as likely to be derived via global implature derivation given the contextually salient alternative, 'most'.

One aim of this thesis will be to establish whether cases of Straight Scalars, where the scalar term is not embedded, might nevertheless be sometimes derived via local pragmatic enrichment. Another aim is to explore some methodological consequences if this were the case.

1.2.3 RSA approach

Standard RSA approach

The standard RSA approach (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) explains Scalar Implicature in a more or less global inference mechanism manner. This approach combines Bayesian probabilistic reasoning with an adaption of the iterated best response approach taken from Game-Theoretic implementations of Grice (Franke,

2009; Jäger, 2012). It postulates a 'literal listener' strategy whereby the listener assumes a speaker with a message, *w*, chooses an utterance at random from the set of possible true utterances. A speaker, faced with such a listener, would choose an utterance, *u*, seeking to maximise the likelihood that such a listener would pair it with the right message. In turn, a listener who faces such a speaker assigns to the utterance the message that has the highest probability. Further iterations of these computations may or may not result in a stable pair of strategies for the speaker and hearer. In basic scalar cases, however, it does.

Let us consider a simple computation in which the speaker uses the expression 'some' and where there are two possible situations the speaker has observed, ∃&¬∀ and ∀, and where 'all' is an alternative utterance. Putting aside some details, the speaker would more likely choose 'all' than 'some' when the observed state is ∀. A listener who knows this can compute that the likelihood that 'some' communicates ∀ is lower than it communicates ∃&¬∀. Further iterations of this mechanism confirm the basic inference that the speaker uses 'some' to communicate ∃&¬∀.

Lexical uncertainty

Bergen, Goodman, & Levy (2012) and Bergen, Levy, & Goodman (2016) outline a development of RSA model that includes compositional lexical uncertainty. As with Relevance Theory, RSA with Lexical Uncertainty (RSA-LU) accounts for the fact that some local adjustment processes are able to manipulate the interpretation of a sub-constituent of an utterance. Bergen and colleagues only look at narrowing of meaning and only of lexical items, but the framework could conceivably be extended to account for loosening of meaning and apply to complex constituents.

Under the lexical uncertainty view, the meaning of a lexical item is uncertain. Consider the scalar term ‘some’. It's literal or encoded meaning is simply existential, consistent with the set of possibilities that includes states where *all* is the case and where *some and not all* is the case, {∃&¬∀, ∀}. Enriched interpretations of 'some' could be such that they exclude possibilities where *all* is the case, {∃&¬∀}, or exclude possibilities where *some and not all* is the case, {∀}. RSA-LU then treats the problem of co-ordination between speaker and hearer as involving more than one possible meaning

at the level of the speaker, given a 'literal listener' for each possible lexicon (i.e. each possible mapping from 'some' to a meaning). I.e. the first-order listener faces a speaker that may be addressing a literal listener who uses any one of the potential meanings of the scalar term in question. Bergen et al. (2016) account for the listener's uncertainty about which lexicon the speaker uses by marginalising (taking the weighted average) over lexicons. The account can best be illustrated by considering the example mentioned above, where 'some' could have its literal meaning {∃&¬∀, ∀} or either of two enriched meanings {∃&¬∀}, {∀}. If the prior probability of these meanings is the same, then the use of unembedded 'some' in an example like ((1)), would result in an implicature mostly due to the effect of the global reasoning outlined above in the standard RSA approach. However, the effect of allowing lexical strengthening is that, after one iteration of speaker and hearer beyond the literal listener, the likelihood attached to the scalar inference (*not all*) is higher. Although the case is not discussed, I should note here that if the prior probability of the enriched meanings differ because it is higher for the {∃&¬∀} enrichment than for the {∀} enrichment, then the inference from even an unembedded use of 'some' to *not all* would be even stronger, in the sense that after one iteration the likelihood of this enrichment would be greater.

In Potts et al. (2015) a wider range of sentences are considered with also cases of embedded SI. Potts et al. observe that computations of likely meanings using (their version of) RSA with Lexical Uncertainty, when all possible strengthenings of sentences are taken into account, lead to some surprising results. By contrast, they observe that, if the range of pragmatic strengthenings is constrained to something akin to meaning strengthened by fixed scales, then the pattern of inferences predicted are more in line with intuition. That is to say, when they only allow constituents to be strengthened only as if they were unembedded, then the outcome of the model is more in line with intuition. For example the constituent [scored] is literally true if the player gets some or all of his shots, while [aced] is true only when all shots are hit. In an umbedded context (e.g. 'Player A scored'), the scalar implicature would be that the player did not ace (i.e. not get all of the shots). Thus |[scored]| could be enriched to be true just in situations where some but not all shots are successful. A computation that takes account of all possible enrichments would include also the enrichment where |[scored]| is true only

when all shots are successful. If the latter case is ruled out, fewer surprising results emerge. One such result occurs in the sentence context, 'No player scored'. An unconstrained RSA-LU system assigns a surprisingly high likelihood to situations where one player gets none but the other gets some but not all shots. This is so because, on the 'scored'-means-scored-all enrichment, this sentence is true in the scenario described.

One conclusion to draw from these considerations is that, if the RSA-LU system computes likelihoods when all possible enrichments of an expression are possible, it can throw up some intuitively strange results. However, when local enrichments of scalar terms are constrained to those meanings that occur in unembedded contexts, then the outcomes of an RSA-LU computation are more in line with intuition.

In addition to observing the likelihood which different versions of RSA-LU attach to example items, Potts et al. constructed a model based on RSA-LU to predict the results of an experiment on embedded and unembedded uses of scalars. They find that the model that only takes account of a more constrained set of enrichments gives the best qualitative picture. It is worth describing this experiment in some detail and the analysis Potts et al. perform since it is relevant to the studies conducted in this thesis.

In each trial of the experiment, participants see an image containing three players and each of these players is presented as having scored none (N), some and not all (S) or all (A) of their shots. Participants read a sentence like, 'Every player got none of his shots'. They then make a binary True/False judgement. The experimental items included a positive ('every'), negative ('no') or non-monotonic ('exactly one') quantifier in subject position. The object noun phrases were, 'some/all/none of his shots'. In a preliminary analysis, they found that participants accessed the embedded enrichment reading of 'hit some of his shots' when the subject quantifier was both non-monotonic and negative. These responses are not predicted to occur according to the standard Gricean approach that can only derive global implicatures.

More relevant to our interest in the factors that lead to local enrichment. Potts et al. compare how well four different models fare at predicting the specific outcomes of the experiment. One of these models is based only on the literal meaning of the experimental sentences; one allows for global pragmatic reasoning, as in the standard

RSA model; and two different models allow for local enrichment as in RSA-LU. One of these LU models allows for all enrichments, the other allows only for a constrained set of enrichments, based on simple alternatives for unembedded cases. Although the literal and global-only models do as well as, if not a little better in cases where no embedded enrichment is required, the LU models perform clearly better when embedded enrichment is required. Finally, as mentioned, among the LU models, the more constrained model does better than the unconstrained one.

Although these RSA-LU papers raise some questions, they do reveal that, in the RSA-LU framework, the presence or absence of a locally enriched interpretation of an expression as a potential interpretation of the utterance can affect both whether a particular inference becomes available at all (e.g. in embedded contexts), and the strength of that inference – i.e. in the sense of the confidence the hearer might have in the inference, or perhaps the activation of that inference in processing. In particular, the existence of upper-bound local enrichments of a scalar term has an impact even on confidence in an unembedded scalar implicature – a Straight Scalar.

Another point that emerges from this discussion is that presumably not all local enrichments are equal. Potts et al.'s results suggest that enrichments that would be obtained when the scalar term is unembedded are favoured as local enrichments over the unattested, 'symmetric' enrichment. For example 'some' used in unembedded contexts often implies 'not all', never, 'all'. In embedded contexts, there is no evidence that 'some' is ever enriched to mean {∀} and there is reason to think that such an enrichment does not enter into consideration as a local pragmatic enrichment. While Potts et al. do not explore why this should be the case, this thesis will hopefully shed some light on what factors would make it so. The fact that not all potential local enrichments of a particular scalar term are equal leads to the question of whether potential local enrichments across scalar terms are also not equal. Specifically, it is interesting to ask if the prior likelihoods of upper-bounded local enrichments of two scalar terms could differ. This question is relevant to our understanding of the scalar diversity phenomena, to be investigated in Chapters 2 and 3.

1.3 THE GRAMMATICAL THEORY

The grammatical theory (GT) is outlined in Chierchia, Fox, & Spector (2012) and Fox (2007). This approach accounts for SS (e.g. ((4)) above) in terms of a covert exhaustification operator being present in the syntactic structure for the sentence. For instance, the syntactic representation of ((4)) is given in ((19)):

(19) *Exh* [Some of the students got an A on the test]

Exh can be inserted in any position of the logical form of a sentence. For the current purpose, *Exh*(*X*) is understood as the conjunction of *X* and the negation of the non-weaker alternative of *X*. *X* can be an entire proposition or a non-propositional constituent. The main motivation for the grammatical view is to offer an account of EE. EE is derived in similar fashion as SS. Consider the example in ((7)) again. The embedded enrichment in ((7)b) could be derived by inserting *Exh* in the scope of non-monotonic operator as shown in ((20)):

(20) [[Exactly one player]$_x$ [*Exh* [t$_x$ hit some of his$_x$ shots]]]

In this way the GT accounts for both SS and EE by the same mechanism. As for Ignorance Inference, Fox (2007) in particular argues that this should be explained by appeal to a Gricean pragmatic system. This division of labour contrasts with the Gricean approaches mentioned above. The latter can provide an account for SS and II using the same global inference mechanism, while EE requires an additional local enrichment mechanism. However, the Gricean approaches that do take local enrichment seriously can also account for SS by the same mechanism as EE. I.e. there are two mechanisms for SS for Gricean approaches.

It is not an aim of this thesis to discuss in detail the competing merits of GT vs. Gricean approaches to scalars. In this thesis, I will work with the assumption that some kind of Gricean approach that can accommodate embedded enrichments is worth pursuing as part of a research project to investigate the cognitive underpinnings of scalar implicature. This being said, I will return to touch on some implications for the GT that follow from studies reported in Chapter 4.

1.4 THESIS OVERVIEW

Chapter2 focuses on the phenomenon of 'scalar diversity', which refers to the fact that different scalar terms give rise to straight scalar implicatures at different rates. I first review different theoretical approaches to scalar implicature and alternatives, and I point out two potential sources of scalar diversity: (i) the relation between a scalar term and its lexical alternatives and (ii) the strength of association between a scalar term and its upper-bounding local enrichment. Although these theoretical approaches implicitly assume that neither relation should differ for different scalar terms, empirical evidence reviewed in section 2.2 suggests otherwise. In section 2.3, I propose that the specificity of the scalar term might influence (ii). As a result, the likelihood that a scalar term gets a locally enriched interpretation (what I called 'local enrichability') varies across different scalar terms and contributes to scalar diversity. In section 2.4 to 2.6, I present a replication of a key study which demonstrates scalar diversity and investigate two factors that might contribute to scalar diversity, one of which is local enrichability. It turns out that local enrichability could explain in part the observed scalar variability, which provides supporting evidence that local pragmatic enrichment takes place in the cases of Straight Scalars.

Chapter 3 further investigates scalar variability in *everyday real use*. I argue that the scalar diversity pattern could be better established in a large-scalar corpus-based study. In section 3.2 to 3.3, I describe a Twitter corpus of sentences containing scalar terms and present a corpus-based paraphrase task that re-examines whether there is variation among scalar terms in terms of how strongly they give rise to scalar implicatures. In section 3.4, I address the question of whether factors established in Chapter 2 could affect scalar implicatures derivation in real use. The data reported in Chapter 3, together with that of Chapter 2, show that local enrichment plays a role in interpreting Straight Scalars and argues for an integrated Gricean system that allows for a local enrichment mechanism in addition to a global inference mechanism.

Chapter 4 is devoted to exploring the mechanisms underlying unembedded and embedded scalar enrichments using priming methodology. Previous studies interpreted priming of pragmatic enrichment as evidence for a shared mechanism. In section 4.2 to 4.3, I present two enrichment priming experiments that investigate whether

unembedded and embedded enrichments could prime each other. The effect of priming found in these experiments provides evidence that local enrichment as a shared mechanism is responsible for both Straight Scalar and Embedded Scalar Enrichment.

Chapter 5 looks at a separate issue related to the time course of scalar implicatures. Visual-world studies have found conflicting evidence regarding whether there is a delay in integrating pragmatic *some* relative to the semantic interpretation of *all* and exact numbers. I argue that prior expectations about the set size associated with quantifying expressions render the interpretation of previous visual world data problematic. In section 5.2, two off-line studies demonstrate that there is a low-level association between set size and quantifier use (some/all). In section 5.3 to 5.4, three visual-world studies show that such prior expectation influence the pattern of target anticipatory looks during online interpretation of scalar quantifiers. I introduce a novel indicator to measure the timecourse of scalar processing, which is less affected by other expectations. By using this new indicator, all three visual-world studies suggest that pragmatically enriched interpretation of *some* is accessed in the same timecourse as literal interpretations of *all*.

Chapter 2 SCALAR DIVERSITY AND ITS CAUSES.

This chapter examines a phenomenon known as scalar diversity. In a series of experiments various authors such as Doran (2009), Doran et al. (2012) and van Tiel et al. (2016) present evidence suggesting that different scalar terms give rise to straight scalar implicatures at different rates. This has become known as scalar diversity. Scalar diversity is related to the question whether certain lexical alternatives have special relations with scalar terms. With respect to the symmetry problem, I will consider neo-Gricean accounts which assume a special status for certain lexical alternatives. I will compare this early account to the structural approach (Fox & Katzir, 2011; Katzir, 2007) and rational speech-act approach (Bergen, Levy, & Goodman, 2016; Goodman & Stuhlmüller, 2013) which both assume no special status to any alternatives. The status of alternatives may affect the availability of scalar implicatures drawn from utterances containing scalar terms. As mentioned, previous experimental research has shown that the rate of scalar implicatures triggered by different scalar terms varies a great deal. These results might be explained by diversity in the strength of relation between scalar terms and alternatives. However, as suggested by Gricean approaches to Embedded Enrichment, I will argue that there is another factor that may affect rates of unembedded scalars – this is what I call 'local enrichability', which is associated with the likelihood that a scalar term gets a locally enriched interpretation.

Here Experiment 1 presents a replication of a key study which demonstrates scalar diversity, by van Tiel et al. (2016), but using a different measurement scale. Experiment 2 and 3 investigate two factors that might affect the availability of lexical alternatives. The overall goal is to explore what factors contribute to scalar diversity and whether any of these could be traced back to a special relation between scalar terms and their alternative and whether these may be related to local enrichability.

2.1 THEORETICAL BACKGROUND

2.1.1 The basic phenomenon

Consider the following examples

(1) John ate some of the cookies.

~> John did not eat all the cookies.

(2) It is possible she will win.

~> It is not certain she will win.

(3) It is warm.

~>It is not hot.

In each case, what follows '~>' would be a plausible implication in easily imaginable situations. These implications are widely discussed as examples of Scalar Implicature (SI) (Horn, 1972, 1984) or Quantity Implicature (Geurts, 2010). The classical Gricean theory explains these as conversational implicatures which could be derived on the basis of the first Maxim of Quantity: Make your contribution as informative as is required (for the current purposes of the exchange) (Grice, 1975). Consider (1), the implicature would be derived as follows:

I. The literal meaning of (1) is that John ate at least some of the cookies. This does not rule out the possibility that John ate many, most, or all of the cookies.
II. Assuming that (i) it would be relevant to the conversational purpose how many cookies did John ate and (ii) the speaker observed the quantity maxim or at least the cooperative principle, ...
III. the speaker has said (1) rather than a more informative alternative such as 'John ate all of the cookies' must because the speaker does not know whether or not John ate all of the cookies.
IV. Assuming that the speaker is in a position to know whether the more informative alternative would be true or not.
V. One could conclude that the speaker believes John did not eat all of the cookies.

The same derivation goes for (2) and (3). In general, these implicatures take the form *not A* where *A* is the relevant alternative. Thus the alternative in (2) is that *It is certain she will win*, while in (3) it is that *It is hot*. Under this account, the only constraints on alternatives are that they be relevant and more informative. These are apparently not sufficient as standard Gricean derivation would run into the well-known symmetry problem.

First noted in Kroch (1972), the symmetry problem is about why one alternative is chosen rather than its symmetric counterpart. To illustrate, consider (1) above. The speaker could have said:

(4) a. John ate all of the cookies.

b. John ate some and not all of the cookies. (≅ John ate just some of the cookies)

Both propositions in (4) are more informative than (1). In the derivation outlined previously, (4a) is considered as the relevant alternative. However, Kroch's point (echoed in Fox & Katzir, 2011; Katzir, 2007 and elsewhere) is that to the extent that (4a) would be relevant in a context, (4b) would be equally relevant. If we consider (4b) as the more informative proposition, then the same reasoning would give rise to a wrong implicature that the speaker believes *John ate all of the cookies*. The problem then is to explain why (4a) is chosen as the scalar alternative and not (4b). In general, for an assertion *W* that gives rise to a scalar implicature ¬*S*, the Symmetry Problem for a Gricean theory is to explain why the 'symmetric alternative', *W&¬S*, is not apparently considered in the derivation.

2.1.2 Neo-Gricean accounts

Later interpretations of Grice's proposals introduce the notion of lexical scales. A lexical scale is an entailment-based scale where the elements of the scale are equally lexicalized items, of the same word class, from the same register, and about the same semantic relations (Atlas & Levinson, 1981; Gazdar, 1979; Horn, 1972, 1984; Levinson, 2000). Examples of lexical scales from different lexical categories are given in (5):

(5) Quantifiers: <all, most, many, some>

Modals: <necessarily, possibly>, <must, should, may>

Adverbs: < always, often, sometimes>

Adjectives: <hot, warm>

Verb: <love, like>, <know, believe>

The role of lexical scales is to provide an ordered set of lexical alternatives which can be the basis of implicature calculation. In the cases of (1-3), scales involved could be <all, some>, <certain, possible>, <hot, warm>. The more informative alternatives are then determined by replacing the scalar term with any term higher in the lexical scale

(subject to satisfying certain conditions regarding grammatical context). Thus, neo-Gricean accounts assume a special status for lexical alternatives that lie on the same scale as the scalar term. That is, when the scalar term is used in an utterance, the alternative proposition is available in virtue of lexical association.

Using the idea of lexical scales, neo-Gricean accounts do not offer a solution for the symmetry problem as much as a description of how it is solved. Consider (1) again, (4a) is the stronger alternative while (4b) is not, simply because 'all' is a lexical alternative for 'some' whereas 'some and not all' is not. Thus, only (4a) is derived as the more informative alternative. However, this proposal also raises the question where these alternatives come from. The answer given by early neo-Gricean literature is somewhat obscure. Gazdar (1979), for instance, comments that scales are "in some sense given to us" (Gazdar, 1979: P58). So far, it is commonly viewed that alternatives are given lexically.

By assuming lexical scales are linguistically given, neo-Gricean accounts distinguished lexical scales from ad hoc scales which are constructed based on what is relevant in specific contexts. Consider (6):

(6) A: Ivan was planning to wash the car and cut the grass on the weekend. I wonder how he got on.
B: He cut the grass.
~> He did not wash the car.

As discussed in Hirschberg (1985), in this context where the conjunction is relevant, an ad hoc scale <cut the grass and washed the car, cut the grass> could be constructed. The use of the less informative expression *cut the grass* would give rise to a scalar implicature because the more informative alternative *cut the grass and washed the car* is made available by the specific context.

Gricean reasoning takes places in both cases like (1)-(3) and ad hoc cases like (6). In the framework outlined in (Grice, 1975), and discussed in detail in (Geurts, 2010), alternatives would have to be such that knowing them would be relevant to the conversational purpose. Implicatures are only available where supported by the contextual goals. Though some researchers, especially Levinson (2000), Horn (1984) and

Gazdar (1979) have seen a contrast between scalar implicatures involving lexical scales and those involving ad hoc scales. The former is seen as a case of generalised implicature while the latter as particularised implicature.

2.1.3 Alternative approaches

More recently, two related proposals argue that the symmetry problems can be solved without assuming special status to any alternative. Both approaches make critical use of the relative complexity, or cost, of the symmetric alternative, in solving the symmetry problem.

The structural approach

The structural approach to alternatives (Fox & Katzir, 2011; Katzir, 2007; Trinh & Haida, 2015) is aligned with a grammatical view of Scalar Implicature, but it could just as well be adopted by a standard Gricean view. According to the structural approach, alternatives for SI can be selected from a set that is constrained by structural, linguistic factors. In particular, alternatives can be derived from the asserted sentence by lexical substitution (replacing one lexical item with any other syntactically equivalent item from the lexicon), replacing constituents of that sentence with sub-constituents of the same constituent. Through these two procedures, alternative structures are no more structurally complex than the original. In addition, structures that are salient in the discourse context are licensed to replace structures in the asserted sentence, regardless of whether these are more complex or not. This latter condition accounts for examples such as those discussed in Matsumoto (1995) where the alternatives are necessarily more complex than the assertion:

(7) Yesterday it was warm. Today it was a little bit more than warm.

~> Yesterday it was not a little bit more than warm.

Given this formally defined set of alternatives, the set of contextually relevant alternatives can then be chosen, subject to the further condition that an alternative cannot be chosen while a symmetric alternative is excluded.[1]

[1] Here I leave some details of the current standard structural approach. In particular, this approach assumes that alternatives for SI are not required to entail the assertion. It is only required that the

Let us see how the structural approach to alternatives gets the right solution to examples discussed above. According to the structural account of alternatives, examples like those in (1)-(3) the alternative is derived by lexical replacement. For example, the lexical item 'all' replaces 'some'. Note that, according to the structural approach, there is no means for the set of formal alternatives for (1) to include [John ate some and not all of the cookies], since the relevant constituent is structurally more complex and unavailable in the discourse context. Similar considerations apply to (2) and (3). When it comes to ad hoc cases, many can also be handled by lexical replacement. For example in (6), one can replace verb and object noun in the asserted sentence to construct the relevant alternative ([wash the car]) or, in the context given, the relevant constituent is also salient.

Thus, according to Katzir (2007) and Fox & Katzir (2011), so-called lexical scales are accorded no special status. They are no more 'given to us' than contextually specific ad hoc scales.

Rational-speech act approach

A similar point of view is presented when one considers the Rational Speech Act (RSA) approach to Scalar Implicature (Bergen, Levy, & Goodman, 2016; Frank & Goodman, 2012; Goodman & Frank, 2016). As mentioned in Chapter 1, this approach is a development of work on Gricean pragmatic reasoning inspired by game-theoretic approaches and employing Bayesian techniques to account for how speakers and hearers reason about each other's behaviour. As recently outlined in RSA approaches, both cost and informativity of alternatives can be used to explain why one alternative is negated in scalar implicature, rather than another. In the case of (1)-(3), the relative cost of symmetric alternatives means that the speaker is less likely to assert the weaker term when they believe, respectively, 'all', 'certain', 'hot' are true than when the symmetric alternatives ('some and not all', 'possible and not certain', 'warm and not hot') are true. As Bergen et al. (2016) observe, an RSA model can include all possible alternatives, including symmetric alternatives in Scalar Implicature reasoning; i.e. the speaker and hearer can assume that both the attested alternative and its symmetric counterpart can

alternative not be entailed by the assertion. See Chierchia, Fox, & Spector(2012), Fox (2007) and Breheny, Klinedinst, Romoli, & Sudo (2018) for a discussion.

be considered in the derivation. The correct scalar implication is still derived. Again, the so-called lexical alternatives (e.g. all, certain, hot) need have no special status in the explanation. The theory can explain why these are chosen as alternatives simply based on assumptions about cost and informativity.

2.1.4 The uniformity assumption

While early neo-Gricean theory tended to assume a special status for lexical alternatives that lie on the same scale as the scalar term, as discussed, recent theory that addresses the question how alternatives are chosen with respect to the symmetry problem do not assume special status for any alternatives. To date, neither proponents of neo-Gricean scales nor of the structural theory of alternatives nor the RSA theorists have said anything about whether scalar terms might vary in the degree to which they are liable to give rise to SI. The assumption that scalar terms do not vary in this way has been termed 'The uniformity assumption' in recent literature (Doran, 2009; Doran et al., 2012; van Tiel et al., 2016). A simple characterization of the uniformity assumption is that experimental results of 'some' and 'or' are assumed to be representative for all scalar expressions. In particular, the rates of SIs generated by different scalar terms are assumed to be the same. The uniformity assumption has been challenged by results of studies reported in these papers. In fact, these studies show that there is 'scalar diversity', whereby different scalar terms apparently give rise to scalar implicature at considerably different rates. In particular, there is a large difference between quantifiers and modals, on the one hand, and scalar adjectives, on the other.

It is possible to read the neo-Gricean literature on Scalar Implicature and see the use of lexical scales as a commitment to a uniformity assumption. As Geurts (2010) points out, the use of neo-Gricean scales suggests a kind of fixed operation resulting in a scalar implicature which is default, as long as the relevant alternative term is on a lexical scale. This understanding of the neo-Gricean approach involving lexical scales does suggest a uniformity of scalar implicature, as long as the scalar terms and alternatives are on the same scale. However, other interpretations of the work of Gazdar, Horn and so on are possible.

With regards to the structural and standard RSA approaches to scalar implicature alternatives, while these do not rely on a special status for alternatives, they do not rule out the possibility of scalar diversity – i.e. different propensities among scalar terms to give rise to scalar implicatures. However, if we adopt the structural approach to alternatives in a general Gricean explanation of scalar implicatures, or if we adopt the standard RSA approach without further assumptions, one can argue that these approaches imply a uniformity assumption, or at least, they do not predict scalar diversity.

On the other hand, as mentioned in Chapter 1, RSA with lexical uncertainty (RSA-LU) allows that even unembedded scalar enrichment can be affected by the fact that scalar terms can be enriched by a local pragmatic process. Where local scalar enrichments are entertained and these go in the same direction as the global enrichment, then these strengthen the likelihood of the inference for the hearer without many interactions of higher-order reasoning. For example, even when both a {∃&¬∀} and a {∀} lexical enrichment are considered to have equal likelihood as the literal meaning of 'some', the fact that one lexical enrichment goes in the same direction as the globally derived implicature means that the scalar enrichment is strengthened – in the sense that the RSA-LU system would attach a higher likelihood to the inference than the standard RSA system (Bergen et al. 2016). Other combinations of bias in lexical interpretation in this system could increase rates of SI in unembedded cases. It was also mentioned that not all enrichments may be considered by an RSA-LU system (Potts et al., 2015). In particular, it seems that entertaining the {∀} enrichment for 'some' gives unintuitive results. In other words, for reasons to be determined, this enrichment is disfavoured. If this is the case then, to the extent that a {∃&¬∀} enrichment is associated with 'some' we should expect a strengthening of the enrichment beyond what would be expected, given standard RSA.

The implications of these considerations impact on how we think about scalar diversity. Given a scalar term, W with meaning *w*, which may be associated with a locally enriched interpretation, *w & ¬ s*, (where *w* includes *s* in its denotation) more than the symmetric local enrichment, *s*, then this will potentially increase the strength of the unembedded scalar implicature. As the strength of association of the upper-bounding

local enrichment (its prior probability) increases, so should the strength of the scalar implicature, even in unembedded contexts[2]. So, here is scope for scalar diversity. If different scalar terms differ in the strength of association of the local enrichment, then we should expect different rates of implicatures in a fair test of these different scales.

However, whether we consider relations between scalar terms and their lexical alternatives (i.e. between, 'W' and 'S') as a source of scalar diversity, or the strength of association between a scalar term 'W' and its upper-bounding local enrichment *w&¬s*, the approaches to scalar implicature and alternatives mentioned above provide no a priori reason why either kind of relation should differ for different scalar terms. Let us consider why this is for the case of alternatives first.

For both approaches, scalar implicatures are only available when alternatives are relevant to the conversational purpose. As will be discussed in greater detail in Chapter 3, all possible contexts of a scalar term could be divided into three types: contexts where the stronger alternative is clearly relevant (upper-bound contexts), contexts where the stronger alternative is clearly irrelevant (lower-bound contexts) and contexts where the relevance of the stronger alternative is uncertain (neutral contexts). Assuming the proportions of each type of context are the same across different scales[3], we would expect if nothing else affects the availability of the alternatives and the mechanism of Gricean scalar implicature derivation, then over all possible contexts, different scalar terms should give rise to roughly similar rates of SIs via Gricean reasoning.

As for local enrichment, the question is whether we should expect rates to differ among different scalar terms. If we consider that local enrichments are sometimes mandated in embedded contexts, there is no prior reason to think one kind of scalar term finds itself embedded any more than the next, although this is open to testing. There is one other kind of factor discussed by Relevance Theorists, mentioned in Chapter

[2] This claims depends on what kinds of linking hypotheses between the Computational Level account (in the sense of Marr) put forward in various RSA systems and what happens at the level of processing. Potts et al. (2015) discuss some such hypotheses. For example, it may be that heuristic strategies approximate the computations described in the RSA literature. See Griffiths, Lieder, & Goodman (2015). One assumption sometimes made in the probabilistic literature is that prior probability and activation are highly correlated (see Jaeger & Snider, 2013). Sometimes here I will work with this assumption.

[3] Although the total number of all possible contexts might be different across scalar terms, there is no reason to expect that the ratio of upper-bound contexts to all possible contexts should be different across scales. The same reasoning goes for the other two context types.

1. This has to do with the specificity of the scalar term. I will elaborate on this point in the introduction to Experiments 2 and 3 below.

2.2 PREVIOUS STUDIES

2.2.1 Doran et al. (2009, 2012)

Doran (2009) and Doran et al. (2012) employed a verification paradigm to investigate the rates at which different scale types generate SIs. They manipulated both scale types and discourse contexts in one experiment. Four scale types were examined including cardinals, ranked ordering, gradable adjectives and quantificational items. A scalar term occurred in three possible contexts that differ in the number of relevant alternatives contained: (i) no other alternatives (e.g. (8)a), (ii) one stronger alternative (e.g. (8)b) and (iii) one stronger alternative and one weak alternative (e.g. (8)c). In their experiment, participants read a dialogue between two characters which began with one of the questions in (8a-8c). Following the dialogue, some relevant information was provided as fact. Participants were asked to take Literal Lucy's perspective (a literal-minded character) when they gave True or False judgment to Sam's response on the basis of the given fact.

(8) Irene: (a) How attractive is Kate? / (b) Is Kate gorgeous? / (c) Is Kate average-looking, pretty, or gorgeous?
Sam: She is pretty.
FACT: Kate was voted "World's Most Beautiful Woman" this year.

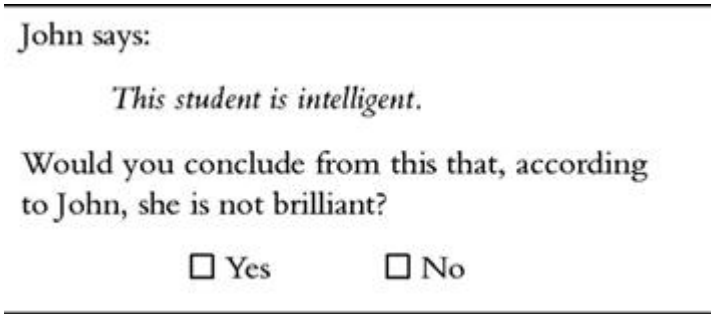
In (8), the literal reading of Sam's response was true based on the fact, but Sam's response might convey implicitly that 'she is pretty but not gorgeous' via a scalar inference. This enriched reading contradicted the fact. Therefore, a 'True' response indicated that the SI was not available or was not considered as part of the truth-conditional content of Sam's utterance. Whereas a 'False' response indicated that the SI had been drawn and was incorporated into the truth-conditional content. Their results showed that SIs drawn from gradable adjective scales were less frequent to arise or to be incorporated into truth-conditional content compared to SIs drawn from other scale types.

Moreover, providing stronger scale-mates in the context significantly increased the inference rates of adjective scales but not the other scales. The authors suggested that the observed variation might be due to two properties of the lexical scale. One is that adjective scales have vague boundaries and no upper bound, so that the weak scalar term from adjective scales is less likely to be interpreted as excluding the stronger term. The other property is that adjective scales are domain specific, so that weak scalar term need more context to evoke the relevant stronger alternative in order to give rise to an implicature. I consider this account in some more depth below.

van Tiel et al. (2016) pointed out several issues with Doran et al.'s experiments. One criticism was that different scale types were categorised in a rather coarse fashion. For instance, quantificational items included quantifiers, quantificational adverbs, and adverbial phrases (e.g. <all, most, some>, <always, frequently, sometimes>, <permanently, a year, a month>). Thus, the behaviour of different scales within one category might vary greatly and the comparisons made among these scale types might not accurately reflect the diversity phenomenon. Other issues were related to the experimental paradigm used. One was that the use of 'Literal Lucy' increased task complexity and rendered the data uninterpretable in terms of participants' own understanding of the sentences. The other was that the relevant fact provided in each dialogue might vary in their impact on participants' truth value judgement.

2.2.2 van Tiel et al. (2016)

In order to get a more fine-grained picture of any scalar diversity, and to overcome the methodological problems associated with Doran et al.'s verification task, van Tiel et al. (2016) used an inference paradigm to further study the variability of inference rates across a wider range of scalar expressions. Figure 1 is an example of a critical item (van Tiel et al., 2016: Experiment 2):


John says:

This student is intelligent.

Would you conclude from this that, according to John, she is not brilliant?

☐ Yes ☐ No

Figure 1 *Sample item in van Tiel et al. (2016) - Experiment 2*

Participants read a statement uttered by a character. Then they were asked whether or not the speaker implied the negation of the stronger statement in which scalar expression was replaced by its stronger scale mate. For example, when the character states that the student is intelligent, participants are asked whether, according to the speaker, the student is not brilliant. A 'Yes' response indicated that participants drew the scalar inference and a 'No' response indicated that the inference was unavailable. 43 scales were tested, including 2 quantifiers, 1 adverb, 2 auxiliary verbs, 6 main verbs and 32 adjective scales. Their results were consistent with those reported in Doran (2009). In particular, they found significant variation in the derivation rates of SIs across different scales, ranging from 4% to 100%. Quantifiers and modal expressions generated SIs more frequently than adjectives and verbs. Moreover, quantifiers and modal expressions consistently gave rise to SIs, but there was much greater variability within adjectives and verbs.

van Tiel et al. explored a wide range of explanations for this variability. They hypothesized that the availability of the stronger alternative and the distinctness of the scale mate could account for some of the variability in inference rates. The availability of the stronger alternative was measured in four parameters:

i. Association strength. The strength of association between the weak and strong scalar term was measured using a modified cloze task. Participants were given sentences containing a weak scalar term such as 'she is intelligent'. Their task was to provide three alternative words that could have occurred instead of the weak term (e.g. intelligent). Association strength was measured as the percentage of participants who mentioned a stronger scalar term.

ii. Grammatical class. Different scales were categorised by whether the scalar pair is from an open or closed grammatical class.
iii. Word frequencies. For each pair of scalar terms, they measured both the absolute frequency of the stronger alternative and the frequency of the weak term relative to that of the stronger alternative (i.e. Freq(W) / Freq(S)).
iv. Semantic relatedness. For each pair of scalar terms, the shared collocation of the weak and strong terms was measured as relatedness values (i.e. how frequently they co-occur with the same words).

The distinctness was measured in two parameters:

i. Semantic distance. For each pair of scalar terms, the semantic distance between the weak and strong term was measured as the difference in the perceived strengths between the two. In a distance task, participants were given pairs of sentences such as (a) She is intelligent. / (b) She is brilliant. Their task was to indicate on a seven-point Likert scale to what extent the sentence containing the weak term is stronger than the sentence containing the stronger alternative from the same scale. The distance value was calculated as the mean perceived strength.
ii. Boundedness. Different scales were categorised by whether the stronger scalar term from a scalar pair referred to an endpoint of the scale. For example, 'all' could be considered to denote an endpoint on a scale containing 'some', while 'hot' does not denote an endpoint on a scale with 'warm'.

The authors obtained the amount of variance explained by each parameter from a generalized linear mixed-effects models. All the parameters described above were included as fixed factors; participants and items were included as random factors. They found that the full model could explain 52% of the variance in rates of SIs, of which 22% was explained by fixed factors and 30% by random factors. Among fixed factors, none of the four parameters related to availability could independently explain the variability. The analysis showed that only semantic distance and boundedness could independently account for a significant amount of variance, where boundedness accounted for over

three times more variance than did semantic distance. van Tiel et al.'s investigation still left a large amount of variation unexplained.

Doran (2009) and van Tiel et al. (2016)'s findings have been viewed as evidence for the 'scalar diversity' phenomenon that different scales give rise to scalar inferences at different rates. Given the observed scalar variability, researchers should be cautious in generlizing the experimental results of 'some' and 'or' to other scalar terms. Apart from this methodological implication, these results are important for theoretical considerations. As discussed Section 2.1, scalar variability was predicted neither by theories that assume certain alternatives are lexically given nor by theories that assume no special status for any alternatives. If we consider these results at face value, they suggest that the likelihood that a scalar implicature is derived is affected by linguistic factors relating to the scalar terms themselves. For instance, in van Tiel et al. (2016), the rates of SIs varied over a wide range from 4% (e.g. <content, happy>) to 100% (e.g. <sometimes, always>). So far, only two factors, semantic distance and boundedness, have been established empirically to explain some of the variances in inference rates among different scalar terms.

It is interesting that no factors that relate to the 'availability' of the alternative explained the variance found in van Tiel et al.'s study. One might think that if a scalar term had a stronger association with its alternative, it might make that alternative term more activated when processing the sentence. Greater activation, or salience, of the stronger alternative might be expected to impact on which alternatives are considered in the derivation of implicatures. For example, in the RSA framework, a selection of alternatives is assumed to be made in the utterance context. One could assume that this process could be affected by strength of association between scalar terms. Similarly, from the perspective of the structural approach, the assumption is that a selection of alternatives is made from the set of formal alternatives. Again, if there is a strong association among terms, this might be expected to impact on whether the stronger alternative is considered in a derivation. It is possible that any effect that the availability of scalar alternative might have on scalar implicature cannot be well detected in the current paradigm. This is so because, for all scales, the stronger alternative is mentioned

in the process of eliciting the response. Since all alternatives are explicitly mentioned, this would raise the activation of that term to the same level in all cases. It remains for a more implicit measure of scalar diversity to determine if availability could be a factor.

By contrast, distinctness of scalar terms clearly has an impact on scalar variability. In particular, semantic distance and boundedness. These results are open to interpretation but it is possible that semantic distance is a factor because some scalar pairs are not clearly in an entailment relation. For example, 'snug' and 'tight' have slightly variable meanings such that, in context, their meanings may overlap. In such cases, one would be reluctant to infer *not tight* from 'snug'. To the extent that semantic distance reflects this aspect of the items, then the results suggest that some 'scalar diversity' remains an artefact of the task. It would not be true scalar diversity since the assumption has to be that the two scalar terms are sufficiently distinct to be in a clear entailment relation. In the studies reported below, I explore a further issue with the items used in establishing scalar diversity – scale homogeneity.

The second measure of distinctness, boundedness, relates less to methodological problems with the scalar-diversity paradigm and more to an apparently genuine factor in scalar diversity. If a scale has an endpoint that is denoted by a single lexeme, then this seems to make the scalar implicature more robust. There may be a number of reasons for this that have to do with the clear distinction in meaning between the weak term and the scale endpoint, or the likelihood that an endpoint is considered relevant by speaker and hearer and so on. Whatever the reasons, it seems that boundedness may be a genuine factor affecting rates of scalar implicature.

Together, these measures of distinctiveness account for only 22% of the variance in van Tiel et al's analysis. The authors speculate that the remaining variance is unsystematic but that simple frequency of implicature rates may impact on Gricean derivations of scalar implicatures. Their idea seems to be that the more a term gives rise to a scalar implicature, speakers will attach a higher prior likelihood that the scalar alternative will be relevant. This account is predicated on the assumption that there are no other real explanatory factors beyond distinctiveness; but that random variation can be amplified by its impact on Gricean reasoning. This suggestion is yet to be tested.

In the next section, I will discuss two potential factors that might affect the extent to which scalar terms give rise to scalar implicatures. One of these is methodological and the other is more substantive.

2.3 POTENTIAL FACTORS AFFECT THE STATUS OF ALTERNATIVES

2.3.1 Scale homogeneity

Scale homogeneity relates to the property of expressions to be underspecified or polysemous in their meanings. Consider the scale <hard, unsolvable> taken from van Tiel et al. (2016). 'hard' has a sense related to difficult. Under this sense, 'unsolvable' could be the hyponym of 'hard' with respect to problem-solving (e.g. this is a really hard question), while 'unbearable' could be the hyponym with respect to suffering (e.g. times were hard at the end of the war). Thus, it is sometimes the case that 'unsolvable' is not construed as being on the same entailment scale as 'hard', and the same happens with other scales such as <low, depleted>, <silly, ridiculous>, and <content, happy>.

When asked to judge whether 'hard' implies 'not unsolvable' or whether 'low' implies 'not depleted', participants in van Tiel et al.'s experiments may have evoked senses of these terms that are not on the same scale. By contrast, consider the scale <sometimes, always>. 'sometimes' and 'always' have fairly homogeneous senses across uses, relating to the frequency of an event. It would be difficult to construe these terms as not being in an entailment relation. Thus, when asked to judge whether 'sometimes' implies not always, participants were more likely to derive an implicature. I hypothesize that other things being equal, the more homogeneous the sense of the items in a pair, the higher the rate of scalar implicature derivation. I will test this hypothesis in Experiment 2.

2.3.2 Local enrichability

Local enrichability relates to the propensity of a scalar term to be locally enriched during utterance comprehension. As discussed in Chapter 1, local enrichment is generally seen as being a product of a separate mechanism from scalar implicature, when the latter is understood from a broadly Gricean perspective (see Carston, 2002; Geurts & van Tiel, 2013; Bergen et al. 2016). It is also a mechanism that is not restricted to embedded sites

when it comes to scalars. Rather, it may operate in a way that derives what looks like a 'global' Gricean scalar implicature.

One potentially important driver of local enrichment concerns a principle that the explicit proposition expressed by an utterance is itself relevant (Magri, 2009; Russell, 2012). Consider the following examples from Russell (2012):

(9) a. #Oh crap! Some of the students passed

b. Oh crap! Only some of the students passed

c. Oh crap! Not all of the students passed

In the context of an interjection like, 'Oh crap!', we can assume that (9b-c) are felicitous due to the fact that *not all* is part of the explicit assertion, as it provides the relevant bad news. The infelicity of (9a) suggests that it is not sufficient for the speaker to merely scalar-implicate *not all* to achieve a basic level of relevance in context. In this case, we can observe that the explicitly expressed proposition has no direct relevance at all in the context (in which there is an expectation of bad news) and this explains the infelicity. Evidence such as this suggests a difference in status of the *some* and *not all* implications between (9a) and (9b).

On my view, if the literal meaning of an utterance is not sufficiently relevant to the context, listeners will locally attempt to enrich the literal meaning to yield an adequately relevant interpretation of the sentence uttered. For example, if a speaker says "the cinema is some distance from the restaurant", the literal meaning of "some" does not make the utterance relevant at all and has to be enriched.

My conjecture is that scalar expressions differ in their susceptibility to local enrichment. For widely discussed scalar quantifiers like 'some' or 'few' and modal expressions (e.g. 'possible'), the unenriched meaning is very non-specific. To become even adequately relevant, utterances containing these expressions frequently get enriched locally. Such enrichment would more typically involve strengthening the meaning of the lower bound but sometimes also the inclusion of an upper bound (i.e. some but not many/most/all, possible but not probable/certain). Whereas for adjectives

and verbs, like 'intelligent' or 'start', the unenriched meaning tends to be more specific in terms of where on the scale to fix the interpretation. So compared to quantifiers and modals, these expressions are less liable to be locally enriched to achieve an adequate degree of relevance. As discussed in section 2.1 above, if scalar terms differ in their liability for local enrichment in the upper bound, then this could feed into the rates at which even sentences with unenbedded scalar terms are understood as including scalar implicature in their meaning.

There is an additional, related, consideration here when it comes to explaining the results scalar diversity experiments. In van Tiel et al's experiment, sentences are presented out of context, leaving participants to wonder how a sentence with these weak scalar terms might be relevant. Lower-bound strengthening requires context to some extent. For example, when the cinema is described as, 'some distance' from the restaurant, it may mean that it is a long walk, that it is too far to walk or that it is a long drive, depending on other background factors. By contrast, upper-bound enrichment may come into play when background information is minimal.

My conjecture is that local enrichability may have been a factor in increasing the rates of scalar inference for weak quantifiers and modals in van Tiel et al.'s study, due to their being less specific than the adjectival terms and possibly also due to lack of context in van Tiel et al.'s experimental items. I hypothesize that other things being equal, the more liable a scalar term is to be locally enriched, the higher the rate of implicature response. I will test this hypothesis in Experiment 3.

2.4 EXPERIMENT 1

2.4.1 Overview

Experiment 1 was more or less a replication of Experiment 2 of van Tiel et al. (2016) using a different measurement scale. I obtained a continuous measure of participants' judgment on the availability of SIs for each scalar pair.

2.4.2 Methods

2.4.2.1 Participants

36 participants were recruited from our university campus via an online psychological subject pool. They participated either for course credit or £2.5. All participants speak English as a native language.

2.4.2.2 Materials and procedure

I tested all 43 scale pairs from van Tiel et al. (2016) in an inference task to measure scalar implicature derivation. The only difference in procedure was that, instead of providing a yes/no response, participants were asked to rate on a 0-100 scale to indicate to what extent they could infer from the speaker's statement that the speaker does not believe the stronger alternative. In van Tiel et al. (2016) Experiment 2, the statements were created based on the results of the sentence completion task, e.g. The __ is attractive but she isn't stunning. Three statements were selected for each scale, partially, based on the completion frequency. Here we selected the two more frequent statements for every scale (see Appendix A.1 for a list of items used). If the statements used in the original study had the same completion frequency, a random selection was made. We also used the exact same control items from van Tiel et al.'s experiment. Four lists were created, each participant judged 21 (22) experimental items and 7 control items. No participant judged the same scale twice. Participants were randomly assigned to one of four lists. A randomized order of presentation of the items was created for each participant.

2.4.3 Results

Two participants were excluded from the analysis for making mistakes in more than four control items. The mean ratings for entailments and non-coherent inferences were 87 and 8. The mean ratings for experimental items for all scalar items are shown in Figure 2 (red bars). The rates of SIs from van Tiel et al. (2016, Experiment 2) are also included in that figure (blue bars).

We carried out one-way ANOVAs with the ratings on the inference task as the dependent variable and lexical categories as the independent variable. The ratings were averaged by items (43 scales) before entering into the analysis. There was a statistically

significant difference among lexical categories (F(3,39)=9.52, p<.001). A Tukey post-hoc test revealed that the ratings of scalar inference for quantifiers (M=76.03) and modals (M=64.35) were significantly higher than for adjectives (M=34.95, p=.001, p=.006) and verbs (M=35.30, p=.004, p=.034), but there were no statistically significant differences between quantifiers and modals (p=.77), and between adjectives and verbs (p=1). These results are in line with those seen in van Tiel et al. (2016). Inspecting the graph, one can see some differences among items, but the general pattern is the same.

To examine whether factors identified by van Tiel et al. (2016) could still explain some of the variation found in Experiment 1, I conducted a multiple linear regression analysis to predict the ratings of scalar inferences in Experiment 1 from all the potential factors reported in van Tiel et al. (2016) including association strength, grammatical class, word frequencies, semantic relatedness, semantic distance and boundedness. The ratings of scalar inferences in Experiment 1 were averaged by item (43 scales) before entering the analysis. The results of the linear regression are summarized in Table 1. The model explained 48.7% of the variance (R2=.562, F(6,35)=7.48, p<.001). As in van Tiel et al. (2016) only semantic distance and boundedness were significant predictors of the inference task results, whereas other factors did not make a significant contribution to the model.

	β	**SE**	**t**	**p**
(Intercept)	8.2649	18.5252	0.45	0.6582
Association strength	0.0238	0.1081	0.22	0.8270
Grammatical class	13.5745	9.4287	1.44	0.1588
Word frequencies	-3.6025	2.6046	-1.38	0.1754
Semantic relatedness	3.0363	14.0848	0.22	0.8306
Semantic distance	7.2344	3.2026	2.26	0.0302*
Boundedness	-20.8023	4.8969	-4.25	0.0002*

Table 1 *Results of multiple linear regression for inference ratings of Experiment 1*

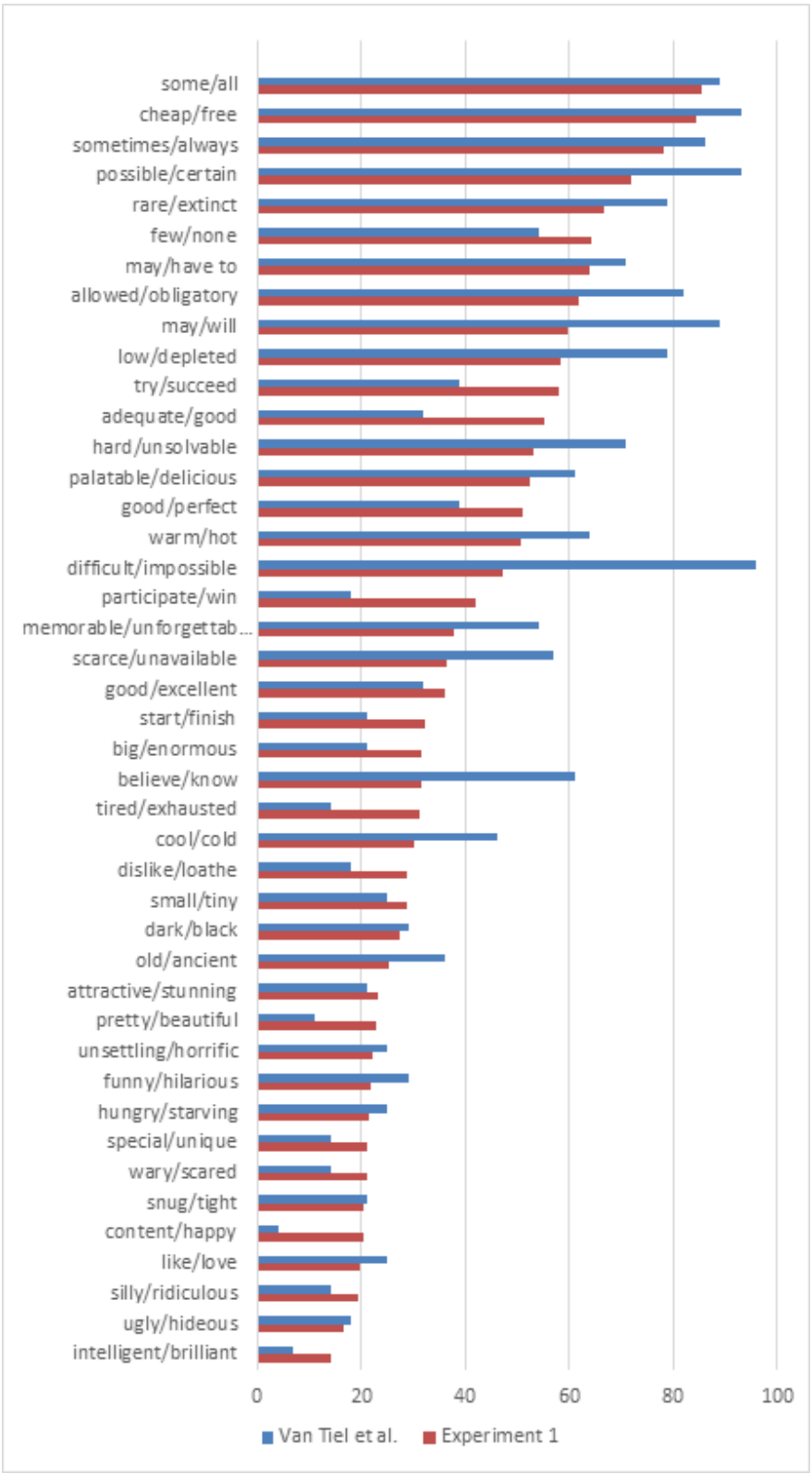


Figure 2 *Mean inference ratings for Experiment 1*

2.4.4 Discussion

Experiment 1 established that there is a considerable amount of variation among scalar terms in terms of how strongly they give rise to scalar implicatures. The general pattern found in van Tiel et al. (2016) was replicated, but with a different measurement scale. These results provide additional evidence that conflicts with the uniformity assumption. Experiment 1 also showed that established factors, semantic distance and boundedness, could still explain some of the variation. But the remaining variance calls for further investigation.

2.5 EXPERIMENT 2

2.5.1 Overview and prediction

The aim of Experiment 2 is to investigate whether scale homogeneity could explain some of the variation in the rates of SIs found in Experiment 1. Scalar homogeneity was operationalised in terms of the naturalness judgment of 'X but not Y' construction where <X, Y> is a scalar pair and X is stronger than Y.

In Experiment 2, a group of participants was asked to rate the naturalness of sentences of the form *X but not Y*, e.g. (10):

(10) a. The student is brilliant but not intelligent. <brilliant, intelligent>
b. The water is hot but not warm. <hot, warm>
c. The dancer finished but she did not start. <finish, start>

'But' has a denial-of-expectation conventional implicature. Thus a sentence, 'X but not Y' is felicitous to the extent that X can be construed to not strictly entail Y but normally or often to imply Y. A scale with high homogeneity is one where the stronger term is interpreted to entail the weaker term. Entailment relations require that if X entails Y, whenever X holds, Y must hold. Therefore these 'X but not Y' sentences should be very unnatural if the contrasting predicates X and Y are on the same entailment scale. So if the naturalness rating for 'but' sentences is low, it suggests a high degree of homogeneity for the given scale; whereas if the rating is high, then the degree of homogeneity is relatively low.

Following the hypothesis outlined in section 2.3.1, that other things being equal, the more homogeneous the sense of the items in a pair, the higher the rate of scalar implicature derivation. I predicted that the naturalness rating for scalar expressions in Experiment 2 should negatively correlate with the results of Experiment 1.

2.5.2 Methods

2.5.2.1 Participants

40 Participants were recruited via Amazon Mechanical Turk. They were asked to indicate their native language and only participants with English as a native language were included in the analysis.

2.5.2.2 Materials and procedure

Figure 3 is an example item. I used the 43 scales investigated in Experiment 1 to construct experimental sentences for Experiment 2. The experimental sentences were of the form *X but not Y*, where according to van Tiel et al. (2016), X and Y are a pair of scalar terms and X is stronger than Y; for example, 'The student is brilliant but not intelligent'. We constructed two experimental sentences for every scale (see Appendix A.2 for a list of items used). The nominal ('student') used in each experimental sentence was the same as for the corresponding statement in Experiment 1. For the auxiliary verb 'may', experimental sentences were constructed differently to make sure that the weaker term was in the scope of negation; for instance, 'The lawyer will appear in person but it is not the case that he may appear in person'. In addition, we constructed 7 filler sentences, which contained clearly felicitous (e.g. 'The banker is rich but not happy') and clearly infelicitous items (e.g. 'The man left the party but he never came'). Participants were asked to rate how natural these constructions are on a 1 (very unnatural) -7(very natural) scale. Each participant judged 43 experimental sentences and 7 fillers. No participant judged the same scale twice. Eight survey versions with pseudo-randomized order of items were created. Participants were randomly assigned to one of eight surveys.

Figure 3 *Sample item in Experiment 2*

2.5.3 Results

Two participants were excluded because their mean ratings for the infelicitous items were above 5. The mean ratings for the clearly felicitous and clearly infelicitous control items were 5.8 (CIs: 5.51-6.09) and 2.31 (CIs: 2-2.62). The mean rating for experimental items ranged from 1.33 (<may, will>) to 4.47 (<special, unique>). Critically, I found that the naturalness ratings of the 'but' sentences correlated negatively with the ratings of SIs in Experiment 1 (r=-.31, p=.04) – see Figure 4. In addition, it also correlated negatively with the results from van Tiel et al. (2016, Experiment 2) (r=-0.36, p=.02). These results confirmed the prediction outlined earlier. I defer discussion of these results until after the combined analysis in Section 2.8.

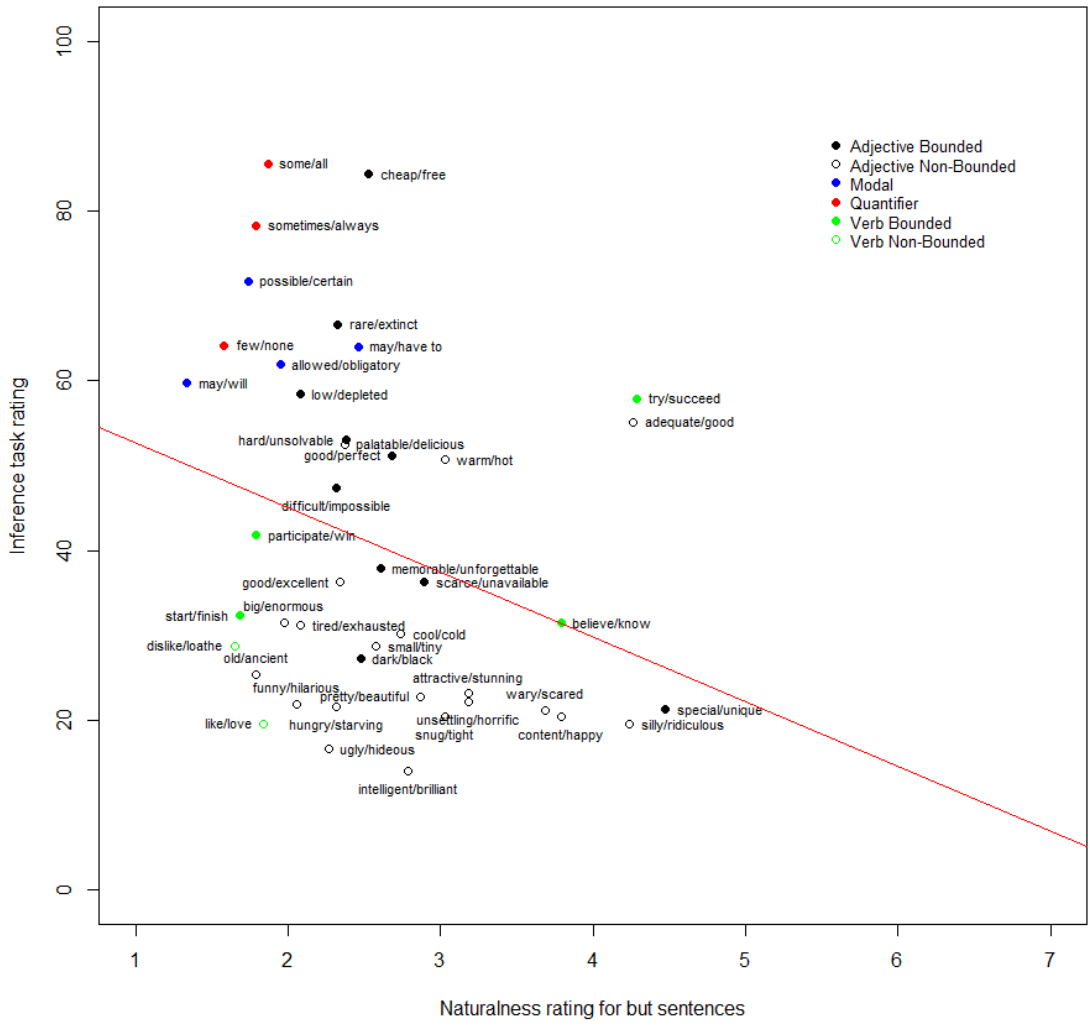


Figure 4 *Negative correlation between the absence of homogeneity and inference rate*

2.6 EXPERIMENT 3

2.6.1 Overview and prediction

The aim of Experiment 3 is to investigate whether local enrichability could account for some of the variation in the rates of SIs found in Experiment 1. Local enrichability is the degree to which a weak scalar term is liable to undergo local enrichment. It was operationalised in terms of the naturalness judgment of 'X so not Y' construction where <X, Y> is a scalar pair and X is stronger than Y.

A separate group of participants rated the naturalness for sentences of the form, 'X so not Y', e.g. (11):

(11) a. The student is brilliant so not intelligent. <brilliant, intelligent>

b. The water is hot so not warm. <hot, warm>

c. The dancer finished so she did not start. <finish, start>

The discourse function of 'so' contrasts with that of 'but' in a number of ways (see Blakemore, 2002). 'So' implies that the second segment follows in some way from the first. While 'X but not Y' suggest that one might expect Y, given X, 'X so not Y' suggests that one might expect not Y, given X. Thus, 'X so not Y' sentences should be more coherent to the extent that the weaker scalar expression can be locally enriched to have an upper bound meaning. For example, to understand (11b) as felicitous, 'warm' must have its meaning locally enriched to have an upper-bound meaning 'warm but not hot'. Notice that this has to involve local enrichment rather than Gricean scalar-implicature reasoning because the weaker term is in the scope of negation. The negation of an un-enriched weaker term is more informative than the negation of an enriched weaker term. In Experiment 3, if the naturalness rating for 'so' sentences is low, it suggests that the scalar expression is less liable to be enriched; whereas if the rating is high, then it is more liable to be locally enriched.

Following the hypothesis outlined in section 2.3.2, that if other things being equal, the more liable a scalar term to be locally enriched, the higher the rate of the seeming implicature response. I predicted that if local enrichability inflated rates on the inference task, then the naturalness rating for scalar expressions in Experiment 3 should positively correlate with the results of Experiment 1.

2.6.2 Methods

2.6.2.1 Participants

40 Participants were recruited from our university campus via an online psychological subject pool. All participants native English speakers. They came into the lab to fill out a paper-based survey.

2.6.2.2 Materials and procedure

Figure 5 is an example item. I used 43 scales investigated in Experiment 1 to construct experimental sentences for Experiment 3. Two experimental sentences were constructed for each scale (see Appendix A.3 for a list of items used). The experimental sentences were of the form *X so not Y*, where X is stronger than Y; for example, 'The student is brilliant so not intelligent'. As in Experiment 2, the nominal ('student') used in each experimental sentence was from statements used in Experiment 1. For the auxiliary verb 'may', experimental sentences were constructed differently; for example, 'The lawyer will appear in person so it is not the case that he may appear in person'. 7 filler sentences were constructed, which contained clearly felicitous (e.g. 'The cup is red so not blue') and clearly infelicitous items (e.g. 'The banker is rich so not happy'). Participants were asked to indicate how natural these constructions are on a 1 (very unnatural) -7 (very natural) point scale. Each participant judged 43 experimental sentences and 7 fillers. No participant judged the same scale twice. Eight paper-based survey versions with pseudo-randomized order of items were created.

Sentence	unnatural ---------- natural
The student is brilliant so not intelligent.	1 2 3 4 5 6 7

Figure 5 *Sample item in Experiment 3*

2.6.3 Results

The mean rating for the clearly felicitous and clearly infelicitous control items were 5.89 (CIs: 5.59-6.20) and 1.53 (CIs: 1.30-1.75). The mean rating for experimental items ranged from 1.13 (<start, finish>) to 5.2 (<few, none>). Critically, I found that the naturalness of the 'so' sentences positively correlated with the ratings of SIs in Experiment 1 (r=.44, p=.004) – see Figure 6. And it also positively correlated with the results from van Tiel et

al. (2016, Experiment 2) (r=0.35, p=.02). Thus, the prediction of Experiment 3 was confirmed. I defer discussion of these results after the combined analysis in Section 2.8.

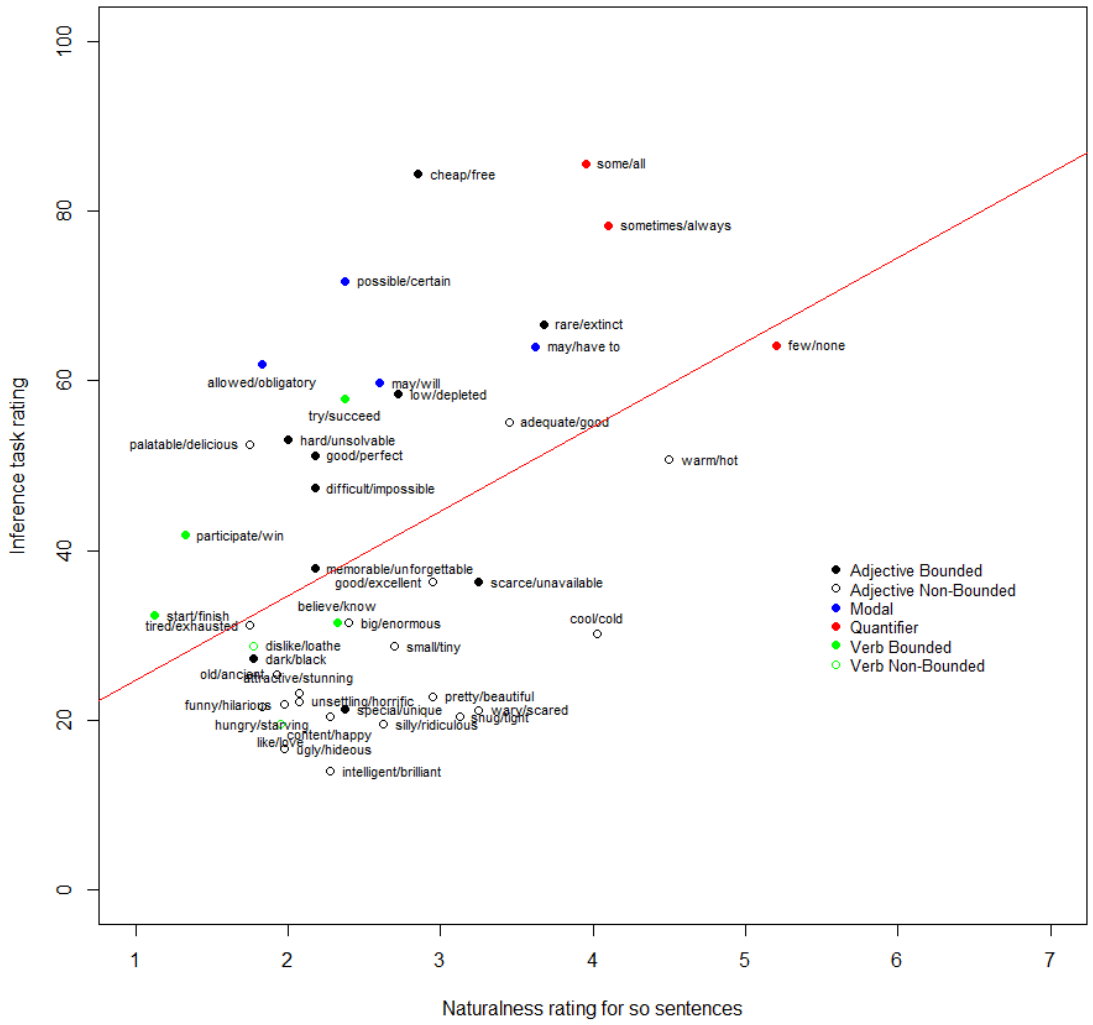


Figure 6 *Positive correlation between the propensity of local enrichment and inference rate*

2.7 COMBINED ANALYSIS

To investigate the proportion of variance explained by all the factors I have looked at so far, multiple linear regression analyses were conducted to predict the ratings of scalar inferences in Experiment 1 from scale homogeneity degree, propensity for local enrichment, and all factors established in van Tiel et al. (2016). The rating of scalar inferences in Experiment 1, and the naturalness rating from Experiments 2 and 3 were

averaged by item (43 scales) before entering the analysis. The results of the linear regression are summarized in Table 2.

I found that the regression model accounted for 63% of the variance (R^2=.70, F(8,33)=9.73, p<.001). In this model, I found the propensity for local enrichment, semantic distance and boundedness were substantial factors, with the propensity for local enrichment explaining 16%, semantic distance explaining 6.5%, and boundedness explaining 28%. None of the other factors significantly accounted for the variation in the rates of SIs. In this model, scale homogeneity did not significantly explain the variance. The variance in inference ratings explained by scale homogeneity was largely overlapped with the variance accounted for by semantic distance. When the semantic distance was omitted from the model, scale homogeneity could explain a significant amount of the variance. In fact, I found that scale homogeneity was highly correlated with semantic distance (r=-0.53, p<.001).

	β	**SE**	**t**	**p**
(Intercept)	-19.7385	23.3367	-0.85	0.4037
Scale homogeneity	-3.1419	3.0077	-1.04	0.3038
Local enrichment	10.4415	2.6684	3.91	0.0004*
Association strength	0.0479	0.0921	0.52	0.6064
Grammatical class	-1.5059	9.0324	-0.17	0.8686
Word frequencies	-2.9258	2.2620	-1.29	0.2048
Semantic relatedness	-8.1512	12.3029	-0.66	0.5122
Semantic distance	8.2910	3.1502	2.63	0.0128*
Boundedness	-21.5644	4.1711	-5.17	<.0001*

Table 2 *Results of combined analysis*

2.8 DISCUSSION OF EXP. 2 AND 3

I adapted the items from van Tiel et al. (2016, Experiment 2) for these two naturalness judgment tasks. Participants were asked to judge the felicity of sentences of the form 'S but/so not W' where 'S' is the stronger term from the scalar inference judgement task ('all', 'hot' etc.) and 'W' is the weaker term ('some', 'warm' etc.). The respective sentences have different felicity conditions due to the function of 'but' and 'so' respectively. I argue that the 'but' sentences probe scale homogeneity, while the 'so' sentences probe what I call local enrichability.

Concerning scalar homogeneity, the degree to which the senses of the items in a scalar pair are homogenous is measured in terms of entailment relation. In Experiment 2, if participants find a way to read a sentence of the form 'S but not W' felicitous, then it indicates that the items of this scalar pair could be constructed as not always being on the same scale, thus relatively low homogeneity of the pair. The results of Experiment 2 showed that the degree of homogeneity varied across different scales. That is, quantificational and modal scales, as well as most verb scales, are in clear entailment relation, but most adjective scales are not. As discussed in Section 2.3.1 (scale homogeneity), this variation in the degree of homogeneity is expected due to factors like underspecification or polysemy. I also found that high homogeneity led to high rates of SIs, whereas low homogeneity led to low rates of SIs.

The results of Experiment 2 are closely related with the hypothesis discussed in Doran (2009). They suggested that there are domain-general scalar expressions such as quantifiers and modals and domain-specific ones such as adjectives. The former are more likely to give rise to SIs, whereas the latter requires more contexts in order to derive SIs. Therefore, the homogeneity difference among different scalar terms found in Experiment 2 is in line with the distinction between domain-general and domain specific scalar expressions. Doran (2009) found that only the derivation of adjective scales was affected by providing stronger scalemate in the context. This result might be due to the low homogeneity in adjective scales. That is, without restriction in the context, the use of scalar adjectives may evoke alternatives that are irrelevant in deriving scalar implicatures.

Since scalar homogeneity is strongly correlated with semantic distance, it raises the question of what the relation between the two concepts is. One possibility is that the naturalness of 'S but not W' measured in Experiment 2 reflects semantic distance rather than homogeneity. If the sentence of the form 'S but not W' was judged to be felicitous, then it might also indicate that for a scalar pair <S,W> S is not necessarily stronger than W in terms of semantic entailment. If this is the case, I argue the naturalness of 'S but not X' is a better way to measure semantic distance compared to van Tiel et al.'s distance task. Since in their task participants were asked explicitly how much stronger one statement is than the other, the results of this distance task might be influenced by participants' own understanding of 'strong', which might be something other than semantic entailment. Another possibility is that the naturalness of 'S but not W' measured both semantic distance and homogeneity. We find that pairs like <silly, ridiculous>, <snug, tight>, and <content, happy> are both semantically close and also rated high on the naturalness of 'S but not W', as well as ranking low on scalar inference availability. In Chapter 3, I will further explore whether it is lack of semantic distance or lack of scale homogeneity that explains low rates of implicature, particularly for adjective items.

Concerning the local enrichability, it is a new factor unexplored in previous studies. In Experiment 3, if participants find a way to read the sentences of the form 'S so not W' as felicitous, then it indicates that 'W' (e.g. 'some') has been locally enriched in the scope of negation to exclude *s* (e.g. *all*). The results of Experiment 3 showed that the naturalness of 'S so not W' varied across different scales, suggesting that weak scalar terms differ in their propensity for being locally enriched. The positive correlation between the naturalness of 'S so not W' and the rates of SIs measured in the inference task suggested that local enrichability is influencing the judgement in van Tiel et al's original inference task. Local enrichment can give rise to what looks like a standard Gricean scalar implicature in the unembedded case and this could have inflated rates measured in the inference task.

2.9 CONCLUSION

This chapter explores the question whether certain lexical scalar alternatives have a special relation with scalar terms. While neo-Gricean theory assumes a special status for lexical alternatives that lie on the same scale as the scalar term, the structural approach and standard rational speech-act approach assume no special status for any alternatives. What these different theoretical approaches to alternatives have in common is the implicit assumption that the status of alternatives should be homogeneous across different scalar terms. The homogeneous status of alternatives leads to a uniformity assumption that the availability of scalar implicatures drawn from utterances containing different scalar terms should be the same.

Experiment 1 provides evidence against this assumption, replicating previous research by van Tiel et al. It showed that there is a considerable amount of variation in the rates of scalar implicatures generated by different scalar terms. These results suggest that scalar terms may have different strength of relations. I also replicated the other results in van Tiel et al. (2016) which suggest, surprisingly, that different levels of association between scalar terms are not responsible for some of the variance in rates of implicature drawn. We find, as did van Tiel et al. that measures of distinctness – semantic distance and boundedness – can account for some of the variation among rates of scalar inference. Of these, I conjectured that semantic distance could point to a methodological problem with the paradigm since certain terms may be so close in meaning that they are liable to overlap. By contrast, I suggest that boundedness may point to a genuine underlying factor affecting rates of scalar implicature.

Van Tiel et al. suggest that remaining variation in rates of scalar implicature may be unsystematic. In Experiment 2 and 3, I explored further factors which might affect the scalar variability. In addition to previously established factors (i.e. semantic distance and boundedness), I found two factors, scalar homogeneity and local enrichability, could also explain a significant amount of the variance. However, scale homogeneity strongly correlated with semantic distance and did not independently explain variance in this inference task. As both semantic distance and scale homogeneity are factors that bear on this inference task method of determining scalar diversity and the choice of actual

items for the experiments, it seems clear that some of the 'diversity' among scalar terms may be an artefact of the experiment.

By contrast, local enrichability did explain a significant amount of the variance but does not bear on methodological issues. This result is in line with assumptions in RSA-LU accounts, that the extent to which a scalar term, 'W' is likely to be enriched to the upper bound, *w&¬s*, impacts on the strength of even an unembedded scalar implicature. The results are also broadly in line with the idea from the Relevance tradition that less specific scalar terms like 'some' are more liable to be locally enriched than more specific terms.

In the next chapter, I investigate the status of alternatives in a corpus-based study. This will allow for a further investigation on whether factors established in this chapter could affect scalar implicatures derivation in real use.

Chapter 3 SCALAR DIVERSITY – A CORPUS STUDY

This chapter further investigates scalar diversity. I argue that the uniformity assumption could be better tested in a large-scalar corpus-based study. Here I construct a Twitter corpus of sentences containing scalar terms and re-examine the uniformity assumption in a corpus-based paraphrase task. Previously established factors are tested to see if they affect the scalar variability in *everyday real use*. The goals of this chapter are to establish scalar diversity properly and to explore whether factors that explain some of the variation in the laboratory-based tasks could account for the variation in the corpus-based study.

3.1 INTRODUCTION

Recall the uniformity assumption discussed in Chapter 2, that is, scalar terms do not vary in the degree to which they are liable to give rise to scalar implicatures. This is an implicit assumption in different theoretical approaches to alternatives on the basis of the homogeneous status of alternatives. However, in the experiment presented in Chapter 2, Experiment 1, I found different scalar terms give rise to scalar implicatures at different rates (in line with previous work, e.g. Doran, 2009; Doran et al., 2012; Van Tiel et al., 2016). One way to interpret the observed scalar variability is that there is variation in the status of alternatives. However, van Tiel and colleagues found that none of the factors related to the availability of alternatives could explain the variation. Rather, around two thirds of the variation are explained by a linguistic factor related to scalar terms (i.e. boundedness) and a novel factor related to local enrichment mechanism (i.e. enrichability). While the boundedness of alternative could be seen as having a bearing on the relation between scalar term and its alternative, I argued that local enrichability is related to a different aspect of scalar implicature – local pragmatic enrichment. That local enrichability can explain some of the variance in the inference task results is predicted by approaches to scalars that allow for two modes of scalar enrichment – global and local. So the results of Chapter 2 add support to such approaches. Alternatively, one could argue that the observed scalar variability may not be a good estimate of the scalar diversity pattern. This will be discussed in more detail below. If

this is the case, it would raise at least two questions: (i) assuming a corpus-based study provides a better estimate of the diversity pattern, whether the rates of SIs found in the corpus-based study vary to the extent found in laboratory studies, and (ii) whether established factors could explain the variation found in *real use*.

In the remainder of this section, I discuss reasons why previous lab-based tasks might not accurately reflect scalar diversity and suggest that the uniformity assumption could be better tested in a large-scale corpus-based study.

3.1.1 The availability of scalar implicatures and the role of context

The theoretical approaches to alternatives outlined in Chapter 2 differ in how alternatives are made available given the use of a scalar term. However, what is in common in these approaches is the application of Gricean reasoning that scalar implicatures are only available when alternatives are relevant to the context. Depending on the relevance of alternatives, all possible contexts of a scalar term could be categorised into three types: upper-bound contexts, low-bound contexts and neutral contexts.

Upper-bound and low-bound contexts

Upper-bound contexts are those where the stronger alternative is clearly relevant to the context. Whereas lower-bound contexts are those where the stronger alternative is clearly irrelevant. Consider two examples adapted from Breheny, Katsos, & Williams (2006).

(1) Mary: Are you going to host all of your relatives?
    John: I will host some of my relatives.
(2) Mary: Why are you cleaning your apartment?
    John: I will host some of my relatives.

In both cases, a stronger alternative for John's utterance could be 'I will host all of my relatives'. (1) is an upper-bound context as the stronger alternative is relevant to answer Mary's question. Thus in ((1)), *not all* implicature is licensed. Whereas (2) is a lower-bound context as the stronger alternative is not relevant to answer Mary's question. Then *not all* implicature is implicitly cancelled or unlikely to arise. Breheny, Katsos, & Williams (2006) show in a reading-time study that participants do derive the scalar

implicature more in contexts like (1) than contexts like (2). In both upper-bound and lower-bound contexts, it is quite explicit whether or not the stronger alternative is relevant, I expect that in such contexts, judgments on the presence or absence of a scalar implicature should be consistent among comprehenders.

Neutral contexts

Neutral contexts are those where the relevance of the stronger alternative is uncertain. Consider (3) (taken from Degen, 2015):

(3) [two people talked about The Civilian Conservation Corps in the United States]
Speaker A: Well, it seems like it would develop pride, you know, in people, in their own country.
Speaker B: It would certainly help them to appreciate **some of the things that we have here**.

((3)) is a neutral context as it is unclear whether the stronger alternative for B's utterance (replacing 'some' with 'all') is relevant. According to Grice, in such contexts, scalar implicatures may arise on the basis of the assumption that the speaker should provide appropriate specification if the hearer is likely to be interested in a certain question (see Grice, 1989: page 38). That is to say, even if there is not a specific question about *all* being addressed in the context, the speaker should know that if they specify that *some* is the case, the hearer, for predictable reasons, may wonder whether or not *all* is the case. In the situation where the hearer is likely to be interested in the stronger proposition, Grice suggests that it is incumbent on the speaker, to some extent at least, to give the information if they can. In cases where it is common ground that a hearer is likely to wonder about *all*, the implicature from *some* to *not all* would become available. However, in neutral contexts, judgments on the availability of scalar implicatures should be less consistent among comprehenders compared to judgments in the upper-bound and low-bound contexts, as the judgment would be based on factors other than explicit contextual questions.

Table 3 summarizes how different contexts affect the availability of scalar implicatures and the agreement among comprehenders' judgments.

Context	Availability of SIs	Agreement among comprehenders
Upper-bound	yes	consistent
Lower-bound	no	consistent
Neutral	mixed	inconsistent

Table 3 *Overview of the role of context in Gricean derivations of scalar implicatures*

3.1.2 Re-examining the uniformity assumption

Given that the derivation of scalar implicatures is affected by contextual constraints, the uniformity assumption is in fact formulated based on a hidden premise in (4).

(4) The proportions of each type of context are the same across different scales.

Under this premise, if nothing else affects the availability of the alternatives and the mechanism of Gricean derivations, then different scalar terms should give rise to roughly similar rates of SIs over all possible contexts.

However, (4) is not satisfied in previous studies that investigated the uniformity assumption. To illustrate, consider the inference tasks in van Tiel et al. (2016). In their studies, the rates of SIs generated by different scalar terms were measured using a very small sample of hand-crafted sentences. These sentences were designed to be bland or to contain little contextual information. However, participants may imagine what context these sentences are in. In the case of scalar quantifiers, for example, participants were presented with “He saw some of them” (exp.1) or “The bartender saw some of the cars” (exp.2). Then they were asked whether or not the speaker believes that “He did not see all of them” or “The bartender did not saw all of the cars”. In these cases, it is likely that participants consider the sentence with ‘some’ as an answer to an implicit ‘how many’ question. This would make the stronger alternative containing ‘all’ relevant and give rise to a scalar implicature. The same reasoning goes for modal expressions. That an implicit question ‘how likely’ would be considered given an utterance with a modal expression. However, in the case of scalar adjectives, for example, participants were presented with “He / The student is intelligent”, and they were asked if the speaker implied that “He / The student is not brilliant”. In these cases, several implicit questions could be raised such as ‘Is he / the student intelligent or dumb?’ or ‘Is he brilliant?’. Depending on what the implicit question is, the relevance of the stronger alternative

varies. Thus, scalar implicatures triggered by adjectives might not be as robust as those triggered by quantifiers and modal expressions.

Therefore, although previous studies provide little contextual information, participants might create their own context by virtue of an inferred contextual question. As a result, the upper-bound and neutral contexts might not be distributed evenly across different scalar terms, which is inconsistent with the premise in ((4)). Assuming ((4)) is indeed the population distribution, then laboratory tasks that measuring the rates of SIs with very small samples might not provide good estimates of the rates generated by different scalar terms.

Results from corpus-based studies also suggest that the derivation rate measured with a small sample of artificial examples could be considerably different from the rate measured with a large sample of naturalistic data. Take the derivation rate from *some* to *not all* as an example. The scalar implicature from *some* to *not all* was traditionally measured in highly controlled experimental settings using hand-crafted examples. Although the exact rate differs across different dependent measures, in general, the *not all* implicature was shown as frequent and context-independent, ranging from two-third of the time to near ceiling performance (Bott & Noveck, 2004; Degen & Goodman, 2014; Geurts & Pouscoulous, 2009; Noveck, 2001). Yet, the prevalence of such inference was challenged by recent corpus-based studies. For example, Degen (2015) measured the rate of *not all* implicatures in naturally occurring utterances containing *some*-NP and found around half the time *some* is used, an SI reading is not judged to be available. Also, a corpus study done by Eiteljörge, Pouscoulous, & Lieven (2016) found that in adult speech, the production of *some* carrying an inference was infrequent (around 15% of adult's uses of 'some'), which was in-line with findings from Degen (2015). These results suggest that the scalar implicature from *some* to *not all* is less frequent and more context-dependent in the naturally occurring language. Following this, it is possible that other scalar terms would also give rise to different rates of SIs depending on the task. Thus, it is reasonable to doubt whether the methods used previously are adequate to provide a fair test of the homogeneity/diversity hypotheses.

I argue that the uniformity assumption would be better tested in a large-scale corpus-based study. I expect that, if one extracts a representative sample of the use of a scalar term from naturalistic data, the distribution of context types in the sample should be the same as that distribution in the population. So the premise in (4) is satisfied. The mean rate of scalar implicatures for a scalar term obtained from such sample should be a good estimate of the rate over all possible contexts. If the uniformity assumption holds, then the mean rates of SIs calculated for different scalar terms should be the same. Furthermore, the average agreement level on comprehenders' implicature judgments should be the same across different scalar terms.

In the following, I describe the collection and annotation of a Twitter corpus of sentences containing scalar terms. I then present a corpus-based paraphrase task that investigates the uniformity assumption with data provided by the Twitter corpus.

3.2 CREATING A TWEET CORPUS

3.2.1 Collection

I aim to create a corpus of texts containing scalar terms that can be used as experimental stimuli in the paraphrase task. 28 scalar terms were selected from 43 of those investigated in van Tiel et al. (2016). There were 2 quantifiers, 1 adverb and 25 adjectives scales (see Appendix A.4 for a full list of scales). I did not select auxiliary verb and verb scales (e.g. <may, will>; <participate, win>) due to implementation issues of the paraphrase paradigm[4]. In addition, 7 adjective scales were not selected either due to the infrequent occurrence of the weak term, like <unsettling, horrific>, or due to the infrequent occurrence of the specific word sense[5]. For each of these 28 scales, I collected texts containing the weak term from Twitter using Twitter API. These texts were retrieved from the United States between December 2016 and January 2017. The maximum length of a text post is 140 characters as this is the restriction of Twitter. To

[4] For target sentences containing these scalar terms, comparison sentences need to be constructed differently. Inserting 'but not <strong term>' would lead to ungrammaticality or would leave the weak term outside the scope of negation. For example, target sentence: I may hate the person I've become; comparison sentence: * I may, but not will hate the person I've become.

[5] For instance, considering <cool, cold>, in the Twitter contexts, for the weak term 'cool', the sense 'slightly cold' occurs much less frequently compared to the sense 'calm'.

ensure a text consisted of at least one single sentence, I filtered out tweets containing less than 30 characters.

3.2.2 Annotation

I first used automatic annotation to filter out texts where scalar terms appear in linguistic environments (such as negative contexts) in which the inferences are unavailable or less likely to arise. I then excluded texts in which the weak term was polysemous and used in a sense that is unrelated to the stronger term, based on crowdsourced annotations. I describe these two steps below.

3.2.2.1 Automatic filtering

Texts were tagged using GATE Twitter part-of-speech tagger (Derczynski et al., 2013). Since many scalar terms can be used as more than one part of speech, I filtered out cases where the scalar term was not used in the part of speech specified by the given scale. For instance, in order to study the adjective scale <hard, unsolvable>, I excluded cases where 'hard' was used as an adverb (e.g. work hard). Moreover, I used regular expressions to exclude cases where the scalar term appeared in environments in which the inference is unavailable or less likely to arise (see Table 4). Furthermore, for each scale, I excluded cases containing certain syntactic constructions that could block the inference. For instance, I adopted Degen (2015)'s criterion that some-NPs headed by singular count nouns should be excluded. Additionally, I filtered out cases where a scalar term is a part of an idiom or a phrase. For instance, the scalar term 'special' forms a number of phrases like 'special force', 'special edition'.

Environment	Example
under negation	I'm not really **hungry**.
conditional antecedents	If the weather was **warm**, I would have some people over for a small party in our backyard.
wh-questions	What type of **intelligent** promoter releases the entire amount before the artists arrive at the venue?
polar questions	Do you get **adequate** vitamin D?
questions with auxiliary verbs	I am a fan and I am trying to make my band can you sent me **some** advice plz…

Table 4 *Environments prohibit the scalar inference (the scalar term was in bold)*

3.2.2.2 Word sense disambiguation task

Scalar terms could be polysemous, especially among adjective scalar terms. I consulted the Merriam-Webster dictionary and found 20 out of 28 scalar terms investigated here have at least two different but related senses. Consider <old, ancient> for example, in (5) the meaning of old is "existing a long time", which could be organized on the same scale with the core meaning of 'ancient'. However, in (6) the meaning of old is "previous", which could not form a scale of informativeness with 'ancient'. Cases like (6) need to be excluded because the weak term is used in the sense that could not be used to investigate the rate of SIs.

(5) I'm in an **old** abandoned train station w/ a translator working on the script.

(6) That means my **old** boss has been approaching a breakdown for the last 2 years.

To perform such exclusion, I conducted a word sense disambiguation task on Amazon Mechanical Turk to obtain human sense annotations on tweets containing polysemous scalar term. In this task, turkers read a text containing a scalar term. Then they were asked to choose the meaning of the scalar term from a number of options. Figure 7 is an example of an item. Among these options, there is always one describing the meaning that could be organized on the same scale as the stronger term, one 'none of the above' option which turkers could opt for if none of the meaning listed is

appropriate, and one 'incomprehensive or offensive' option[6]. For instance, in Figure 7, the test contained the scalar term 'warm'. The first option is the sense that shares the temperature scale with the strong term 'hot', the second and the third options are other relatively common senses of 'warm' listed in the dictionary. In this case, the second option is most appropriate.

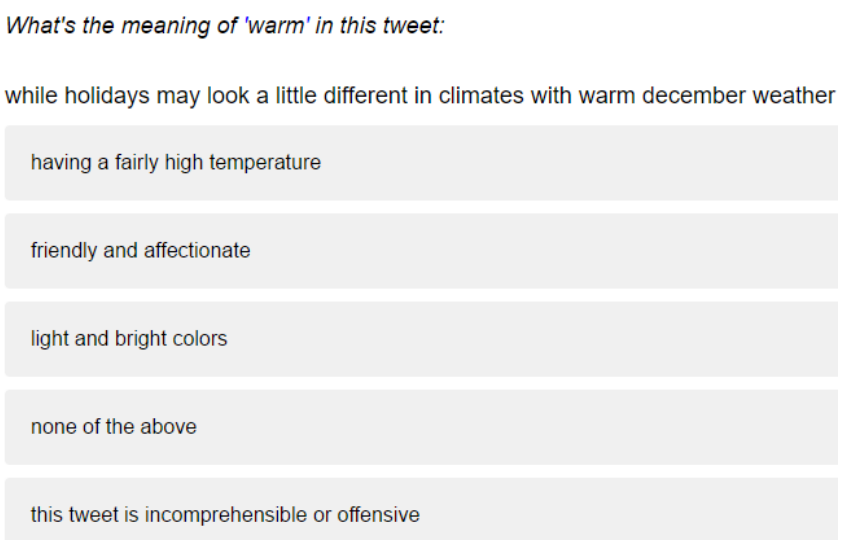


Figure 7 *Word sense disambiguation task example item*

There were 4000 texts in total, 200 texts per scale. 80 Mechanical Turk workers were recruited and each annotated 50 texts of a particular scalar term. Based on the results of the disambiguation task, I excluded texts where the use of the weak scalar term was annotated as the sense that was irrelevant to the strong scale mate. After this final exclusion, I created a tweet corpus consisted of 3075 texts. Each text contained a scalar term. Then I randomly selected 50 texts for each scale and measured the inference rating for each case in the paraphrase task.

3.3 PARAPHRASE TASK

3.3.1 Overview and prediction

Using the same paradigm as Degen (2015), the aim of the paraphrase task is to properly establish whether there is variation among scalar terms in terms of how strongly they give rise to scalar implicatures. In the paraphrase task, participants read a sentence (a

[6] Offensive tweets were mainly filtered out using regular expressions.

tweet) containing a scalar term, such as 'Sometimes I forget that I'm vegan b/c I think that everyone else is vegan too', and a comparison sentence of which the negation of the stronger scale mate was inserted, for example, 'Sometimes, but not always, I forget that I'm vegan b/c I think that everyone else is vegan too'. Participants were asked to rate how similar these two sentences are in meaning on a 1 to 7 scale. A high similarity rating indicated strong support for the implicature, a medium rating medium support and a low rating weak support. In this study, 50 sentences were randomly selected for each scalar term. Assuming each group of 50 sentences is a representative sample of that scalar term, the uniformity assumption would predict that no difference in the mean implicature rating[7] among different scalar terms. In this study, each item was rated by around 11 participants. This allows for measuring the level of agreement among participants' judgments. Assuming different types of context are distributed evenly across different scalar terms, if nothing else affects the availability of the alternatives and the mechanism of Gricean derivations, I would expect no difference in the average agreement level of participants' judgments among scalar terms.

3.3.2 Methods

3.3.2.1 Participants

550 participants from the United States were recruited via the Amazon Mechanical Turk. They were asked to indicate their native language and only participants with English as a native language were included in the analysis.

3.3.2.2 Materials and procedure

Figure 8 is an example item. On each trial, participants read a target sentence that contained a scalar term 'X' (in red) and a comparison sentence, in which the negation of its stronger lexical alternative 'but not Y' (in green) was inserted after 'X'. Participants were asked how similar these two sentences are on a 1 (very different meaning) -7 (same meaning) scale. Similar to Degen (2015)'s study, I included two practice trials at the beginning to ensure that the participants understood the instruction and to encourage them to use the full scale range. Two practice trials are shown in ((7)-(8)). ((7)a) would

[7] The higher the similarity rating, the more likely the implicature would be drawn. Thus, I refer to the similarity rating in this study as the implicature rating.

normally be interpreted as ((7)b), whereas ((8)a) would not be understood as ((8)b). Accordingly, I instructed the participants to choose a high value in (7) and a lower value in (8).

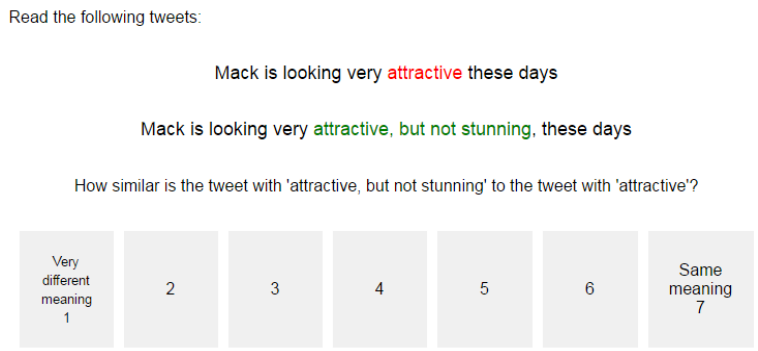


Figure 8 *Paraphrase task example item*

(7) a. And **sometimes** my German shepherd just growls at my empty bathroom.

b. And **sometimes, but not always**, my German shepherd just growls at my empty bathroom.

(8) a. Yes, but the fundamental issue is the need to provide **adequate** funding and joined up thinking.

b. Yes, but the fundamental issue is the need to provide **adequate, but not good**, funding and joined up thinking.

In total, there were 1400 items, 50 items per scale. Each participant judged 28 items, one item per scale. Each item received 8 to 15 judgments (mean 11). This variation was due to the randomization of assigning participants to items using Qualtrics survey platform.

3.3.3 Results

Figure 9 shows the hierarchical structure of the dataset. In this study, each scale had 50 items and each item was rated by 8 to 15 participants. I had individual ratings nested in items nested in scales. I selected the scale as the unit of analysis in order to examine whether mean implicature rating and mean agreement level varied across scalar terms. I calculated the mean implicature rating and mean entropy[8] for each scale by averaging

[8] Entropy is a measure of dispersion of a response distribution, it is also a measure of uncertainty within the distribution (Shannon, 1948). The entropy value was calculated with the formula - , where p(x) is the probability of occurrence of a rating (Manning & Schütze, 1999).

mean ratings and entropies of its 50 items. The implicature rating reflects the availability of the scalar implicature. The entropy value is a measure of participants' level of agreement on their implicature judgments. Given that I used a 7-point rating scale[9], the theoretically maximum entropy for a response distribution is 2.81 that is when each of the responses is chosen equally often. Whereas an entropy of 0 is when only one of the seven responses is chosen each time. For a single item, the lower the response entropy, the higher the degree of participant agreement. Thus, for a given scale, the lower the mean entropy, the higher the average agreement of participants' judgments on items of that scale.

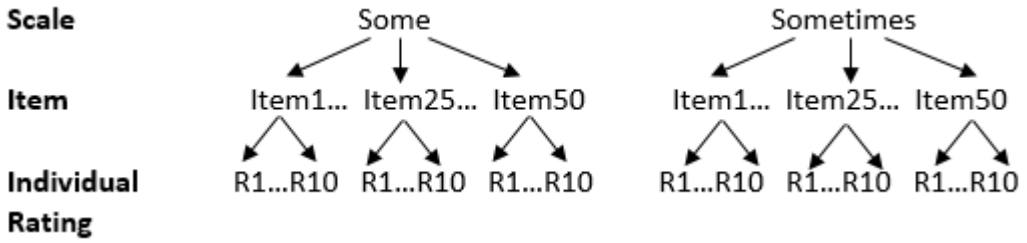


Figure 9 *The hierarchical structure of the dataset*

The overall mean implicature rating was 4.34 (median: 4.32). The mean implicature rating varied across scales, ranging from 3.3 < adequate, good> to 5.5 <sometimes, always>. The overall mean entropy value was 2.17 (median: 2.17), and the mean entropy ranged from 1.95 <sometimes, always> to 2.27 <small, tiny>. I found that mean implicature rating negatively correlated with mean entropy value (r=-0.72, p<.001), as shown in Figure 10. This suggested that as the implicature rating increased, the level of agreement among judgments increased as well.

[9] This was treated as a discrete scale in order to compute entropy values.

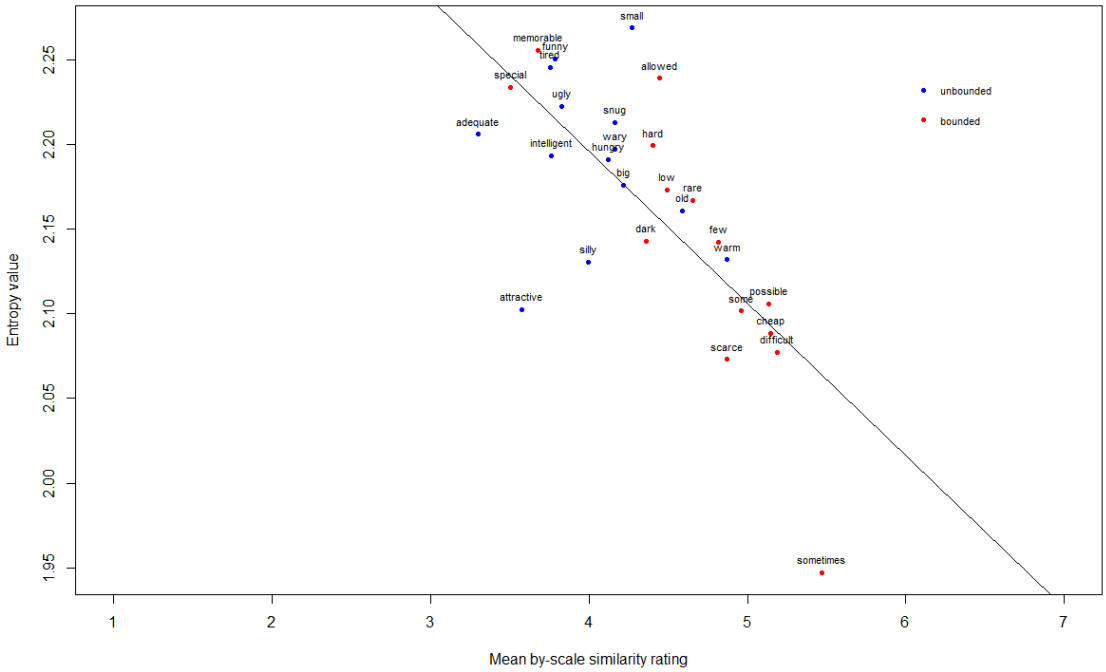


Figure 10 *Negative correlation between the mean rating and the mean entropy*

Previous work treated scalar implicature as an all-or-none phenomenon and only obtained binary responses on the availability of scalar implicatures. In order to compare with van Tiel et al.'s findings, I coded the responses collected from the paraphrase task into three categories: low (ratings were 3 or lower), median (ratings were 4), and high (ratings were 5 or above). I considered high ratings as an indicator of SIs being drawn. The percentages of implicature response for each scale are visualized in Figure 11 (blue bars). The rates of SIs from van Tiel et al. (2016, Experiment 2) are also included (orange bars).

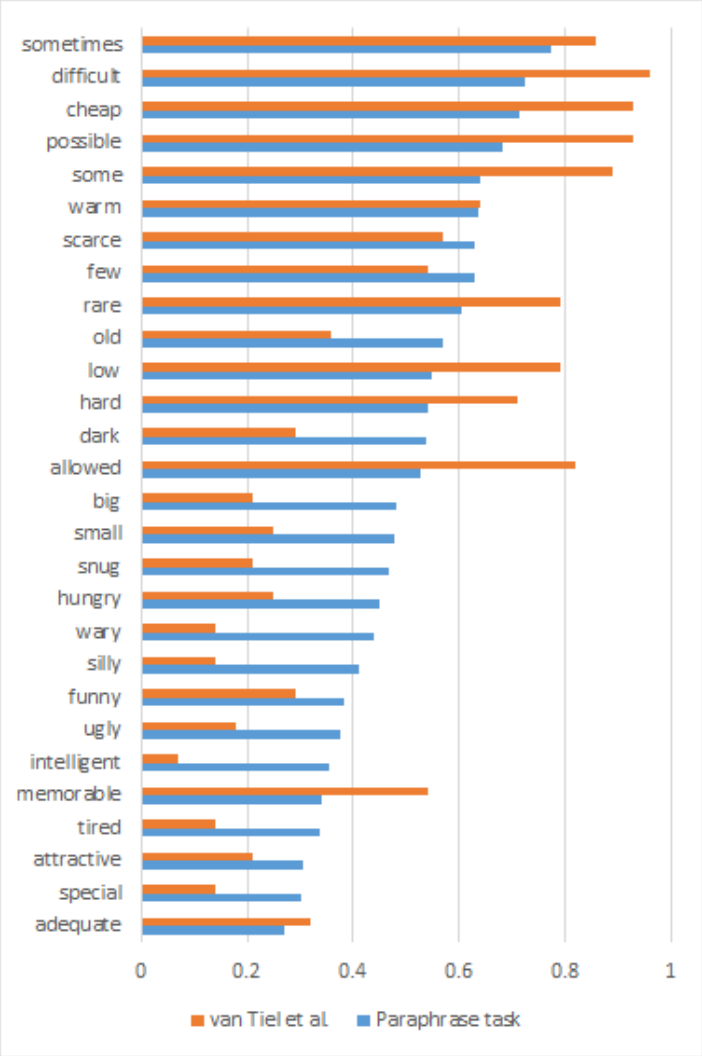


Figure 11 *Mean inference ratings for the paraphrase task*

I found the percentages of implicature response from the paraphrase task correlated with van Tiel et al. (2016)'s results ($r=0.81$, $p<.001$). Kendall's W (Kendall's coefficient of concordance) showed that these two studies have significant agreement on the ranking of implicature rates across scalar terms ($w=.905$, $p=.006$). This suggests that the results yield from the inference task based on a small sample of hand-crafted sentences could reflect, to some extent, rates of SIs measured with large-scale naturalistic data.

However, Levene's test for equality of variances suggested that the variances of these two studies are not equal ($F(1,54)=14.69$, $p<.001$). Visual inspection suggests that

there is less variation on the paraphrase task. The rates of SIs for quantifier[10] and modal expressions in the paraphrase task are far lower than rates found in the inference task, whereas many adjective expressions give rise to SIs more frequently in the paraphrase task. So overall there appears to be less variation in implicature rates measured in the corpus-based paraphrase task.

3.3.4 Discussion

The results of the paraphrase task indicate that there is considerable amount of variation in the implicature ratings across scales. These results provide further evidence against the uniformity assumption. In addition, the average degree of agreement level of participants' judgments on the availability of implicature varies across scales. Thus, the null hypothesis that predicts no difference in the average agreement level of participants' implicature judgments among scales is rejected. Note that scalar terms show less variation in generating SIs in this work compared to previous investigations (e.g. van Tiel et al., 2016). Before exploring what factors could explain the observed variation, I discuss three methodological points that might contribute to the differences in results between the current study and van Tiel et al.'s study.

First, the differences in the rate of SIs for adjective scales between two studies may be due to the effect of negative strengthening – as discussed in Benz, Ferrer, & Gotzner (2017). Negative strengthening refers to the phenomenon that the negation of the stronger scale mate would give rise to an inference that negates the weaker. For example, if I take a scale of intelligence, there is a point above which I say a person is intelligent and below which they are not intelligent. However, as widely discussed in the literature,[11] the meaning of 'intelligent' can become strengthened to *intelligent** (via pragmatic processes) to inhabit a space on the intelligence scale from the minimum level of intelligence up to the point where the stronger term (*brilliant*) would be judged to begin. In a similar process, 'not brilliant' is strengthened in meaning to inhabit the scale

[10] Concerning the similarity ratings collected for 'some' cases, in our study, 64% of ratings were higher than midpoint. This result did not replicate exactly Degen's result (44.7%). One possibility is that compared with Degen's dataset, our items contained higher percentage of partitive some (27% vs.34%). Previous studies suggested that the partitive is a strong cue for computing SIs. Another possibility is that our exclusion of irrelevant cases was stricter. In particular, Degen (2015) did not filter out cases where the scalar term occurred in the scope of negation, in questions or in the antecedents of conditionals.

[11] See Horn (1989), Levinson (2000), Kritka (2007)

from just below the point where *intelligent* begins to the point where *stupid* begins. To the extent that this process occurs, because in the inference task, participants read 'John says: 'This student is intelligent'. Would you conclude from this that, according to John, she is not brilliant?' negative strengthening may impact on judgments, since 'not brilliant' now excludes intelligent. If this is the case, Benz et al. argue, that a 'no' response could indicate either the absence of the scalar inference or the presence of the strengthening of the negative.

However, such negative strengthening is unlikely to happen in the paraphrase task. In this task, participants were asked to indicate how similar the sentence 'she is intelligent' to a comparison sentence 'she is intelligent but not brilliant'. Given that the use of 'but' carries a denial-of-expectation conventional implicature and 'she is intelligent' is restated in the comparison sentence, the inference 'she is dumb' is unlikely to be triggered.

A recent study by Benz, Ferrer and Gotzner (2017) has shown that the likelihood of negative strengthening negatively correlated with the rate of SIs found in van Tiel et al. (2016) and such correlation was mainly detected among adjective scales. Therefore, it may be that fewer SIs were drawn for adjective scales in van Tiel et al. (2016) than our corpus-based task because negative strengthening only appears in the inference task, not in the paraphrase task.

Secondly, the difference in the rate of SIs for quantifiers and modal expressions between the two studies may be due to the distribution of context types. As discussed in the Introduction, in the inference task, participants were more likely to create upper-bound contexts for quantifiers and model expressions and this would lead to high rates of SIs. However, in the paraphrase task, more contextual information was provided. Quantifiers or modal expressions occurred in different types of context, and the relevance of the stronger alternative varied. To illustrate, consider two examples from the case of quantifiers:

(9) In other news, some of our Electoral College members are teenagers.

(10) Got invited to paint some of my own artwork onto clothing!

Given (9) participants might consider the utterance as an answer to a 'how many' question, then the stronger alternative 'all' is relevant. Whereas given (10), participants might wonder what happened, then it is unclear whether or not the stronger alternative is relevant. Cases like (10) would lower the average implicature ratings. Note that differences in the distribution of context types between two studies might have a smaller effect on the rates of SIs for adjectives. This is because, unlike quantifiers and modals, participants might not always create upper-bound contexts for adjectives in the inference task.

Thirdly, the difference in the rates of SIs between two studies might also be due to the difference in the sample size. In the paraphrase task, I randomly selected 50 items per scale. I expect that the distribution of each type of context (upper-bound; lower-bound; neutral) in the sample is the same as the distribution of context types in the population. Then the inference rate of each scale measured in the paraphrase task is an estimation of the inference rate for that scale over all possible context types. By contrast, in van Tiel et al.'s inference task, the authors hand selected three items per scale. Thus, the rate of SI for each scale measured in this task may not be able to generalise to a larger set or the population. It is likely that the sampling error is larger in the inference task than in the paraphrase task.

3.3.5 Combined analysis

To investigate whether factors discussed in the last chapter could explain some of the variation found in the corpus-based paraphrase task, I fitted two multiple linear regressions. The first regression analysis was conducted to predict the implicature rating in the paraphrase task from scale homogeneity degree, propensity for local enrichment and factors from in van Tiel et al. (2016). The inference rating in the paraphrase task was averaged by scale before entering into the analysis. The degree of scalar homogeneity and propensity for local enrichment were measured in Experiments 2 and 3 presented in Chapter 2. The model included all of the factors explored in van Tiel et al. (2016). Table 5 summarizes the result. I found that the regression model accounted for 67% of the variance (R^2=.77, F(8,19)=7.99, p<.001). In this model, I found scale homogeneity degree, propensity for local enrichment, boundedness, and grammatical class are substantial factors, with scale homogeneity explaining 11%, propensity for local enrichment

explaining 18%, boundedness explaining 13% and grammatical class explaining 7%. None of the other factors significantly accounted for the variation in the rates of SIs.

	β	**SE**	**t**	**p**
(Intercept)	3.4473	0.8134	4.24	0.0004 ***
Scale homogeneity	-0.3295	0.1181	-2.79	0.0117 *
Local enrichability	0.3278	0.0936	3.50	0.0024 **
Association strength	-0.0043	0.0029	-1.45	0.1642
Grammatical class	-0.7784	0.3401	-2.29	0.0337 *
Word frequencies	-0.1077	0.0766	-1.41	0.1760
Semantic relatedness	0.5696	0.3758	1.52	0.1461
Semantic distance	0.1448	0.1094	1.32	0.2014
Boundedness	0.4047	0.1355	2.99	0.0076 **

Table 5 *Results of combined analysis with the inference rating from the paraphrase task as dependent variable*

The second regression analysis was conducted to predict the agreement level of participants' judgments in the paraphrase task from scale homogeneity degree, propensity for local enrichment and factors from in van Tiel et al. (2016). The entropy value was averaged by scale as the dependent variable. Table 6 summarizes the results of the analyses. I found that the regression model accounted for 43% of the variance (R^2=.60, F(8,19)=3.50, p=.01). In this model, I found semantic distance significantly accounted for the variance (21%). Propensity for local enrichment and boundedness are marginally significant predictors (p=.06; p=.09), explaining 8.6% and 6.5% of the variance respectively.

	β	**SE**	**t**	**p**
(Intercept)	2.5841	0.1345	19.21	<.0001 ***
Scale homogeneity	-0.0134	0.0195	-0.68	0.5017
Local enrichability	-0.0310	0.0155	-2.00	0.0601
Association strength	-0.0003	0.0005	-0.52	0.6070
Grammatical class	0.0838	0.0562	1.49	0.1526
Word frequencies	0.0203	0.0127	1.60	0.1262
Semantic relatedness	-0.0601	0.0621	-0.97	0.3458
Semantic distance	-0.0526	0.0181	-2.91	0.0090 **
Boundedness	-0.0405	0.0224	-1.81	0.0868

Table 6 *Results of combined analysis with the entropy value from the paraphrase task as dependent variable*

3.4 GENERAL DISCUSSION

The rating results from the paraphrase task suggest that the uniformity assumption does not hold, as scalar terms tend to vary in the degree to which they are liable to give rise to scalar implicatures. Using a very different methodology, the general pattern found in previous laboratory-based tasks was replicated, albeit without such radical variation. To explain the variance found in the paraphrase task, previously established factors were tested. Among factors from van Tiel et al. (2016), I found that factors related to the availability of the alternatives did not explain the variance, expect for the grammatical class. For factors related to the distinctness of scalar terms, boundedness remained a substantial factor, whereas semantic distance did not. Once again, local enrichability affected the variability and interestingly scale homogeneity also accounted for a substantial amount of the variance.

Among the 28 scales investigated in the paraphrase task, only two quantifier expressions (i.e. ‘some’ and ‘few’) are from the closed class whereas the rest are from the open class. That grammatical class is a factor might reflect this aspect of the items. Concerning boundedness, it could explain the variance found in both laboratory-based and corpus-based tasks. This suggests that boundedness is a genuine underlying factor affecting the scalar diversity. Thus at least some aspects of the relation between scalar term and stronger alternative can impact the rates of scalar inference.

Regarding measures of association between scalar term and alternative, we replicated the finding from inference tasks that no such measure explains a significant amount of variance. However, this is not to say that the association between scalar term and alternative may not be a factor in real daily use. As observed in Chapter 2, the design of the inference task would tend to obscure this as a factor since the task always mentions the alternative term – making that term maximally salient or accessible, independently of any underlying association. The same considerations apply to the paraphrase task used in this experiment since the paraphrase also mentions the alternative in all cases. Thus, it remains for a different kind of study to properly assess whether association between scalar term and alternative can impact on rates of scalar implicature.

Concerning scale homogeneity, the fact that it has an impact on scalar variability may be because participants are more likely to evoke multiple senses of a scalar term when presented with a natural sentence than when presented with an artificial sentence. In the paraphrase task, for instance, participants were asked how similar two sentences in (11) are. In this case, given ((11)a), participants may have evoked multiple senses of 'hard', then the stronger alternative could be 'unbearable' rather than 'unsolvable'. If this is the case, it will lead to a low implicature rating.

(11) a. It is going to be extremely **hard** because my entire family eats meat.
b. It is going to be extremely **hard, but not unsolvable**, because my entire family eats meat

By contrast, in the inference task used in van Tiel et al. (2016), participants were presented with 'The puzzle is hard', and they were asked if the speaker implied that 'The puzzle is not unsolvable'. It is unlikely that they would evoke other senses of 'hard' as those senses are irrelevant when applies to 'puzzle'. Therefore, scale homogeneity explains more variance in the corpus-based task than in the laboratory-based task.

Concerning local enrichability, it has been found to affect the scalar variability in both laboratory-based and corpus-based tasks. These results suggest that the locally enriched interpretation of the scalar term has an impact on both kinds of task and that variability in local enrichability is having an impact on scalar variability. These results provide indirect support for the RSA-LU framework and also the RT approach, which see

local enrichment as active in both embedded and unembedded environments. However, such support is conditioned to some extent on whether the theoretical approach can predict variability in the likelihood of local enrichment. On its own, RSA-LU does not make predictions about variability in upper-bound enrichments, *w&¬S*, across scalar terms. This is not to say that variability is ruled out by this approach – on the contrary. In Chapter 2, I mentioned one factor that might give rise to variability in rates of local enrichment that stems from the constraint that the explicit utterance has to have some relevance. This constraint implies that scalar terms lacking specific content might be more liable to local enrichment, putting terms like 'some' and 'intelligent' in contrast. The results of the inference task in Chapter 2, Experiment 1 showed a significant difference between quantifiers and modals on the one hand and adjectives on the other, with the 'so' task results reflecting this trend. As mentioned, some of the stark differences between these two classes of terms might have been an artefact of the design of the inference task. The corpus-based task presented in this chapter remedies this potential confounding factor and correspondingly finds less variation between quantifiers and modals on the one hand and adjectives. Still the pattern remains the same and so we find indirect support for this conjecture about the source of scalar diversity.

The results of the paraphrase task showed that the entropy of participants' judgments for each scale also varied across different scales. The entropy results provide further evidence against the uniformity assumption. I found only semantic distance could explain a certain amount of the variation. One possible interpretation of this finding is that semantic distance may provide a source of disagreement. For example, 'attractive' and 'stunning' could be organized on the same scale with the core meaning of appealing to look at. Given a sentence 'That dress is unflattering, especially for an attractive woman', the literal meaning of 'attractive' can be strengthened to *attractive** via lower-bound raising enrichment. Since the semantic distance between the two terms is not large to start with, the stronger term 'stunning' should not be very far away from *attractive** on the scale. If so, participants who enrich the weak term frequently would not see the 'not stunning' implicature. Thus, depending on participants' interpretation

of the weak term, the agreement on the implicature rating is lower when the semantic distance between two terms is small.

3.5 CONCLUSION

The phenomenon of scalar diversity has interested linguists since it first emerged in work by Doran and colleagues and later substantiated in work by van Tiel and colleagues. That there might be scalar diversity is relevant on many levels. On a simple methodological level, it is important because it means that results of experiments or surveys involving often-used terms like 'some' or 'or' cannot necessarily be taken as representative of all cases of scalar implicature. On a theoretical level, there is interest because it suggests that not all lexical scales are equal when it comes to generating scalar implicature. In this chapter and the last, I introduced a new perspective to the theoretical side of the scalar diversity phenomenon. This has to do with the fact that local enrichment of scalar terms has an impact on unembedded scalar implicature – the case of Straight Scalars.

A more thorough investigation of factors that impact on diversity by van Tiel et al. (2016) suggested that relations among scalar terms play only a limited role in explaining variance – through the notion of boundedness. Apart from semantic distance, which relates to methodological limitations of the design and items used in past experiments, van Tiel and colleagues suggest that the rest of the variance is unsystematic. In these last two chapters, I have presented results that replicate the diversity effect in an inference task, and put the diversity phenomenon on a firmer footing via a corpus based task. I have shown that diversity may not be as great as previous lab-based tasks suggest, but it is nevertheless real. I have also shown that not all of the remaining variance is unsystematic but that mechanisms for local enrichment are involved. The results linking local enrichability and scalar diversity provide indirect support for an integrated Gricean system that allows for scalar phenomena to be explained by two mechanisms, a global inference mechanism and a local enrichment mechanism.

One tempting interpretation of the role of local enrichment in unembedded 'Straight Scalars' lies in the idea that at the level of language processing, a memory trace of previous local enrichments can impact on decisions about how to interpret an

utterance. Terms like ‘some’ and ‘might’, that are more frequently locally enriched might activate these upper-bounded enriched meanings in a way that impacts on comprehension processes, giving rise to more enriched representations of the intended meaning. As discussed above, there are a variety of ways one can conceptualise linking hypotheses between the computation (or inferential) level and the level of processing (see Griffiths, Lieder, & Goodman, 2015; Potts et al., 2015). However, the results in this chapter are suggestive that locally enriched meanings might be subject to priming. This is something I shall explore in the next chapter.

Chapter 4 SHARED MECHANISM UNDERLYING UNEMBEDDED AND EMBEDDED ENRICHMENTS

This chapter uses a priming paradigm to explore the mechanisms underlying unembedded and embedded scalar enrichments. In particular, the aim is to see if local pragmatic enrichment could be a shared mechanism, involved in both. Two experiments presented in this chapter adopt Bott & Chemla (2016)'s enrichment priming paradigm and tests whether unembedded and embedded enrichments could prime each other. The goal is to investigate whether local pragmatic enrichment is indeed being accessed for interpreting the unembedded scalar and whether local enrichments, like other lexical semantic phenomena, are susceptible to priming.

4.1 INTRODUCTION

The basic idea of local enrichment processes is that a sub-constituent of the sentence contributes more than its literal meaning to the truth-conditional content of the sentence. In scalar implicature literature, the main argument for applying local enrichment to the interpretation of scalar term is to account for embedded enrichments (EE), as many cases of EE cannot be explained by global implicature derivation. However, little has been established regarding whether local enrichment applies to cases of apparently unembedded scalar implicatures. Experiments presented in the past two chapters showed that local enrichability affects the interpretation of unembedded scalar terms, such that the more liable a scalar term to be locally enriched on the upper bound, the higher the rate of enriched responses. This finding provides supporting evidence that local enrichment as a separate mechanism from global inference derivation can influence the interpretation of unembedded scalar terms. In this chapter, by using an enrichment priming paradigm, I more directly explore whether unembedded scalar implicatures might nevertheless be sometimes derived from local pragmatic enrichment. In addition, I address a more fundamental question, whether there is a shared mechanism for EE and SS. In this chapter, I re-introduce the Grammatical Approach (GT) to scalar implicature since the experiments here are relevant also to that approach. As mentioned in the introduction, this thesis is more focused on the

mechanisms of scalar implicature on the assumption that they are a pragmatic phenomenon and are to be explained on general Gricean principles. There are some points of discussion emerging from the results of this chapter that bear on the merits of a Gricean vs. GT approach, but this is not the focus here.

In this chapter, I first outline different accounts of Scalar Implicatures which hold different views about the mechanisms underlying unembedded and embedded scalar enrichments. I then introduce enrichment priming paradigm developed by Bott & Chemla (2016). Finally, I discuss the rationale for testing whether unembedded and embedded enrichments could prime each other and predictions made by different accounts.

4.1.1 One mechanism or two?

As discussed in Chapter 1, embedded pragmatic effects are widespread in language use, such that all kinds of implicatures can be embedded under linguistic operators. For the purpose of this chapter, the embedded phenomenon is restricted to embedded scalar enrichments. Recall the following example where 'some' is embedded under the non-monotonic quantifier 'exactly one':

(1) Exactly one player hit some of his shots.
    a. Exactly one player hit some and possibly all of his shots.
    b. Exactly one player hit some but not all of his shots.

The literal reading of ((1)) is given in ((1)a). Potts et al. (2015) reported that the enriched interpretation in ((1)b) is optional but indeed available. Since ((1)b) is logically independent from ((1)a), ((1)b) cannot be derived by conjoining the literal meaning with the negation of some other alternative proposition. Thus, the standard Gricean account cannot explain embedded scalar enrichments in non-monotonic environments in general.

However, EE like ((1)b) can be explained by theories that allow for local enrichment processes. There are different approaches on the implementation of local enrichment. Here I focus on the grammatical theory and the Gricean approaches, Relevance Theory and the RSA approach with Lexical Uncertainty. Both accounts are able to derive EE in cases like ((1)b) via local enrichment processes, yet they make

different predictions on the derivation mechanisms underlying unembedded scalar enrichments like (2).

(2) John hit some of his shots.

~> John hit some but not all of his shots

The grammatical theory

The grammatical theory (GT) accounts for both unembedded and embedded scalar enrichments in terms of a covert exhaustification operator being present in the LF for the sentence (see Chierchia, 2006; Chierchia, Fox, & Spector, 2012; Fox, 2007). LFs for (1) and (2) that give rise to enrichments discussed above are given in (3) and (4).

(3) [Exactly one player]$_x$ [*Exh* [t$_x$ hit some of his shots]].

(4) *Exh* [John hit some of his shots].

Thus, according to the grammatical theory, there is a single mechanism for deriving both unembedded and embedded scalar enrichments.

Gricean Approaches

Relevance Theory

As outlined in the introduction, from its inception, Relevance Theory has attempted to accommodate local pragmatic effects into its pragmatic framework. The theory simply assumes that the linguistic information (lexical semantics, syntactic structure, rules of composition) attached to an utterance are not fully determinant of what proposition is explicitly expressed, but a starting point. As discussed in Noveck & Sperber (2007) scalar implicatures, even straight scalars may be a result of either a global inference process (where the speaker intends the hearer to see that they were not in a position express a contextually salient alternative), or via local enrichment.

RSA approach with Lexical Uncertainty

RSA approach with Lexical Uncertainty (RSA-LU) allows two mechanisms to explain scalar phenomena, the global derivation mechanism and the lexical adjustment mechanism. In cases where the scalar term occurs in the scope of non-monotonic operator like ((1)), only lexical adjustment mechanism can be applied to the scalar term to derive the local reading in ((1)b). Under this mechanism, the meaning of the scalar

term ‘some’ is uncertain. By using ‘some’, the speaker could intend to mean (i) {∃&¬∀, ∀}, (ii) {∃&¬∀} or (iii) {∀}. (i) is the literal meaning, whereas (ii) and (iii) are enriched meanings. As noted in Potts et al. (2015), meaning narrowing in embedded contexts is affected by the enriched meaning used in the unembedded contexts. In other words, since ‘some’ in unembedded contexts is often interpreted as ‘not all’ rather than ‘all’, in embedded contexts the ‘not all’ enriched meaning might be favoured over ‘all’.

In contrast to embedded scalars, in cases where the scalar terms occurs in unembedded position like ((2)), both global and local mechanisms can be used for deriving the enriched interpretation. Therefore, according to the RSA approach with Lexical Uncertainty, there is a shared mechanism for deriving both unembedded and embedded scalar enrichments, namely the lexical adjustment mechanism. However, unembedded scalar enrichments could also be the results of global inference mechanism which, however, cannot be used for deriving many embedded enrichments.

In summary, the standard Gricean account is deficient in dealing with embedded scalars in the scope of non-monotonic operator. Both GT and RT or RSA-LU could account for this kind of embedded effect by allowing local enrichment. The sub-constituent is enriched via a grammatical operation in GT, however, in RT and RSA-LU it is enriched via pragmatic reasoning (so-called ‘local pragmatic enrichment’ in Chapter 1). When it comes to unembedded scalars, both accounts agree that unembedded scalar implicatures might nevertheless be sometimes derived from local enrichment. GT derives unembedded enrichments in similar fashion as embedded enrichments, whereas RT/RSA-LU allows unembedded enrichments to be derived via two routes, of which lexical adjustment mechanism is also responsible for embedded enrichments. For either account, it follows that there should be a shared mechanism underlying unembedded and embedded enrichments.

4.1.2 Enrichment priming

Bott & Chemla (2016) developed an enrichment priming paradigm for the purpose of obtaining empirical evidence for shared mechanisms within and across different categories of unembedded scalar enrichments (i.e. quantifiers, numerals, ad hoc). In this task, each sentence is presented with two pictures and participants are asked to click on

the picture that is a better match for the given sentence. The critical items of within-category priming are illustrated in Figure 12.

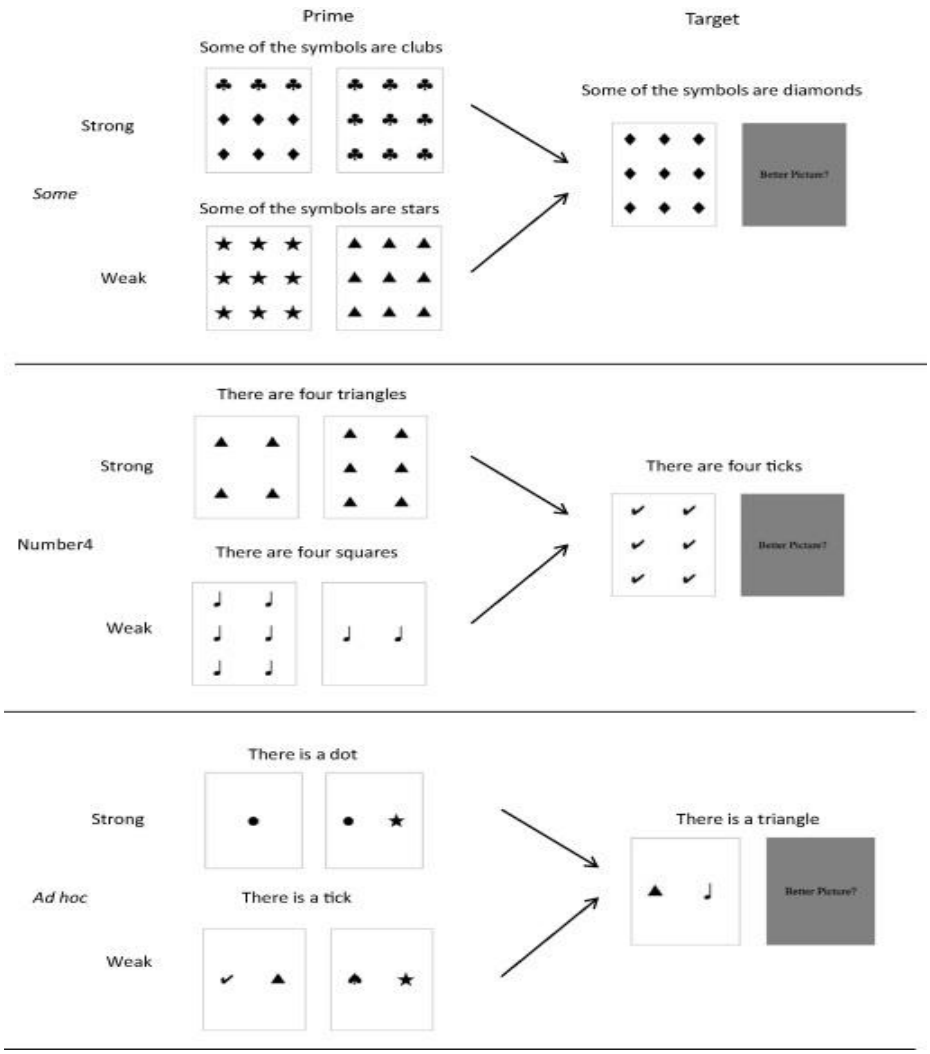


Figure 12 *Example items in Bott & Chemla (2016)*

In this condition, the target and prime trials involve the same enrichment category. For instance, a target trial with 'some' is preceded by prime trials also with 'some'. The same goes for number and ad hoc cases. There are two types of prime trials. In strong primes, the scalar implicature is true; in weak primes, the scalar implicature is false. For example, consider *some* → *some* in the top panel of Figure 12. In the strong prime trial, given the sentence *Some of the symbols are clubs*, one picture shows some and not all symbols are clubs, and the other shows all symbols are clubs. The 'some-not-all' picture makes the scalar implicature (*some and not all symbols are clubs*) true. Whereas in the weak prime trial, given the sentence *Some of the symbols are stars*, one picture contains all stars and the other contains only non-stars. Neither picture makes

the interpretation that includes scalar implicature true. Participants who choose the strong image are thus primed by the SI-enriched reading in the strong prime, and they are forced to access the unenriched reading in the weak prime.

For the target trial, Bott & Chemla (2016) adopted the 'Better-picture' method used in Huang, Spelke, & Snedeker (2013). Participants are shown one of two images that make the unenriched reading true, while the other is covered. Participants are told that if they think that there is a picture that would be a better match for the sentence, they can choose the covered picture. Since the visible picture is consistent with the unenriched reading and inconsistent with the SI-enriched reading of the target sentence, choosing the covered picture indicates that participants access the SI-enriched reading.

In addition to within-category priming, the other condition is between-category priming where the target and prime trials involve different enrichment categories. For instance, a target trial with number term (e.g. 'four') is preceded by prime trials with 'some'. Bott & Chemla included all between-scale combinations in this condition, such as *some* ↔ *number*, *some* ↔ *ad hoc*, and *number* ↔ *ad hoc*.

The logic behind this paradigm is that, if there is a shared derivation mechanism which is subject to priming, then for both conditions it is more likely for participants to access the enriched reading of the target sentence (i.e. choosing the covered picture) after strong prime trials than after weak prime trials. Their results are shown in Figure 13. The y-axis is the rate of covered-picture responses. There is a within-category priming. For instance, the leftmost column of Figure 13 shows that participants were more likely to interpret 'some' to imply 'not all' after strong primes where they accessed 'some and not all' interpretation than after weak primes where they were forced to access the 'some and possibly all' interpretation. They also found a between-category priming.

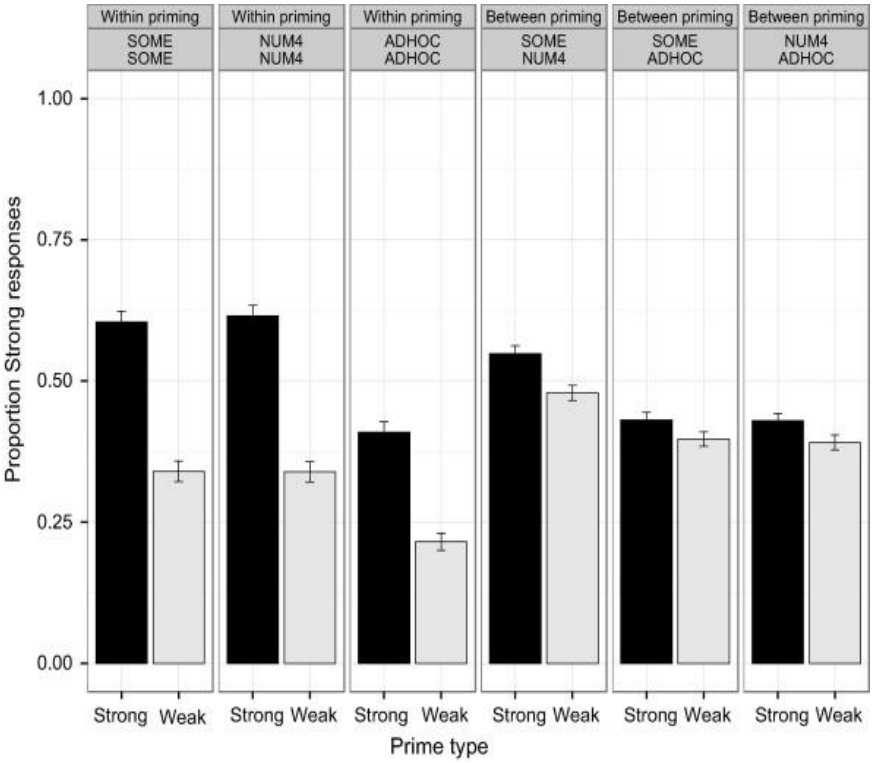


Figure 13 *Proportion of strong responses for within-category and between-category priming in Bott & Chemla's Experiment 1*

Bott & Chemla (2016) interpreted the between-category priming effect as evidence for activation of shared mechanisms in deriving enrichments involving different scales. As for the within-category priming effect, they suggested that along with the activation of the derivation mechanism, there could also be a lexical priming which is an association between the stimulus, the derivation mechanism and specific alternative. For Bott & Chemla, the between-category priming effect is the result of interest, because it shows that general SI derivation mechanism can be primed. However, it follows that the enrichment priming paradigm could be employed to investigate whether local enrichment can be primed as a shared mechanism between unembedded and embedded enrichments.

4.1.3 Rationale and predictions

The first goal of experiments in this chapter is to determine whether embedded and unembedded scalar phenomena have a shared mechanism. A related aim is to explore whether unembedded scalar terms are sometimes enriched through local enrichment. I investigate the mechanisms underlying unembedded scalar enrichment using the same

paradigm as in Bott & Chemla (2016). The rationale is that, if unembedded scalar implicatures are derived from local pragmatic enrichment, then participants should be more likely to access embedded enrichments that require local enrichment after strong primes with unembedded scalar implicature than after weak primes with no implicature. The critical items are illustrated in **Figure 14**.

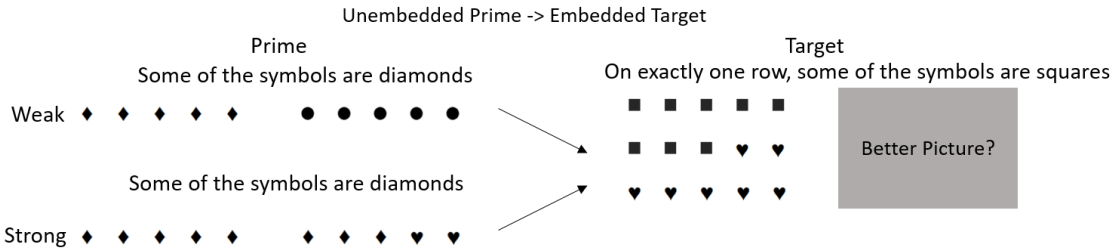


Figure 14 *Critical items for Embedded Target trials in Experiment 1 and 2*

In the embedded target condition, the target trial involving embedded 'some' is preceded by prime trials involving unembedded 'some'. In strong primes, the unembedded scalar implicature is true, while in weak prime trials, the unembedded scalar implicature is false. For example, given a prime sentence *Some of the symbols are diamonds*, in strong primes, the sentence is presented with one picture depicting a row with some but not all symbols being diamonds and another picture depicting a row with all symbols being diamonds. The 'some-not-all' picture makes the SI-enriched reading true. For the same sentence, in weak primes, it is presented with one picture depicting a row with all symbols being diamonds and one picture depicting a row of non-diamonds symbols. Neither picture makes the SI-enriched reading true. Thus, participants are primed by the SI-enriched reading in strong primes and the unenriched reading in weak primes.

As in Chemla & Bott, we employ the covered picture paradigm in the target trials. We have experimental trials when a sentence with an embedded scalar term is target. We also include a set of trials where an unembedded sentence is the target, following embedded prime trials. For target trials in the embedded target condition, a target sentence like 'On exactly one row, some of the symbols are squares', is presented with a visible picture and a covered picture. The visible picture makes the locally enriched reading true and other available readings false. The image in Figure 14 shows the visible

image having two rows containing squares. One of those has some and not all square, the other has all squares. Only if the sentence is understood as, *On exactly one row, some and not all of the symbols are squares* would a participant not choose the covered card. If the literal meaning of the target sentence is accessed, or even an interpretation that includes a global implicature, the participant should choose the covered square.

This is a change from Bott & Chemla's procedure. As previously mentioned, the visible picture used in Bott & Chemla's paradigm makes the literal reading true and SI-enriched reading false. The motivation for changing their design comes from the availability of the global-SI reading. The global-SI reading of the target sentence is that in exactly one row, some symbols are squares and it's not true that on exactly one row all symbols are squares. If the target sentence is presented with a visible picture that makes the literal reading true, as shown in Figure 15 below, then participants might choose the covered picture because they derive the global reading of the sentence and expected a better match such as Figure 16. If this is the case, then choosing the covered picture in Figure 15 might reflect a mixture of local reading and global reading.

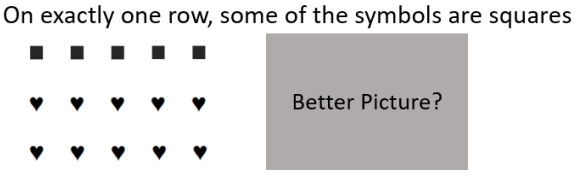


Figure 15 *Discarded displays*

Figure 16 *Example display where the global reading is true*

Thus, in order to properly measure the rate of locally enriched reading, in both Experiments 1 and 2 below, the embedded target sentence is paired with a visible picture false on any available reading except for the local one. In this case, choosing the visible picture indicates that participants access the locally enriched reading, whereas choosing the covered picture indicates that they access either the literal reading or the global reading.

Regarding whether unembedded enrichments could prime embedded enrichments, the grammatical account predicts a priming effect, as there is a single mechanism for both prime and target trials involving *Exh* operator in LF. On the other hand, the RT/RSA-LU approach predicts priming between the two based on the mechanism of lexical adjustment which can be used in both prime and target trials. However, these approaches do not rule out the possibility that there is no priming effect. This is so since RT/RSA-LU argue that there are two mechanisms underlying scalar enrichments rather than a single one. It is possible that the lexical adjustment mechanism is not used very much in target trials. If this is the case, then there might not be a priming effect between unembedded and embedded enrichments.

In addition to the embedded target condition, both experiments also included an unembedded target condition. In the unembedded target condition, the target trial involving unembedded 'some' is preceded by prime trials involving embedded 'some'. Experiment 1 and 2 differ in the prime items used in unembedded target condition, which will be discussed in more details below. Regarding whether embedded enrichments could prime unembedded scalar implicature, the grammatical account again predicts a priming effect on the basis of a single shared mechanism. The RT/RSA-LU approaches also predict a priming effect, as the lexical adjustment mechanism is needed for embedded prime trials (especially in exp.2) and the target trial can be enriched in the same way.

Putting aside competing predictions of GT and Gricean approaches, these experiments should give us some more concrete insights into the question addressed in the previous two chapters. This is whether unembedded scalars are sometimes understood via the activation of a locally enriched meaning. Recall that in previous chapters, we found that local enrichability could explain variance in rates of scalar implicature for different scalar terms. This provided indirect evidence for Gricean approaches that allowed for local enrichment. However, as outlined in Chapter 1, for the RSA-LU approach at least, it is the fact that a local upper-bound enrichment has a certain likelihood that can impact on the rates of scalar implicature. That inference is neutral whether the speaker intended a local enrichment or not. The priming paradigm

used in this chapter will potentially allow us to observe more direct evidence of an effect of local enrichment in unembedded scalars.

4.2 EXPERIMENT 1

4.2.1 Overview and prediction

In prime trials, participants were presented with a sentence paired with two pictures. Their task was to click on the picture that makes the sentence true. The sentences contained a scalar term 'some' which could occur in either unembedded or embedded position. Three types of pictures were available for each sentence: (i) false pictures which make all possible readings false, (ii) weak pictures which make the literal reading true but the enriched reading false, and (iii) strong pictures which make both the literal and enriched readings true. Two types of priming effects were examined, (i) unembedded prime -> embedded target, as shown in Figure 14, and (ii) embedded prime -> unembedded target, as shown in Figure 17. There were two types of prime trials. Participants were primed by the literal reading in weak primes and the enriched reading in strong primes. Following the procedure in Bott & Chemla (2016) and Raffray & Pickering (2010), each target trial was preceded by two prime trials, in order for the priming effect to be given a better chance of having an effect. For target trials, the sentence was presented with an open picture and a covered picture. Participants were instructed to click on the covered picture ('Better Picture?') if they thought there was a picture that would be a better match for the given sentence.

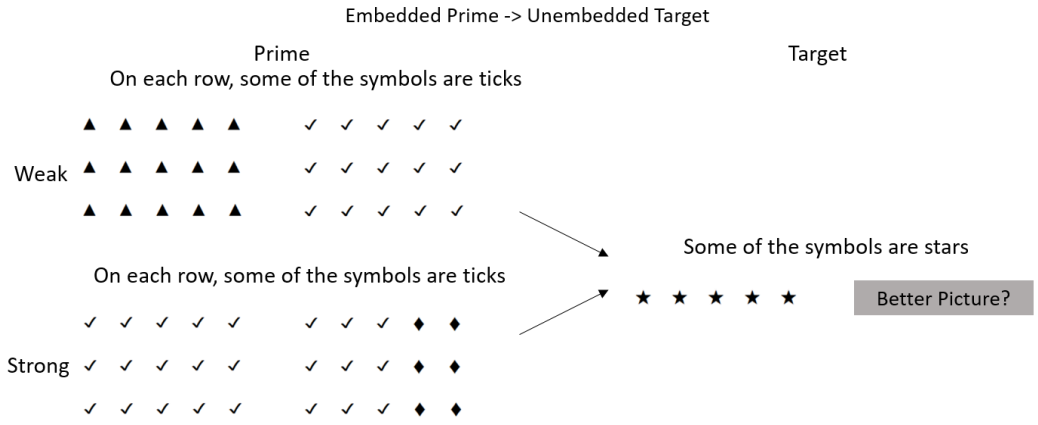


Figure 17 *Sample items of Experiment 1*

The embedded target condition has been discussed in detail in Section 4.1.3. Here I focus on the unembedded target condition. The critical items of this condition are illustrated in Figure 17. In the unembedded target condition, the target trial involving unembedded 'some' was preceded by prime trials involving embedded 'some'. For embedded prime trials, given the prime sentence like *On each row, some of the symbols are ticks*, in strong primes, the sentence was presented with a weak picture depicting all symbols being ticks and a strong picture depicting rows of symbols with some but not all being ticks. The strong picture made the locally enriched reading of the sentence true (i.e. On each row, some but not all of the symbols are ticks). For the same sentence, in weak primes, it was presented with a weak picture and a false picture depicting all symbols being non-ticks. Neither picture made the local reading true. Participants were thus forced to access the literal reading in weak primes.

Note that the sentences used for embedded target trials like '*on exactly one row, some of the symbols are squares*' were not used in embedded prime trials. This is because when 'some' is embedded under a non-monotonic quantifier, the literal reading and local enriched reading is logically independent. Thus, if non-monotonic cases are used as embedded primes, there is no better picture (in the sense of entailment) between a picture that makes the literal reading true and a picture that makes the enriched reading true.

As for unembedded target trials, the target sentence was the same as the one used for unembedded prime trials. Unlike embedded target trials, here the unembedded target sentence was presented with a visible picture that made the literal reading true. In this case, choosing the visible picture indicates that participants access the literal reading, whereas choosing the covered picture indicates that they access the SI-enriched reading.

In general, both the grammatical account and the RT/RSA-LU approach predict priming effects between unembedded and embedded enrichments since both approaches assume there is a shared mechanism between unembedded and embedded enrichments. Overall, the rate of enriched-reading responses to target trials should be higher after strong primes than after weak primes. However, as mentioned above, there

is a subtle difference between the two approaches in terms of the potential strength of priming in the different target conditions. The GT says that there is only one mechanism of exhaustification and it is present in both unembedded and embedded scalar enrichments. Thus, whether unembedded trials or embedded trials are primes, the subsequent target should receive more enriched responses after strong prime trials. For the RSA-LU approach, this prediction holds for the embedded prime --> unembedded target trials. However, for the case where the prime is unembedded, there are two routes to an enriched response. Only if enriched responses in unembedded primes involve a local pragmatic enrichment should there be substantial priming in the embedded target conditions. We shall return to this difference below.

4.2.2 Method

4.2.2.1 Participants

20 participants were recruited via Prolific Academic (http://prolific.ac). All participants were native English speakers.

4.2.2.2 Materials

This experiment had a two by two within-participant design. The two independent variables were the embeddedness of the target and the type of the prime. These two variables generated four prime-target combinations, as shown in Table 7. Sixteen experimental prime-target triplets were constructed. In each triplet, one target trial was preceded by two prime trials. Each trial consisted of a single sentence and two pictures. Eight triplets formed the unembedded prime → embedded target trials, the other eight formed the embedded prime → unembedded target trials. In half of the unembedded prime → embedded target trials, the target was preceded by two weak primes, while in the other half the target was preceded by two strong primes. This was the same for the embedded prime → unembedded target trials.

Target embeddedness	Prime type	Number of sets	Number of trials
embedded target	weak	4	12
	strong	4	12
unembedded target	weak	4	12
	strong	4	12
			48

Table 7 *Design of experimental items in Experiment 1*

For unembedded prime and unembedded target trials, the sentence was of the form *Some of the symbols are [symbol]*. For embedded prime trials, the prime sentence was of the form *On each row, some of the symbols are [symbol]*, whereas for embedded target trials, the target sentence was of the form *On exactly one row, some of the symbols are [symbol]*. The symbols were one of the circles, crosses, diamonds, hearts, squares, stars, ticks, or triangles.

48 filler trials were constructed. As with experimental trials, each consisted of a single sentence and two pictures. The sentence either contained 'some' as in *Some of the symbols are [symbol]* or *On each row, some of the symbols are [symbol]*, or contained 'all' as in *All of the symbols are [symbol]* or *On each row, all of the symbols are [symbol]*. Following the design in Bott & Chemla (2016), each type of the filler sentences occurred in three situations: (i) the sentence was presented with a strong picture and a 'Better Picture?', (ii) the sentence was presented with a false picture and a 'Better Picture', and (iii) the sentence was presented with a false picture and a strong picture. (i) and (ii) were included to counterbalance the times when in the target trials the covered picture ('Better Picture') was always paired with the weak picture. These trials also counter-balanced the extra times when in prime trials the sentence was always paired with two visible pictures. (iii) was included so that all possible pair combinations of three picture types (false, weak, strong) had equal occurrence. Examples of filler items were given in Appendix A.5.

In total, Experiment 1 contained 48 experimental trials (i.e. 16 prime-target triplets) and 48 fillers. The triplets of trials and the fillers were presented in a randomized order created for each participant. For prime trials, the position of the correct choice

was counterbalanced across trials so that half the trials the correct choice was on the left and half was on the right[12]. Furthermore, for half the experimental triplets the correct choice was on the same side for the first and the second prime, while for the other half it was on the opposite side. For target trials, the covered picture was always on the right. In addition, in one dual prime-target triplet, a different symbol was used as the predicate for each sentence. There were 8 symbol types. Each was used as the predicate in an equal number of times.

4.2.2.3 Procedure

Participants were instructed to click on the picture that made the sentence true. On occasions where one of the two pictures was covered, the task was the same. But participants were told that "if you think that there is a picture that would be a better match for the sentence, click on the 'Better Picture' option". Two examples were given. One involved 'many', in which the sentence 'There are many stars' was presented with one picture containing six stars and the other containing two. Participants were told to click on the picture containing six stars. The other example involved an ad hoc enrichment, in which the sentence 'There is a spade' was presented with one covered picture and one picture containing a spade and a diamond. In this case, participants were instructed to click on the 'Better Picture' option.

There were four practice trials to familiarise participants with the task. In these trials, the sentence was either presented with a false picture and a strong picture or with a false picture and a covered picture. No feedback was given in both practice and experimental trials. The whole experiment lasted approximately 10 minutes.

4.2.3 Data treatment and analysis

The analysis was performed on the responses of target trials. Only target responses that were preceded by two correct prime responses were included in the analysis. 35 out of 320 target responses were removed. Of the 35, 19 were embedded targets and 16 were non-embedded targets. For the remaining target responses, I coded the enriched response as 1, and the unenriched response as 0. Note that the enriched response for

[12] For weak primes, the correct response was the weak picture. For strong primes, although both pictures made the sentence true, we coded the strong picture as the correct response.

embedded target trials was choosing the visible picture, whereas the enriched response for unembedded target trials was choosing the covered picture.

I fitted a logistic mixed-effect model to predict the log odds of choosing an enriched over unenriched response from fixed effects of embeddedness (embedded targets / non-embedded targets) and prime type (weak/ strong). Embeddedness and prime type were deviation coded (embedded = 0.5, non-embedded = -0.5; strong = 0.5, weak = -0.5). The model contained maximal random effects structure supported by the data, which included random intercepts and slopes for subjects and random intercepts only for items, e.g. (1 | Item). All fixed effects and their interactions were included as random slopes, e.g. (1+Embeddedness * Prime type | Subject). Statistical analyses were carried out using R (version 3.3.3, R Core Team, 2017) with lme4 package (Bates et al., 2015) and lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2014).

4.2.4 Results and discussion

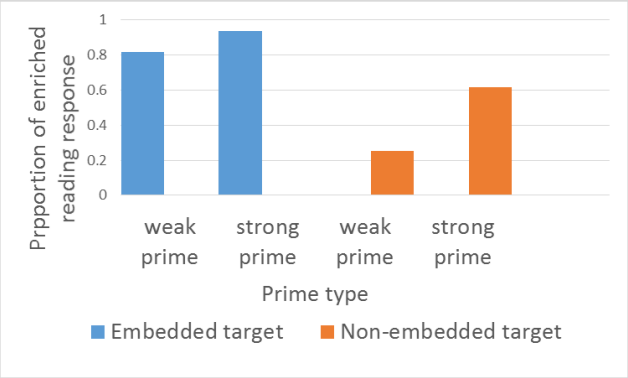


Figure 18 *The proportions of enriched responses across conditions in Experiment 1*

Figure 18 shows the proportions of enriched responses across conditions. We found a main effect of priming (β = 1.84, SE=0.62, p=.003). However, planned comparisons on each level of prime type showed that the rate of enriched responses was significantly higher after strong primes than after weak primes only in unembedded target conditions (β = 3.48, SE=1.36, p=.01) but not in embedded target conditions (β = 4.55, SE=3.87, p=.24). Thus, the observed priming effect was mainly driven by the priming in unembedded target condition. There was a main effect of embeddedness (β =4.81, SE=1.22, p<.001), suggesting that the overall rate of enriched responses was higher for

embedded target trials than for unembedded target trials. The interaction between embeddedness and prime type was not significant (β = -2, SE=1.42, p=.16).

The main effect of embeddedness in the present study is inconsistent with findings from previous research that demonstrate unembedded scalar enrichments are more robust than embedded cases (e.g. Benz & Gotzner, 2014; Geurts & Pouscoulous, 2009). However, it is difficult to read too much into this result since the enriched response in the embedded target condition is the open card, while the enriched response in the unembedded target condition is the covered card.

Regarding whether unembedded enrichments could prime embedded enrichments, the results of this experiment are difficult to interpret. On the one hand, there is a main effect of prime type and we found no significant interaction. On the other hand, we failed to find a significant difference between Strong and Weak conditions in the embedded target condition. The main effect was driven by the significant difference between Strong and Weak trials in the unembedded target condition. This latter result is supportive of the idea that there is a shared mechanism in EE and SS. However, an alternative explanation for this priming effect could be given without appealing to local enrichment. Consider the items in Figure 17 again. As long as participants access the reading 'On each row some of the symbols are ticks and it is not the case on each row all of the symbols are ticks', they would choose the strong picture. This means that local enrichment is not required in deriving this reading. It could be the result of global inference mechanism. Then what seems to be a local → local priming turns out to be a global → global priming. Thus, the priming effect in unembedded target condition cannot be taken as conclusive evidence for a shared mechanism in deriving unembedded and embedded enrichment.

4.3 EXPERIMENT 2

In order to properly explore whether embedded and unembedded enrichments could prime each other, I conducted Experiment 2 which addressed the problems of interpreting the results of Experiment 1.

4.3.1 Method

4.3.1.1 Participants

30 participants were recruited via Prolific Academic (http://prolific.ac). All participants were native English speakers.

4.3.1.2 Materials, procedure

The materials were similar to Experiment 1 with one key difference, that for the embedded prime trials, the prime sentence was of the form *On exactly one row, some of the symbols are [symbol]*. As illustrated in Figure 19, in strong primes, the sentence was presented with a picture that made the literal reading true and a picture that made only the local reading true. If the participants access the local enriched reading, *On exactly one row, some but not all of the symbols are ticks*, then the only picture that made the sentence true is the 'local' picture. Since embedded enrichments in the non-monotonic environment can only be explained by local enrichment, in Experiment 2, participants who choose 'local' picture must access local enrichment.

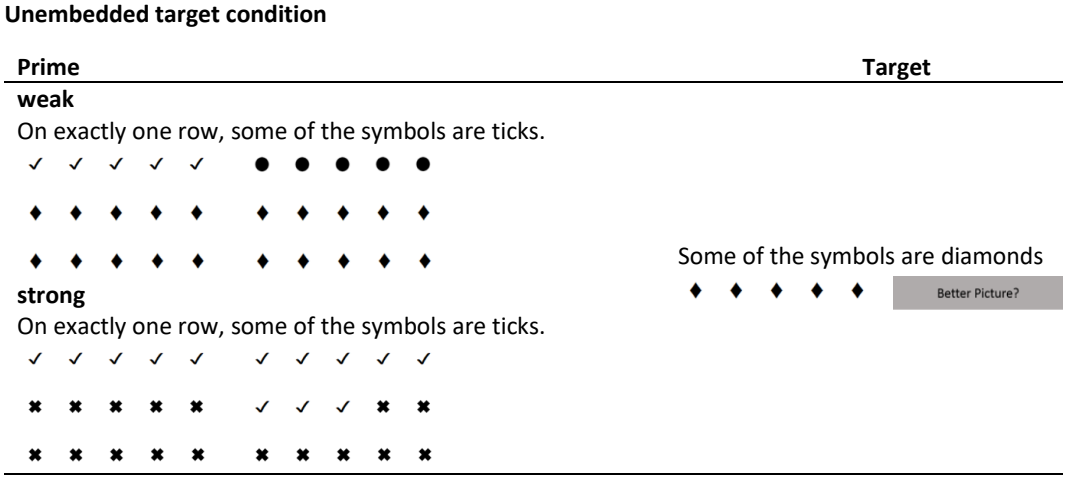


Figure 19 *Sample items of unembedded target condition in Experiment 2*

As with Experiment 1, 48 filler trials were constructed. The filler sentence was of the form *All of the symbols are [symbol]* or *On exactly one row, all of the symbols are [symbol]*. Like in Experiment 1, each type of the filler sentences occurred in three situations: (i) the sentence was presented with a strong picture and a 'Better Picture?', (ii) the sentence was presented with a false picture and a 'Better Picture', and (iii) the sentence was presented with a false picture and a strong picture. Examples of filler items

were given in Appendix A.6. All the other materials and the procedure were the same as Experiment 1.

4.3.2 Data treatment and analysis

As in Experiment 1, the analysis was performed on target responses that were preceded by two correct prime responses. 84 out of 480 target responses were removed. Of the 84, 24 were embedded targets and 60 were non-embedded targets. For the remaining target responses, we coded the enriched response as 1, and the unenriched response as 0.

Again I fitted a logistic mixed-effect model to predict the log odds of choosing an enriched over unenriched response from fixed effects of embeddedness (embedded/ non-embedded) and prime type (weak/ strong). The model contained random intercepts and slopes for subjects and random intercepts only for items. All fixed effects were included as random slopes.

4.3.3 Results and discussion

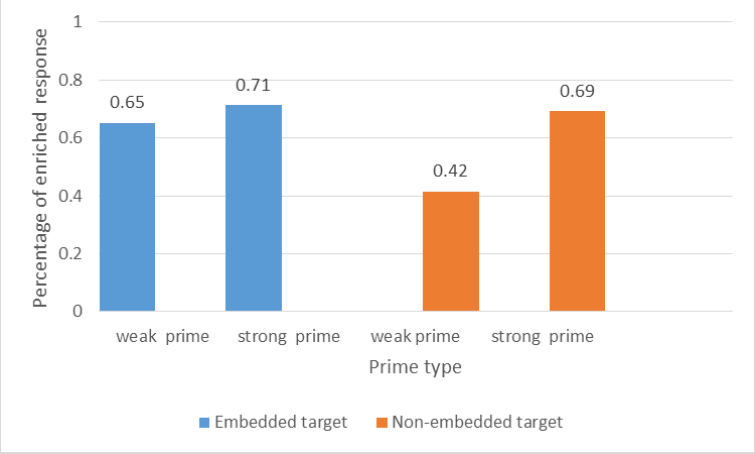


Figure 20 *The proportions of enriched responses across conditions in Experiment 2*

Figure 20 shows the proportions of enriched responses across conditions. There was a main effect of priming (β = 1.33, SE=0.39, p<.001). Again, planned comparisons on each level of prime type showed that the rate of enriched responses was significantly higher after strong primes than after weak primes only in unembedded target conditions (β = 1.56, SE=0.54, p=.004) but not in embedded target conditions (β = -1, SE=1.71, p=.56).

There was no main effect of embeddedness (β =2.07, SE=1.23, p=.07) and the interaction between embeddedness and prime type was not significant (β = -0.75, SE=0.77, p=.33).

In this experiment, enriched responses in both embedded prime and embedded target trials could not be the product of a global enrichment. Thus, the main effect of prime types provides clear evidence that embedded and unembedded scalar implicature share a mechanism. In particular, if we assume the general Gricean approach to scalars, the priming of the enriched response in the unembedded target by the embedded prime provides somewhat more direct evidence that unembedded scalar enrichments can be derived by the mechanism for local enrichment.

Overall, the main effect of prime provides support to both GT and RT/RSA-LU accounts. In terms of discriminating between the two approaches, once again, the results are difficult to interpret, although suggestive. On the one hand, we found a priming effect in the unembedded target condition but not the embedded target condition; on the other hand, the interaction did not reach significance. It is also worth noting that the items in the embedded target condition were identical across both experiments and in both cases no effect was found. As mentioned above, the RT/RSA-LU approach predicts that, if there would be an asymmetry in the priming effect, it would occur in the direction found. This is because, while embedded prime trials involve mandatory enrichment, unembedded prime trials do not. Thus the Gricean approaches suggest a stronger priming effect in the unembedded target condition.

4.4 Inverse Preference and Frequency of Local Enrichment

In this section, I will relate the results of Experiment 2 to the so-called 'Inverse Preference Effect'. Inverse preference is the phenomenon whereby a less frequent parse of a word or structure gives rise to a larger priming effect than more frequent parses (Hartsuiker & Kolk, 1998; Hartsuiker, Kolk, & Huiskamp, 1999; Hartsuiker & Westenberg, 2000; Scheepers, 2003). For example studies that manipulate active and passive syntactic structure find passives, which are the less frequent construction, give rise to larger priming effects than actives (Bock, 1986). Currently favoured explanations of this effect turn on the idea that priming itself is a result of implicit learning (Pickering &

Ferreira, 2008) and that inverse preference results from error correction (Jaeger & Snider, 2013).

Inverse preference is relevant to the results in Bott & Chemla (2016) because it potentially helps to explain a surprising result in their main experiment. Let us reconsider Figure 13, which shows rates of enriched response for trials where the primes were the same as the target (Within trials) and those where they differed (between trials). Of interest here is the fact that in this experiment, there was a main effect of Within/Between such that there were more enriched responses in the Between condition than Within, even though there was a significantly bigger effect of prime in the Within condition. This can be explained in terms of inverse preference if it is assumed that the unenriched response in prime trials is the less frequent or somehow unexpected one. This means that for Weak prime trials, there is a big priming effect for the unenriched response, causing participants to select the open picture in target trials. Bott & Chemla observe that indeed the large priming effect in Within trials is mostly due to a below baseline response in Weak trials. I.e. compared to a condition where the prime was unrelated to the target in terms of scalar implicature, participants made fewer enriched responses in the Weak prime condition.

Let us now turn back to the results of Experiment 2 of this chapter to consider where there might be an inverse preference effect. When we consider the unembedded target condition, it could be that because unenriched 'some' in Weak prime trials is unexpected, this primes the unenriched interpretation in the target. However, if the priming effect in Unembedded target trials is because of below baseline rates in weak trials, it would not explain why a similar effect is not obtained in the Embedded target condition. Of course, it could be that, again, we simply failed to find the same below-baseline effect in this condition. Alternatively, it could be that, if there are two mechanisms involved in scalar implicature, that the literal interpretation of 'some' is intermediate in its expectedness between a more frequent globally enriched and a less frequent locally enriched. This would explain the large priming effect in Unembedded target trials, because the Strong primes in this condition require local enrichment and, by hypothesis, local enrichment is a less frequent response than no enrichment.

When it comes to the Embedded target condition, if global enrichment is more often used to respond to Strong unembedded prime trials than local enrichment, and literal unenriched meanings are used in Weak trials, then we should not expect to see such a great priming effect because the target trials require local enrichment. The latter possibility would mean that, although both global and local processes may be responsible for unembedded Scalar Implicature, the global process may be the more common route.

At present, we have too little data to discriminate among these possibilities. Further studies would be required to shed light on the relation between global and local scalar enrichments in terms of their frequency. At a minimum, we would need to include an unrelated control condition here to get a better baseline.

4.5 CONCLUSION

The primary aim for this chapter was to use the enrichment priming paradigm to determine whether Embedded Scalar Enrichments and Unembedded Scalar Enrichments involved a shared mechanism. In two experiments, we found supporting evidence that there is a shared mechanism. In particular, Experiment 2 showed clearly that embedded prime trials where local enrichments are mandated lead to more Unembedded Scalar Implicature in Target than when only the literal meaning of 'some' is used in primes. This latter result in particular highlights that activation of locally enriched meanings of 'some' can impact on rates of Straight Scalar implicatures.

Although the focus of this thesis is to investigate the underlying mechanisms for scalar implicature from a Gricean perspective, it is also noted that the main results of this chapter are broadly supportive of the Grammatical Theory approach to scalars. Although there are relevant differences between the Gricean and Grammatical Theories, the data in this chapter does not conclusively favour one or the other. However, a twice-replicated lack of effect in the Embedded Target condition fit better with the Gricean picture than the Grammatical. Again, more studies would be needed to pursue this matter further. For instance, a similar kind of study that mixed lexical triggers in an

Unembedded target condition might provide such a test. I leave this question open for future research.

Finally, a speculative discussion about whether the results reported in Experiment 2 might be the result of an inverse preference effect led to the suggestion that perhaps the locally enriched interpretation of 'some' is less frequent/more surprising than either the globally enriched or literal interpretation. While this suggestion is highly speculative and requires further study, it is in line with a claim in Geurts & van Tiel (2013) that local pragmatic enrichment is a marked operation, and so less frequent. In the next chapter, I present a paper on the time course of access to local pragmatic enrichment of 'some'. To date, research on that topic has had mixed results, but an influential finding reported in Huang & Snedeker (2009, 2011) suggests that locally enriched readings of 'some' emerge at a significant delay, compared to the literal meaning of other quantifiers. If that is right, it would support Geurts & van Tiel's position.

Chapter 5 WHAT WOULD A COMPOSITIONAL LISTENER DO? – ANOTHER LOOK AT THE TIME COURSE OF SCALAR IMPLICATURES.

5.1 INTRODUCTION

Visual-world eye-tracking studies have been used to investigate the processing of pragmatic enrichments associated with *some*. Some of these studies found a delay in integrating the pragmatic interpretation of *some* relative to the semantic interpretation of *all* and exact numbers. For example, using a visual world paradigm, Huang & Snedeker (2009) compared the time course of referential disambiguation based on pragmatic *some* and conditions where no pragmatic enrichment was involved (interpreting *all*, *two*, and *three*). They presented participants a visual display depicting one girl with a total set of three soccer balls, another girl with a subset of two of four socks and some distractors, while listening to a sentence of the form "Point to the girl that has *some/all/two/three* of the *noun*". Upon hearing *some*, if participants interpret it as *two or more*, then they will not be able to disambiguate the referent until the *noun*, because both girls are compatible with the semantic interpretation of the referential phrase. However, if participants interpret *some* as *some and not all* rapidly, then they should be able to disambiguate the referent in the same timecourse as that for *all*. The results showed that participants were slower to anticipate the reference in *some* compared to *all*, *two* and *three* conditions. In particular, visual biases to the target in *all* and *three* conditions were significant from 400ms after the determiner onset, whereas the bias in *some* condition was not significant until approximately 800ms. The authors interpreted the delay in *some* as evidence that deriving the pragmatic interpretation is preceded by accessing the semantic interpretation in the early stage. We shall call this view on on-line pragmatic enrichment the *slow pragmatic* view.

However, other similarly constructed visual-world studies reported rapid integration of the pragmatic *some* (Richard Breheny, Ferguson, & Katsos, 2013; Grodner et al., 2010). These papers argue that no delay in pragmatic enrichment vis a vis semantic interpretation is consistent with a large body of previous research showing that effects of contextual inference are not necessarily delayed relative to the effects of semantic

composition, even where the contextual inference is based on Gricean reasoning (G. Altmann & Steedman, 1988; Sedivy, 2003; Sedivy et al., 1999; Tanenhaus et al., 1995). We shall call this the *fast pragmatic* view.

So far, it is unclear what factors lead to these conflicting results. One possibility is that rapid integration is the result of a pre-coding strategy. Comparing the studies that found little or no delay to those that found a delay, in the former, the referent was only described by the quantificational expressions, *all/some*, whereas in the latter, the referent was described by both quantificational and numerical scales *three*/*two*. According to the pre-coding account, when there is only one way of referring to the target item, participants may implicitly label the girl with the total set and the one with the subset with *all* and *some* respectively prior to the quantifier onset. This pre-coding strategy would facilitate target identification in both conditions and lead to no differences in the timecourse. By contrast, when there are two ways of referring to each target item (e.g. using *some*/*two* or *all*/*three*), a pre-coding strategy is unlikely to facilitate target identification because it is less efficient to label the potential referents with quantifier expressions. Therefore, according to a pre-coding account, the contradictory results are explained by a pre-coding strategy rather than fast pragmatic processing.

Another possibility is that the observed delay in implicature calculation is due to the availability of number alternatives which influences listeners' expectation of quantifier use. Several offline studies have shown that number terms are preferred over *some* when referring to a small set (Degen & Tanenhaus, 2011; van Tiel, 2014). Degen & Tanenhaus (2016) suggested that when the number terms are available in the experimental context, participants would expect to use number terms instead of *some* when referring to a set in the subitizing range. Hence, the delay in the *some* condition is due to a mismatch between the quantifier in use and participants' expectation.

Degen & Tanenhaus (2016) presented two visual world studies that provide a test for these hypotheses. They examined the time course of access and integration of scalar inferences while manipulating the presence and absence of number terms. In both studies, *all* and *some* were used equally often to refer to both big and small sets (of 4 or 5 items vs. 2 or 3 items). In the number-absent study, they replicated the results

of Grodner et al. (2010). This result is inconsistent with the slow pragmatic processing account and it cannot be explained by pre-coding strategy because verbal pre-coding is inefficient due to the use of additional garden path trials (using *some* to refer to a total set and *all* to refer to a subset set)[13]. In the number-present study, Degen & Tanenhaus replicated the results of Huang & Snedeker (2009) that in the quantifier time window, the target bias in *all* trials where the target was a larger set was greater than in *some* trials where the target was a small set. This delay could be explained by the availability of number terms and its effects on the listeners' expectations rather than slow pragmatic processing. Interestingly however, in the number-present study, Degen & Tanenhaus (2016) also found an interaction between quantifier use and set sizes, such that in trials where the target set size was big, the target bias was greater in the *all* than in the *some* condition, and in trials where the target set size was small, no difference was found. This interaction poses problems for the slow pragmatic processing account as the temporary delay was only observed when the target was a big set but not when it was a small set. It is of interest also, however, that this result was not predicted by Degen & Tanenhaus' account concerning the effect of the presence of numbers in the subitizable range. When *all* and *some* were used to refer to a big set (i.e. out of the subitizing range), the observed delay cannot be explained by the effect of number terms on listeners' expectation. Conversely, when *all* and *some* were used to refer to a small set that is in the subitizing range, no difference between timecourse in *all* and *some* was found, contrary to what we might expect. Therefore, the interaction between quantifiers use and target set size opens a new dimension in exploring the timecourse question by investigating how set sizes affect eye movements during online interpretation of scalar quantifiers.

In this paper, we aim to account for the effects that have been found in previous studies in part by proposing that participants in these studies have prior expectations about the relative set size of the target in a display given the quantifier (*all*/*some*) and such expectations about set sizes influence the online measures. Considering Huang & Snedeker (2009)'s visual displays, the *all* referent was always an agent with a larger set

[13] The possibility of pre-coding could not be completely ruled out, as only 16 out of 64 trials were garden path trials.

of three items, and the *some* referent was always an agent with a smaller set of two. If a prior association between *all* and the larger set is stronger than between *some* and the smaller set, then the delay in pragmatic *some* could be explained in part as the result of prior expectations rather than the slower pragmatic processing.

In addition to the effect of relative set size on target bias, there are several further issues that make the interpretation of previous visual world data problematic. First, in studies with number terms, the timecourse of target identification in *number* and *all* conditions does not reflect the difference between numbers and non-numbers that should take place in the verification processing. Considering Huang & Snedeker's study, the process of identifying the referent of the description 'the girl that has *some/all/two/three* of the…' involves verifying the quantificational NP against the sets of objects associated with the characters in the display. In the case of *number* trials, it is sufficient to only inspect the cardinality of the sets that each girl has. However, for *all* trials, to establish that the girl with the three soccer balls is the *all* referent, the whole display need to be checked to ensure that no other characters obtained any soccer ball. Thus, given the difference in region of inspection required to anticipate the referent, we expect that target identification should be faster in *numbers* than in *all* and *some*. Yet, Huang & Snedeker (2009) found no difference in looks to the target between *three* and *all* during the critical time window.

Secondly, experiments supporting the no-delay account were set up to establish a null effect. Working on the assumption that pragmatic interpretation should be available in the same timecourse as the semantic interpretation, many experimental studies, for instance Grodner et al. (2010), predicted no difference in the looking pattern between quantifiers. In order to accept the null hypothesis, Grodner et al. (2010) demonstrated in a post-analysis that an effect of quantifier is mostly small. So far, no online studies formulated an alternative hypothesis based on a no-delay account.

To address the problem of finding positive evidence to establish if pragmatic enrichment of *some* occurs in the same timecourse as the interpretation of *all*, we introduce a new, less problematic measure. This involves a 'residue set' of objects, where participants can inspect which type of objects are in the partitioned sets. For example, in our displays, when a character has some but not all of the set of stripy circles,

the stripy circles not in her possession are located in a separate visual region of interest. In order to determine whether a character has *all* or *some but not all* of a set of objects, a participant should check this residue set region to ascertain for a given character if she has a total set of objects or a partitioned set. Thus, we would predict an increase in visual search of the residue set region after determiner onset for *all* and pragmatically enriched *some*. By contrast, to establish if a character has *two/three* or *some and possibly all* of a set of objects, it is sufficient to only check the objects in the character's target region – not the residue region. So, for both numbers and un-enriched *some*, there should be no increase in visual search in the residue set after determiner onset. Using this measure in a design with number items allows us to test for an effect of pragmatic *some* items against numbers – rather than rely on null effects as evidence for the fast-pragmatic account.

Lastly, an initial visual preference affects the timecourse of target identification in previous visual world studies. Many studies have found participants tend to look at the big set of objects in the display before the onset of the quantifier (Richard Breheny et al., 2013; Grodner et al., 2010; Huang & Snedeker, 2009). This visual preference results in an increase in the time it takes to make a saccade to an area with fewer objects. In these studies, the big set is often the referent of *all* description and the small set is often the referent of *some* descriptions. Thus, *all* trials, relative to *some* trials, accrued an advantage from this initial visual preference. So far, the initial looking preference was controlled for only in the analyses, not in the experimental design.

In the first two sets of experiments, we present a pair of off-line studies (Experiments 1a,b) that allow a preliminary exploration of any association between quantificational determiners (*some* and *all*) and relative set size. We follow up these studies with a pair of visual-world eye-tracking studies (Experiments 2a,b) in which the relative set-sizes from Experiments 1a,b are used. Experiment 3 provides a final test of the idea that relative set size is a factor in these visual world experiments and serves as a further test of the fast-pragmatic and slow-pragmatic views.

5.2 EXPERIMENT 1

Our hypothesis is that participants have expectations that an agent with all of something would possess a relatively large set of objects and this expectation gives an unfavourable advantage in *all* trials over *some* if the not controlled for. To measure prior expectations, we conducted Experiment 1(a,b) in which participants were presented with a statement containing a quantifier, e.g. 'The girl has *all/some* of her sister's flowers' and a slider scale with one image located on each end (see Figure 21). The slider of the scale is initially placed in the centre. We asked participants to indicate which image fits better with the statement by moving the slider toward the chosen image. Both images depicted a girl possessing a set of flowers. One image is a girl possessing a larger set (e.g. a girl with three roses), the other is a girl possessing a smaller set (e.g. a girl with two daisies). Since neither of the images are clear on whether the sets of flowers had been partitioned or not, participants may find the task ambiguous and leave the slider in the centre. In this case, they ignore any prior expectations. By contrast, if after reading the quantifier, participants integrate a prior expectation for set size, then we would expect that for an *all* statement, the slider should be moved further towards the image depicting an agent with a larger set than for the *some* statement. In Experiment 1a the large/small set sizes are 3 and 2 respectively. In Experiment 1b, they are 4 and 3 respectively.

5.2.1 Experiment 1(a)

5.2.1.1 Participants

52 Participants were recruited via Amazon Mechanical Turk. They were asked to indicate their native language and only participants with English as a native language were included in the analyses.

5.2.1.2 Materials and procedure

The two target sentences were "The girl has all of her sister's flowers" and "The boy has some of his cousin's candies". Each target sentence was paired with a slider scale that has one image located on each end (Figure 21a). Both images depicted a character with a set of objects. The objects fall into the same category (e.g. flowers) but belong to different kinds (e.g. rose, daisy). One character possesses a set of three objects and the

other character possesses a set of two. The slider's starting point was placed at the centre of the scale. Participants were instructed to use the slider to represent their intuition about the correspondence between the sentence and the image; such that, if one image is a lot better than the other, they move the slider right over toward that image, and if one image is only a little bit better, then they may move the slider a little bit toward that image. In addition, there were 4 fillers, of which two were clearly unambiguous on which image it described (e.g. The boy has less oranges than lemons), and two were clearly ambiguous (e.g. The girl has red and green apples). In total, each participant judged 2 experimental items and 4 fillers. 8 lists with the pseudo-randomized order of trials were created. This was to ensure that overall, for the quantifiers *all* and *some*, big set and small set were equally often to appear at two ends of the slider scale. Objects (e.g. rose, daisy) contained in big and small set were counterbalanced within each quantifier. Participants were randomly assigned to one of eight lists.

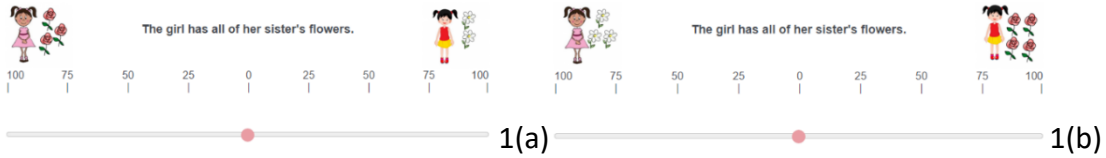


Figure 21 *Examples for experimental items used in Experiments 1(a) and 1(b)*

5.2.1.3 Results and Discussion

Responses from non-native English speakers and participants who made mistakes on the clear true or clear false fillers were excluded. 39 participants were analysed. For each trial, participants' rating was mapped on to a 0-200 continuous scale, on which 100 corresponds to the slider starting anchor point. A rating of 100 indicates no preference for one image over the other. When the rating is above 100, it indicates a preference for the big set over the small set, the stronger the preference for the big set, the higher the rating. When the rating is below 100, it indicates a preference for the small set over the big set, the stronger the preference for the small set, the lower the rating. We found that the mean rating for *all* trials was 131.69 and the mean rating for *some* trials was 120.85. A one-sample t-test showed that mean rating for both quantifiers were significantly higher than 100 (all: $t(38) = 6.69$, $p < .001$; some: $t(38) = 3.763$, $p = .001$). Thus for both quantifiers, participants preferred the agent with the larger set of three objects as the referent of the sentence. A paired sample t-test showed that the preference for

one image over the other in *all* trial was stronger than in *some* trial (t(38) =2.35, p =.024). This suggested that the preference for *all* to be used with the larger set was stronger than the preference for *some* to be used with the larger set.

Our conjecture that there is an association between the larger-set target and 'all' is confirmed. What we found regarding *some* was not predicted but strengthens the case that factors beyond simple composition of meanings may be driving anticipatory looks in previous visual world studies, such as Huang & Snedeker (2009). That is, because participants in the *some* condition of this experiment showed a preference for the girl with the set of three things over the one with the set of two things, low-level associations are not only driving looks toward the correct target in *all* trials, but also driving looks away from the correct target in *some* trials.

Regarding our hypothesis that it is an association between the larger set and *all* that explains the eye-tracking results, we should consider an alternative hypothesis regarding Experiment 1(a). This is that there should be a strong dis-preference for the agent with two objects as the referent of the description 'the girl with *all* of the flowers' due to what is termed an anti-presupposition (Heim, 1991). Anti-presuppositions block the use of a term when an alternative term has a stronger, more informative presupposition. For the items at hand, 'the girl with all of the flowers' is infelicitous for the two-flower target given that the speaker could say, 'the girl with both of the flowers'. Note that, the presence of this anti-presupposition should have an effect also on target bias in Huang & Snedeker's visual world study, where one female agent has two objects, while the other has three. Thus if anti-presuppositions affect anticipation, this would impact unfairly on the comparison between *all* and *some* conditions in Huang & Snedeker's world study. However, the effect of anti-presuppositions would be lower in Degen & Tanenhaus' study, where the competitor in half of the big-set *all* trials is a set of three objects. Thus, in order to more properly explore our hypothesis and to get some sense of the effect of anti-presuppositions, we conducted Experiment 1(b). In this study, we used the quantities, 4/3. A secondary motivation for changing the quantities in this follow-up relates to the result in the *some* condition in Experiment 1(a). We were interested to see if the preference for the girl with three objects in Experiment 1(a) had to do with the specific quantities involved.

5.2.2 Experiment 1(b)

5.2.2.1 Participants

52 Participants were recruited via Amazon Mechanical Turk. They were asked to indicate their native language and only participants with English as a native language were included in the analyses.

5.2.2.2 Procedure and materials

The procedure was identical to Experiment 1(a). The materials were similar to Experiment 1(a) with one key difference. We increased set sizes so that in experimental trials, the big set contained four objects and the smaller set contained three objects (See Figure 21 (b) for the example *all* display).

5.2.2.3 Results and Discussion

Responses from non-native English speakers and participants who made mistakes on the clear true or clear false fillers were excluded. 38 participants were analysed. Participants' ratings were mapped on to a 0-200 continuous scale as in Experiment 1a. The mean rating for *all* trials was 115.32 and the mean rating for *some* trials was 97.50. A one-sample t-test showed that the meaning rating for *all* was significantly higher than 100 ($t(37) = 4.99$, $p < .001$), but the mean rating for *some* was not significantly different from 100 ($p = .56$). Paired sample t-test showed that on average, the preference for one image over the other in *all* trial was stronger than *some* trial ($t(37) = 3.29$, $p = .002$).

As in Experiment 1a, we found the big set preference for *all* trials. This result cannot be explained as an effect of anti-presupposition since the cardinality of the smaller set was greater than two. In Experiment 1b, we found no clear preference for *some* trials. This suggests that the *some*-result in Experiment 1a was not due to a preference for the set of three items, but perhaps due to a dis-preference for the set of two items, relative to the set of three.

Taken together, the results of Experiment 1 provide evidence that people expect an agent with a larger set in a display as the *all* referent. The expectations about the set size given the use of 'all' is stronger than expectations (if there are any) given the use of 'some'. While there is no clear evidence for prior expectations in the case of 'some'

based simply on relative set size, the results of Experiment 1a suggest that 'some' is disprefered to be used with a set of two objects. To investigate whether prior expectations of the set size given the quantifier use affect online measures, we conducted Experiment 2.

5.3 EXPERIMENT 2

5.3.1 Experiment 2(a)

One aim of Experiment 2(a) is to explore how prior expectations of set sizes interact with scalar processing using a visual-world paradigm. In the visual display, there are always two agents that each has a large set of objects and another two agents that each has a small set of objects. The large set contained three objects and the small set contained two objects. Based on offline preferences found in Experiment 1a, we predict that if prior expectations influence anticipatory processing, bias to the target in *all* and *some* should increase faster when the target is a big set compared to when it is a small set. We also predict that when the target is a big set, bias to the target should increase faster in the *all* condition than the *some* condition. Additionally, the design of this experiment involves clearly distinct residue sets and hence enables us to test predictions about the search procedure for determining the target in *all* and *some* conditions.

5.3.1.1 Methods

5.3.2 Participants

36 participants were recruited from our university campus via an online psychological subject pool. All participants speak English as a native language. They have uncorrected or corrected to normal vision.

5.3.2.1.1 Materials

The experiment employed a three by two within-subject design. The two independent variables were *Determiner* (All, Some, Number) and *Target size* (Big, Small), which generated six experimental conditions: big *all*, small *all*, big *some*, small *some*, big *number* (i.e. three), small *number* (i.e. two). Auditory instructions were of the form "Click on the girl with [Det] of the [modifier] [shape]". [Det] was one of *some*, *all*, *two*, *three.*

[modifier] was one of *dotted*, *stripy*, *checked*; and [shape] was one of *circles*, *squares*, *triangles*. 36 experimental displays were constructed and paired with an audio instruction containing one of the determiners. Each experimental display was preceded by an initial display which consisted of four identical agents and four sets of different objects, as in Figure 22 (a) or Figure 23 (a). In the subsequent experimental display, four sets of objects were distributed among the agents, as in Figure 22 (b) or Figure 23 (b). There were always two agents that had a total set of one kind of object and the other two had a proper subset. The residues of two partitioned sets remained in the centre. In terms of set sizes, two of the four agents always possessed a set of three objects and the other two possessed a set of two objects. We counterbalanced the target set size for *all* referent and *some* referent by adopting two types of initial display and changing objects in the residue set. In particular, starting from Figure 22 (a) in which each set contained three objects, Figure 22 (b) can be used in a small-set *some* or big-set *all* trial. Starting from Figure 23 (a) in which two sets contained two objects and the other two contained four objects, Figure 23 (b) can be used in a big-set *some* or small-set *all* trial.

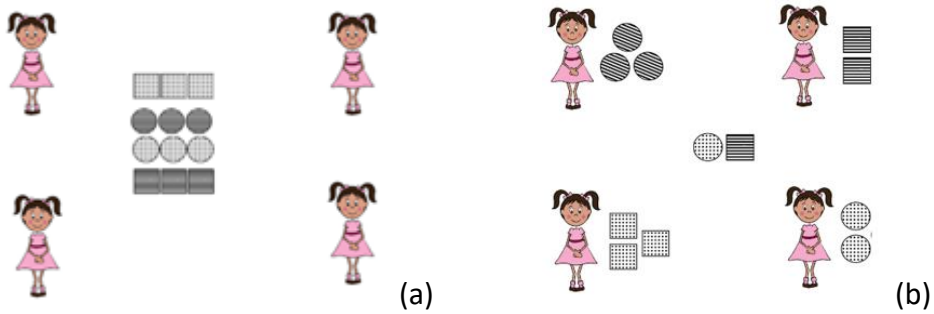


Figure 22 *Example displays in Experiment 2(a) Presented after (a), (b) can be paired with instructions 'Click on the girl that has all/three of the stripy circles' or 'Click on the girl that has some/two of the stripy squares'.*

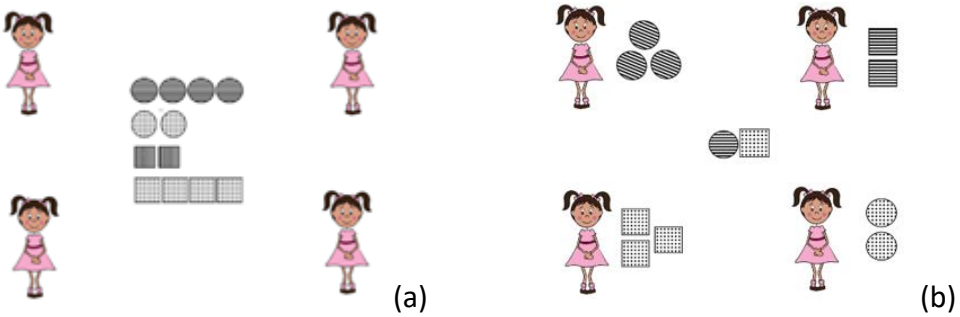


Figure 23 *Example displays in Experiment 2(a) Presented after (a), (b) can be paired with instructions 'Click on the girl that has some/three of the stripy circles' or 'Click on the girl that has all/two of the stripy squares'.*

Three lists were created (see Table 8). Each list contained 36 experimental items, 12 items per determiner. Each experimental display only appeared once in each list in one of the six conditions. In addition, each list contained 18 fillers. Fillers were similar to experimental items but contained different determiners (*One*, *Four*, *None*) in the instruction. Of these fillers, 12 number fillers were included to counter-balance the extra times that the target was referred by quantifiers in experimental items.

Exp./Filler	Determiner	Target size	Number of items	Total
Exp.	some	big	6	12
		small	6	
	all	big	6	12
		small	6	
	number	big	6	12
		small	6	
Filler	four	big	6	18
	one	small	6	
	none		6	
				54

Table 8 *Experimental Design of Experiment 2(a)*

The audio descriptions and instructions were recorded in a single session by a male native British English speaker. The speaker was instructed to record all of the sentences with a neutral intonation. The audio instructions were cross-spliced in order to avoid co-articulation information in favour of any condition.[14] Across stimuli, the onset of the determiner was the same. The durations of critical time windows were adjusted using phonetics analysis software Praat (Boersma & Weenink, 2017). The average duration for the determiner time window was 773ms (*all of the*: 741ms, *some*

[14] Each audio instruction was first recorded individually. Then we spliced determiner, modifier and shape words into an instruction schema created from a recording of, 'Click on the girl with most of the orange oblongs', which provides no advantage to any condition in terms of co-articulation information prior to critical words.

of the: 793ms, *three of the*: 784ms, *two of the*: 773ms), the average duration for modifier window was 596ms (*stripy*: 597ms, *dotted*: 594ms, *checked*: 596ms).

The shape (circles, triangle, square), pattern (stripy, dotted, checked) and location of the target were counterbalanced within each condition. All pictures of an agent with a set of objects measure 336*315 pixels. Picture of items in the middle measure 168*210. The screen resolution is 1680*1050 pixels.

5.3.2.1.2 Procedure

Each trial began with a display in which four agents surrounded four sets of objects, as in Figure 22 (a) and Figure 23 (a). Participants heard a description of the types of objects in the middle, for example, “There are stripy squares, dotted squares, stripy circles and dotted circles”. Six seconds after the onset of the description, the next display appeared, as in Figure 22 (b) and Figure 23 (b). The objects were distributed to four identical agents. After 2.5 seconds, participants were given an auditory instruction, for example, “Click on the girl with some of the stripy circles”. Participants’ task was to click on the image according to the instruction, and they were asked to respond as quickly as possible. The average length of the instruction was 3.8s. The session was set to jump to the next trial 5.5 seconds after the onset of the instruction. There were six practice trials in the beginning to ensure that participants understood the instruction, display and procedure. They then completed 54 trials, divided into 36 critical trials and 18 fillers. Randomized order of presentation of the items was created for each participant.

The experiment was conducted using E-Prime software and a Tobii TX300 eye-tracker. Fixations were sampled every 17ms. Participants were calibrated at the beginning of the experiment using a nine-point display. Before each trial, there was a fixation cross in the centre of the screen, and participants' eye gaze had to be fixed on this point for a continuous 1 seconds before the trial started. Eye movements were recorded from the onset of the instruction for 5.5 seconds for each trial. The whole experiment lasted approximately 20 minutes.

5.3.2.2 Data analyses and Results

We excluded trials which participants clicked on the wrong target (0.93%). A fixation that landed within the coordinates (in pixels) of an agent with a set of objects or the

residue set was coded as a look to that area, otherwise, it was coded as background. Any fixations shorter than 80 milliseconds were excluded, as extremely short fixations are often due to false saccade planning (Rayner & Pollatsek, 1989). Eye movements and auditory input have been synchronized according to the onsets and offsets of individual words on an item-by-item basis. Note that for all plots and data analyses, word regions have been offset by 200ms, as it takes around 200ms to launch an eye-movement (Hallett, 1986). Statistical analyses were carried out using R (version 3.3.3, R Core Team, 2017) with lme4 package (Bates et al., 2015) and lmerTest package (Kuznetsova et al., 2014).

5.3.2.2.1 Analyses of target anticipatory eye movements

In the first set of analyses we examined the timecourse of target identification after hearing quantifiers and numerical determiners. In particular, we were interested in whether target bias formation in the *all* and *some* conditions was influenced by prior expectations. We defined two critical time windows: the determiner window ([Det] onset-'the' offset, e.g. during 'some of the') and the modifier window (modifier onset-modifier offset, e.g. during 'stripy'). During the determiner window, the two agents that had an incomplete collection of one object type were targets for *some* and competitor for *all*, the two agents that had a complete collection were targets for *all* and competitors for *some*. During the modifier window, the target was the agent of the description, the competitor was the agent that had the objects with the same pattern.

To represent the time course of target bias formation, we calculated the natural log ratio of percentage of looks to the target over that of the competitor as a function of time (Ln (P(T)/P(C)). When the log ratio is 0, there is an equal percentage of looks to target and competitor. When the log ratio is above 0, there is a bias towards the target, and when the log ratio is below 0, there is a bias towards the competitor. We visualised the data for the determiner window and the modifier window separately because regions of interest changed as the sentence unfolded. We plotted log ratios over each 17ms sample for the average length of each time window. Curves in these log ratios plots were resynchronized at time window onsets to ensure the target bias formation was visually represented accurately (G. T. M. Altmann & Kamide, 2009).

For statistical analyses, due to the eye-movement based dependencies, the data were aggregated over 50ms time bin (every three samples). The natural log ratio of percentage of looks to the target over competitor was calculated for each 50ms bin as dependent measure. For each critical time window, the determiner window and the modifier window, we fitted a mixed effect model to predict log ratios from fixed effects of Determiner (All, Some, Number), Target size (Big, Small), a continuous time variable and their interactions. Determiner was coded as a dummy variable (0 = number), target size was coded using contrasts codes (Small, -0.5; Big, 0.5), and Time was centred before entering the model. The model contained maximal random effects structure supported by the data, which included random intercepts for subjects and items (e.g. (1|Subject)) and uncorrelated random slopes (e.g. (0+Determiner)|Subject). All fixed effects and their interactions were included as random slopes. Model comparisons were conducted to test if a model with the effect of interest fits the data significantly better than a maximally similar model without the effect. Both significant main effects and interactions were followed up by planned comparisons, in which models were recoded with different reference level. We first report the analyses in the determiner window, then the analyses in the modifier window.

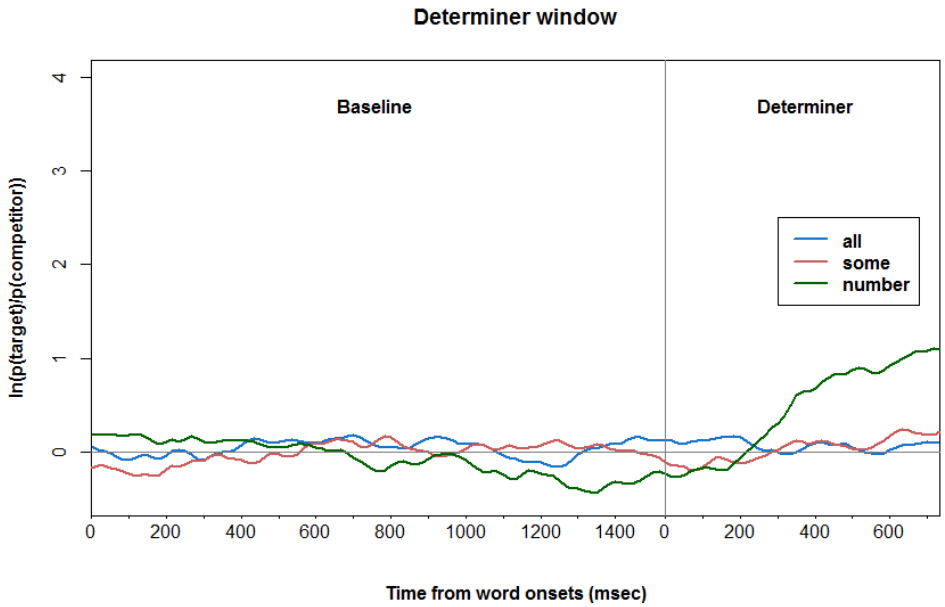


Figure 24 *Log ratios of percentage of looks to target over competitor by Determiner from the instruction onset to the determiner window offset in Experiment 2(a) (e.g. 'Click on the girl with some of the')*

Figure 24 shows the time course of target bias formation by *Determiner* from the instruction onset to the determiner window offset (i.e. 'the' offset). Inspecting the graph, the target bias in numbers increased immediately after the onsets of the determiner, whereas for both quantifiers, biases toward target developed much later. By the end of the determiner window, there was a clear difference in gaze bias between numbers and quantifiers, but not within quantifiers. The results of mixed effects analyses were consistent with the visual inspection. We found a main effect of *Time* ($\chi2(1) = 16.59$, $p<.001$), indicating that biases toward target increased overall. Importantly, there was a significant main effect of *Determiners* ($\chi2(2) = 12.68$, $p=.002$) and a significant interaction of *Determiners* and *Time* ($\chi2(2) = 31.35$, $p<.001$). The results from the planned comparisons showed that the target bias was greater in the number condition than both quantifier conditions (all: b=-0.51, SE=0.17, p=.003; some: b=-0.61, SE=0.18, p=.001), also it increased faster over time in numbers than both quantifiers (all: b= -2.97, SE=0.55, p<.001; some: b= -1.55, SE=0.55, p=.007). Within quantifiers, we found no difference between *some* and *all* in the overall target bias (b=-0.06, SE=0.17, p=.71), and as the time increased, target biases did not increase at different rate (b=0.94, SE=0.54, p=.09).

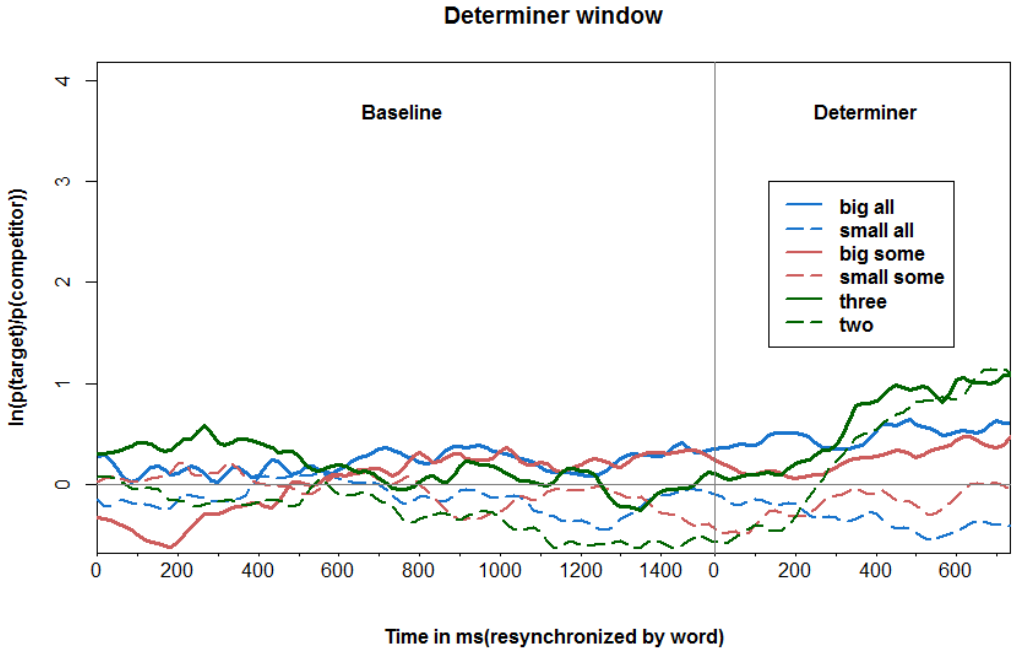


Figure 25 *Log ratios of percentage of looks to target over competitor by Determiner and Target size from the instruction onset to the determiner window offset in Experiment 2(a)*

Figure 25 depicts how target bias developed over time by *Determiner* and *Target size* from the instruction onset to the determiner window offset. There was a significant main effect of *Target size* (b=1.00, SE=0.25, p<.001), indicating that the overall target bias was stronger for the big-set target than for the small-set target. We also found a significant interaction of *Determiner* and *Target size* ($\chi2(2) = 7.89$, p=.02). Planned comparisons on each level of *Determiner* showed that the simple main effect of *Target size* holds for *all* and *some* but not in *numbers* (b=0.21, SE=0.21, p=.32). Target biases in big *all* and big *some* were stronger than biases in small *all* and small *some* respectively (all: b= -0.96, SE= 0.27, p<.001; some: b= -0.60, SE=0.22, p=.007). Planned comparisons on each level of *Target size* showed when the target set was big, no difference was found in overall target bias among *all, some* and *numbers* (i.e. three). When the target size was small, we found again no difference between small *some* and small *all* (b=0.14, SE=0.21, p=.51), but the target bias in small *number* (i.e. two) was greater than both quantifier conditions (all: b=-0.94, SE=0.23, p<.001; some: b=-0.82, SE=0.19, p<.001).

There was also a significant three-way interaction of *Determiner, Target size* and *Time* ($\chi2(2) = 9.12$, p=.01). At each level of *Determiner*, *Target size* influenced bias formation in *numbers* but not in *all* and *some* (all: b=-1.22, SE=0.63, p=.06; some: b=-0.13, SE=0.62, p=.84). In particular, the target bias increased more quickly in the *two* than in the *three* (b= -1.54, SE=0.69, p=.03). Considering each level of *Target Size*, when the target size was big, neither the difference between big *all* and big *some* nor the difference between big *number* and big *some* was significant (b=0.28, SE=0.70, p=.69; b=0.87, SE=0.67, p=.20, respectively) , we only found the bias in big *number* increased faster than in big *all* (b= 1.35, SE=0.61, p=.03). When the target size was small, we found a marginally significant advantage in the slope for small *some* over small *all* (b=1.3, SE=0.66, p=.055), and the target bias in small *number* increased faster than both

quantifiers (all: b=-4.17, SE=0.71, p<.001; some: b=-2.76, SE=0.69, p<.001).

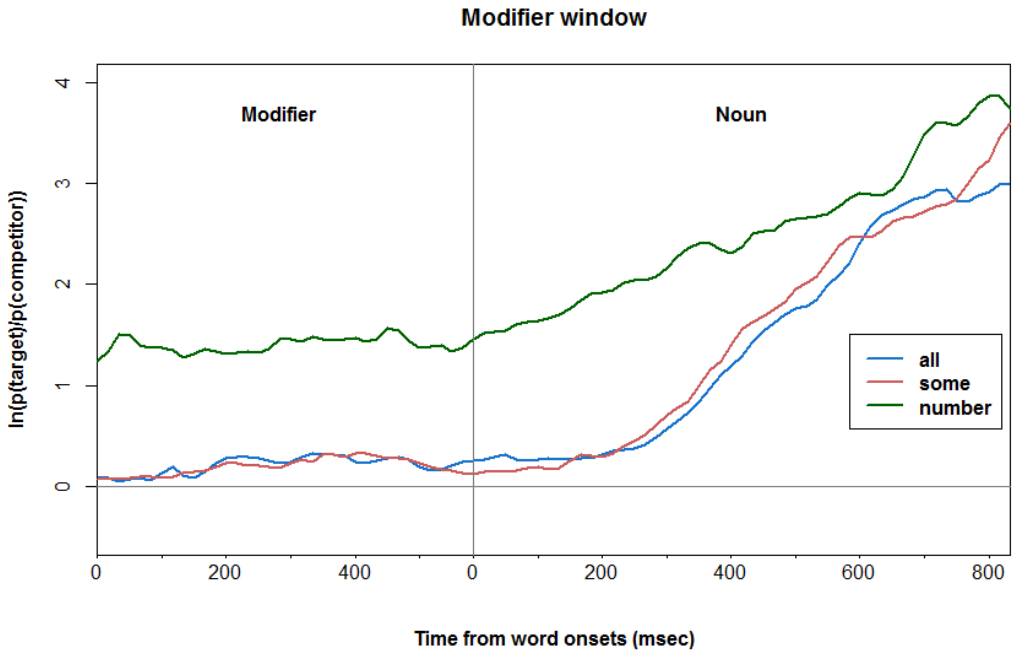


Figure 26 *Average ln((Target)/(Competitor) by Determiner from the modifier onset to the instruction offset (e.g. 'stripy squares') in Experiment 2(a)*

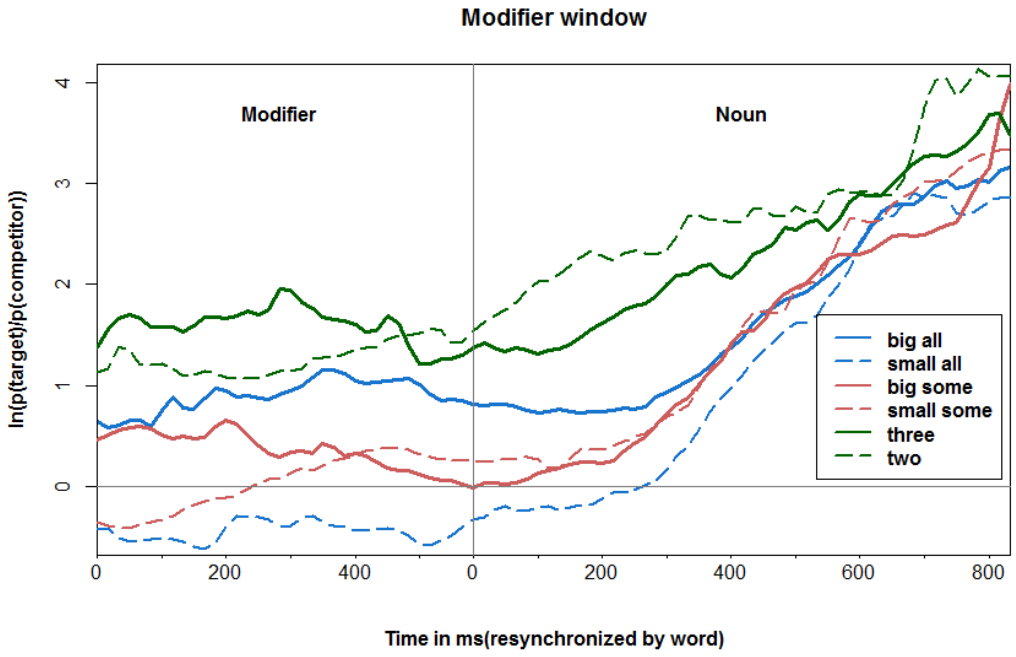


Figure 27 *Average ln(P(Target)/P(competitor) by Determiner and Target size from the modifier onset to the instruction offset in Experiment 2(a)*

For the modifier window, as shown in Figure 26, the main effect of *Determiner* continued ($\chi2(2) = 30.02$, $p<.001$), there was a stronger target bias in number conditions than in both quantifier conditions (all: b= -0.94, SE=0.21, p<.001; some: b=-0.86, SE=0.20, p<.001), and within quantifiers, the overall target bias was not different (b=0.05, SE=0.20, p=.80). We again found a significant main effect of *Time* ($\chi2(1) = 3.98$, p=.046), but we found no significant interaction of *Determiners* and *Time.*

Figure 27 depicts how target bias developed over time by *Determiner* and *Target size* during the modifier window. The main effect of *Target size* continued (b=0.44, SE=0.19, p=.02). We also found a significant interaction of *Determiner* and *Target size* ($\chi2(2) = 10.12$, p=.006) and a significant three-way interaction of *Determiner, Target size* and *Time* ($\chi2(2) = 6.44$, p=.04). Planned comparisons on each level of *Determiner* showed that for *all*, the target bias was greater and increased faster when the target set was big compared to when it was small (b=-1.09, SE=0.27, p<.001; b=-1.35, SE=0.56, p=.02). For *some*, the overall bias to the target did not differ between two set sizes (b=-0.16, SE=0.24, p=.51), but the target bias increased faster in small-*some* than in big-*some* (b=1.65, SE=0.60, p=.008). For numbers, *Target size* influenced neither overall looks to the target nor the changes in bias over time (b=-0.08, SE=0.22, p=.71; b=-0.59, SE=0.58, p=0.31).

Comparison within *Target size* showed that when the target size was big, the overall target bias in *all* was marginally greater than in *some* (b=-0.44, SE=0.25, p=.09), and the target bias increased faster in big *all* than in big *some* (b=1.68, SE=0.63, p<.001). In addition, we found no difference in looking pattern between big *number* and big *all*, and the target bias was greater and increased faster in big *number* than in big *some* (b=-0.70, SE=0.25, p=.006; b=1.39, SE=0.47, p=.004). When the target size was small, although the overall target bias did not differ between small *all* and small *some* (b=0.46, SE=0.28, p=.10), we found the target bias increased faster in small *some* than in small *all* (b=1.46, SE=0.63, p=.025). We also found that the overall target bias was greater in small *number* than in both quantifiers (all: b=-1.50, SE=0.28, p<.001; some: b=-1.14, SE=0.25, p<.001). The bias in small *number* increased faster than in small *all* (b=1.46, SE=0.63, p=.025), though there was no difference between small *number* and small *some* in terms of the changes in target bias over time (b=0.52, SE=0.59, p=.66).

To summarize, in both time windows, we found that if we disregard the distinction between set size, target bias emerged earlier and stronger in number conditions than in *all* and *some*, whereas between *all* and *some*, bias formation did not differ. These results reflect the difference in the verification process between numbers and non-numbers and are consistent with fast-pragmatic view that accessing and integrating 'not all' inference triggered by *some* should be rapid.

Concerning the effect of prior expectations on scalar processing, during the determiner window, the main effect of *Target size* together with significant interaction of *Determiner* and *Target size* reflect prior expectations found in Experiment 1(a). We found overall more target looks in *all* and *some* when the target was a big set compared to when it was a small set. And this difference cannot be explained away simply by the big set bias as overall anticipatory looks were not different between big and small sets in number conditions. During the modifier window, we see a difference emerge between the bias in big *all* vs. big *some* – reflecting the perhaps the stronger association, found in Experiment 1(a), between the larger set (of 3 objects) and *all* than between the larger set and *some.* Correspondingly, in this window, we see a stronger bias forming for small targets in *some* trials than *all*.

Interestingly, in both time windows, we found almost no difference in overall bias to the target between *three* and big *all*, which is comparable to Huang & Snedeker (2009)'s results. While this result was unexpected given the wider region of inspection in establishing *all* referent compared to *three* referent, it is possible that when the target size is big, the advantages that *all* trials accrue in virtue of prior expectations compensate for the disadvantages, relative to *number* trials, from the verification processes.

We noticed that for both quantifiers, the target bias did not increase rapidly even by the end of the modifier window. To determine whether the target referent was identified before the noun onset, we preformed one sample t-test to compare target proportions to chance (50%) over the modifier window. Results showed that the target proportion for number conditions was significantly above chance ($t_1(35)=4.70$, $p<.001$; $t_2(35)=6.66$, $p<.001$), whereas bias in neither *all* nor *some* conditions was significantly above chance. Our conjecture is that slower target identification may due to the

complexity of the visual display and the instruction format. We address this issue further in Experiment 3, below.

5.3.2.2.2 Analyses of visual search to the residue set

In the second set of analyses, we re-examine the timecourse question by comparing looks to the residue set after hearing quantifiers and numerical determiners. We analysed visual search to the residue set in two critical time windows: the determiner window and the modifier window. For visualization and statistical analyses, we calculated the empirical logit as the log-odds of looks to the residue over other looks with a constant value $\ln(\frac{\text{residue looks}+0.5}{\text{total looks}-\text{residue looks}+0.5})$.

Figure **28** plots empirical logits over each 17ms sample for the average length of each time window. The curves in

Figure **28** were resynchronized at each time window onset.

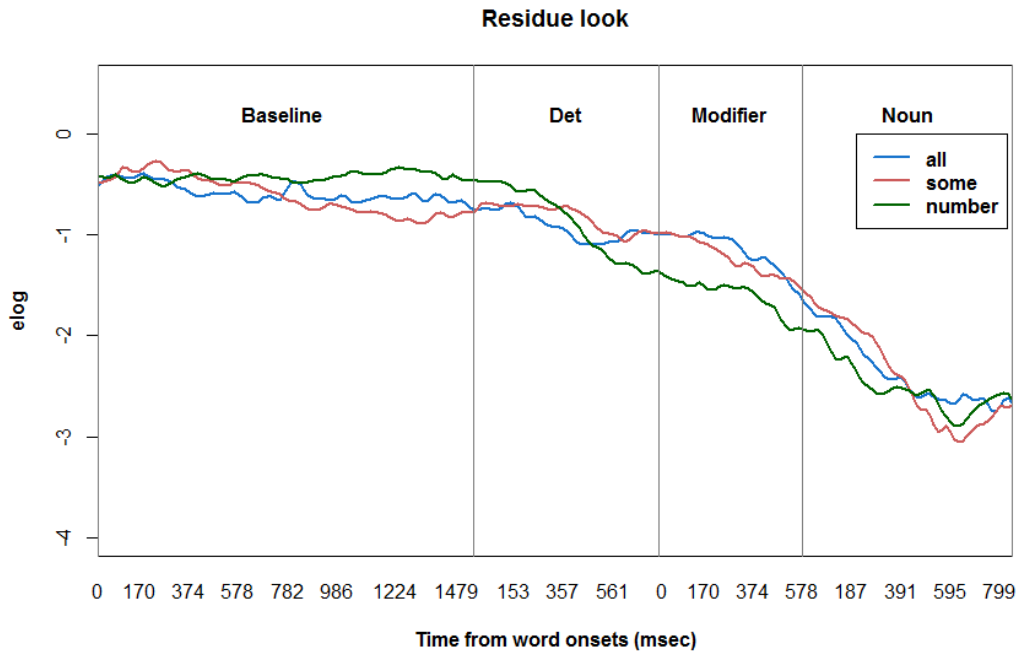


Figure 28 Bias to residue set (empirical logits) by Determiner from the instruction onset to the instruction offset in Experiment 2(a)

The eye movement data was aggregated over 50ms time bin and the empirical logit was calculated for each 50ms bin as dependent measure. For each time window, we fitted a mixed effect model to predict empirical logits from fixed effects of

Determiner (All, Some, Number), *Target size* (Big, Small), a continuous variable *Time* and their interactions. The model contained maximal random effects structure, which included random intercepts for subjects and items and uncorrelated random slopes. Fixed effects were coded in the same way as in the first analyses. Significant main effects and interactions were followed up also in the way outlined in the first analyses.

For the determiner window, we found a significant main effect of *Time* ($\chi2(1) = 5.34$, p=.02), indicating that overall looks to the residue set decreased over time. We also found a significant interaction of *Determiners* and *Time* ($\chi2(2) = 7.96$, p=.019). Planned comparisons revealed that, as shown in

Figure **28**, looks to the residue set decreased faster in numbers than both *all* (b=0.44, SE=0.19, p=.02) and *some* (b=0.66, SE=0.21, p=.003), between *all* and *some*, we found no difference in looking patterns (b=0.13, SE=0.24, p=.57). Other main effects and interactions were not significant in the determiner window. For the modifier window, the main effect of *Time* continued ($\chi2(1) = 16.07$, p<.001). We also found a significant main effect of determiner ($\chi2(2) = 6.26$, p=.04), overall looks to the residue set were significantly less in numbers than in *all* (b=0.25, SE=.09, p=.01) and *some* (b=0.21, SE=.10, p=.04). Between *all* and *some*, we found no difference in overall looks to the residue set (b=-0.02, SE=0.09, p=.85). Other main effects and interactions were not significant.

Therefore, we found in the determiner window, looks to the residue set decreased rapidly in *numbers* compared to *all* and *some*. This trend resulted in a clear difference in overall looks in the modifier window. That is, there were significantly less looks to the residue set in *numbers* trials than in both *all* and *some* trials. The lower bias to the residue set in *numbers* relative to *all* was predicted by the difference in verification processes. With regard to the verification process, we also predicted that if accessing pragmatically enriched *some* is rapid, rate of bias formation to the residue set should differ between the *number* and *some* condition, but not for unenriched *some*. We found lower bias to the residue set in *numbers* relative to *some*, suggesting that accessing the pragmatic interpretation of *some* is occurring in the determiner time window, in the same timecourse as accessing the semantic interpretation of *all*. In addition, in both time windows, we found no effect of *Target size*, which suggested that

looks to the residue set were less influenced by prior expectations of set sizes and mainly driven by the online interpretation of quantificational determiners.

5.3.2.2.3 Growth Curve Analysis

The residue set results provide evidence for the fast-pragmatic account of *some*, which predicts that visual bias to the residue set should be greater for both *some* and *all* trials, compared to numbers. However, to explore further whether the results disconfirm the slow-pragmatic account, we need to consider what that account entails for *some* trials. If the participant only accesses the weaker 'some or all' meaning of *some* on encountering the determiner, then the referential expression remains ambiguous up until the shape noun onset. This means that, other than the stimulus *girl*, which occurs in the baseline period, the linguistic input during determiner and modifier windows would not drive participants' gaze to any region in particular. By contrast, because mention of the number term (*two*/*three*) would drive looks to the relevant target as early as the determiner window, this should result in a decline in looks to the residue set. As mentioned, *all* trials should result in shifts in gaze to the residue set. So the alternative unenriched-*some* account would predict a pattern of fixations to the residue set for *some* trials which is somewhere between those in *number* trials and *all* trials. The results of the above analysis do not conclusively rule out this hypothesis, since we may have simply failed to find a difference between *some* and *all*.

In order to determine whether the differences in bias between *some* and *all* vs. *number* conditions are a result of shifts toward the residue set in the former case, we conducted a growth-curve analysis. The growth-curve analysis was conducted in the determiner window using the method described in Mirman, Dixon, & Magnuson (2008). We fitted a model to predict empirical logits from fixed effects of *Determiner* (All, Some, Number), *Time* and their interactions. The interaction of *Determiner* and *Time* was treated as nested within each level of Determiner. Time was represented using a 2nd-order orthogonal polynomial (Time1, Time2). Time1 is the linear representation of Time and Time2 is the quadratic representation of Time. According to Mirman, Dixon, and Magnuson (2008), the coefficient of Time 1 reflects a single change of focus, i.e. from a neutral start to target, whereas the coefficient of Time2 reflects two changes in focus, i.e. from neutral start to competitor, from competitor to target. In our case, a positive

coefficient of Time1 indicates an increase in looks to the residue set over time and a negative coefficient of Time1 indicates a decrease in looks to the residue set. The positive coefficient of Time2 indicates the curve is convex, that there are looks to other areas followed by shifts in looks to the residue set. Whereas the negative coefficient of Time2 indicates the curve is concave, that there are looks to the residue set followed by shifts in looks to other areas. The predictions for *all* and pragmatically enriched-*some* are that participants will initially shift away from the residue set, due to an association between set size and determiner and this will be followed by shifts toward the residue set. In other words, we should see a positive Time2 term. The unenriched-*some* account predicts that there should be no positive Time1 or Time2 term. By contrast, on both accounts, *number* items should see a negative coefficient for Time1.

The model contained maximal random effects structure supported by the data, which included random intercepts for subjects and items and uncorrelated random slope. All fixed effects were included as random slopes, interactions were not included as random slope because the model did not converge.

We found a significant interaction of *number* and Time 1 (b=-0.79, SE=0.15, p<.001), indicating looks to the residue set decreased significantly in *numbers* over time. Neither the interaction of *all* and Time1 nor the interaction of *some* and Time1 was significant (all: b=-0.08, SE=0.16, p=.63; some: b=-0.28, SE=0.15, p=.07). We found a significant interaction of *all* and Time2 (b=0.64, SE=0.10, p<.001), a significant interaction of *some* and Time2 (b=0.23, SE=0.09, p=.01), and a significant interaction of *number* and Time2 (b=0.48, SE=0.09, p<.001), all with a positive coefficient, indicating that for each condition, there was an upward curving quadratic component in looks to the residue set over time. That is to say, after hearing the determiner, participants first looked towards other viewing region and then shifted their looks to the residue set.

To summarize, the growth-curve analysis revealed that in the determiner window, as time increased, in the linear term, looks to the residue set decreased significantly in *numbers* but not in *some* and *all*; in the quadratic term, looking pattern of each condition contained two changes in focus (looked away from the residue set and looked towards the residue set). These results support the *enriched*-some account and provide negative evidence for the unenriched-*some* account.

However, we also find that, in spite of there being an overall bias away from the residue set in *number* conditions, we also found some 'return' to the residue set in this period. This is not predicted on any account of the on-line interpretation of these items. We suspect that shifts to the residue set in the *number* condition were due to the use of subitizable sets. When the number of objects in a set is within subitizable range (1-3), the quantity of the set is rapidly available and salient to participants through pre-attentive visual recognition processes (Dehaene, 1997). Thus, in number conditions, after identifying two referents that had the correct amount of objects, participants might have time to fixate any viewing region before the modifier onset. The shifts toward the residue set might reflect such 'noise' event. If our suspicion is correct, the 'return' effect in number trials is less likely to occur in the next experiment, where the targets may be less straightforward to recognise and distinguish, since one of the numbers (4) is on the outer edge of the subitizable region while the pair (3/4) involve a smaller difference (larger Weber fraction) and are less discriminable than 2/3.

5.3.2.3 Discussion

In Experiment 1(a), we explored the idea that participants associate determiners like, *all* and *some* with certain set sizes. Our results suggested that when participants are shown agents with either a set of three objects or two, there is an expectation that the agent with three objects is the agent with *all* and there is also an expectation that the agent with three objects is the agent with *some*, but that the association between *all* and the set of three is greater than the association with *some*. If these results tap into underlying expectations about the relative set size of an agent with *all* and *some*, then the prediction was that in a visual world experiment similar to Huang & Snedeker (2009), we would see evidence for a stronger anticipatory target bias in the *all* trials when the agent has three objects (the big set) than when she has two objects (the small target). We also predicted a similar, though not as marked pattern for *some* trials. The results of Experiment 2(a) bear these predictions out. In particular, in the earlier determiner time window, for both *all* and *some* we found a greater target bias emerge during big-set trials, compared to small; in the latter modifier window, we found only an advantage in big-set trials for *all*. We also found that target bias was greater in *all* trials than *some* when target set was large. In particular, in the modifier window, the target bias was

greater and increased faster in big-*all* trials than in big-*some* trials. This result can be attributed to the greater strength of association between *all* and big sets than *some* and big sets. In the modifier window, we also found the reverse pattern occurred when targets were small: bias formed faster in *some* trials than *all*. A big-set association for a given determiner means that in small-set trials, bias formation should be penalised as the target cannot be identified by reliance on a simple association between determiner and set size. Thus a greater big-set association for *all* means a greater small-set penalty for *all* – leading to an advantage in *some* trials when target set is small.

In contrast to target bias formation, bias to the residue set was predicted to be unaffected by set size, and this is what we found. Residue set search should only be driven by accessing the meaning of the determiner and was predicted to be different between *all* and *number* conditions, given the lexical semantics involved. The question of interest was whether the rate of visual search of the residue set changes after determiner onset for *some* compared to *number* trials. Consistent with the fast-pragmatic view we found that during the earlier Determiner time window, the rate of residue set bias differed from numbers both for *all* and *some* conditions. The growth-curve analysis added further support to the fast-pragmatic view, revealing an initial shift away from the residue set in both *some* and *all* trials, followed by a return. These results are unexplained by the slow-pragmatic account. A similar pattern in the numbers condition however was unexpected. We attribute this to the relative set-sizes and will explore this effect in the next experiment.

In Experiment 2b, we used the same design and procedure as for Experiment 2a, except that we used the proportions of objects from Experiment 1b. According to our hypothesis, we expect to no longer see a big set bias for *some* trials but to replicate our results for big sets in the *all* condition.

5.3.3 Experiment 2(b)

5.3.3.1 Methods

5.3.3.1.1 Participants

36 participants were recruited from our university campus via an online psychological subject pool. All participants speak English as a native language. They have uncorrected or corrected to normal vision.

5.3.3.1.2 Procedure and materials

The materials were the same as Experiment 2a except that in Experiment 2b we changed the target set sizes. In experimental displays, the larger sets contained four objects and the smaller sets contained three objects. As in Experiment 2a, three lists were created. Each list contained 36 experimental items, 12 items per determiner. Accordingly, determiners used in the number condition were changed to 'three' and 'four'. Again, there were 18 fillers, of which 12 were number fillers. The determiner used in the filler item was one of *none*, *two* and *one*. The audio instructions were cross-spliced and adjusted. The average length of the instruction was 4.1s. The average duration for determiner window was 718ms (all of the: 703ms, some of the: 725ms, four of the: 720ms, three of the: 729ms), the average duration for modifier window was 632ms (stripy: 638ms, dotted: 638ms, checked: 622ms). The procedure was identical to Experiment 2a except one difference. That is, given the complexity of the display, participants were given more time to respond in each trial. It was set to jump to the next trial 6 seconds after the instruction onset.

5.3.3.2 Data analyses and Results

We excluded trials in which the response was missing (3.4%), and trials which participants clicked on the wrong target (2%). We also excluded one participant whose accuracy rate is lower than 2 standard deviations away from the mean accuracy. We again first analysed the timecourse of target identification for quantifiers and numerical determiners in each critical time window. In particular, we examined whether prior expectations found in Experiment 1b influenced bias formations in *all* and *some*.

5.3.3.2.1 Analyses of target anticipatory eye movement

As in Experiment 2a, we defined two critical time windows: the determiner window ([Det] onset-'the' offset, e.g. during 'some of the') and the modifier window (modifier onset-modifier offset, e.g. during 'stripy'). The regions of interests in each window were also defined in the same way as in Experiment 2a. The eye movement data were aggregated over 50ms time bin and natural log ratio of percentage of looks to the target over competitor was calculated for each 50ms bin as dependent measure. For each time window, we fitted a mixed effect model to predict log ratios from fixed effects of *Determiner* (All, Some, Number), *Target size* (Big, Small), a continuous variable *Time* and their interactions. The model contained maximal random effects structure, which included random intercepts for subjects and items and uncorrelated random slopes containing all fixed effects and their interactions.

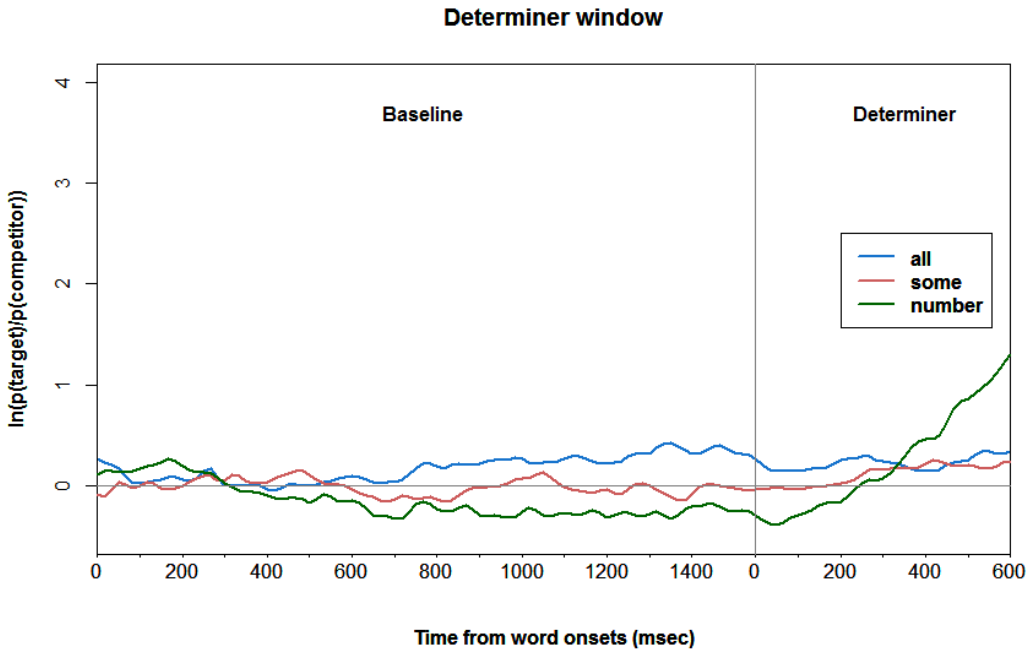


Figure 29 Log ratios of percentage of looks to target over competitor by Determiner from the instruction onset to the determiner window offset (e.g. 'Click on the girl with some of the') in Experiment2 (b)

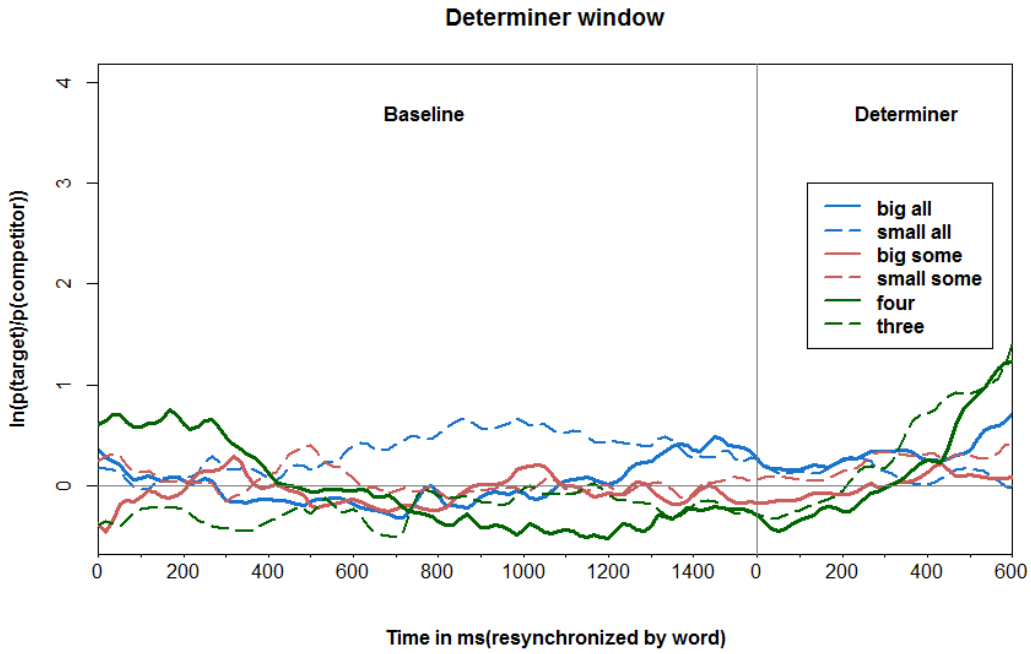


Figure 30 Log ratios of percentage of looks to target over competitor by Determiner and Target size from the instruction onset to the determiner window offset in Experiment 2(b)

In the determiner window, we again found a significant main effect of *Time* (b=1.37, SE=0.29, p<.001), indicating that biases toward target increased over time. While there was no main effect of *Determiners* in this window, we found a significant interaction of *Determiners* and *Time* ($\chi 2(2) = 24.33$, p<.001). As shown in

Figure **29**, the target bias increased more rapidly in numbers than both *all* and *some*, (all: b=-2.89, SE=0.63, p<.001; some: b=-2.97, SE=0.67, p<.001). Between *all* and *some*, target biases did not increase at different rates (b=.04, SE=0.61, p=.95).

Figure **30** depicts how target bias developed over time by *Determiner* and *Target size* from the instruction onset to the determiner window offset. In contrast to Experiment 2a, we found no main effect of *Target size* in the determiner window. Also, unlike Experiment 2a, neither interaction of *Determiner* and *Target size*, nor three-way interaction of *Determiner, Target size* and *Time* was significant in the determiner time

window.

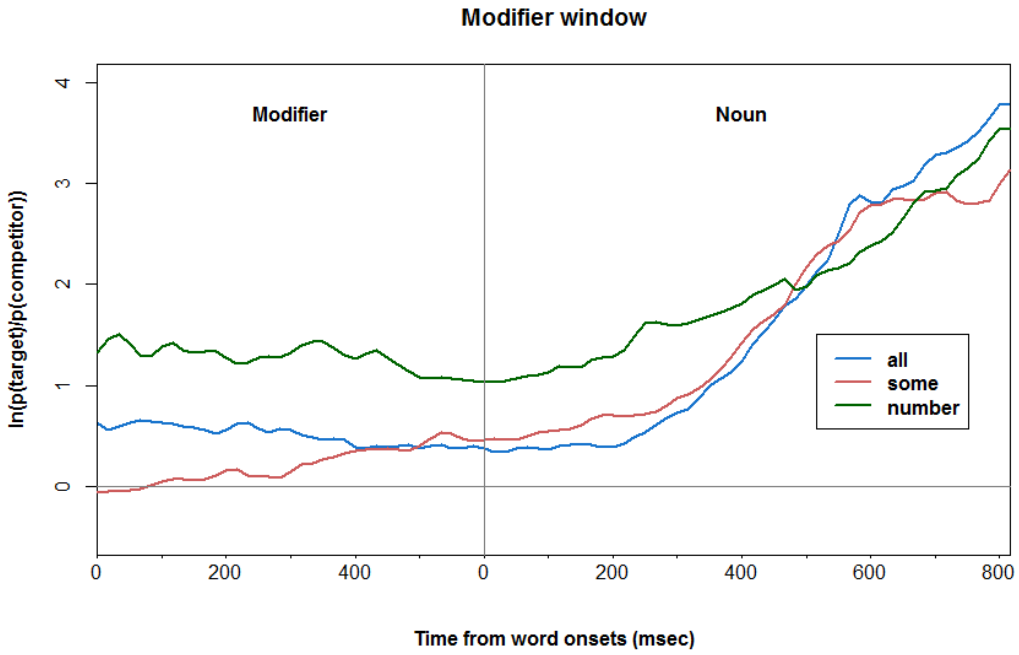


Figure 31 *Log ratios of percentage of looks to target over competitor by Determiner from the modifier onset to the instruction offset in Experiment 2(b)*

For the modifier window, we found a significant main effect of Determiner ($\chi2(2) = 12.91$, p=.002). As shown in Figure 31, the target bias was greater for *numbers* than for both quantifiers (all: b=-0.66, SE=0.22, p=.005; some: b=-0.64, SE=0.19, p=.002), and the overall target bias was not different between *all* and *some* (b=-0.07, SE=0.21, p=0.75). There was a marginal main effect of *Time* ($\chi2(1) = 3.66$, p=.056), like Experiment 2a, target biases did not increase rapidly in this window and we found no significant

interaction of *Determiners* and *Time*.

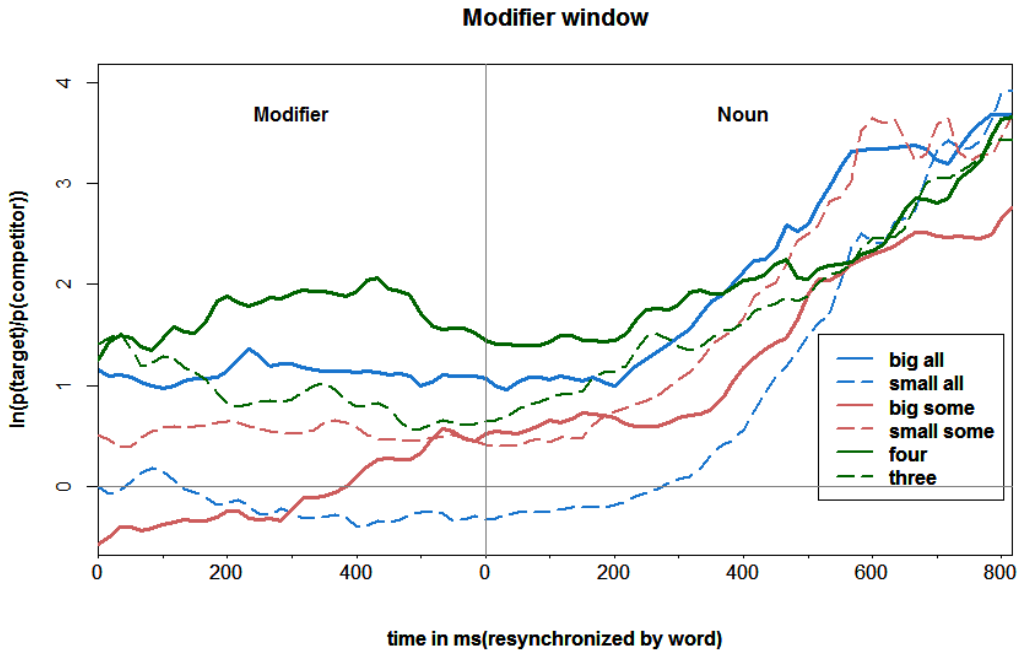


Figure 32 *Log ratios of percentage of looks to target over competitor by determiner and target sizes from the modifier onset to the instruction offset in Experiment 2(b)*

During the modifier window, we found a significant main effect of *Target size* ($\chi2(1) = 4.34$, p=.04) and a significant interaction of *Target size* and *Time* ($\chi2(1) = 4.76$, p=.03). More importantly, we found a significant interaction of *Determiner and Target size* ($\chi2(2) = 9.12$, p=.01). Figure 32 depicts how target bias developed over time by *Determiner* and *Target size* during the modifier window. Planned comparisons within the level of *Determiner* showed that for *all*, the target bias was greater when the target was a big set compared to when it was a small set (b=-1.23, SE=0.28, p<.001), but for *some* and *numbers*, the simple main effect of *Target size* was not significant (some: b=0.41, SE=0.26m p=.12; numbers: b=0.53, SE=0.27, p=.05). Planned contrast within the level of *Target size* showed that when the target was a big set, the target bias was greater in *all* than in some (b=0.93, SE=0.29, p=.002). The target bias was greater in *number* (i.e. four) than in *some* (b=1.18, SE=0.27, p<.001), but there was no difference in the target bias between *all* and *number* (b=0.29, SE=0.29, p=.33). However, when the target was a small set, the target bias was greater in *some* than in *all* (b=0.66, SE=0.28, p=.02). The target bias was greater in *number* (i.e. three) than in *all* (b=0.89, SE=0.32,

p=.008), and there was no difference between *some* and *number* (b=0.25, SE=0.27, p=.36). Other main effects and interactions were not significant.

So in Experiment 2b, we replicated the findings from Experiment 2a that in both time windows, target bias emerged earlier in number conditions than in *all* and *some*, whereas between *all* and *some*, bias formation did not differ. However, we note that the target bias in the number condition emerged later in Experiment 2(b) compared to Experiment 2(a). This may be due to the increasing number of objects in the visual display, and the change in the Weber ratio, which made specific quantities more difficult to discriminate in this experiment.

Regarding prior expectations, we found the predicted looking pattern for *all* trials: there was a stronger target bias when the target was a big set compared to when it was a small set. Recall that in the off-line study, Experiment 1b, in which participants chose between a character with four objects and one with three, there was only a bias to the larger set in the *all* condition, not the *some* condition. This confirms that prior expectations facilitate target identification during online processing. Correspondingly, we also found that when the target was a small set, there was a stronger target bias after hearing *some* than after hearing *all*.

We again examined whether the target referent was identified before the noun onset, we performed one sample t-test to compare target proportions to chance (50%) over the modifier window. Results showed that the target proportion for number and *all* conditions was significantly above chance (number: $t_1(34)=7.38$, $p<.001$; $t_2(35)=7.35$, $p<.001$; all: $t_1(34)= 3.10$, $p=.004$; $t_2(35)=2.39$, $p=.02$), whereas *some* condition was not significantly above chance. This raises an issue which is addressed in Experiment 3.

5.3.3.2.2 Analyses of visual search to the residue set

As we did for Experiment 2(a), we then examined looks to the residue set in each time window by fitting a mixed effects model to predict empirical logits from fixed effects of *Determiner* (All, Some, Number), *Target size* (Big, Small), a continuous *Time* variable and their interactions. The model contained maximal random effects structure, which

included random intercepts for subjects and items and uncorrelated random slopes.

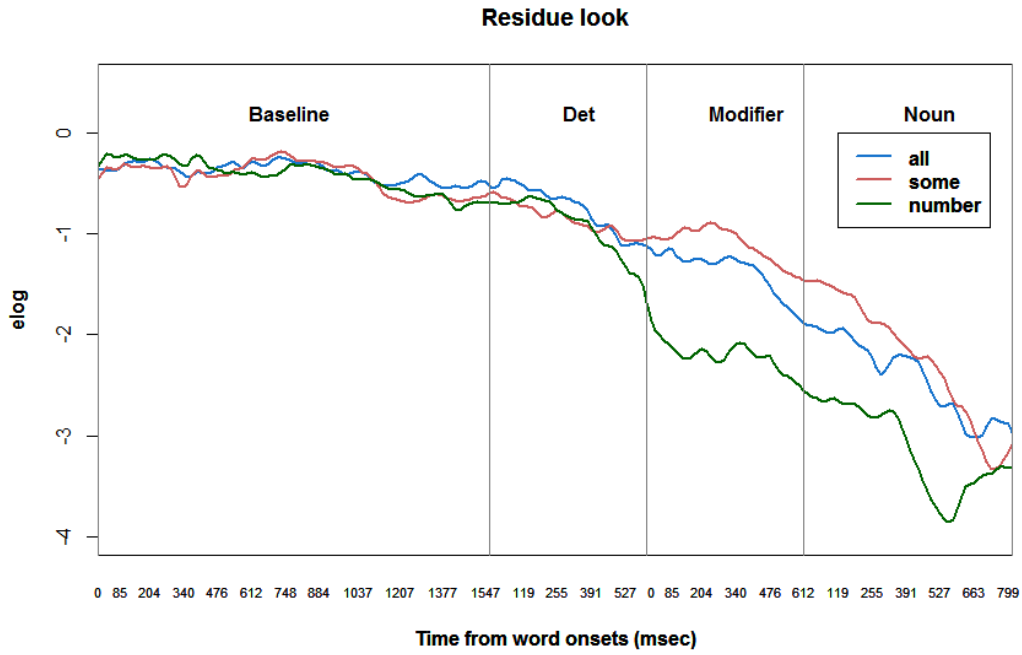


Figure 33 *Bias to residue set (empirical logits) by Determiner from the instruction onset to the instruction offset in Experiment 2(b)*

For the determiner window, we found a significant main effect of *Time* ($\chi2(1)$ = 21.18, p<.001) and a significant interaction of *Determiners* and *Time* ($\chi2(2)$ =6.18, p=.046). As shown in Figure 33, looks to the residue set decreased faster in numbers than *some* (b=0.62, SE=0.20, p=.003) and there was a similar trend for *all* (b=0.33, SE=0.19, p=.09). We found no difference in looking pattern between *all* and *some* (b=0.28, SE=0.21, p=.18). Other main effects and interactions were not significant in the determiner window. For the modifier window, the main effect of *Time* ($\chi2(1)$ = 8.41, p=.004) and the interaction of *Determiners* and *Time* ($\chi2(2)$ =6.34, p=.04) continued. We also found a significant main effect of determiners ($\chi2(2)$ =28.48, p<.001), indicating overall there were more looks to the residue set in *all* and *some* compared to numbers (all: b= 0.33, SE=.07, p<.001; some: b=0.42, SE=.09, p<.001). Again there was no difference in looking pattern between *all* and *some* (b=0.12, SE=0.08, p=.14). Other main effects and interactions were not significant. These results showed that even when target set sizes were increased, Experiment 2b was able to replicate the findings from Experiment 2a. This provides further evidence that compared with anticipatory looks to

the intended referent, visual search to the residue set is relatively unaffected by prior expectations and it is a better indicator to explore the timecourse of scalar processing.

5.3.3.2.3 Growth Curve Analysis

Once again a growth-curve analysis was conducted in the determiner window. We fitted a model to predict empirical logits from fixed effects of *Determiner* (All, Some, Number), *Time* and their interactions. The interaction of *Determiner* and *Time* was treated as nested within each level of Determiner. Time was represented using a 2nd-order orthogonal polynomial (Time1, Time2). The model contained maximal random effects structure supported by the data, which included random intercepts for subjects and items and uncorrelated random slope. All fixed effects were included as random slopes, interactions were not included as random slop because the model did not converge.

We found a significant interaction of *number* and Time1 (b=-1.01, SE=0.14, p<.001), a significant interaction of *all* and Time1 (b=-0.62, SE=0.14, p<.001) and a significant interaction of *some* and Time1 (b=-0.38, SE=0.14, p=.009), indicating looks to the residue set decreased significantly in all three conditions over time. More important, we found a significant interaction of *all* and Time2 (b=0.30, SE=0.10, p=.005) and a significant interaction of *some* and Time2 (b=0.31, SE=0.10, p=.003), both with a positive coefficient, indicating that for both conditions, there was an upward curving quadratic component in looks to the residue set over time. Thus, after hearing quantifiers, *all* and *some*, participants first looked towards other viewing region and then shifted their gaze to the residue set. By contrast, no significant interaction of *number* and Time2 was found (b=-0.05, SE=0.10, p=.65), indicating that there was no quadratic component in changes of looks to the residue set over time. This result is in contrast to the finding of Experiment 2(a) that shifts in gaze to the residue set were found in *number* condition as well.

We suggest two reasons for the difference in results between Experiments 2a and 2b. First, compared to Experiment 2a, the duration of the determiner window happens to have been shorter in Experiment 2b. In particular, in Experiment 2b, the duration for 'four of the' and 'three of the' was 720ms and 729ms respectively. By contrast, in Experiment 2a, the duration for 'two of the' and 'three of the' was 784ms and 773ms respectively. Second, as we increased the number of objects in each set and

the corresponding Weber fraction, the task of identifying the numerical quantities in Experiment 2b is more complex than in Experiment 2a. This assumption is borne out by the fact that bias to target in number conditions did not become significantly greater than quantifiers in this time window – in contrast to the result in Experiment 2a. Thus we would expect less 'return' away from the target to other areas in number condition in Experiment 2b. And this seems to be what we found.

In spite of factors leading to a lower likelihood of 'return' effects, we did find significant positive quadratic components in *some* and *all* conditions. We thus feel confident in attributing these effects to participants searching the residue set to determine which girl has *some and not all* and which girl has *all* of the relevant targets. Thus, taken together, these residue set analyses tend to disconfirm the slow-pragmatic account, suggesting that the enriched, 'some and not all' understanding of *some* is accessed in the same timecourse as the literal understanding of *all*.

5.4 Summary of Experiments 1 and 2

Previous visual-world investigations that compare the timecourse of pragmatically enriched *some*'s interpretation to that of *all* have had mixed results. To control for the possibility of pre-coding, our experiment included number items and our discussion here focuses on previous studies that also included number items. Huang & Snedeker (2009, 2011) find a clear delay in terms of visual bias in *some* vs. *all* conditions, and no delay between *all* and number conditions. Degen & Tanenhaus (2016) found a delay for *some* compared to *all* when each quantifier was paired with a larger set (of 4 or 5 objects), but they found no delay between *some* and *all* when each quantifier was used to refer to a smaller set (of 2 or 3 objects). To date, no account has been given for all of these facts. We proposed that the mixed results found in Huang & Snedeker (2009, 2011) and Degen & Tanenhaus (2016) could be partly explained by prior expectations about set size and quantifier. Our hypothesis was that participants have expectations about the relative size of sets used to identify a girl with all of something, and a girl with some of something. Results from Experiment 1 support our assumption about *all* trials: that participants expect a girl with a larger set to be the target. For the *some* cases, we found an interesting combination of results. When the relative set sizes are 2 and 3, participants

favoured the girl with the larger set (which is the same preference as for *all*), when the relative set sizes were 3 and 4, there was no overall preference. In both experiments 1a, and 1b, we found a greater preference for the larger set for the stimulus in the *all* case than *some.*

In Experiments 2a and 2b, we established that the preferences shown off-line had an impact on on-line target-gaze formation. When the ratio of set sizes is 2/3 (Experiment 2a), during the earlier determiner time window, we found that bias to target was greater when the character had the big set in both *all* and *some* trials. In the later modifier time window, we found a continued advantage for big-set over small-set trials for *all,* but not for *some.* When the ratio of set sizes is 3/4 (Experiment 2b), we found only in the modifier window, there was a greater target bias for big-set *all* trials than for small-set *all* trials. There was no advantage for big-set *some* over small-set *some* in either window. In both Experiment 2a and 2b, during the modifier window, there was an advantage for *big-all* over *big-some* conditions and a corresponding advantage for *small-some* over *small-all* conditions. Note that the latter inversion of results in the small set condition is expected if prior associations concerning set size are the only thing that makes a difference in bias between the *all* and *some* condition.

In the course of experiments 2a and 2b, we broadly replicated effects found in previous visual word studies that also include number items. Specifically, in Experiment 2a, we showed that bias in big-all did not differ from *big-number* (*three*) conditions in the determiner window. This is basically the pattern of results reported in Huang & Snedeker (2009, 2011). Degen & Tanenhaus (2016) used a mixture of trials where the ratio of set sizes was 2/4 or 3/5 and reported a difference between big-all and big-some conditions, but no difference between small-all and small-some. The difference between big-*all* and big-*some* was predicted by a stronger prior association between the larger set and *all*. We replicated this result in Experiment 2b where the effect of anti-presuppositions was well controlled (i.e. the small target was not a set of two).

In Experiment 2, we fully counterbalanced set size and found the predicted results that bias in number trials overall would be greater than *all*. Identifying the target in *all* trials involves verifying a clause containing *all* and this requires a wider visual search than verifying the relative clause containing a number. We attribute the greater

bias and lack of difference between *all* trials and *three* trials in Huang & Snedeker (2009) to the 'boost' obtained in *all* trials by prior expectations.

In contrast to previous visual-world research on the timecourse of pragmatically enriched *some*, we designed our items to test for a visual reflex of the compositional understanding of the quantifiers. This involved a visual region of interest where the residues of the partitioned sets of objects are located. For both *all* trials and enriched *some*, but not for number trials or unenriched *some*, we predicted that participants should also search this residue area to verify the relative clause with respect to given target regions, to the extent that they relied on the compositional meaning of the linguistic input, rather than prior expectations about set size and quantifier. In both Experiments 2a and 2b, we found the predicted difference in looks to the residue set between *all* trials and number trials. We also found that residue set search was equivalently greater in *some* trials compared to number trials. A growth curve analysis in both Experiments 2a and 2b revealed a pattern of shifting from outside the residue region in *some* and *all* trials. This pattern provides disconfirming evidence for an alternative slow-pragmatic account of the gaze data.

There are issues with the results from Experiment 2, however. In particular, although big-set *all* bias was comparable to the corresponding numeral trial bias, overall, bias to target in *all* and *some* conditions did not rise significantly above chance prior to the disambiguating shape-noun onset in Experiment 2a. Similarly, in Experiment 2b, although we found a significant bias to target in *all* trials, we failed to find the effect in *some* trials. We attribute the absence of these effects to a number of mitigating factors: Compared to previous visual world research on quantifiers embedded in referential phrases, our items had more complex stimuli. The relative clause involved composing the meanings of a quantifier and modifier with the noun, rather than simply quantifier and noun. There were four identical agents in the display, two of which are still potential targets during the determiner window. Since the positioning of these two targets was randomly allocated across trials, participants who attempt to visually anticipate the target would be required to search around the whole display more during the instruction. Given the complexity of the items and the difficulty in anticipation in the determiner region, it may be that some participants were discouraged from attempting to compose

determiners, *some* and *all* with the modifier prior to the noun and rather opted to exploit the modifier-noun composition – which completely disambiguates the referring expressions. To address this issue, Experiment 3 uses less complex stimuli; it is designed so that there is only one target from the onset of the determiner region; it makes the modifier non-informative (a genitive phrase that applies to all images in the display) so that the determiners become the focus of any anticipation by participants. In addition, in order to encourage anticipation and discourage participants waiting to hear the disambiguating noun, we allowed participants to control the pace of experiment such that at any point of an ongoing trial, as long as participants clicked on the image, the next trial began.

5.5 EXPERIMENT 3

Experiment 3 was set up to re-examine the timecourse question in a visual world paradigm that address the issue of slower target identification in general in Experiment 2. Target identification in Experiment 3 was simpler in the sense that the region of inspection needed to anticipate the referent was narrower compared to that in Experiment 2. We predicted that, target identification should be faster in *numbers* compared to both *all* and *some.* Also, independent of set sizes, the timecourse of target identification based on enriched *some* should not be different to that for *all*. We also predicted that, looks to the residue set should reflect scalar processing, such that there should be lower bias to the residue set in *numbers* compared to *some* condition; also a contrast in shift to residue set target from outside this region between *some* and number trials.

5.5.1 Method

5.5.1.1 Participants

36 participants were recruited from our university campus via an online psychological subject pool. All participants speak English as a native language. They have uncorrected or corrected to normal vision.

5.5.1.2 Procedure

There were six practice trials in the beginning. Each practice trial began with a display depicting a character in the centre of the display who was about to distribute four sets of objects to two boys and two girls. Participants heard a background story describing the situation, for example, "This is Susan. She gives out fruits to children every day. Here is what she has on Monday. She has apples, pears, bananas, and oranges. She always brings more than enough. The leftover fruits are put in the middle". On the next display, the objects were distributed to boys and girls with the residue set in the centre. 1 second after the display onset, participants were given an auditory instruction, for example, "Click on the girl that has some of Susan's apples". Participants' task was to click on the image according to the instruction. The experimental script was set to jump to the next trial after participants clicked on the image. These six practice trials familiarised participants with the three characters (Susan, Amy, Michael) to be used in the experiment and the types of object each character brings (fruits, stationary, kitchenware respectively).

After the practice session, we ensured that that participants understood the story, instruction, display and procedure. Then the experiment began. The procedure was identical to the practice session except for one difference: On each trial, background stories and the starting display were not presented again. Participants were presented with the experimental display directly (see **Figure 34**). There were 48 trials, divided into 36 critical trials and 12 fillers. Randomized order of presentation of the items was created for each participant. The experiment was conducted using E-Prime software and a Tobii TX300 eye-tracker. Fixations were sampled every 17ms. Calibrations were performed in the same way as in Experiment 2. For each trial, eye movements were recorded from the onset of the display to the point when the click occurred. The whole experiment lasted approximately 15 minutes.

5.5.1.3 Materials

The experiment employed three by two within-subject design. The two independent variables were *Determiner* (All, Some, Number) and *Target size* (Big, Small), which generated six experimental conditions: big *all*, small *all*, big *some*, small *some*, big *number* (i.e. three), small *number* (i.e. two). The auditory instructions were of the form

"Click on the [gender] that has [Det] of [name's] [object]". [gender] was either *boy* or *girl*, [Det] was one of *some*, *all*, *two*, *three*, [name's] was one of *Susan's*, *Amy's*, *Michael's*. 36 experimental displays were constructed and paired with an audio instruction containing one of the determiners (e.g. 'Click on the girl that has some of Susan's apples'). The experimental display contained four agents, two boys and two girls. As in Huang & Snedeker (2009), we arranged the display so that vertically adjacent agents were in the same gender and the horizontally adjacent characters were not. This means that participants could expect to locate boys and girls in the same locations on each trial. Four sets of objects were distributed among the agents. For two agents in the same gender, there was always one agent that had a total set of one kind of object and the other one had a proper subset. The residues of the two partitioned sets remained in the centre. In terms of set sizes, as in Experiment 2a, two agents always had a set of three objects and another two had a set of two objects. We counterbalanced the target set size for *all* and *some* by changing objects in the residue set. **Figure 34**(a) can be used on a small-set *some* or big-set *all* trial and **Figure 34**(b) can be used on a big-set *some* or small-set *all* trial.

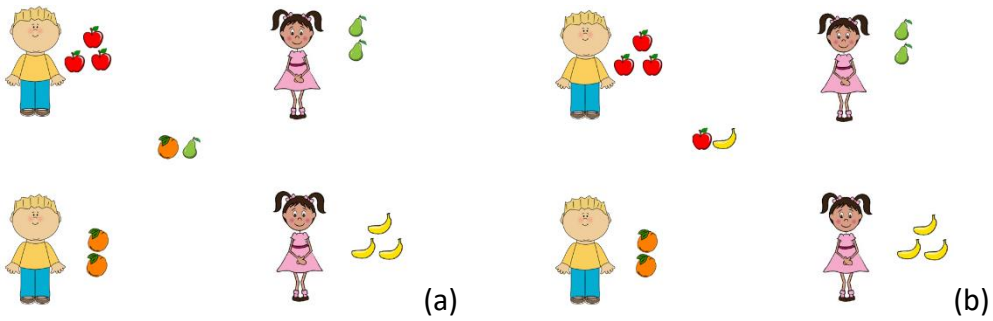


Figure 34 *Example displays in Experiment 3 (a) can be paired with instructions 'Click on the boy that has all/three of Susan's apples' or ''Click on the girl that has some/two of Susan's pears', (b) can be paired with instructions 'Click on the boy that has some/three of Susan's apples' or ''Click on the girl that has all/two of Susan's pears'.*

Again three lists were created. Each list contained 36 experimental items, 12 items per determiner. In addition, each list contained 12 fillers. Fillers were similar to experimental items but contained different determiners (One, Four) in the instruction. The audio instructions were cross-spliced and adjusted as in Experiments 2a-b. The

average duration for determiner window was 708ms (all of: 709ms, some of: 711ms, three of: 713ms, two of: 700ms), the average duration for name window was 550ms (Susan's: 551ms, Michael's: 547ms, Amy's: 552ms). Each gender (boy/girl) was referred to an equal number of times within each condition. The scenario, the object and location of the target were counterbalanced within each condition. All pictures of an agent with a set of objects measure 336*315 pixels. Picture of items in the middle measure 168*210. The screen resolution is 1680*1050 pixels.

5.5.2 Data analyses and Results

We excluded trials on which participants clicked on the wrong target (1.9%) and two participants whose accuracy rate is lower than 2 standard deviations away from the mean accuracy.

5.5.2.1 Analyses of target anticipatory eye movements

We first examined the timecourse of target identification after hearing quantifiers and numerical determiners. We defined two critical time windows: the determiner window ([Det] onset-'of' offset, e.g. during 'some of') and the name window (name onset-'s' offset, e.g. during 'Susan's'). In both time windows, the target was the agent of the description, the competitor was the agent in the same gender. As in Experiments 2a-b, the eye movement data were aggregated over 50ms time bin and log ratio was calculated for each 50ms bin as dependent measure. For each time window, we fitted a mixed effect model to predict log ratios from fixed effects of *Determiner* (All, Some, Number), *Target size* (Big, Small), a continuous variable *Time* and their interactions. The model contained maximal random effects structure, which included random intercepts for subjects and items and uncorrelated random slopes containing all fixed effects and their interactions.

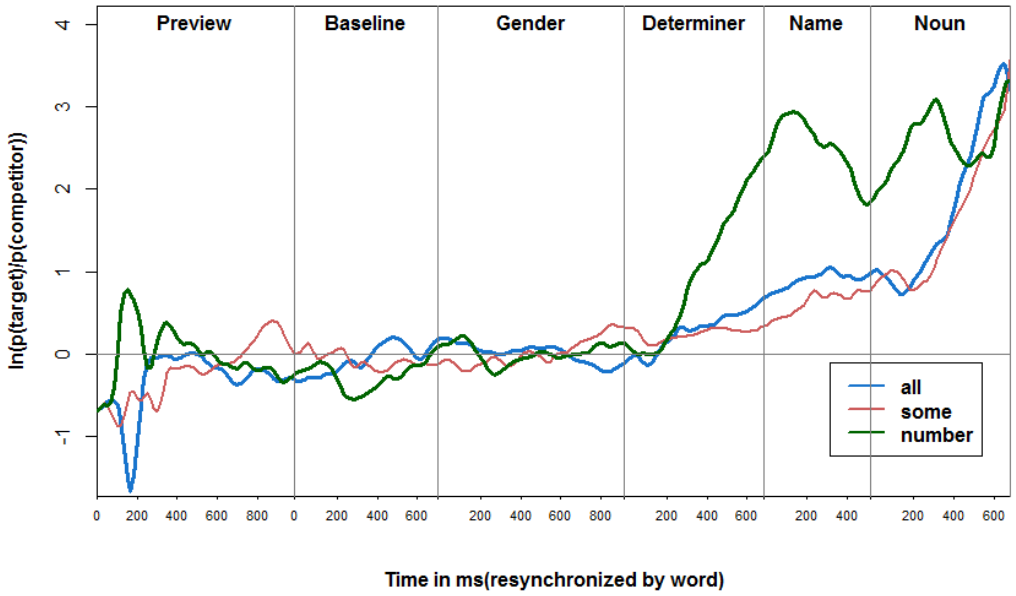


Figure 35 *Log ratios of percentage of looks to target over competitor by Determiner from the display onset to the instruction offset in Experiment 3*

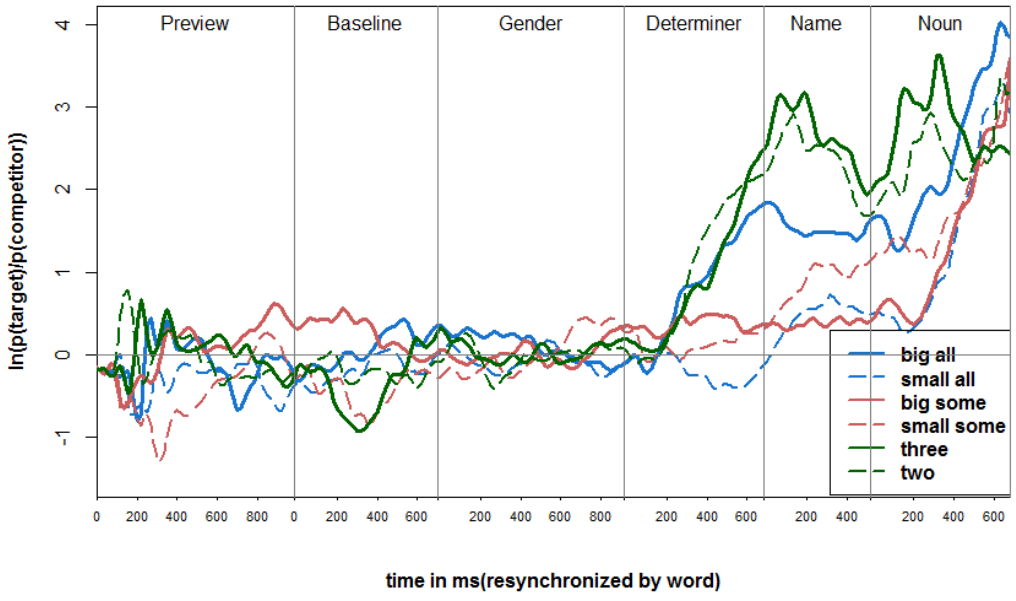


Figure 36 *Log ratios of percentage of looks to target over competitor by Determiner and Target size from the display onset to the instruction offset in Experiment 3.s*

In the determiner window, there was a main effect of *Time* ($\chi2(1) = 30.46$, $p<.001$). Critically, there was a significant main effect of *Determiners* ($\chi2(2) = 29.04$, $p<.001$) and a significant interaction of *Determiners* and *Time* ($\chi2(2) = 43.54$, $p<.001$). Planned comparison revealed that the overall target bias was greater in number condition than in both quantifier conditions (all: b=-0.89, SE=0.18, p<.001; some: b=-0.95, SE=0.19, p<.001), also bias to the target increased faster in *numbers* than in *all* and *some* (all: b= -3.70, SE=0.67, p<.001; some: b= -4.42, SE=0.74, p<.001). There was no difference in looking pattern between the *all* and *some* trials (overall looks: b=-0.02, SE=0.14, p=.87; bias changes over time: b=-0.62, SE=0.34, p=.07).

During this time window, we also found a significant main effect of *Target size* ($\chi2(1) = 8.06$, $p=.004$). Critically, there was a significant interaction of *Determiner* and *Target size* ($\chi2(2) = 14.30$, $p<.001$) and a significant three-way interaction of *Determiner, Target size* and *Time* ($\chi2(2) = 9.47$, $p=.009$). Planned comparisons on each level of *Determiner* revealed that the overall target bias in big *all* was stronger than in small *all* (b= 0.96, SE= 0.19, p<.001), but the difference in the overall target bias between big *some* and small *some* did not reach significance (b=0.30, SE=0.20, =-0.15). There was no reliable difference between big *number* and small *number* either (b=-0.12, SE=0.17, p=0.51). In terms of the changes in target bias over time, we found the target bias increased more quickly in the big *all* condition than in the small *all* condition (b= 3.20, SE=0.99, p=.002), but no such pattern was found in neither the *some* nor *number* trials (some: b=0.06, SE=0.90, p=.94; number: b=-0.39, SE=0.82, p=.63).

Comparison within the *Target size* found that when the target size was big, although the overall target bias did not differ between big *all* and big *some* (b=-0.02, SE=0.14, p=.87), we found the target bias increased faster in big *all* than in big *some* (b=-2.35, SE=.96, p=.02). In addition, we found the target bias was greater in big *number* (i.e. three) compared to both *all* (b=-0.39, SE=0.19, p=.04) and *some* (b=-0.69, SE=0.21, p=.002), also the bias increased faster in big *number* than in both quantifiers (all: b=-2.12, SE=0.96, p=.03; some: b=-4.17, SE=0.97, p<.001). When the target size was small, notably, there was a marginally significant advantage in the overall target bias in small *some* over small *all* (b=0.34, SE=0.19, p=.08), though the changes in bias over time between small *some* and small *all* did not differ (b=0.67, SE=0.85, p=.44). We also found

the target bias was greater in small *number* (i.e. two) than in both quantifiers (all: b=-1.42, SE=0.25, p<.001; some: b=-1.17, SE=0.25, p<.001), and the bias increased faster in small *number* trials as well (all: b=-5.19, SE=0.89, p<.001; some: b=-4.84, SE=0.84, p<.001).

During the name window, the predicted significant main effect of *Determiners* continued ($\chi2(2) = 94.12$, p<.001). Planned comparison revealed a stronger target bias in the number condition than in quantifier conditions (all: b=-1.81, SE=0.18, p<.001; some: b=-2.08, SE=0.25, p<.001). Within quantifiers, the overall target bias was not different (b=-0.32, SE=0.21, p=.14). We found a marginal main effect of *Time* ($\chi2(1) = 3.61$, p=.057), but there was no significant interaction of *Determiners* and *Time*.

In the name window, we also found a significant interaction of *Determiner* and *Target size* ($\chi2(2) = 10.12$, p=.006). Planned comparison on each level of *Determiner* revealed that, as in the previous window, the target bias in big *all* was stronger than in small *all* (b=-1.09, SE=0.27, p<.001), but not for *some* (b=0.41, SE=0.29, p=0.16) or *numbers* (b=0.21, SE=0.22, p= 0.32). Comparison with *Target size* found that when the target size was big, the target bias was greater in big *all* than in big *some* (b=-1.05, SE=0.33, p=.002), whereas when the target size was small, no difference was found between small *all* and small *some* (b=0.32, SE=0.32, p=.31). In addition, regardless of set size, bias to the target was always greater in *numbers* than in *all* and *some* (*three* vs. *all*: b=-1.39, SE=0.25, p<.001, *three* vs. *some*: b=-2.55, SE=0.33, p<.001; *two* vs. *all*: b=-2.07, SE=0.32, p<.001, *two* vs. *some*: b=-1.87, SE=0.30, p<.001). There was no significant three-way interaction of *Determiner, Target size* and *Time*.

Visual inspection of Figure 36 suggested that in Experiment 3, the target bias in *all* and *some* increased steadily before the onset of disambiguating noun. To determine whether the target referent was identified before the noun onset, we performed one sample t-test to compare target proportions to chance (50%) over the combined window (from the determiner window onset to the name window offset). Results showed that the target proportion was significantly above chance for *all*, *some* and *numbers* (all: $t_1(34)=3.26$, p=.003; $t_2(35)=5.65$, p<.001; some: $t_1(34)=4.81$, p<.001; $t_2(35)=5.29$, p<.001; numbers: $t_1(34)=12.35$, p<.001; $t_2(35)=19.40$, p<.001). Thus, we found a significant bias to target in all three conditions before the disambiguating noun.

5.5.2.2 Analyses of visual search to the residue set

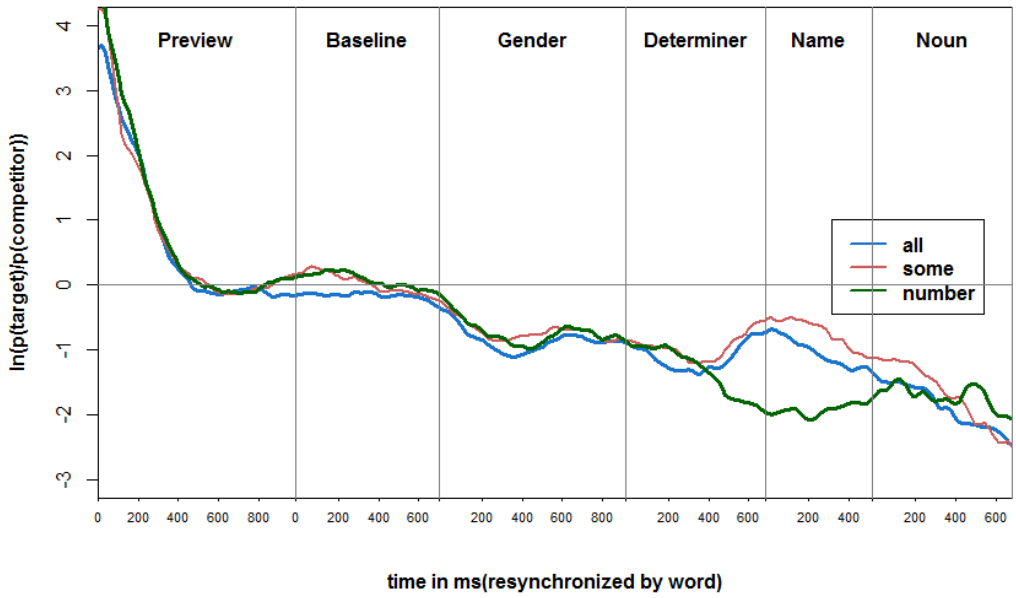


Figure 37 *Bias to residue set (empirical logits) by Determiner from the instruction onset to the instruction offset from the display onset to the instruction offset in Experiment 3*

We then examined looks to the residue set in each time window by fitting a mixed effects model to predict empirical logits from fixed effects of Determiner (All, Some, Number), Target size (Big, Small), a continuous Time variable and their interactions. The model contained maximal random effects structure, which included random intercepts for subjects and items and uncorrelated random slope containing all fixed effects and their interactions.

For the determiner window, we found a main effect of *Determiner* ($\chi2(2) = 7.86$, p=.02) and a significant interaction of *Determiners* and *Time* ($\chi2(2) = 32.42$, p<.001). Planned comparisons showed that overall looks to the residue set were significantly less in *numbers* than in *some* (b=0.35, SE=0.10, p<.001), but the difference between *numbers* and *all* did not reach significance (b=0.13, SE=0.09, p=.16). Between quantifiers, there were more looks to the residue set in *some* than in *all* (b=0.16, SE=.07, p=.04). In terms of changes in residue fixations over time, we found looks to the residue set decreased faster in numbers than in both *all* (all: b=1.47, SE=0.26, p<.001) and *some* (b=1.49,

SE=0.28, p<.001), and no difference was found between *all* and *some* (b=0.04, SE=0.24, p=.86). Other main effects and interactions were not significant.

For the name window, we again found a main effect of determiner ($\chi2(2) = 53.22$, p<.001) and a significant interaction of *Determiners* and *Time* ($\chi2(2) = 19.07$, p<.001) . Planned comparisons showed that overall looks to the residue set were significantly less in *numbers* than in both quantifiers (all: b=0.58, SE=0.10, p<.001; some: b=0.89, SE=0.10, p<.001). There were still more looks to the residue set in *some* than in *all* (b=0.29, SE=.10, p=.004). In terms of changes in residue fixations over time, again no difference was found between *all* and *some* (b=0.10, SE=0.2, p=.62), but looks to the residue set decreased faster in *all* and *some* trials than in *numbers* (all: b=-0.97, SE=0.22, p<.001; some: b=-0.86, SE=0.22, p<.001).

As shown in Figure 37, during the determiner window, the time course of looks to the residue set for *all* and *some* developed in a non-linear fashion. To determine that the differences in bias between *some* and *all* vs. *number* conditions were indeed due to the shifts towards the residue set in the former case, we conducted a growth-curve analysis in the determiner window. We fitted a model to predict empirical logits from fixed effects of *Determiner* (All, Some, Number), *Time* and their interactions. To capture the bend, Time was represented using a 2nd-order orthogonal polynomial (Time1, Time2). The interaction of *Determiner* and *Time* was treated as nested within each level of Determiner. The model contained maximal random effects structure supported by the data, which included random intercepts for subjects and items and uncorrelated random slope. All fixed effects were included as random slops, interactions were not included as random slope because the model did not converge.

We found a significant interaction of *number* and $Time^1$ with a negative slope (b=-0.79, SE=0.17, p<.001), indicating that looks to the residue set decreased significantly in number trials over time. However, we found both the interaction of *all* and $Time^1$ and the interaction of *some* and $Time^1$ had a positive slope (all: b=0.41, SE=0.17, p=.02; some: b=0.44, SE=0.17, p=.01), indicating that looks to the residue set increased significantly over time. These results suggested that the overall angle of the curve for numbers was different from that for quantifiers. Critically, we found a significant interaction of *all* and $Time^2$ (b=0.66, SE=0.10, p<.001) and a significant

interaction of *some* and Time^2 (b=0.57, SE=0.10, p<.001), both with a positive coefficient. These indicated that for both *all* and *some*, participants initially looked away from the residue set and then this was followed by looks towards the residue set. By contrast, we found no significant interaction of *number* and Time^2 (b=0.04, SE=0.10, p=.68), indicating that there was no quadratic component when looks to the residue set decreased over time.

5.5.2.3 Discussion

In Experiment 3, we broadly replicated the findings from Experiment 2. In the analysis of target anticipatory eye movements, we found, in both time windows, the target bias emerged earlier and stronger in the *number* condition than in *all* and *some*, whereas between *all* and *some*, bias formation did not differ. Thus, when the effect of target set size was explicitly controlled, accessing and integrating 'not all' inference triggered by *some* is rapid. Concerning the impact of offline preferences on anticipatory processing, we found, in both time windows, for *all* condition, the target bias was greater in big-set trials than in small-set trials. This result indicated that the target bias formation in *all* trials was influenced by the prior association between *all* and the larger set. In addition, we found when the target size was big, the target bias increased faster after hearing 'all' than hearing 'some', and this trend led to an overall greater target bias in big-set *all* trials compared to big-set *some* trials. Whereas when the target size was small, we found a reverse pattern during the determiner window: the overall target bias in small-set *some* were marginally greater than in small-set *all*. These differences in bias formation between the *all* and *some* trials reflected the stronger big-set association for *all* than for *some*.

However, our analyses of target anticipatory eye movements found one unexpected result. In both time windows, there were no significant differences in overall target bias between the big-set *some* and small-set *some* trials. The changes in bias over time between the two did not differ either. These results were not predicted on the basis of the prior association between *some* and the larger set when the relative set size was 2 and 3. One possible explanation is that given the simpler stimuli used in Experiment 3, when identifying the *some* target, participants relied more on the linguistic processing and relied less on the big-set association. The pattern of residue

fixations in *some* trials supports the explanation, as we found that in both time windows, there were more looks to the residue set in *some* than in *all* and *numbers*.

In the analyses of visual search to the residue set, we found, after the determiner onset, looks to the residue set decreased rapidly in *numbers* compared to *all* and *some*, which resulted in significantly less overall looks to the residue set in numbers than in *all* and *some* in the name window. Therefore, these results again confirmed the fast-pragmatic account of *some*, which predicts that an increase in visual search of the residue set in *all* and pragmatically enriched *some* relative to *numbers*. More importantly, the growth-curve analysis in the determiner window revealed that in fact, as time increased, in the linear term, looks to the residue set increased significantly in both *all* and *some* and only decreased significantly in *numbers*. An increase in bias to this region for *all* and *some* provides clearer evidence that looks to the residue set were driven by the compositional meaning of the linguistic input. In addition, the growth-curve analysis also showed in the quadratic term, there was a decrease followed by an increase in looks to the residue set in both *all* and *some* but not in *numbers*, indicating a pattern of shifting towards the residue set from other region in *all* and *some*, but no such trend was found in the *numbers*. Taken together, the results from the growth-curve analysis provide negative evidence for the slow-pragmatic account, which predicts no increase in looks to the residue set and no shifts in fixation focus in *some* trials.

In Experiment 3, we did not find shifts to the residue set in the *number* condition, which is in contrast to the finding of Experiment 2(a). We suspect that no 'return' effect might be due to the relatively short duration of the determiner window in Experiment 3. As in this experiment, the determiner window (i.e. *all of/some of/two of/ three of*) did not contain the definite article 'the'. This result also confirmed that the 'return' effect in *numbers* found in Experiment 2(a) might largely due to the noise caused by specific design issue.

In this study, the overall bias to target in *all* and *some* conditions rose significantly above chance prior to the disambiguation noun onset. These results suggested that we successfully addressed the issue of slow target identification by using simpler stimuli and a modified procedure. In Experiment 3, after the gender window (i.e. girl/boy that has), only two characters matched the gender information. Also

participants could identify the correct target in the determiner window. Since the name window was non-informative (i.e. Susan's), instead of waiting to hear the disambiguating noun, participants were more willing to exploit the determiners to disambiguate the referent at the earliest point so that they could proceed to the next trial as soon as possible.

5.6 General Discussion

This paper explored factors that could partly account for the mixed results in the timecourse of access and integration of pragmatically enriched *some*. In previous visual-world investigations, Huang & Snedeker (2009, 2011) reported a delay in pragmatic *some* compared to *all*, whereas Degen & Tanenhaus (2016) found a temporary delay only when the target was a big set but not when it was a small set. We hypothesize that people have prior expectations about set size and quantifier. In Experiments 1a,b, we demonstrated that there is a low-level association between the larger-set target and *all*, and it is stronger than the association between any of the set sizes and *some*. In Experiments 2a,b, we showed such prior expectations influenced the pattern of target anticipatory looks during online interpretation of scalar quantifiers. In particular, the prior association between the larger set and *all* favour the target identification in *all* compared to *some*. These findings render the interpretation of previous visual world data problematic. When set size is not controlled, as in Huang & Snedeker (2009), the delay in target identification for *some* relative to *all* could be partly due to the stronger prior association between the larger set (of 3 objects) and *all* rather than the slow pragmatic calculation. To address the issue that target anticipatory looks were interfered by prior expectations, in Experiment 2 and 3, we fully counterbalanced the target set size for the *all* and *some* referents and explicitly modeled the target size in the analyses. We found that, when set size was controlled, the timecourse of looks to the target based on enriched-*some* is not different to that for *all*.

In order to tease apart effects of prior expectation and effects meaning composition in the incremental interpretation of quantifiers, in Experiment 2 and 3, we introduced a novel indicator to measure the timecourse of scalar processing. This was looks to the region where the residues of the partitioned sets are located. We

hypothesized that visual search to the residue set should only be driven by the meaning assigned to the quantificational determiner and should be unaffected by other expectations. The data support our hypothesis that across three experiments involving different quantity ratios, there was no effect of set size on the visual bias to the residue set. Critically, visual search to the residue set allowed us to test different predictions made by the fast-pragmatic account and the slow-pragmatic account. Using visual search to the residue set in number trials as a baseline, after the determiner onset, the fast-pragmatic account predicts an increase in shifts in visual bias to this region in enriched-*some* compared to *number* trials, whereas the slow-pragmatic account predicts no difference in shifts in visual bias to the residue set between unenriched-*some* and *number* trials. Both accounts predict a difference between *all* and *numbers*. Our results show positive evidence for the fast-pragmatic account and negative evidence for the slow-pragmatic account. In particular, for all three experiments, we found a greater bias to the residue set in *some* and *all* compared to *number* trials. In addition, the pattern of shifting from outside the residue region in *some* and *all* trials provide further evidence that the enriched-some is accessed and integrated rapidly.

Therefore, across three visual-world studies, both the pattern of target anticipatory looks and the pattern of residue fixation have shown that pragmatically enriched interpretation of *some* is accessed in the same timecourse as literal interpretations of *all*. Regarding different models of language processing, these findings are not compatible with a literal first model, which suggests that establishing the semantic interpretation precedes deriving the pragmatic interpretation. By contrasts, these results are consistent with an interactive model, which suggests that establishing the semantic interpretation and the pragmatic interpretation could operate in parallel.

More generally, our results also showed that incremental interpretation of quantificational determiners could be impacted by factors that are independent of the integration of the compositional meaning. Recall that independent of set size, we found target bias emerged earlier and stronger in *numbers* than in *all* across three visual-world studies. Referential disambiguation in both *numbers* and *all* rely on only the semantic interpretation of the expression. Such latency of target identification in *all* compared to *numbers* was expected considering that in these visual-world studies, verifying the *all*

referent involves a wider region of inspection compared to verifying the number referent. Thus, this finding suggested that factors such as the difference in the verification process between number and non-number affects the online measure of the incremental processing. The rapid target identification in *all* found in previous studies indicates other factors such as the simple association between a larger set and *all* affect the incremental interpretation as well. These findings suggest when using visual-world paradigm to investigate the quantifier processing, future studies should be aware of factors that affect comprehension processes for quantificational determiners in general, regardless of whether the pragmatic interpretation is involved or whether the number terms are present.

Chapter 6 CONCLUSIONS

This thesis looked at Scalar Implicature from a broadly Gricean perspective. It focussed mostly on proposals for an integrated Gricean system that explains scalar phenomenon via two routes, a global inference mechanism and a local enrichment mechanism. A series of experiments were conducted to investigate the interpretation of scalar term in unembedded and embedded positions. I argue that the phenomena considered can be understood in terms of such an integrated Gricean system and that the results lend support to the two-mechanism view. I summarise the findings of Chapter 2, 3, 4 and 5 as follow:

I started out by investigating the phenomenon of scalar diversity phenomenon from a novel perspective that local enrichment of scalar terms might have an impact on unembedded scalar implicature. In both laboratory-based and corpus-based tasks, I replicated the results of previous research that there is a considerable amount of variation in the rates of scalar implicatures generated by different scalar terms. As for the source of scalar diversity, critically, I found local enrichability, which is associated with the likelihood that a scalar term gets a locally enriched interpretation, could independently explain a significant amount of the variance. This finding is in line with RSA-LU accounts that the prior likelihood of a local upper-bound enrichment of a scalar term could influence the rates of unembedded scalar implicatures. The observed variation in such likelihood is consistent with the idea from the Relevance tradition that less specific scalar terms are more liable to be locally enriched than more specific terms.

In Chapter 4, I obtained more direct evidence that local pragmatic enrichment is indeed being accessed for interpreting the unembedded scalar. In particular, using the enrichment priming paradigm, I found priming effects between embedded and unembedded scalar enrichments, suggesting local enrichment as a shared mechanism is involved in both.

In Chapter 5, I investigated the timecourse of access and integration of locally enriched *some*. I demonstrated that there is a low-level association between the relative set sizes of the target in a display and the quantifier use. Such prior expectations influenced participants' anticipatory looks during online interpretation of scalar

quantifiers. These findings render the interpretation of previous visual world data problematic. By using a novel indicator which is unaffected by other expectations, I found the pragmatically enriched *some* is accessed and integrated rapidly.

Among other results, studies presented in this thesis showed that (i) local pragmatic enrichment is not restricted to embedded scalar implicatures, it applies to the cases of Straight Scalar as well, (ii) scalar terms differ in prior likelihoods of being locally enriched at the upper-bound end and such difference might influence comprehension processes of future encounters and (iii) there was rapid access to (local) pragmatic enrichments.

REFERENCES

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*(3), 191–238. https://doi.org/10.1016/0010-0277(88)90020-0

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55–71. https://doi.org/10.1016/j.cognition.2008.12.005

Atlas, J., & Levinson, S. (1981). It-clefts, informativeness and logical form: radical pragmatics (revised standard version). In P. Cole (Ed.), *Radical pragmatics* (pp. 1–61). New York: Academic Press. Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:66725/component/escidoc:532196/1981_Atlas_Levinson.pdf

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benz, A., Ferrer, C. B., & Gotzner, N. (2017). Scalar diversity and negative strengthening. In *Proceedings of Sinn und Bedeutung 22*.

Benz, A., & Gotzner, N. (2014). Embedded implicatures revisited: Issues with the Truth-Value Judgment Paradigm. *Proceedings of the Formal & Experimental Pragmatics Workshop, European Summer School for Language, Logic and Information (ESSLLI)*, (March 2016), 1–6. https://doi.org/10.3765/sp.2.4.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*, 120–125.

Bergen, L., Levy, R., & Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*. https://doi.org/10.3765/sp.9.20

Blakemore, D. (2002). *Relevance and linguistic meaning – The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140. https://doi.org/10.1016/j.jml.2016.04.004

Bott, L., & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457. https://doi.org/10.1016/J.JML.2004.05.006

Breheny, R. (2018). Scalar implicatures and a Gricean cognitive system. In N. Katos & C. Cummins (Eds.), *Handbook of Experimental Pragmatics*. Oxford University Press.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation.

Language and Cognitive Processes, *28*(4), 443–467. https://doi.org/10.1080/01690965.2011.649040

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463. https://doi.org/10.1016/j.cognition.2005.07.003

Breheny, R., Klinedinst, N., Romoli, J., & Sudo, Y. (2018). The symmetry problem: Current theories and prospects. *Natural Language Semantics*.

Carston, R. (1988). Language and cognition. In *NEWMEYER, F. Linguistics: the Cambridge survey* (pp. 38–68).

Carston, R. (1998). Informativeness, relevance and scalar implicature. Retrieved from http://discovery.ucl.ac.uk/22922/

Carston, R. (2002). *Thoughts and utterances : the pragmatics of explicit communication*. Blackwell Pub.

Chierchia, G. (2006). Broaden your views. Implications of domain widening and the "logicality" of language. *Linguistic Inquiry*, *37*(4), 535–590.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (p. Vol. 3 pp. 2297–2331). Berlin: Mouton de Gruyter.

Cohen, L. J. (1971). The logical particles of natural language. In Y. Bar-Hillel (Ed.), *Pragmatics of natural language* (pp. 50–68). Dordrecht: Reidel.

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(0), 11-1–55. https://doi.org/10.3765/sp.8.11

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 397–402.

Degen, J., & Tanenhaus, M. K. (2011). Making Inferences: The Case of Scalar Implicature Processing. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 3299–3304.

Degen, J., & Tanenhaus, M. K. (2016). Availability of Alternatives and the Processing of Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science*, *40*(1), 172–201. https://doi.org/10.1111/cogs.12227

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition*. New York: Oxford University Press.

Derczynski, L., Ritter, A., Clarke, S., & Bontcheva., K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, ACL*.

Doran, R. (2009). On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, *1*, 211–248.

https://doi.org/10.1163/187730909X12538045489854

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, *88*(1), 124–154. https://doi.org/10.1353/lan.2012.0008

Eiteljörge, S. F. V., Pouscoulous, N., & Lieven, E. (2016). Implicature production in children: a corpus study. In *Pre-proceedings of Trends in Experimental Pragmatics* (p. pp.51-58).

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics* (pp. 71–120). Basingstoke: UK: Palgrave Macmillan.

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*(1), 87–107. https://doi.org/10.1007/s11050-010-9065-3

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Franke, M. (2009). *Signal to act : game theory in pragmatics*. Institute for Logic, Language and Computation. Retrieved from http://dare.uva.nl/search?metis.record.id=313416

Gazdar, G. (1979). *Pragmatics: Presupposition, implicature, and logical form*. Academic Press. Retrieved from https://scholar.google.co.uk/scholar?q=Gazdar%2C+G.+%281979%29.+Pragmatics%3A+Implicature%2C+Presupposition%2C+and+Logical+Form.+Academic+Press.&btnG=&hl=en&as_sdt=0%2C5

Geurts, B. (2009). Scalar implicatures and local pragmatics. *Mind and Language*, *24*(1), 51–79.

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press. https://doi.org/10.1016/j.neuroimage.2009.02.016

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, *2*(4), 1–34. https://doi.org/10.3765/sp.2.4

Geurts, B., & van Tiel, B. (2013). Embedded scalars. *Semantics and Pragmatics*, *6*(9), 1–37. https://doi.org/10.3765/sp.6.9

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184. https://doi.org/10.1111/tops.12007

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388.

Grice, H. P. (1967). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics III: Speech acts* (pp. 43–58).

Grice, H. P. (1969). Utterer's Meaning and Intention. *The Philosophical Review*, *78*(2), 147–177.

Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41–58. https://doi.org/10.1111/j.1365-2664.2006.01229.x

Grice, H. P. (1989). *Studies in the Way of Words*. *Philosophy* (Vol. 65). Harvard University Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. https://doi.org/10.1111/tops.12142

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55. https://doi.org/10.1016/j.cognition.2010.03.014

Hallett, P. E. (1986). Eye movements. In & J. P. T. K. B. Boff, L. Kaufman (Ed.), *Handbook of perception and human performance I: Sensory processes and perception* (pp. 10–102). New York: Wiley.

Hartsuiker, R. J., & Kolk, H. H. J. (1998). Syntactic Persistence in Dutch. *Language and Speech*, *41*(2), 143–184. https://doi.org/10.1177/002383099804100202

Hartsuiker, R. J., Kolk, H. H. J., & Huiskamp, P. (1999). Priming word order in sentence production. *The Quarterly Journal of Experimental Psychology*, *52*(1), 129–147. https://doi.org/10.1080/713755798

Hartsuiker, R. J., & Westenberg, C. (2000). Persistence of word order in written and spoken sentence production. *Cognition*, *75*, B27–B39. https://doi.org/10.1016/S0010-0277(99)00080-3

Heim, I. (1991). Artikel und definitheit. In A. von Stechow and D. Wunderlich. (Ed.), *Semantik: ein internationales Handbuch der* (pp. 487–535). Berlin: de Gruyter.

Hirschberg, J. (1985). *A Theory of Scalar Implicature*. University of Pennsylvania.

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English.* University of California Los Angeles.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, Form, and Use in Context: Linguistic*. Retrieved from http://www.princeton.edu/~harman/Courses/PHI534-2012-13/Oct8/horn1984.pdf

Horn, L. R. (1989). *A Natural History of Negation*. University of Chicago Press.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415. https://doi.org/10.1016/j.cogpsych.2008.09.001

Huang, Y. T., & Snedeker, J. (2011). *Logic and conversation* revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*(8), 1161–1172. https://doi.org/10.1080/01690965.2010.508641

Huang, Y. T., Spelke, E., & Snedeker, J. (2013). What Exactly do Numbers Mean? *Language Learning and Development*, *9*(2), 105–129.

https://doi.org/10.1080/15475441.2012.658731

Hurford, J. R. (1974). Exclusive or Inclusive Disjunction. *Foundations of Language*, *11*(3), 409–411.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83. https://doi.org/10.1016/J.COGNITION.2012.10.013

Jäger, G. (2012). Game theory in semantics and pragmatics. In C. Maienborn, P. Portner, & K. Heusinger, von (Eds.), *Semantics: An international handbook of natural language meaning* (pp. 2487–2425). De Gruyter Mouton. https://doi.org/10.1111/j.1749-818X.2008.00053.x

Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, *30*(6), 669–690. https://doi.org/10.1007/s10988-008-9029-y

Kritka, M. (2007). Negated Antonyms: Creating and Filling the Gap. In U. Sauerland & P. Stateva (Eds.), *Presupposition and Implicature in Compositional Semantics* (pp. 163–177). London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230210752_6

Kroch, A. (1972). Lexical and inferred meanings for some time adverbials. *Quarterly Progress Reports of the Research Laboratory of Electronics*, (104).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests for random and fixed effects for linear mixed effect models. *R Package Version*. https://doi.org/http://CRAN.R-project.org/package=lmerTest

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press. Retrieved from http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2000.27.3.462

Magri, G. (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics*, *17*(3), 245–297. https://doi.org/10.1007/s11050-009-9042-x

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Matsumoto, Y. (1995). The conversational condition on horn scales. *Linguistics and Philosophy*, *18*(1), 21–60. https://doi.org/10.1007/BF00984960

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Noveck, I. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. https://doi.org/10.1016/S0010-0277(00)00114-1

Noveck, I., & Sperber, D. (2007). The why and how of experimental pragmatics : The case of " scalar inferences ." In N. Roberts (Ed.), *Advances in Pragmatics*. Palgrave.

Paul Boersma & David Weenink. (2017). Praat: doing phonetics by computer. Retrieved

from http://www.praat.org/

Pickering, M. J., & Ferreira, V. S. (2008). Structural Priming: A Critical Review. *Psychological Bulletin*, *134*(3), 427–459. https://doi.org/10.1037/0033-2909.134.3.427.

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2015). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *0*(1975), 1–48. https://doi.org/10.1093/jos/ffv012

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A Developmental Investigation of Processing Costs in Implicature Production. *Language Acquisition*, *14*(4), 347–375. https://doi.org/10.1080/10489220701600457

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Raffray, C. N., & Pickering, M. J. (2010). How Do People Construct Logical Form During Language Comprehension? *Psychological Science*, *21*(8), 1090–1097. https://doi.org/10.1177/0956797610375446

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Retrieved from http://psycnet.apa.org/psycinfo/1989-97284-000

Russell, B. (2012). Probabilistic reasoning and the computation of scalar implicatures. *Doctoral Dissertation, Brown University*. Retrieved from http://semarch.linguistics.fas.nyu.edu/Archive/WY1YTRhM/russell2012probabilistic.pdf

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367–391. https://doi.org/10.1023/B:LING.0000023378.71748.db

Scheepers, C. (2003). Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. *Cognition*, *89*(3), 179–205. https://doi.org/10.1016/S0010-0277(03)00119-7

Sedivy, J. C. (2003). Pragmatic Versus Form-Based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. https://doi.org/10.1023/A:1021928914454

Sedivy, J. C., K. Tanenhaus, M., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. https://doi.org/10.1016/S0010-0277(99)00025-6

Shannon, C. E. (1948). A mathematical theory of communication. *Ell System Technical Journal*, (27), 379–423, 623–656.

Sperber, D., & Wilson, D. (1998). The mapping between the mental and the public lexicon. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 184–200). Cambridge: Cambridge University Press.

Sperber, D., & Wilson, D. (2004). Relevance theory. In *Handbook of Pragmatics* (pp. 607–632).

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language

Comprehension. *Science*. American Association for the Advancement of Science. https://doi.org/10.2307/2888637

Trinh, T., & Haida, A. (2015). Constraining the derivation of alternatives. *Nat Lang Semantics*, *23*, 249–270. https://doi.org/10.1007/s11050-015-9115-y

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177. https://doi.org/10.1093/jos/fft002

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175. https://doi.org/10.1093/jos/ffu017

Wilson, D. (1975). *Presuppositions and Non-Truth-Conditional Semantics*. Academic Press.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Netherlands Graduate School of Linguistics.

APPENDIX A: LIST OF EXPERIMENTAL ITEMS

A.1 ITEMS USED IN EXPERIMENT 1 (CHAPTER 2)

Experimental sentences

• adequate/good: The food | The salary is adequate. • allowed/obligatory: Copying | Drinking is allowed. • attractive/stunning: The singer | This model is attractive. • believe/know: The mother | The teacher believes it will happen. • big/enormous: That elephant | The house is big. • cheap/free: The food | The water is cheap. • content/happy: The homemaker | This child is content. • cool/cold: The air | The weather is cool. • dark/black: That fabric | The sky is dark. • difficult/impossible: The problem | The task is difficult. • dislike/loathe: The doctor dislikes coffee. The teacher dislikes fighting. • few/none: The biologist saw few of the birds. The cop saw few of the children. • funny/hilarious: This joke | This movie is funny. • good/perfect: The layout | This solution is good. • good/excellent: That movie | The food is good. • hard/unsolvable: The problem | The puzzle is hard. • hungry/starving: The boy | The dog is hungry. • intelligent/brilliant: The professor | This student is intelligent. • like/love: The actress likes the movie. The princess likes dancing. • low/depleted: The energy | The gas is low. • may/will: The teacher may come. This lawyer may appear in person. • may/have to: The boy may watch television. The child may eat an apple. • memorable/unforgettable: This movie | This party is memorable. • old/ancient: That house | That mirror is old. • palatable/delicious: The food | The wine is palatable. • participate/win: The runner | The skier participated. • possible/certain: Failing |Success is possible. • pretty/beautiful: The girl | The model is pretty. • rare/extinct: The plant | This bird is rare. • scarce/unavailable: This recording | This resource is scarce. • silly/ridiculous: That joke | That song is silly. • small/tiny: The car | This fish is small. • snug/tight: That dress | The shirt is snug. • some/all: The bartender saw some of the cars. The nurse saw some of the signs. • sometimes/always: The director is sometimes late. The doctor is sometimes irritable. • special/unique: That dress | That painting is special. • start/finish: The dancer | The runner started. • tired/exhausted: The runner | The worker is tired. • try/succeed: The athlete | The candidate tried. • ugly/hideous: That painting | The wallpaper is ugly. • unsettling/horrific: The movie | The news is unsettling. • warm/hot: The soup | The weather is warm. • wary/scared: The dog | The victim is wary.

Control sentences

• clean/dirty: The table is clean. • dangerous/harmless: The soldier is dangerous. • drunk/sober: The man is drunk. • sleepy/rich: The neighbor is sleepy. • tall/single: The gymnast is tall. • ugly/old: The doll is ugly. • wide/narrow: The street is wide.

A.2 Items used in Experiment 2 (Chapter 2)

Experimental sentences

• adequate/good: The food | The salary is good but not adequate. • allowed/obligatory: Copying | Drinking is obligatory but not allowed. • attractive/stunning: The singer | This model is stunning but not attractive. • believe/know: The mother | The teacher knows it will happen but doesn't believe it will happen. • big/enormous: That elephant | The house is enormous but not big. • cheap/free: The food | The water is free but not cheap. • content/happy: The homemaker | This child is happy but not content. • cool/cold: The air | The weather is cold but not cool. • dark/black: That fabric | The sky is black but not dark. • difficult/impossible: The problem | The task is impossible but not difficult. • dislike/loathe: The doctor loathes coffee but he does not dislikes coffee. The teacher loathes fighting but he does not dislike fighting. • few/none: The biologist saw none of the birds but it is not the case that he saw few of the birds. The cop saw none of the children but it is not the case that he saw few of the children. • funny/hilarious: This joke | This movie is hilarious but not funny. • good/perfect: The layout | This solution is perfect but not good. • good/excellent: That movie | The food is excellent but not good. • hard/unsolvable: The problem | The puzzle is unsolvable but not hard. • hungry/starving: The boy | The dog is starving but not hungry. • intelligent/brilliant: The professor | This student is brilliant but not intelligent. • like/love: The actress loves the move but she doesn't like the movie. The princess loves dancing but she doesn't like dancing. • low/depleted: The energy | The gas is depleted but not low. • may/will: The lawyer will appear in person but it is not the case that he may appear in person. • may/have to: The boy has to watch television but it is not the case that he may watch television. The child has to eat an apple but it is not the case that he may eat an apple. • memorable/unforgettable: This movie | This party is unforgettable but not memorable. • old/ancient: That house | That mirror is ancient but not old. • palatable/delicious: The food | The wine is delicious but not palatable. • participate/win: The runner | The skier won but he did not participated. • possible/certain: Failing |Success is certain but not possible. • pretty/beautiful: The girl | The model is beautiful but not pretty. • rare/extinct: The plant | This bird is extinct but not rare. • scarce/unavailable: This recording | This resource is unavailable but not scarce. • silly/ridiculous: That joke | That song is ridiculous but not silly. • small/tiny: The car | This fish is tiny but not small. • snug/tight: That dress | The shirt is tight but not snug. • some/all: The bartender saw all of the cars but not some of the cars. The nurse saw all of the signs but not some of the signs. • sometimes/always: The director is always late but he is not sometimes late. The doctor is always irritable but he is not sometimes irritable. • special/unique: That dress | That painting is unique but not special. • start/finish: The dancer | The runner finished but she did not start. • tired/exhausted: The runner | The worker is exhausted but not tired. • try/succeed: The athlete | The candidate succeed but he did not try. • ugly/hideous: That painting | The wallpaper is hideous but not ugly. • unsettling/horrific: The movie | The news is horrific but not unsettling. • warm/hot: The soup | The weather is hot but not warm. • wary/scared: The dog | The victim is scared but not wary.

Control sentences

The banker is rich but not happy. | The technology is sustainable but not affordable. | The assistant is busy but not effective. | The task is urgent but not important. | John left the party but he never came. | The woman has four children but not three children. | The man divorced his wife but he was never married.

A.3 Items used in Experiment 3 (Chapter 2)

Experimental sentences

• adequate/good: The food | The salary is good so not adequate. • allowed/obligatory: Copying | Drinking is obligatory so not allowed. • attractive/stunning: The singer | This model is stunning so not attractive. • believe/know: The mother | The teacher knows it will happen so doesn't believe it will happen. • big/enormous: That elephant | The house is enormous so not big. • cheap/free: The food | The water is free so not cheap. • content/happy: The homemaker | This child is happy so not content. • cool/cold: The air | The weather is cold so not cool. • dark/black: That fabric | The sky is black so not dark. • difficult/impossible: The problem | The task is impossible so not difficult. • dislike/loathe: The doctor loathes coffee so he does not dislikes coffee. The teacher loathes fighting so he does not dislike fighting. • few/none: The biologist saw none of the birds so it is not the case that he saw few of the birds. The cop saw none of the children so it is not the case that he saw few of the children. • funny/hilarious: This joke | This movie is hilarious so not funny. • good/perfect: The layout | This solution is perfect so not good. • good/excellent: That movie | The food is excellent so not good. • hard/unsolvable: The problem | The puzzle is unsolvable so not hard. • hungry/starving: The boy | The dog is starving so not hungry. • intelligent/brilliant: The professor | This student is brilliant so not intelligent. • like/love: The actress loves the move so she doesn't like the movie. The princess loves dancing so she doesn't like dancing. • low/depleted: The energy | The gas is depleted so not low. • may/will: The lawyer will appear in person so it is not the case that he may appear in person. • may/have to: The boy has to watch television so it is not the case that he may watch television. The child has to eat an apple so it is not the case that he may eat an apple. • memorable/unforgettable: This movie | This party is unforgettable so not memorable. • old/ancient: That house | That mirror is ancient so not old. • palatable/delicious: The food | The wine is delicious so not palatable. • participate/win: The runner | The skier won so he did not participated. • possible/certain: Failing |Success is certain so not possible. • pretty/beautiful: The girl | The model is beautiful so not pretty. • rare/extinct: The plant | This bird is extinct so not rare. • scarce/unavailable: This recording | This resource is unavailable so not scarce. • silly/ridiculous: That joke | That song is ridiculous so not silly. • small/tiny: The car | This fish is tiny so not small. • snug/tight: That dress | The shirt is tight so not snug. • some/all: The bartender saw all of the cars so not some

of the cars. The nurse saw all of the signs so not some of the signs. • sometimes/always: The director is always late so he is not sometimes late. The doctor is always irritable so he is not sometimes irritable. • special/unique: That dress | That painting is unique so not special. • start/finish: The dancer | The runner finished so she did not start. • tired/exhausted: The runner | The worker is exhausted so not tired. • try/succeed: The athlete | The candidate succeed so he did not try. • ugly/hideous: That painting | The wallpaper is hideous so not ugly. • unsettling/horrific: The movie | The news is horrific so not unsettling. • warm/hot: The soup | The weather is hot so not warm. • wary/scared: The dog | The victim is scared so not wary.

Control sentences

The woman has four children so not three children. | The window is open so not closed. | The cup is red so not blue. | John left the party so he never came. | The train arrived so it never departed. | The man divorced his wife so he was never married. | The banker is rich so not happy.

A.4 Scales used in Corpus and Paraphrase task (Chapter 3)

<adequate, good>, <allowed, obligatory>, <attractive, stunning>, <big, enormous>, <cheap, free>, <dark, black>, <difficult, impossible>, <few, none>, <funny, hilarious>, <hard, unsolvable>, <hungry, starving>, <intelligent, brilliant>, <low, depleted>, <memorable, unforgettable>, <old, ancient>, <possible, certain>, <rare, extinct>, <scarce, unavailable>, <silly, ridiculous>, <small, tiny>, <snug, tight>, <some, all>, <sometimes, always>, <special, unique>, <tired, exhausted>, <ugly, hideous>, <warm, hot>, <wary, scared>

A.5 Filler items used in Experiment 1 (Chapter 4)

(i) The sentence was presented with a strong picture and a 'Better Picture?':

All of the symbols are diamonds.

Some of the symbols are diamonds.

On exactly one row, all of the symbols are diamonds.

On each row, some of the symbols are diamonds.

(ii) The sentence was presented with a false picture and a 'Better Picture'.

All of the symbols are diamonds.

Some of the symbols are diamonds.

On exactly one row, all of the symbols are diamonds.

On each row, some of the symbols are diamonds.

(iii) The sentence was presented with a false picture and a strong picture.

All of the symbols are diamonds.

Some of the symbols are diamonds.

On exactly one row, all of the symbols are diamonds.

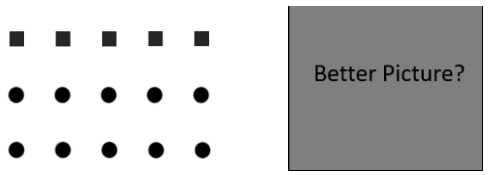
On each row, some of the symbols are diamonds.

A.6 Filler items used in Experiment 2 (Chapter 4)

(i) The sentence was presented with a strong picture and a 'Better Picture?':

All of the symbols are diamonds.

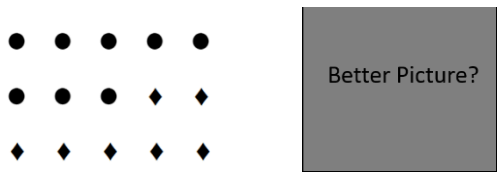
On exactly one row, all of the symbols are squares.



(ii) The sentence was presented with a false picture and a 'Better Picture'.

All of the symbols are diamonds.

On exactly one row, all of the symbols are squares.



(iii) The sentence was presented with a false picture and a strong picture.

All of the symbols are diamonds.

On exactly two rows, all of the symbols are crosses.