Quantity matters

Implicatures, typicality and truth

Bob van Tiel

Quantity matters

Implicatures, typicality and truth

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus
prof. mr. S.C.J.J. Kortmann
volgens besluit van het college van decanen
in het openbaar te verdedigen op
woensdag 11 juni 2014
om 16.30 uur precies
door

Bob Johannes Maria van Tiel

geboren op 19 mei 1986
te Eindhoven

Promotoren:
Prof. dr. Bart Geurts
Prof. dr. Rob van der Sandt

Manuscriptcommissie:
Prof. dr. Helen de Hoop
Dr. Corien Bary
Dr. Michael Franke (Universiteit van Amsterdam)

Contents

Acknowledgements

First of all, I would like to thank my supervisors. Rob's lectures inspired me to study language philosophy and he has been a supportive critic ever since. Bart and I have spent countless hours discussing all of the topics in this thesis and much more. This thesis owes a tremendous debt to his comments and criticisms. I could not have wished for a better pair of supervisors.

There are a number of people who have contributed to this thesis or otherwise helped to make the past four years more bearable. To start off, I would like to thank Helen de Hoop, Michael Franke, and Corien Bary for plowing through the manuscript of this thesis and providing helpful comments along the way. I am glad at least some people took to reading the product of four years work.

Most of this work was done within the confines of the Philosophy department of the Radboud University Nijmegen. I would like to thank all of my colleagues for making my stay in Nijmegen fruitful and pleasant; in particular, Corien Bary, Emiel van Miltenburg, Maddalena Tacchetti, Natalia Zevakhina, Noor van Leusen, and Sammie Tarenskeen for their help in shaping the thoughts that constitute this thesis and for making the past four years much more enjoyable than they would have been without them.

I also owe special thanks to Jacques Jayez, Ye Tian, and Richard Breheny, with whom I have collaborated on two XPRAG projects. This thesis has benefited from discussions with many people, including Anneke Neijt, Benjamin Spector, Chris Cummins, Dan Grodner, Emmanuel Chemla, Ira Noveck, Jacques Jayez, Judith Degen, Katja Jasinskaja, Katrijn Pipijn, Matthijs Westera, Michael Franke, Monique Lamers, Nausicaa Pouscoulous, Paula Rubio-Fernández, Richard Breheny, Uli Sauerland, Walter Schaeken, Xavier Villalba, and Ye Tian.

On a more personal level, I should not forget to thank my parents and sister, who were always there for me. My final and foremost thanks are for Anke. I hope you realise how much you mean to me.

Preface

Parts of this thesis have been published before and/or present collaborative research. In particular, this holds for the following sections:

- Section 2.2: This is a revised version of van Tiel (2014).
- Sections 2.2-2.3: These sections present results from Geurts & van Tiel (2013).
- Section 2.4: This is a revised version of van Tiel (2012).
- Sections 3.2-3.3: The work reported in these sections has been done in collaboration with Katrijn Pipijn and Walter Schaeken.
- Section 4.2: This is a revised version of van Tiel & Geurts (2014).
- Section 4.3: The work reported in this section has been done in collaboration with Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts.

The work presented in this dissertation was conducted within the NWO-funded project “Quantity matters: Building a theory of Q-implicature”.

1
Introduction

1.1 Semantics and pragmatics

Sharing information is one of the functions of linguistic communication. The information that is conveyed by an utterance comes in different varieties and flavours. Consider the following dialogue:

(1) A: Was the exam difficult?
B: Most of the students failed.

Speaker B's utterance carries several pieces of information (or *propositions*): for example, it implies...

i. that more than half of the students failed the exam.
ii. that not all of them did.
iii. that the exam was difficult.

Although these are all aspects of the information expressed by B's utterance, they differ in several respects. Intuitively, proposition *i* is closely connected to the meaning of the words that constitute his utterance, whereas *iii* depends importantly on the context in which it was uttered: somebody who says that most of the students failed does not always imply that the exam was difficult. Information that is closely connected to the meaning of words is called *semantic*, whereas information that largely depends on the context is called *pragmatic*. So proposition *i* is semantic, while *iii* is pragmatic in nature.

This thesis focuses on one specific kind of information: *quantity inferences*. Quantity inferences are propositions that follow from the assumption that the speaker provides as much information as is relevant for the purpose of the discourse. A good example is B's suggestion that not all of the students failed the exam. This proposition follows from the assumption that if he believed that all of the students failed the exam, it would have been informative and relevant to mention. Since he did not, we conclude that B does not believe that all of the students failed the exam. In the next section, I discuss the notion of quantity inferences in greater detail.

The main purpose of this thesis is to determine the position of quantity inferences in the semantics/pragmatics spectrum. Pragmatic theories about quantity inferences view them as a species of *conversational implicature*. The notion of conversational implicature will be discussed at some length in the section 1.3. In particular, I discuss the tenability of one of the diagnostics that are often used to delineate the class of conversational implicatures, namely *cancellability*. In recent times, the pragmatic account has been challenged by a number of theorists who have argued that quantity inferences are semantic rather than pragmatic in nature. These *conventionalist* theories will be discussed in section 1.4.

In order to compare these theories, I first focus on the issue of '*embedded implicatures*'. (It will later become apparent why the scare quotes are there.) As I will explain in chapter 2, 'embedded implicatures' offer a criterion for deciding whether quantity inferences are closer to the semantic or pragmatic end of the spectrum, although the details of this criterion are often misunderstood. First, I evaluate the theoretical evidence for the existence of embedded implicatures. Afterwards, in sections 2.2 to 2.4, I experimentally analyse the interpretation of 'some' and 'or' in various kinds of embedding. One of the lessons of this chapter is that the interpretation of expressions like 'some' and 'or' is affected by various meaning aspects. Aside from literal meaning and conversational implicature, the interpretation of these expressions is also influenced by considerations of *typicality*.

In chapter 3, I approach the question of whether quantity inferences are semantic or pragmatic in nature by investigating their processing cost. It has often been observed that some varieties of quantity inferences are associated with a processing cost. In section 3.2, I determine whether this processing cost occurs during sentence comprehension, as expected if quantity inferences are semantic information, or afterwards. In section 3.3, I address the more general question of what factor is responsible for the processing cost by comparing four kinds of quantity inferences. It turns out that not all quantity inferences are associated with a processing cost. The data that is discussed in chapters 2 and 3 provide broad support for a pragmatic account of quantity inferences.

Based on the findings of the previous chapters, I address two outstanding issues in chapter 4. First, I take a closer look at how quantity inferences, literal meaning, and typicality interact in the interpretation of quantified statements. Second, I provide evidence that quantity inferences differ in the likelihood with which they occur, and analyse a number of potential explanations for this variability. Both of these sections introduce new issues that should be addressed by any adequate account of quantity inferences.

1.2 Quantity inferences

1.2.1 What are quantity inferences?

The protagonists of this thesis are quantity inferences. Quantity inferences occur whenever a speaker utters a sentence in a context in which another *alternative* sentence would have been both relevant and more informative. Informativeness is usually defined in terms of logical strength: a sentence ψ is more informative than φ if ψ entails φ but not vice versa. By uttering a less informative sentence, the speaker implies that he does not believe that the stronger alternative is true. If, moreover, the speaker is assumed to know whether or not the alternative is true, it follows as a matter of logical consequence that he believes the alternative is false. The computation of quantity inferences can be schematised as follows (cf. Horn 1972, Gazdar 1979, Soames 1982, Geurts 2010):

i. The speaker said φ.
ii. He could have been more informative by saying ψ. Why didn't he?
iii. He does not believe that ψ is true. $(\neg \textsc{Bel}_S \psi)$
iv. The speaker knows whether ψ is true or false. $(\textsc{Bel}_S \psi \vee \textsc{Bel}_S \neg \psi)$
v. Hence, he believes that ψ is false. $(\textsc{Bel}_S \neg \psi)$

Note that the *ignorance inference* that is obtained in step *iii* is weaker than the quantity inference obtained in step *v*. If the speaker does not believe that ψ is true, it does not necessarily follow that he believes that ψ is false. Perhaps he does not know whether the sentence is true or false. In contexts in which the competence assumption is implausible, the listener will only compute the ignorance inference.

In the next section, we consider four hypothesised cases of quantity inferences. Afterwards, we discuss various theories about the nature of these inferences.

1.2.2 Four kinds of quantity inferences

Exhaustivity inferences

Answers to *wh*-questions are often interpreted as being exhaustive (Groenendijk & Stokhof 1984, van Rooij & Schulz 2006, Spector 2003). An example is provided by B's answer in the following dialogue:

(2) A: Who was at the door?
B: John.

In this context, B's answer implies that, aside from John, there was no one at the door. Had there been someone else at the door, the speaker should have said 'John and ...' mentioning whoever else was at the door. Since he did not, the listener concludes that he does not believe that such alternatives are true. Moreover, it is plausible that B knows who was at the door. It follows that he believes that such alternatives are false. In other words, he believes that only John was at the door.

Conditional perfection

As the name suggests, conditional perfection involves the interpretation of conditional statements, such as the following:

(3) If you mow the lawn, I will give you five dollars.

An utterance of this sentence implies, among other things, that the speaker will not give five dollars unconditionally (Geis & Zwicky 1971). This inference can be computed on the basis of the alternative 'I will give you five dollars'. By not using this alternative, the speaker implies that he does not believe it is true. Since, furthermore, it is plausible that the speaker knows who he will give five dollars and under which circumstances, it follows that he believes it is false that he will give the listener five dollars unconditionally.

Free choice inferences

The status of exhaustivity inferences and conditional perfection as a kind of quantity inferences is relatively uncontroversial. The same cannot be said of free choice inferences (cf. Kamp 1973). Free choice inferences often occur if 'or' is embedded under a quantifier or modal expression. In the following sentence, 'or' is embedded under a deontic modal operator:

(4) You may have coffee or tea.

Someone who utters this sentence implies that the listener is allowed to have coffee and that she is allowed to have tea. These free choice inferences appear to compromise the widely held view that 'or' is the natural language equivalent of logical disjunction, since logical disjunctions do not entail their disjuncts. To make this a bit clearer, consider sentences in which 'or' occurs in an unembedded clause:

(5) She had coffee or tea.

Someone who utters this sentence does not convey that she had coffee or that she had tea. So there is an asymmetry between 'or' in unembedded clauses, whose meaning can be equated to that of logical disjunction, and 'or' embedded under a quantifier or modal expression, which communicates the truth of both disjuncts.

The problem of free choice inferences has provoked a wide range of possible solutions. Some of these locate the source of free choice inferences in the semantics of the existential operator (Barker 2010, Merin 1992) or the lexical item 'or' (Geurts 2005, Zimmermann 2000). Others adopt a different approach, explaining free choice inferences as a kind of quantity inferences (Franke 2011, Fox 2007, Geurts

2010, Kratzer & Shimoyama 2002, Schulz 2005). According to these accounts, the reasoning process that underlies the computation of free choice inferences can be summarised along the following lines.

The speaker could have been more informative by saying 'You may have coffee' or 'You may have tea'. (Observe that these sentences are more informative than the speaker's utterance, since a disjunction is entailed by either of its disjuncts.) Why didn't the speaker say, e.g., 'You may have coffee'? If he had used this alternative, he would have communicated not just that the listener is allowed to have coffee but also that she is not allowed to have tea. The reason that he did not use this alternative, then, is that he believes it is false that the listener is allowed to have coffee but not tea. The same reasoning goes, mutatis mutandis, for 'You may have tea', which implies that the speaker believes it is not the case that the listener is allowed to have tea but not coffee. Together with the literal meaning of (5), this entails the free choice inferences that the listener is allowed to have coffee and that she is allowed to have tea.

There is an important difference between the computation of free choice inferences and that of other kinds of quantity inferences, such as exhaustivity inferences and conditional perfection. In the case of exhaustivity inferences and conditional perfection, the listener reasons about the literal meaning of the alternatives whereas, in order to compute free choice inferences, it is necessary to also reason about what the speaker would have implied pragmatically had he used one of the alternatives.

Scalar inferences

Scalar inferences are the kind of quantity inferences that have received the most attention in the theoretical and experimental literature. It is also the kind of quantity inferences that will receive the most attention in this thesis. B's suggestion that not all of the students failed the exam, as discussed in the previous section, provides a clear example. Another one is given by the following sentence:

(6) I ate some of the cookies.

Someone who utters this sentence is likely to convey that he did not eat all of the cookies. This scalar inference is computed on the basis of the alternative 'I ate all of the cookies'. By not using this alternative, the speaker implies that he does not believe it is true. Assuming that the speaker is competent, it follows that he believes the alternative is false.

Scalar inferences receive their name for being associated with expressions that form a lexical scale. In (6), the scalar inference is triggered by 'some' which evokes the scale ⟨some, all⟩. Alternatives can be generated by substituting scalar expressions with their *scalemates*. In the case at hand, generating the alternative involves replacing 'some' with 'all'. Table 1.1 provides a sample of lexical scales consisting of elements from various grammatical categories. In section 4.3, we take a closer look at some of the differences between these and other lexical scales.

Category	Examples	
Adjectives	⟨intelligent, brilliant⟩	⟨difficult, impossible⟩
Adverbs	⟨sometimes, always⟩	⟨possibly, necessarily⟩
Connectives	⟨or, and⟩	
Determiners	⟨some, all⟩	⟨few, none⟩
Nouns	⟨mammal, dog⟩	⟨vehicle, car⟩
Verbs	⟨might, must⟩	⟨like, love⟩

Table 1.1: Sample scales for various grammatical categories.

An important assumption in the computation process that underlies scalar inferences is that replacing a scalar term in a simple sentence with a stronger scalemate leads to a more informative sentence. To illustrate, suppose it is inferred from the following sentence that around here we do not love coffee, based on the lexical scale ⟨like, love⟩:

(7) Around here we like coffee.

In order to compute this inference, it has to be assumed that the alternative ‘Around here we love coffee’ is more informative than the literal meaning of the sentence. In other words, it has to be assumed that the literal meaning of ‘like’ is ‘like and perhaps love’ rather than ‘like but not love’. There are a number of observations that corroborate the assumption that the literal meaning of scalar expressions is lower-bounded rather than circumbounded.

A first argument in favour of the view that the literal meaning of scalar expression is lower-bounded is that there are instances where the upper bound is clearly not excluded. The following sentences are cases in point:

(8) a. Nobody around here likes coffee.
    b. If she likes coffee, we’ll get her a coffee mug.
    c. Everyone who likes the coffee will receive a free refill.

Someone who says (8a) implies that everyone dislikes coffee. But if ‘like’ meant ‘like but not love’, the sentence should be compatible with a situation in which some people love coffee. Similarly, (8b) implies that we will also get her a coffee mug if it turns out she loves coffee, which is unexpected if ‘like’ literally meant ‘like but not love’. Lastly, (8c) implies that someone who loves the coffee will receive a free refill too, again leading to the conclusion that ‘like’ is not interpreted as ‘like but not love’.

These observations can be explained on the assumption that the literal meaning of ‘like’ is lower-bounded: in the sentences in (8), the alternatives that are generated by replacing ‘like’ with ‘love’ are less informative than the original sentences, since ‘like’ occurs in a *downward-entailing environment*. In such environments, replacing an expression with one that is logically stronger leads to a weaker claim. For example, (8c) offers a refill to a group of people that is at least as large as the group of people its alternative with ‘love’ offers a refill to:

(9) Everyone who loves the coffee receives a free refill.

In this respect, downward-entailing environments differ from upward-entailing environments such as (7). Here, the corresponding alternative with ‘love’ makes a stronger claim.[1]

Since the statements in (8) are more informative than the corresponding alternatives, no upper-bounding inferences occur. But this explanation only works on the assumption that the literal meaning of ‘like’ is lower-bounded rather than circumbounded.

A second observation that supports the view that the literal meaning of scalar expressions is lower-bounded is that scalar inferences are *cancellable*, while aspects of the literal meaning usually are not. (See section 1.3.2 for more on cancellability.) To illustrate, someone who says (10a) felicitously cancels the scalar inference that he did not eat all of the cookies, whereas it is infelicitous to cancel the literal meaning that the speaker ate more than one cookie, as shown in (10b):

(10) a. I ate some of the cookies—in fact, all of them.
b. I ate some of the cookies—in fact, none of them.

A final argument in favour of the view that scalar expressions have a lower-bounded literal meaning is that the upper bound associated with scalar expressions has a different informational status than the lower bound: the upper bound is less prominent than the lower bound. This difference in prominence manifests itself in various ways. First, it is not possible to directly affirm or reject the upper bound associated with scalar terms. To illustrate, consider the following pair of dialogues:

(11) A: Sue ate some of the cookies.
B: No, she didn’t.

(12) A: Sue ate some but not all of the cookies.
B: No, she didn’t.

[1] There are also environments that are neither upward-entailing nor downward-entailing. In the following pair of sentences, for example, neither sentence entails the other one.

(i) a. Exactly three people like coffee.
b. Exactly three people love coffee.

Such *non-monotone* quantifiers will be discussed in section 2.3.

In (11), the information that Sue did not eat all of the cookies is communicated by means of a scalar inference. This information cannot be rejected directly: B's utterance unambiguously implies that Sue ate none of the cookies. The situation is different if the upper bound is made explicit. In (12), B's utterance is also compatible with a situation in which Sue ate all of the cookies.

Second, the upper bound associated with 'some' cannot be referred to by means of expressions like 'so' or 'because' (Horn 2004). Consider:

(13) a. Some of my friends eat meat so I'm not the only vegetarian.
b. Some but not all of my friends eat meat so I'm not the only vegetarian.

In the case of (13a), 'so' cannot refer to the scalar inference that not all of my friends eat meat. Therefore the sentence is pragmatically infelicitous: if at least some of my friends eat meat it does not follow that not all of them do. 'So' is only able to refer to the upper bound if it is made explicit. This is illustrated by the felicitous counterpart of (13a) in (13b).

Third, it is not possible to exhaustively answer a question by means of the upper bound associated with 'some'. This is illustrated by the following pair of dialogues:

(14) A: I hope not all of your friends are coming.
B: Some of my friends are coming so you needn't worry.

(15) A: I hope not all of your friends are coming.
B: Some but not all of my friends are coming so you needn't worry.

In the dialogue in (14), B tries to resolve A's question by means of the scalar inference that not all of his friends are coming. However, due to the special informational status of scalar inferences, this leads to an infelicitous exchange. If the upper bound is made explicit, like in (15), it is possible to resolve the question with the upper-bounded inference. Taken together, these three observations indicate that the upper bound associated with 'some' has a less prominent status than the corresponding lower bound, in support of the conclusion that the literal meaning of scalar expressions is lower-bounded.

Conclusion

We have discussed four kinds of quantity inferences: exhaustivity inferences, conditional perfection, free choice inferences, and scalar inferences. In all of these cases, the speaker uses an underinformative utterance and thereby implies that a more informative alternative is unassertable, which means that either its literal meaning (exhaustivity inferences, conditional perfection, scalar inferences) or one of its concomitant inferences (free choice inferences) is false.

What is the nature of these quantity inferences? *Pragmatic* accounts explain them as a variety of conversational implicatures. According to these accounts, quantity inferences are pragmatic in nature. More recent *conventionalist* accounts have proposed that quantity inferences are caused by lexical or syntactic ambiguities, which would assign them a position on the semantic side of the semantics/pragmatics divide. In the next section, I introduce the notion of conversational implicature and the pragmatic explanation of quantity inferences. Afterwards, I consider its conventionalist alternatives.

1.3 The pragmatic account

1.3.1 Conversational implicatures

Linguistic communication is usually a *cooperative* enterprise: interlocutors do their best to further the purpose of the conversation by means of their utterances. Based on the assumption of cooperativity, it is sometimes possible to extract more information from an utterance than what is encoded by the words that constitute it. Consider the following dialogue (borrowed from Grice 1975):

(16) A: Where does C live?
B: Somewhere in the South of France.

Suppose it is clear that A wants to find out in which city C lives. In that case, B's contribution falls short of achieving this goal because it is not specific enough. This might lead A to reason as follows: assuming that B does his best to further the purpose of the conversation, why didn't he say in which city C lives? A plausible reason is that he does not have this information. In other words, B does not know exactly where C lives. This is a conversational implicature: a proposition that must be attributed to the speaker in order to reconcile the literal meaning of his utterance with the assumption of cooperativity (Grice 1969, 147).

Conversational implicatures depend on the assumption that the speaker is cooperative. As I mentioned earlier, a cooperative speaker tries to further the goal of the conversation by means of his utterances. Grice (1975) operationalises the notion of cooperativity by positing a number of *maxims* that a cooperative speaker is expected to uphold:

Quality:
1. Don't say anything you believe to be false.
2. Don't say anything for which you lack evidence.

Quantity:
1. Provide sufficient information for the goal of the conversation.
2. Don't provide too much information.

Relation: Make sure your utterance is relevant for the goal of the conversation.

Manner: Be as clear as possible.

Conversational implicatures occur whenever assumptions have to be made in order to reconcile the literal meaning of an utterance with the belief that the speaker is cooperative. In the dialogue in (16), for example, speaker B violates Quantity-1: he does not provide sufficient information for the purpose of the conversation. In order to reconcile this violation of Quantity-1 with the general assumption of cooperativity, the listener assumes that there must be a good reason for the speaker to have been underinformative. One such reason is that providing more specific information would have been at odds with Quality-2: the speaker thus sacrifices informativeness in order to avoid untruthfulness.

This example shows that the conversational maxims are not on a par in terms of their importance. In this case, the maxim of Quality outranks the maxim of Quantity. Intuitively, this makes sense: it is more detrimental for the purpose of the conversation to be untruthful than it is to be underinformative. There are similar importance orderings between other maxims. For example, it is less problematic to give too much information or to be less than perfectly clear than it is to give too little information or to say something that is entirely beside the point.

In what respects do conversational implicatures differ from other kinds of information? This question is the topic of the next section.

1.3.2 Diagnostics

Calculability and nondetachability

What distinguishes conversational implicatures from other kinds of information? We have already seen one characteristic feature, namely *calculability*: conversational implicatures can be calculated on the basis of the literal meaning of an utterance and the assumption of cooperativity. Calculability is a necessary but not a sufficient condition for conversational implicatures. That is, some propositions are calculable but not conversational implicatures. To illustrate, consider information that logically follows from the literal meaning of an utterance:

(17) Sue is eating a steak.

This utterance implies, among other things, that Sue is not a vegetarian. This proposition can be calculated on the basis of the conversational maxims: assuming that the speaker is truthful, it follows that Sue is not a vegetarian. Nevertheless, most authors are hesitant to call it a conversational implicature. One reason for this hesitance is that the speaker may not have had the intention to inform the listener of the proposition that Sue is a vegetarian. Another is that it would mean that all information conveyed by an utterance is an instance of conversational implicature. The use of such a watered-down notion is limited. So the proposition that she is a vegetarian is calculable but not a conversational implicature.

Another observation suggesting that calculability is not a sufficient condition of conversational implicatures involves idiomatic expressions. To illustrate, consider the expression ‘in a nutshell’:

(18) Here, in a nutshell, is the source of the current disagreement…

The figurative meaning of the expression can be calculated as follows: it is obvious that the speaker does not intend me to interpret his utterance literally. Perhaps he refers instead to a salient property of things that fit into a nutshell; for example, that such things must be really small. Presumably, then, the speaker is going to explain the source of the current disagreement in brief.

Although the figurative meaning of ‘in a nutshell’ is calculable along these lines, it seems unlikely that listeners use this reasoning process in practice to interpret the expression: its interpretation is conventionalised to such a degree that it has become part of its literal meaning. Furthermore, the proposed reasoning process can be applied to many other expressions like ‘in a pocket’ or ‘in a wallet’ even though these cannot be used to express ‘in brief’. In these cases, too, the proposition is calculable even though it is not a conversational implicature.

Grice (1975) formulated several other properties to help delineate the class of conversational implicatures. Aside from calculability, the two most prominent of these are *nondetachability* and *cancellability*. Suppose that a speaker utters a statement and thereby implies a proposition. That proposition is said to be nondetachable if it would have been implied by any utterance with the same literal meaning as the one used by the speaker. To illustrate, someone who says (19a) implies that not all of the children are sick. This proposition is nondetachable because (19b) and (19c), which have the same literal meaning as (19a), also suggest that not all of the children are sick.

(19) a. Some of the children are sick.
b. A few of the children are sick.
c. More than one of the children is sick.

There are several problems with the criterion of nondetachability. First, it fails to distinguish conversational implicatures from information that is part of the literal meaning of an utterance. For example, (19a) also implies that one or more of the children are sick. This proposition is nondetachable, too, since (19b) and (19c) also suggest that one or more of the children are sick. However, it is clear that it is not a conversational implicature but rather the literal meaning of these sentences.

Second, the criterion of nondetachability is circular. Consider the following pair of sentences (borrowed from Grice 1975):

(20) a. He is an Englishman; he is, therefore, brave.
b. He is an Englishman, and he is brave.

The difference between these sentences is that (20a) but not (20b) implies that his being brave is a consequence of his being an Englishman. According to Grice, this proposition is not nondetachable because the literal meaning of both sentences is identical, and someone who says (20b) does not imply that his being brave is a consequence of his being an Englishman. However, this explanation presupposes that it is possible to determine whether a proposition is a conversational implicature or whether it is part of the literal meaning of an utterance. But this is exactly what the criterion of nondetachability is supposed to determine. To put it differently, we might also have assumed that the proposition in question is nondetachable, and that the literal meaning of (20b) is therefore not identical to that of (20a).

This brings us to the last distinguishing feature of conversational implicatures: cancellability. I devote some more attention to this property in the next section.

Cancellability

Grice (1978) distinguishes two kinds of cancellability: *explicit* and *contextual* cancellability. A proposition that is implied by an utterance is explicitly cancellable if the speaker can felicitously contradict it. To illustrate, consider the dialogue in (16) once again. B's answer implies, among other things…

i. that C lives somewhere in the South of France.
ii. that B does not know exactly where C lives.

Proposition *i* is not explicitly cancellable: it would be infelicitous for B to continue with (21a). On the other hand, proposition *ii* is explicitly cancellable, since the speaker can felicitously continue with (21b), thereby contradicting the belief that he does not know exactly where C lives:

(21) a. In fact, she lives in Canada.
b. In fact, she lives in Toulon.

These observations provide evidence for the assumption that proposition *ii* is a conversational implicature, and that proposition *i* is not.

A proposition that is implied by an utterance is contextually cancellable if there is a context in which the same utterance would not have implied that proposition. The contextual cancellability of proposition *ii* can be illustrated by changing the question that A asks:

(22) A: Does C live somewhere in the North or South of France?
B: Somewhere in the South of France.

In this context, B might not imply that he does not know exactly where C lives, which shows that the proposition is contextually cancellable.

In what follows, I focus on explicit cancellability (which I refer to simply as ‘cancellability’). Cancellability seems to offer a relatively straightforward way of delineating the class of conversational implicatures. However, there are several reasons to suppose that cancellability is not a sufficient condition for conversational implicatures. Sadock (1991) shows that there are cases where a proposition is cancellable but not a conversational implicature. The cases he discusses involve sentences that are syntactically ambiguous. An example is:

(23) Everyone speaks one language.

This sentence can mean that there is single language that everyone speaks, or that everyone speaks one language, but not necessarily the same one for everyone. If someone arrives at the first reading of the sentence, this proposition can be cancelled; for example, by continuing with ‘Although no one language is spoken by everyone’. Nevertheless, most theorists agree that the proposition that there is a single language that everyone speaks is not a conversational implicature but that it is caused by a syntactic ambiguity. Apparently, not every proposition that is cancellable is also a conversational implicature.

A similar point can be illustrated with *presuppositions*. Consider:

(24) Harry doesn’t know he got promoted.

This sentence presupposes that Harry got promoted, and this piece of information is cancellable: the speaker can felicitously continue with ‘Because he did not; it was just a rumour’. For that reason, some authors have concluded that presuppositions must be a variety of conversational implicatures (e.g., Thomason 1990). There is, however, an important difference between the cancellability of presuppositions and that of conversational implicatures: presuppositions can only be cancelled when the trigger (‘know’ in the above case) occurs in a nonentailed position, e.g., in the scope of a negation or a modal operator. This is one of the reasons that the consensus nowadays seems to be that presuppositions and implicatures are distinct phenomena (cf. Geurts 1999).

In summary, cancellability is not a sufficient condition for conversational implicatures. Is it a necessary condition? Grice (1978, 116) seems to think so, stating that he believes all implicatures to be cancellable.[2] However, the hypothesis that all implicatures are cancellable has been challenged by Weiner (2006) and van Kuppevelt (1996). Weiner shows that there are cases where cancelling a conversational implicature does not cause the listener to suspend his belief in that proposition; van Kuppevelt provides cases where cancellation is altogether infelicitous. In the next section, I discuss their arguments. Afterwards, I explain how these apparently anomalous cases can be analysed by considering the role of conversational implicatures in explaining the speaker’s behaviour.

[2] As Jaszczolt (2009) notes, it is not clear if Grice’s remark implies that he believes all conversational implicature to be both explicitly and contextually cancellable.

Uncancellable implicatures

Weiner sketches the following scenario to show that some conversational implicatures resist cancellation: Alice and Sarah are riding on a crowded train. Sarah has to stand up, while Alice is sprawled across two seats. Thereupon Sarah asks:

(25) I'm curious as to whether it would be physically possible for you to make room for someone else to sit down.

Alice obviously infers that Sarah is being sarcastic: she knows perfectly well that Alice is physically able to make room, hence her question must be a request to actually make room. Sarah might attempt to cancel this implicature by continuing with 'Not that you *should* make room, I'm just curious'. Weiner observes that this cancellation attempt has the effect of strengthening, rather than suspending, Alice's belief that Sarah wants her to make room for someone else to sit down.

Some authors have suggested that cases like this one are restricted to utterances that involve sarcasm (Borge 2009, Colonna Dahlman 2013), but this is certainly not the case. Consider the following dialogue:

(26) A: Who stole the silver candle?
B: C left remarkably early today.

Speaker A presumably interprets B's utterance as being relevant to the question asked. That is, A will conclude that, according to B, C is a potential suspect. B might attempt to cancel this conversational implicature; for example, by adding 'But I'm not suggesting that she had anything to do with it'. However, it seems unlikely that this cancellation attempt is going to convince A to suspend his belief that, according to B, C is a potential suspect. Instead, his cancellation attempt will be construed, for example, as conveying that he does not want to go on record as implicating C in the crime.

In the foregoing dialogues, the speaker felicitously contradicts a conversational implicature but fails to convince his listener that the cancellation attempt is sincere. Van Kuppevelt has shown that, in some cases, cancelling a conversational implicature is altogether infelicitous. In order to appreciate his argument, I will first introduce the distinction between *topic* and *focus* (which van Kuppevelt calls 'comment'). According to van Kuppevelt, statements usually serve to answer an implicit or explicit question. To illustrate, consider the following dialogue:

(27) A: Who is laughing?
B: [Alan]$_F$ is laughing.

Based on the question, an utterance can be subdivided into a focused and a topicalised part. The focus of an utterance is the part that addresses the question under discussion, i.e., the constituent that corresponds to the *wh*-phrase in the question. So in this dialogue, 'Alan' is the focused constituent. Changing the question can also change which part of the utterance is focused. For example, if B's utterance were a response to the question 'What is Alan doing?', the constituent 'is laughing' would have been focused instead of 'Alan'. Here and in the following, I will sometimes mark focused constituents with a subscript F.

Van Kuppevelt observes that some conversational implicatures are impossible to cancel when they are evoked by words that occur in the focus of an utterance. The following dialogue provides an especially clear example:

(28) *Context*: Harry did a lot of shopping this afternoon.
A: How many books did he buy?
B: He bought $[\text{four}]_F$.

It is part of the literal meaning of B's answer that Harry bought at least four books. His answer furthermore implies he bought at most four books. This proposition is often explained as a conversational implicature. However, it cannot felicitously be cancelled, since it is impossible for B to continue with 'In fact, he bought $[\text{seven}]_F$'. Cancelling a conversational implicature is altogether inadmissible in this situation.

An alternative conclusion might be that van Kuppevelt's example shows that the upper bound associated with numerals is not a conversational implicature in the first place. Indeed, this is a position that has been defended by several authors (e.g., Spector 2013). But similar examples can be constructed without numerals:

(29) A: Do you live in Los Angeles?
B: Well, I live in California.

B's indirect answer carries the conversational implicature that he does not live in Los Angeles. It is impossible to felicitously cancel this implicature by continuing with 'In fact, I live in Los Angeles'. Since this example is not susceptible to the same counterargument as van Kuppevelt's dialogue, it has to be concluded that there are cases where it is infelicitous to cancel a conversational implicature.

In summary, cancelling a conversational implicature can have the following consequences: it is felicitous and leads the listener to suspend his belief in that proposition; it is felicitous but fails to convince the listener to suspend his belief in that proposition; or it results in an infelicitous exchange. In the next section, I argue that the presence of these three possibilities can be explained by recognising the role of conversational implicatures in explaining the speaker's behaviour.

An analysis

Understanding human behaviour often involves *abductive* inferencing (cf. Hobbs 2004, Hobbs et al. 1993, and Lipton 2004). Suppose that you see B pushing a light switch. There are many possible explanations for this observation:

i_1 B wants to turn on the light.

i_2 B is checking if the light works.
i_3 B is about to disassemble the light switch.
i_4 B switched the light on by accident.
…

In most situations, the interpreter will opt for i_1. But further observations can make one of the other explanations more plausible. So in order to explain an observation, the interpreter constructs a hypothesis about someone's intentions that can be confirmed or disconfirmed by further observations. Such inferences from an observation to an explanation are called abductive inferences.

Conversational implicatures are also abductive inferences. In the case of conversational implicatures, the observation is the speaker's utterance and the hypothesis is a speaker's intention. To illustrate, consider the following dialogue once again:

(30) A: Where does C live?
B: Somewhere in the South of France.

B's answer is less informative than the listener expects it to be. There is a large number of possible explanations for this observation:

i_1 B does not know in which city C lives.
i_2 B cannot immediately remember in which city C lives.
i_3 B thinks his answer is sufficiently informative.
i_4 B is unwilling to divulge more specific information.
…

Among these potential explanations, the listener picks out the one that best explains the speaker's utterance, given the assumption that he is cooperative. Presumably, i_1 offers the best explanation for B's underinformativeness. Suppose that B continues with 'In fact, she lives in Toulon'. This utterance forces the listener to suspend his belief in i_1, since that explanation does not account for B's second utterance. He then has to find an explanation that accounts for both of B's utterances. Explanation i_2 is a good candidate. Adopting this explanation causes the conversational implicature to be cancelled.

Let us make this more precise. Let u_1 be B's utterance in the foregoing dialogue and $i_1, \ldots, i_n$ the set of possible intentions. Then we can formulate a function that associates each possible intention i_x with its likelihood given that u_1:

(31) $f(i_x) = p(i_x \mid u_1)$

To simplify matters, we might assume that the intention attributed to the speaker is the maximum of this function. In the foregoing dialogue, this will be intention i_1. Suppose that B attempts to cancel the implicature by continuing with 'In fact, she lives in Toulon'. This utterance will change the likelihood distribution as follows. First, the likelihood that i_1 obtains becomes significantly lower because i_1 fails to explain u_2: $p(i_1 \mid u_1 + u_2) \ll p(i_1 \mid u_1)$. Second, the likelihood that i_2 obtains becomes significantly higher because that intention explains both utterances: $p(i_2 \mid u_1 + u_2) \gg p(i_2 \mid u_1)$.

The examples of uncancellable implicatures that we discussed in the previous section are caused by the absence of one or both of these consequences. Consider the example discussed by Weiner. When Sarah asks Alice if it is physically possible for her to make room for someone else to sit down, there are two prominent explanations for her question:

i_1 Sarah's question is a sarcastic request for Alice to make room for someone else.
i_2 Sarah is genuinely interested in learning about Alice's physical abilities.

Since Alice assumes that Sarah knows that she is able to make room, she adopts explanation i_1. Suppose that Sarah attempts to cancel the conversational implicature by saying 'Not that you *should* make room, I'm just curious'. Unlike the previous example, this utterance does not force the listener to suspend her initial explanation, since Sarah's continuation can be accounted for by assuming that she is again sarcastic: $p(i_1 \mid u_1 + u_2) \approx p(i_1 \mid u_1)$. Correspondingly, the cancellation attempt will not make the likelihood of i_2 significantly higher.

The analysis is similar in the case of the silver candle. How can B's answer be explained? At least three possible explanations suggest themselves:

i_1 B suspects that C might be involved in the disappearance of the candle.
i_2 B's remark is entirely unrelated to the question A asked.
i_3 B suspects that C might be involved in the disappearance of the candle and he does not want C to find out about his suspicion.

Based on the assumption of cooperativity, the listener prefers i_1 over i_2. Suppose that B cancels the conversational implicature by saying 'But I'm not suggesting that she had anything to do with it'. If this utterance is interpreted literally, the listener has to replace his explanation with another one. But i_2 is not a satisfying alternative because it is at odds with the assumption of cooperativity. Therefore the speaker might stick to the content of i_1 and come up with an additional explanation for B's cancellation: for example, that B does not want to go on record as implicating C in the crime. This causes him to converge on i_3 . Since this explanation subsumes i_1, the conversational implicature is not cancelled.

There are two features that distinguish both of the foregoing examples from cases where cancellation is successful. First, the cancellation attempt is consistent with the interpretation that assumes a conversational implicature. Second, the cancellation attempt does not increase the likelihood of the literal interpretation. Blome-Tillmann (2008) shows that the conversational implicature in Weiner's scenario can be made cancellable by changing these two factors. He sketches a scenario in the distant future where Sarah is developing a tractor beam. She wants to test if it is strong enough to prevent Alice, who is sitting alone on the recreation deck, from moving over. In this situation, Sarah's utterance might not be interpreted sarcastically, and if it inadvertently is Sarah will be able to convince Alice that she was not sarcastic, as the following dialogue illustrates:

Sarah: I'm curious as to whether it would be physically possible for you to make room for someone else to sit down.

Alice: Why should I? There's nobody else here who wants to sit down…

Sarah: Oh, I'm sorry. I did not mean to imply that you should make room. We are testing a new tractor beam on you and we are curious as to whether you can do it. This would give us an important indication as to how strong the beam really is.

Blome-Tillmann's scenario differs from Weiner's in the two respects mentioned above. First, the interpretation according to which Sarah wants Alice to make room for someone else to sit down is much less plausible than before, since Alice sits all alone on a recreation deck. Correspondingly, the literal interpretation is more plausible because there is an explanation for why Sarah would be interested in finding out if Alice is able to make room for someone else to sit down. These differences render the cancellation attempt successful in this scenario.

Based on the observation that the conversational implicature in Weiner's scenario is cancellable in a different context, Blome-Tillmann proposes a new diagnostic which he calls *cancellability** (i.e., with an asterisk). A proposition is cancellable* if there is a context in which it is cancellable. The present discussion explains why cancellability*might be a necessary condition for conversational implicatures: whether a cancellation attempt is successful depends on the plausibility of reconciling the cancellation attempt with the conversational implicature and the plausibility of the literal interpretation. Adjusting these plausibility measures by making changes to the context can thus ensure the success of a cancellation attempt.

What about the scenario described by van Kuppevelt where the cancellation attempt is altogether infelicitous? I focus here on the scenario that does not involve numerals. Several explanations can be given for B's answer:

i_1 B lives in California but not in Los Angeles.
i_2 B lives in Los Angeles.
i_3 B cannot immediately remember where he lives.

Explanation i_1 is the most plausible one: i_2 fails to explain why B did not give a direct answer to A's question, and i_3 is at odds with the commonsense assumption that people tend to know where they live. Suppose that B continues with 'In fact, I live in Los Angeles'. This forces the listener to suspend his belief in i_1. In this scenario, however, there is no plausible alternative explanation: i_2 still leaves unexplained why the speaker did not give a direct answer to the question, and i_3 is as implausible as ever.

In other words, the cancellation attempt meets the first requirement for success, since the likelihood that i_1 obtains becomes significantly lower. However, there is no alternative explanation whose likelihood becomes higher. This leads to a situation of entropy in which there is no intention that stands out as the most plausible explanation for the speaker's behaviour. Consequently, the speaker's behaviour becomes uninterpretable without further clarification.

The conversational implicature associated with B's utterance can be cancelled if B has a good reason for stating, first, that he lives in California and, second, that he lives in Los Angeles. This is the case in the following scenarios:

(32) A: Do you live in Los Angeles?
B: I live in California. In fact, I live in Los Angeles if you include its greater metropolitan area.

(33) A: Do you live in California?
B: I live in California. In fact, I live in Los Angeles.

In (32), B's behaviour can be explained as follows: B believes that the listener wants to know if he lives in the city proper of Los Angeles. He does not, so he answers that he lives in California. Since he is not sure about the listener's intentions, he adds that he does live in the greater metropolitan area of Los Angeles. In (33), the reason B indicates that he lives in California is that he wants to provide a direct answer to A's question: the continuation is added to provide more detailed information that B believes the listener might be interested in. In either case, there is a convincing alternative explanation for both of B's utterances.

In summary, consider a speaker who makes an utterance u_1 that leads to an interpretation with a conversational implicature i_c instead of a literal interpretation i_l. If the speaker continues with an utterance u_2 whose literal meaning contradicts the initial explanation, the following three consequences can occur:

- If $p(i_c \mid u_1 + u_2) \approx p(i_c \mid u_1)$: The listener incorporates the cancellation attempt in his explanation, which entails that the cancellation is interpreted nonliterally.
- If $p(i_c \mid u_1 + u_2) \ll p(i_c \mid u_1)$ and $p(i_l \mid u_1 + u_2) \gg p(i_l \mid u_1)$: The listener suspends his initial explanation and adopts a literal explanation that accounts both for the speaker's utterance and his subsequent cancellation.
- If $p(i_c \mid u_1 + u_2) \ll p(i_c \mid u_1)$ and $p(i_l \mid u_1 + u_2) \approx p(i_l \mid u_1)$: The listener suspends his initial explanation but fails to find one that explains both the utterance that triggered the conversational implicature and the cancellation.

The first consequence occurs in Weiner's scenario, and the third one in van Kuppevelt's. The second consequence corresponds to a successful case of cancellation.

Conclusion

Conversational implicatures are abductive inferences about the speaker's behaviour. Such inferences are normally uncertain and therefore cancellable. However, the success of a cancellation attempt depends on several factors:

- How plausible is the listener's explanation?
- Can this explanation also account for the cancellation?
- Is there a better explanation that accounts for the speaker's behaviour?

Not all conversational implicatures are cancellable in the sense that it is possible to suspend the listener's belief in that proposition, or in the sense that it is felicitous to contradict this proposition (which is presumably the sense in which Grice interpreted the term). But all conversational implicatures behave in a systematic and predictable way when they are cancelled: the listener tries to incorporate the cancellation in his initial explanation if he believes it is a plausible one. If that is impossible, he rejects the explanation and looks for one that accounts both for the speaker's first utterance and the subsequent cancellation.

Cancellability is sometimes invoked as a diagnostic to determine if a proposition is a conversational implicature. To illustrate, consider free choice inferences (see section 1.2.2). Someone who says (34) implies that the listener is allowed to choose between taking an apple and a pear:

(34) You may eat an apple or a pear.

Several authors have argued that these free choice inferences are conversational implicatures (Franke 2011, Geurts 2010, Kratzer & Shimoyama 2002, Schulz 2005). Barker (2010), however, dismisses this hypothesis based on the observation that free choice inferences cannot be cancelled (see also Aher 2012 for a similar argument). Someone who says (34) cannot continue with 'Although in fact you may not eat an apple'. The foregoing discussion shows that this argument is fallacious: the observation that free choice inferences are not cancellable does not preclude them from being a kind of conversational implicature.

Cancelling the free choice inference that the listener is allowed to eat an apple raises the question of why the speaker did not just say 'You may eat a pear'. The most plausible explanation is that the speaker first did not realise that the listener is allowed to choose between eating an apple and a pear, but found this out afterwards. Indeed, it seems possible to cancel free choice inferences in these special circumstances:

(35) You may eat an apple or a pear, I cannot remember which one—let me ask someone. Alright, you may eat an apple but not a pear.

In normal circumstances, however, there is no plausible explanation for saying (34) and then cancelling the free choice inferences with 'Although in fact you may not eat an apple'. It is therefore possible that free choice inferences are conversational implicatures despite being uncancellable in most situations.

In summary, we have discussed three possible diagnostics for recognising conversational implicatures: calculability, nondetachability, and cancellability. The calculability criterion states that conversational implicatures can be worked out on the basis of the literal meaning of an utterance and the assumption of cooperativity. Calculability is a necessary but not a sufficient condition: all conversational implicatures are calculable even though not all calculable propositions are conversational implicatures. Nondetachability turns out to be a nonstarter because it is circular. It presupposes that it is possible to delineate the literal meaning of an utterance. Cancellability offers a further guide for deciding if a calculable proposition is a conversational implicature. When evaluating whether a cancellation attempt is successful, however, it is important to recognise that conversational implicatures serve to explain the speaker's behaviour. Several extralinguistic factors can therefore influence the ease with which such an explanation can be cancelled.

In the next section, I return to the case of quantity inferences and discuss how pragmatic accounts explain them as a variety of conversational implicatures.

1.3.3 Quantity implicatures

Speakers share information by what they say, but also by what they do not say. The dialogue in which B says that C lives somewhere in the South of France, repeated below, provides an example:

(36) A: Where does C live.
B: Somewhere in the South of France.

By not providing more specific information, B informs the listener that he does not know exactly where C lives. Theorists agree that this inference is a conversational implicature whose derivation involves reasoning with the first maxim of Quantity, which states that the speaker should provide as much information as is required for the purpose of the conversation. For that reason, these inferences are referred to as *quantity implicatures*.

According to the pragmatic account, quantity inferences are also a kind of quantity implicature. To illustrate, consider:

(37) I ate some of the cookies.

The listener might infer that B did not eat all of the cookies. This scalar inference can be calculated as a conversational implicature by reasoning as follows: B could have been more informative if he had used the alternative 'I ate all of the cookies'. Given that the speaker tries to be as informative as possible and that the alternative would have been both more informative and relevant for the purpose of the conversation, why didn't the speaker use the alternative? The most plausible explanation is that he does not believe he ate all of the cookies. This ignorance inference can be strengthened on the assumption that the speaker knows whether or not he ate all of the cookies, in which case it follows that the speaker believes that he did not eat all of the cookies. An analogous explanation can be given for the other kinds of quantity inferences discussed in section 1.2.2.

The pragmatic account has several desirable properties: it bases its explanation of quantity inferences on general principles of cooperativity; it uses a single mechanism to compute ignorance inferences and quantity inferences; and it explains in a straightforward manner why quantity inferences are cancellable. Despite these advantages, several theorists have posited conventionalist alternatives to the pragmatic account. The main tenet of these conventionalist accounts is that quantity inferences are part of the literal meaning of utterances rather than being derived on the basis of pragmatic reasoning.

1.4 The conventionalist account

1.4.1 The lexicalist account

Current lexical accounts have been developed to explain one specific kind of quantity inferences: scalar inferences. Consider the following sentence once again:

(38) I ate some of the cookies.

Someone who utters this sentence implies that he did not eat all of the cookies. Lexical accounts argue that the upper-bounding inference is caused by scalar expressions being ambiguous between a lower-bounded and a circumbounded meaning. For example, 'some' is ambiguous between meaning 'at least some and possibly all' and 'at least some but not all' much like 'bank' can mean 'side of a river' or 'financial institution' (Ariel 2004, Chierchia 2004, Grodner et al. 2010, Levinson 2000, Storto & Tanenhaus 2005).

The main problem with the lexical account is its limited scope. Other kinds of quantity inferences cannot be explained as the result of a lexical ambiguity. Consider the case of exhaustivity inferences:

(39) A: Who was at the door.
B: John.

In most situations, B's utterance implies that only John was at the door. Nonetheless, it is implausible to conclude from this observation that 'John' must be ambiguous between meaning 'John and perhaps someone else' and 'John and no-one else'.

Grammatical accounts provide a more general mechanism for explaining quantity inferences as semantic pieces of information.

1.4.2 The grammatical account

Grammatical accounts of quantity inferences assume that underinformative sentences are ambiguous between different syntactic structures (e.g., Chierchia 2004, 2006, Chierchia et al. 2012, Fox 2007, Landman 1998). These accounts introduce a covert syntactic operator $\mathbb{O}$ that can be appended to any sentence node in the syntactic tree. The meaning of this operator is essentially that of overt 'only': given a set of alternatives, it excludes all those that are at least as informative as the proposition expressed by the sentence without the operator (Geurts 2010, 144):

(40) $\mathbb{O}(\varphi)$ is true iff φ is true and all alternatives not entailed by φ are false.

Appending this operator to (38) yields the proposition that the speaker ate some of the cookies and the negation of the proposition that the speaker ate all of them. This interpretation can be paraphrased as 'I ate only some of the cookies'.

As emphasised by Magri (2011), the stipulation of covert syntactic operators within grammatical accounts raises a recoverability problem: under what circumstances is the $\mathbb{O}$ operator appended to a sentence? Grammaticalist authors have variously proposed that the operator is appended…

i. to all sentences. (Chierchia 2004)
ii. if it yields a more informative sentence. (Chierchia et al. 2012)
iii. if it does not yield a less informative sentence. (Chemla & Spector 2011)
iv. if the ignorance inference is implausible. (Fox 2007)
v. if the utterance would otherwise be contradictory. (Sauerland 2004a)

There is no consensus among grammaticalists which of these assumptions is correct. Perhaps the most dominant account of grammaticalism assumes that $\mathbb{O}$ is appended whenever this results in a more informative sentence (Chierchia et al. 2012). Most arguments that pertain to this brand of grammaticalism are also relevant for grammatical accounts that adhere to *i* and *iii*. Grammatical accounts that adopt position *v* are indistinguishable from pragmatic accounts, as will become apparent in the next section. The predictions of grammatical accounts that adhere to *iv* have yet to be specified.

1.5 A comparison

Conventionalist accounts have a number of a priori disadvantages compared to their pragmatic competitors. One such disadvantage is that the operations needed to compute quantity inferences are stipulative instead of being derived from general considerations of cooperative behaviour. Conventionalist accounts also lack an explanation for why quantity inferences tend to be cancellable.

A more fundamental issue concerns the role of ignorance inferences. Pragmatic accounts assume that scalar inferences are computed in two steps. First, the listener arrives at an ignorance inference, which, if the speaker is assumed to be competent, is then strengthened to a scalar inference. Conventionalist accounts propose that scalar inferences are computed directly. This means that an additional mechanism is required to explain the existence of ignorance inferences. Fox (2007) and Chierchia et al. (2012) assume that ignorance inferences are computed by means of pragmatic reasoning. But this proposal seems to import a measure of redundancy into their theories: if an ignorance inference is computed by means of pragmatic reasoning, and the competence assumption is plausible, a scalar inference can logically be inferred without having to assume that the corresponding sentence is ambiguous. Most conventionalist theorists do not explicitly address this redundancy. Two possible ways of doing so suggest themselves:

- The pragmatic route to scalar inferences is somehow closed off. Even though it is in principle possible to compute the scalar inference by means of pragmatic reasoning, listeners do not do this in practice.
- Listeners use both routes to compute scalar inferences. If possible, a scalar inference is computed pragmatically. If not, it is computed through ambiguity resolution. This seems to be the option advocated by Sauerland (2010).

Most conventionalists implicitly propagate the first option. Therefore that is the position I attribute to conventionalism in general.

Despite these issues, conventionalists have argued that there are pressing reasons for assuming that quantity inferences are part of the literal meaning of utterances. The first of these concerns the behaviour of scalar expressions under embedding. If scalar inferences are part of the literal meaning of an utterance, scalar expressions are predicted to be interpreted with an upper bound in the scope of embedding expressions like negations, conditionals, and quantifiers. Although the examples in (8) suggest that this prediction is mistaken, the matter turns out to be more subtle than that because there are different mechanisms that variously underwrite scalar inferences, as I explain in the next section.

In sections 2.2 and 2.3, I review the experimental data on the interpretation of ‘some’ embedded under universal and non-monotone quantifiers. I demonstrate that there are at least three related but distinct meaning aspects that affect the interpretation of scalar expressions. Once the role of these meaning aspects is taken into account, the experimental record offers no evidence that ‘some’ is interpreted as ‘some but not all’ when it is embedded under universal or non-monotone quantifiers. In section 2.4, I extend the argument to free choice inferences. Although often explained as a kind of quantity inferences, Chemla (2009) provides evidence that free choice inferences interact with the meaning of embedding operators. I argue that this observation poses a problem for both pragmatic and conventionalist accounts.

The second difference between pragmatic and conventionalist accounts focuses on the processing of scalar expressions. Bott & Noveck (2004) found that the computation of scalar inferences is time-consuming. But it is unclear from their experiments at which stage this processing cost is incurred. Conventionalist accounts predict that it occurs during the parsing stage when the scalar inference is computed. Pragmatic accounts, on the other hand, predict a post-parsing processing cost. In chapter 3, I tease these predictions apart on the basis of the results from a sentence-picture verification task.

2

‘Embedded implicatures’

2.1 Introduction

2.1.1 What’s the issue?

The goal of this thesis is to determine the position of quantity inferences in the semantics/pragmatics spectrum. To address this issue, I first investigate the interpretation of scalar expressions in embedded environments. As I explain presently, different views on the nature of scalar inferences entail different predictions about the interpretation of embedded scalar expressions.

In the following sentence, the scalar expression ‘some’ is embedded in the antecedent of a conditional:

(1) If she eats some of the cookies, her mother will be furious.

Pragmatic accounts predict that ‘some’ in this sentence will be interpreted as ‘at least some and possibly all’. So her mother will be furious even if she eats all of the cookies. As noted in section 1.2.2, a scalar inference is absent because ‘some’ occurs in a downward-entailing environment, which means that replacing ‘some’ with ‘all’ results in a less informative sentence. Hence the speaker did not use the alternative because he tries to be as informative as possible.

A critical assumption in this explanation is that listeners do not reason about the content of embedded sentences. For example, (1) contains the embedded sentence ‘…she eats some of the cookies…’. If it were possible to reason in terms of the conversational maxims about this sentence, the listener might come to the conclusion that ‘…she eats all of the cookies…’ is false and consequently that her mother will not be furious if she eats all of the cookies.

This kind of reasoning about embedded sentences is not possible within a pragmatic framework because the conversational maxims are tailored to full-fledged assertions. To illustrate, consider the maxim of Quality, which exhorts the speaker to make their contribution a truthful one. It is clear that someone who utters (1) is not committed to the claim that 'she eats some of the cookies' is true. Although the speaker includes the sentence in his utterance, he does not intend to communicate its content. This also holds for the other maxims: in cases like (1), using the less informative sentence 'she eats some of the cookies' leads to a more informative global utterance. Most pragmatic theorists therefore assume that quantity implicatures can only be computed by reasoning about whole utterances.[1]

Conventionalist accounts, on the other hand, predict that it is possible to interpret 'some' as 'some but not all' in sentences like (1). In lexical varieties of conventionalism, 'some' can be disambiguated as 'some but not all' regardless of whether it occurs in an embedded or unembedded environment. In grammatical varieties of conventionalism, the $\mathbb{O}$ operator can be appended to any sentence node in the syntactic tree, including embedded sentence nodes. Whether and where an $\mathbb{O}$ operator is predicted to occur in the syntactic tree of (1) depends on the constraints on the distribution of this operator, as discussed in section 1.4.2.

For cases such as (1), the traditional pragmatic account seems to make the correct predictions. But there are also sentences in which embedded scalar expressions are interpreted with an upper bound. The following sentence is a case in point. (In this and the following examples, small caps denote prosodic marking.)

(2) If she eats SOME of the cookies, her mother will be furious; but if she eats ALL of them, her mother might not even notice.

If 'some' is interpreted as 'at least some and possibly all', this statement is contradictory because in that case the first sentence implies that her mother will be furious if she eats all of the cookies, which is then denied in the second sentence. In order to resolve the contradiction, it has to be assumed that 'some' is interpreted as 'some but not all' and thus licenses the embedded scalar inference that her mother will be furous if she eats some but not all of the cookies. These embedded scalar inferences are sometimes referred to as 'embedded implicatures'. But the foregoing discussion makes it clear that this is a misnomer: conversational implicatures involve reasoning about whole utterances rather than embedded sentences. So the presence of 'embedded implicatures' indicates that what might seem to be a conversational implicature is actually an aspect of the literal meaning of the utterance.

In summary, then, there are utterances that license embedded scalar inferences

[1] See Simons (2010, 2014) for a dissenting point of view. She argues that certain conversational principles, such as the maxim of Quantity, also pertain to embedded sentences. This move, however, marks a major departure from the traditional pragmatic account that I discuss here. The proposal is also fraught with several complications that have not been properly discussed or addressed. For one, using a less informative embedded sentence in downward-entailing environments results in a more informative global sentence. In order to accommodate such dependencies, one would need to create much more complex conversational principles than simply extending the maxim of Quantity to embedded clauses.

and utterances that do not. Embedded scalar inferences cannot be explained as quantity implicatures. But this does not mean that, as a matter of principle, pragmatic accounts cannot account for these inferences. According to pragmatic accounts, embedded scalar inferences are caused by a process of *truth-conditional narrowing* (Geurts 2010, Horn 2006). Truth-conditional narrowing is a general process that is not restricted to scalar expressions. To illustrate, consider the following sentences (Wilson & Carston 2007, 235):

(3) a. I'm not drinking tonight.
  b. Buying a house is easy if you've got money.
  c. Churchill was a man.

On its literal meaning, (3a) implies that the speaker is not drinking anything tonight. But a more plausible interpretation is that the speaker is not drinking any alcoholic beverages. So the meaning of 'drinking' is narrowed to 'drinking alcoholic beverages'. In (3b), the meaning of 'money' is similarly narrowed to mean 'a substantial amount of money'. The case of (3c) is interesting because it indicates that the narrowed meaning can vary with the context: 'a man' can be construed either as 'an ordinary man' or 'an ideal man' depending on the situation.

Pragmatic accounts argue that a similar process of truth-conditional narrowing is responsible for the upper-bounded construal of scalar expressions in sentences like (2). Scalar inferences that are caused by truth-conditional narrowing differ in a number of respects from scalar inferences that are computed by means of pragmatic reasoning:

i. Narrowing is a lexical process that can affect scalar terms in embedded clauses.
ii. The narrowed meaning of a scalar expression is part of its literal meaning.
iii. Narrowing of scalar expressions is marked by prosodic cues or by a contrastive relation between scalar expressions. In the case of (2), for example, there is a contrastive relation between 'some' and 'all'. In the absence of these cues, embedded scalar expressions are not narrowed.

Conventionalists, on the other hand, analyse cases like (2) on a par with regular scalar inferences. So lexical accounts assume that 'some' is disambiguated as 'some but not all', and grammatical accounts propose that an $\mathbb{O}$ operator is appended to the embedded sentence '…she eats some of the cookies…'. Observe that the latter explanation is incompatible with grammatical accounts according to which the $\mathbb{O}$ operator is appended only if it leads to a more informative sentence, since the resulting interpretation is less informative than the interpretation without an operator. In order to salvage such an account, one would need to stipulate further constraints on the distribution of the $\mathbb{O}$ operator; for example, that the operator is also appended in order to avoid contradictions.

To summarise, pragmatic and conventionalist accounts agree about the limiting cases of truth-conditional narrowing and ignorance inferences (Bach 1994, Chierchia et al. 2012, Fox 2007, Geurts 2009, 2010, Horn 2006, 2009, Ippolito 2010, Noveck & Sperber 2007, Recanati 2003). Truth-conditional narrowing, as exemplified in (3), determines the literal meaning of an expression, whereas ignorance inferences are the result of reasoning in terms of the conversational maxims. The debate between the two kinds of theories focuses on the position of scalar inferences relative to these limiting cases. Conventionalist accounts analyse all scalar inferences as resulting from a process similar to that of truth-conditional narrowing. According to pragmatic accounts, unmarked scalar expressions acquire a scalar inference on the basis of pragmatic reasoning. But when a scalar expression is marked by means of prosodic stress or a contrastive relation, the double-sided meaning can become part of its literal meaning.

We have seen that embedded scalar inferences cannot be explained in terms of pragmatic reasoning. Since, furthermore, pragmatic accounts assume that the alternative route to scalar inferences—truth-conditional narrowing—is only available if the scalar expression is marked, the interpretation of unmarked scalar expressions in embedded environments offers a means to distinguish between pragmatic and conventionalist theories. Pragmatic accounts predict that unmarked embedded scalar expressions are not interpreted with an upper bound, whereas conventionalist accounts predict embedded scalar inferences to occur “systematically and freely in arbitrarily embedded positions” (Chierchia et al. 2012, 2297).

In the remainder of this section, I discuss a number of arguments that have been advanced in support of either pragmatic or conventionalist accounts based on the behaviour of scalar expressions under embedding. Three other arguments will receive a more extensive discussion in sections 2.2, 2.3, and 2.4.

2.1.2 A brief discussion of some of the arguments

‘Believe’

In his defense of a grammatical account, Chierchia (2004) provides several examples to show that scalar inferences can occur in embedded positions. One of these involves sentences in which a scalar expression is embedded under the verb ‘believe’:

(4) John believes that some of his students are waiting for him.

I will formalise this sentence as ‘$\textsc{bel}_{\text{John}}$(some of his students are waiting for him)’. Chierchia intuits that this sentence sometimes implies that, according to the speaker, John believes that not all of his students are waiting for him (i.e., ‘$\textsc{bel}_{\text{S}}$ $\textsc{bel}_{\text{John}}\neg$(all of his students are waiting for him)’).

Conventionalist accounts can straightforwardly explain this interpretation: lexicalist accounts assume that ‘some’ is disambiguated as ‘some but not all’; according to a grammatical account, the $\mathbb{O}$ operator is appended to the embedded clause ‘some of his students are waiting for him’ which yields a more informative sentence than either appending the operator to the main clause or not appending it at all. On the other hand, pragmatic accounts seem to predict a weaker reading according to which the speaker does not believe that John believes that all of his students are waiting for him (i.e., ‘$\textsc{bel}_{\text{S}}\ \neg\textsc{bel}_{\text{John}}$(all of his students are waiting for him)’). This interpretation is computed by reasoning as follows (where ‘$\varphi[x]$’ stands for ‘x of his students are waiting for him’).

i. The speaker said ‘$\textsc{bel}_{\text{John}}\varphi[\text{some}]$’.
ii. He could have been more informative with ‘$\textsc{bel}_{\text{John}}\varphi[\text{all}]$’. Why didn’t he?
iii. Presumably because he does not believe it is true. $\neg\textsc{bel}_{\text{S}}\ \textsc{bel}_{\text{John}}\varphi[\text{all}]$
iv. The speaker is competent. $\textsc{bel}_{\text{S}}\ \textsc{bel}_{\text{John}}\varphi[\text{all}] \vee \textsc{bel}_{\text{S}}\ \neg\textsc{bel}_{\text{John}}\varphi[\text{all}]$
v. It follows that he believes the alternative is false. $\textsc{bel}_{\text{S}}\ \neg\textsc{bel}_{\text{John}}\varphi[\text{all}]$

While this interpretation might be adequate in some situations, most authors agree that there are also situations in which the stronger interpretation is called for.

This challenge to pragmatic accounts was addressed by Russell (2006) (see also Spector 2006). According to Russell, what is missing in the aforementioned reasoning process is the observation that, according to the listener, John might be competent about the proposition that all of his students are waiting for him. In other words, it might be the case that, according to the listener, either John believes that all of his students are waiting for him, or that not all of his students are waiting for him (i.e., ‘$\textsc{bel}_{\text{John}}\varphi[\text{all}] \vee \textsc{bel}_{\text{John}}\neg\varphi[\text{all}]$’). Together with the conclusion that, according to the speaker, it is not the case that John believes that all of his students are waiting for him, this entails the desired interpretation that John believes it is not the case that all of his students are waiting for him.

One of the merits of this explanation is that it accounts for the observation that whether (4) implies that John does not believe all of his students are waiting for him depends on whether John is assumed to be competent about the proposition that all of his students are waiting for him. If he is not, the weaker interpretation arrived at by standard pragmatic reasoning might be adequate. In summary, then, the behaviour of scalar expressions under ‘believe’ is compatible with pragmatic and conventionalist accounts, first impressions notwithstanding.

Disjunction

A second example Chierchia discusses in support of his grammatical account involves disjunctive sentences:

(5) Mary is either working at her paper or seeing some of her students.

Chierchia notes that someone who says (5) might convey that Mary did not see all of her students. Conventionalist accounts can explain this interpretation without problems. If ‘some’ is disambiguated as ‘some but not all’, (5) entails that Mary did not see all of her students—assuming that ‘or’ receives an exhaustive interpretation. Similarly, using a grammatical account, the $\mathbb{O}$ operator can be appended to the disjunct ‘Mary is seeing some of her students’, which is then interpreted as ‘Mary is seeing some but not all of her students’.

According to Chierchia, pragmatic accounts provide the wrong predictions for disjunctive sentences with a scalar expression in one of the disjuncts. He assumes that pragmatic accounts can only generate alternatives by substituting scalar expressions, and that the scalar expressions ‘or’ and ‘some’ evoke the scales ⟨or, and⟩ and ⟨some, all⟩. Substitution then generates the following alternatives:

(6) a. Mary is either working at her paper or seeing all of her students.
b. Mary is either working at her paper and seeing some of her students.
c. Mary is either working at her paper and seeing all of her students.

The main problem is (6a). Pragmatic reasoning licenses the inference that the speaker does not believe that this alternative is true, and, if the competence assumption is warranted, the inference that he believes that it is false. But the negation of (6a) implies that Mary is not working at her paper, which is manifestly not implied by someone who says (5).

The solution to this conundrum, from a pragmatic perspective, is to recognise that alternatives are not just generated by substituting scalar expressions. Katzir (2007), for example, proposes that alternatives can also be generated by deleting instead of replacing lexical items. Importantly, this mechanism generates, besides the already mentioned alternatives in (6), the following alternatives to (5):

(7) a. Mary is working at her paper.
b. Mary is seeing some of her students.
c. Mary is seeing all of her students.

The second ingredient needed to deal with sentences like (5) is to assume a stepwise procedure for reasoning with alternatives (cf. Sauerland 2004b). Given an utterance, the listener first computes ignorance inferences for all alternatives. Afterwards, he strengthens those alternatives for which the competence assumption holds. This procedure entails, for example, that the ignorance inference associated with (7a) cannot be strengthened to a scalar inference because that would entail that the speaker believes that (7b) is true, which is inconsistent with the previously computed ignorance inference for this alternative.

Similarly, the ignorance inference associated with the problematic alternative (6a) cannot be strengthened because it would contradict the ignorance inference associated with (7b). The ignorance inference that the speaker does not believe that (7c) is true, however, can be strengthened without contradicting any of the previously computed ignorance inferences, yielding the desired inference that, according to the speaker, Mary is not seeing all of her students.

Scalar expressions embedded under disjunction pose a problem for pragmatic accounts only if it is assumed that alternatives are generated by substituting scalar expressions. But it seems plausible to assume that listeners also take into account relevant statements that are shorter than the speaker's utterance and therefore do not involve substitution. Once this assumption is made, the problem with scalar expressions in disjunctions dissipates.

Hurford's constraint

Another issue with disjunctions involves *Hurford's constraint*. Hurford (1974) noted that sentences like the following are infelicitous:

(8) a. John is an American or a Californian.
    b. That painting is of a man or a bachelor.
    c. The value of x is greater than or not equal to 6.

Based on this observation, Hurford concluded that disjunctive statements are infelicitous if the disjuncts stand in an entailment relation. In the case of (8a), for example, the disjunct 'John is a Californian' entails 'John is an American'. A plausible explanation for this generalisation is that someone who uses a disjunction implies that he does not know which of the disjuncts is true: if the disjuncts stand in an entailment relation, however, the speaker's utterance entails that the weaker disjunct is true. Chierchia et al. (2012) put forward the following example that seems to contradict Hurford's constraint:

(9) Mary read SOME or ALL of the books.

In this example, the second disjunct entails the first one. Nonetheless, this sentence is much more felicitous than the ones in (8). Hurford's constraint can be salvaged, according to Chierchia et al., if it is assumed that an $\mathbb{O}$ operator is appended to the first disjunct and that as a consequence 'some' is interpreted as 'some but not all'. Lexicalist accounts potentially offer a similar explanation by assuming that 'some' is disambiguated as 'some but not all'.

Chierchia et al. conclude that examples like (9) speak against a pragmatic view: conversational implicatures are computed globally so it is impossible to locally narrow the meaning of 'some' as 'some but not all'. Therefore, according to Chierchia et al., pragmatic accounts cannot explain why the sentences in (9) are felicitous but the sentences in (8) are not.

As Sauerland (2012) observes, this argument is based on the mistaken assumption that pragmatic accounts cannot account for embedded scalar inferences. As noted in section 2.1.1, pragmatic accounts assume that there is another method for generating upper-bounding inferences: truth-conditional narrowing. So pragmatic accounts might argue that the meaning of 'some' in (9a) is narrowed to 'some but not all', thus salvaging Hurford's constraint without adopting a conventionalist paradigm. In support of this explanation, there is a contrastive relation between the expressions ‘some’ and ‘all’. Examples like those in (8) remain infelicitous because these involve expressions whose meanings cannot straightforwardly be narrowed to exclude the stronger disjunct: for example, prosodically emphasising ‘man’ does normally not lead to an interpretation whereby it excludes ‘bachelor’.

However, Sauerland also provides a number of examples that seem unamenable to the same pragmatic explanation:

(10) a. Either every student solved MOST of the problems, or every student solved ALL of them.
b. Either she must read at least THREE of the books or she must read at least FOUR of them.

I will restrict my discussion to (10a). First of all, note that the literal interpretation of this sentence violates Hurford’s constraint, since if every student solved all of the problems, it follows that every student solved most of them. Pragmatic accounts thus have to assume that ‘most’ is interpreted as ‘most but not all’, which means that the sentence is false if some students solved most of the problems, and the rest solved all of them. According to Sauerland, this interpretation is inadequate. He intuits that the sentence is true in such a situation. This intuition can be explained by assuming that the $\mathbb{O}$ operator is appended to the disjunct ‘every student solved MOST of the problems’, which is then interpreted as ‘every student solved most of the problems, but not every student solved all of them’. The resulting interpretation of the whole disjunction is then compatible with a situation in which some students solved most of the problems and the rest solved all of them.

First of all, note that Sauerland’s solution is at odds with the view that the $\mathbb{O}$ operator is appended whenever it leads to a more informative statement. On that view, it should also be appended to the embedded sentence ‘x solved most of the problems’, which would generate an interpretation that is equivalent to the one predicted by pragmatic accounts. Indeed, it seems difficult to provide a principled account of the distribution of the $\mathbb{O}$ operator in order to ensure that it is appended to the first disjunct but not to the sentence that is embedded under ‘every’.

Even if such a principled account were given, Sauerland’s argument stands in need of empirical evidence. Specifically, two assumptions in his argument require empirical underpinning: first, that (10a) is true in a situation in which some students solved most of the problems, and the rest solved all of them, and, second, that sentences like (10) are more felicitous than their counterparts in (11):

(11) a. Everyone in the room is an American or a Californian.
b. All of the paintings are of a man or a bachelor.
c. The values of x are greater than or not equal to 6.

If the second assumption is unwarranted, it might be argued that the force of Hurford's constraint is diminished if the disjuncts become structurally more complex. In that case, (10a) might in fact be a violation of Hurford's constraint, and it is unnecessary to append the 𝕆 operator to salvage this constraint.

A fundamental issue that besets all of these arguments is that it is not at all clear how general Hurford's constraint is. A cursory search on the internet already provides countless naturally occurring and intuitively felicitous disjunctive sentences in which the disjuncts stand in an entailment relation:

(12) a. The way you read a book or a text depends on your reasons for reading it.
b. You don't have to be a genius to be a politician or a president.
c. Becoming a doctor or a surgeon is an honorable goal.
d. Stay home if you are feverish or sick.
e. In the United States today, only the Maryland Court of Appeals wears scarlet or red robes when hearing arguments.

In summary, further theoretical and empirical work is needed to establish if Hurford's constraint provides evidence in favour of a conventionalist account of scalar inferences.

Contrast

The discussion of Hurford's constraint shows that some of the apparent counterarguments against a pragmatic account of scalar inferences can be explained by invoking the notion of truth-conditional narrowing. This also applies to the following examples:

(13) a. Exactly three students did MOST of the exercises; the rest did them ALL.
b. It is not just that you CAN write a reply. You MUST.
c. If you take salad OR dessert, you pay $20; but if you take BOTH there is a surcharge.

According to Chierchia et al. (2012), these sentences indicate that one needs to postulate the 𝕆 operator and append it to the sentences that are embedded in the antecedent of a conditional, under the quantifier 'exactly three', and under negation, respectively. An alternative explanation, one that is compatible with a pragmatic account, is that the meaning of the scalar expressions is narrowed to exclude the stronger expression in the second sentence. In support of this explanation, there is in each case a contrastive relation between scalar expressions. Cases like (13) are therefore orthogonal to the debate between pragmatic and conventionalist accounts (see Asher 2012 for a more extensive discussion).

Multiple scalar expressions

Sentences containing multiple scalar expressions offer a challenge to a number of conventionalist accounts. Consider the following sentence (cf. Greenhall 2008):

(14) Some of the boys danced with some of the girls.

Someone who utters this sentence will sometimes convey that…

i. Not all of the boys danced with some of the girls
ii. None of the boys danced with all of the girls.

In other words, the utterance can imply that the following alternatives are false:

(15) a. All of the boys danced with some of the girls.
b. Some of the boys danced with all of the girls.

Pragmatic accounts can generate these alternatives without problems and therefore arrive at the desired interpretation. The situation is different for conventionalist accounts. Consider a lexical account: depending on how 'some' is disambiguated, the utterance can receive one of the following interpretations:

(16) a. At least some of the boys danced with at least some of the girls.
b. At least some of the boys danced with some but not all of the girls.
c. Some but not all of the boys danced with at least some of the girls.
d. Some but not all of the boys danced with some but not all of the girls.

None of these readings captures the desired interpretation: (16a) and (16d) imply neither *i* nor *ii*, (16b) fails to imply *i*, and (16c) fails to imply *ii*.

Sentences like (14) also challenge a number of grammatical accounts. Assuming that the $\mathbb{O}$ operator is appended whenever this results in a more informative sentence, it should be appended to the embedded sentence 'x danced with some of the girls' which is then interpreted as 'x danced with some but not all of the girls'. The $\mathbb{O}$ operator is subsequently appended to the main clause, and the sentence is interpreted as being equivalent to (16d). This interpretation is too weak, since it is compatible with a situation in which all of the boys danced with some of the girls.

To explain the interpretation of sentences like (14) within a grammatical account, it is necessary to allow for alternatives to be generated by replacing multiple scalar expressions occurring at different levels of embedding. While this might seem an innocent assumption, Fox (2007, footnote 35) argues that it has detrimental consequences for an explanation of sentences like:

(17) You're required to talk to Mary or Sue.

This sentence implies that you are not required to talk to Mary and that you are not required to talk to Sue. These scalar inferences can be computed on the basis of the alternatives:

(18) a. You're required to talk to Mary.
b. You're required to talk to Sue.

But conventionalist accounts also allow alternatives that are not more informative than the speaker's utterance to be involved in the computation of scalar inferences. This leads to the inclusion of the alternatives:

(19) a. You're allowed to talk to Mary.
b. You're allowed to talk to Sue.

The ignorance inferences associated with these alternatives block the possibility of computing scalar inferences on the basis of the alternatives in (18). Pragmatic accounts solve this conundrum by assuming that alternatives have to be more informative than the speaker's utterance (see also footnote 7). Given, however, that this option is unavailable to conventionalist accounts, these will have to stipulate that alternatives are generated by substituting at most one scalar expression, which then prohibits the generation of the correct alternatives in the case of (14).[2]

The upshot is that grammatical accounts face a dilemma: either accounting for cases like (14) by allowing alternatives to be generated by substituting multiple scalar expressions, or not doing so and thereby opening up the possibility of explaining the scalar inferences associated with sentences like (17).

Infelicitous sentences

Magri (2011) notes that sentences like (20) are somehow odd:

(20) Some Italians come from a warm country.

Presumably, the cause of this oddness is that the sentence licenses a scalar inference that not all Italians come from a warm country, which is strange since all Italians

[2] Substituting scalar expressions that occur at different levels of embedding is explicitly prohibited in the grammatical framework developed by Chierchia (2004), which generates scalar inferences incrementally starting with the most embedded sentence:

> Now, since the plan is to deal with each implicature as soon as possible, I will define α^{ALT} in such a way that it yields the alternatives induced solely by the last scalar element in the tree (i.e. the highest or topmost one). The rationale for this is that scalar terms below the topmost (if present) will have already been taken care of (by the terms below the topmost). This is a locality constraint driven by the guiding idea of our attempt (namely, that implicatures are processed locally in the order in which their triggers appear). (p. 60)

This implies that the $\mathbb{O}$ operator that is appended to the main clause only generates alternatives by substituting the first occurrence of 'some', and is therefore incapable of generating the required alternative 'Some of the boys danced with all of the girls'.

come from the same country. This example shows that the computation of scalar inferences is blind to common knowledge: even though it is known that all Italians come from a warm country, a scalar inference is computed that contradicts this piece of common knowledge. A similar observation was made by Geurts (2010) based on the following example:

(21) Cleo threw all her marbles in the swimming pool. Some of them sank to the bottom.

Even though it is highly unlikely that some but not all of the marbles sank to the bottom, this is exactly what is implied by (21). Geurts explains this blindness to world knowledge as follows:

> What these examples show is that *genuine* scalar inferences are not so easy to cancel, at all (cf. Geurts 1999). From a Gricean perspective, this is to be expected: scalar implicatures arise because the speaker goes out of his way to make a statement that is weaker than what he could have said with equal effort, so it stands to reason that it should require special circumstances to preempt a scalar implicature. But in particular, lack of plausibility will generally be insufficient for doing so. (p. 158)

According to Magri, by contrast, blindness to common knowledge poses a challenge to pragmatic accounts of scalar inferences, since pragmatic reasoning is motivated by principles of rationality, and there "seems to be nothing rational in blindness to common knowledge" (p. 17). The tacit assumption that underlies this argument is that it is more rational to ignore what is conversationally implicated than to ignore common knowledge. There are several problems with this assumption.

Consider the case of (21). If Magri is right, a rational listener should be guided by common knowledge and therefore interpret the second sentence as equivalent to 'All of them sank to the bottom'. But this interpretation leaves unexplained why the speaker used 'some' instead of 'all'. What could explain this behaviour? Either the speaker does not believe that the stronger scalar expression would have been appropriate or his behaviour is altogether irrational. Put differently, although computing a scalar inference leads to a puzzling interpretation that calls for further investigation, not doing so implies that the speaker is viewed as altogether irrational. It seems reasonable to suppose that the second option is a last resort and that it is therefore rational to choose the first one.

Magri uses the oddness of sentences like (21) as an indicator that a scalar inference is present. He subsequently investigates whether sentences with an embedded scalar expression engender a similar pattern of oddness, assuming that this would indicate that an embedded scalar inference is present. The first example he discusses goes as follows:

(22) Every year, the dean has to decide: if the college has made enough profit that

year, he gives a pay raise to every professor who has assigned an A to at least some of his students; if there is not enough money, then no one gets a raise.
a. This year, every professor who assigned an A to some of his students got a pay raise.
b. #This year, every professor who assigned an A to all of his students got a pay raise.

(Note that the sentence with 'some' is more informative than the sentence with 'all' since the scalar expression occurs in a downward-entailing environment.)

The sentence with 'all' is odd, whereas the sentence with 'some' is not. Despite Magri's suggestion to the contrary, this difference can readily be explained within a pragmatic framework. Someone who says (22b) could have been more informative by saying (22a). Why didn't he? Presumably because he does not believe that this sentence is true. Assuming that the speaker is competent, it follows that he believes that (22b) is false. However, common knowledge implies that (22a) is true whenever (22b) is. So the interpretation contradicts common knowledge, and the sentence becomes pragmatically infelicitous just like (21).

The second example seems more problematic for a pragmatic account:

(23) In this department, every professor assigns the same grade to all of his students.
a. #This year, every professor of this department who assigned an A to some of his students got a prize from the dean.
b. This year, every professor of this department who assigned an A to all of his students got a prize from the dean.

In this case, it is the sentence with 'some' that is odd. Since (23a) is more informative than (23b), this difference in oddness cannot be explained by means of a scalar inference. Magri explains it as follows: the $\mathbb{O}$ operator is appended to all sentences. Therefore it is also appended to the embedded sentence 'x assigned an A to some of his students', which is subsequently interpreted as 'x assigned an A to some but not all of his students'. The remainder of Magri's account remains implicit. Strictly speaking, the information that every professor who assigned an A to some but not all of his students got a prize from the dean is not at odds with the common knowledge assumption that every professor assigns the same grade to all of his students. Magri needs a further assumption according to which it is common knowledge that prizes are always given to at least one person. The interpretation caused by the embedded scalar inference contradicts this further assumption, and the sentence is subsequently judged infelicitous.

Magri tacitly assumes that this is the only possible explanation for the difference in oddness between (23a) and (23b), and that it therefore proves that embedded scalar inferences can occur in contexts in which the scalar expression is not marked by prosodic cues or by a contrastive relation. However, a pragmatic account is equally able to account for Magri’s observation. While Grice’s first maxim of Quantity exhorts speakers to be as informative as possible, his second maxim tells the speaker not to be too informative. Assuming that the speaker is cooperative, why did he utter the more informative (23b) rather than (23a)? Presumably, he does not believe that (23a) is adequate: that is, he does not believe that the dean awards the prize just to professors who assign an A to all of their students. This belief contradicts the assumption that professors assign an A either to none or all of their students, and the sentence becomes pragmatically odd.

One of the advantages of this explanation is that it does not entail that (23a) implies that professors who assign an A to all of their students do not receive a prize. According to Magri, this counterintuitive inference is valid, since the presence of the $\mathbb{O}$ operator causes an interpretation according to which only professors who assign an A to just some of their students receive a prize.

2.1.3 Summary

I have discussed a number of apparent problem cases that have been adduced to disprove pragmatic or conventionalist accounts of scalar inferences. The arguments against pragmatic accounts either misconstrue the pragmatic account as being unable to account for truth-conditional narrowing (e.g., Hurford’s constraint) or can be incorporated on closer inspection (e.g., scalar expressions embedded under ‘believe’). The sole argument that cannot straightforwardly be accommodated in a pragmatic account involves complex sentences that seem to violate Hurford’s constraint (cf. Sauerland 2012):

(24) Either every student solved MOST of the problems, or every student solved ALL of them.

However, this argument stands in need of further theoretical and empirical discussion. Even if it turns out that cases like (24) challenge a pragmatic account, one might wonder if this is sufficient evidence to altogether dismiss a pragmatic explanation of scalar inferences. First, conventionalist frameworks have problems of their own, in particular when it comes to providing a principled account of the distribution of the $\mathbb{O}$ operator. Second, the interpretation of ‘or’ has long been a moot point. There are a number of observations that seem to contradict the conventionalist assumption that ‘or’ is the natural language equivalent of logical disjunction. For example, Jennings (1994) provides:

(25) a. I must go now, or I’ll be late.
b. Has Beulah arrived? Or is that Myrtle’s voice?

In neither case can ‘or’ be equated to logical disjunction. Zimmermann (2000) and Geurts (2005) provide further arguments against this equivalence. These observations suggest that the behaviour of ‘or’ might not be the ideal testcase to decide between theories of scalar inferences, at least until a generalised account of its interpretation is given.

The arguments against pragmatic accounts also challenge grammatical accounts to provide a principled account of when the $\mathbb{O}$ operator should be appended to a sentence. None of the options listed in section 1.4.2 are adequate. If grammaticalists were to attempt to provide a predictive account of scalar inferences, this problem stands in need of a solution. In addition, I discussed sentences with multiple scalar expressions that present grammatical accounts with a dilemma: either accounting for these cases or for a number of problem cases noted by Fox (2007).

In the remainder of this chapter, I will discuss three more arguments that have been adduced against pragmatic accounts. In section 2.2, I focus on the interpretation of scalar expressions embedded under universal quantifiers, in section 2.3, on scalar expressions embedded under quantifiers that are neither upward nor downward-entailing such as 'exactly three', and in section 2.4, I deal with embedded free choice inferences. Unlike the previously discussed examples, the interpretation of these sentences is not straightforward. Since moreover linguists' intuitions are likely to be influenced by the theories that they advocate, naive participants have to be probed for their judgements. But, as I will point out, this change in methodology brings with it some problems of its own.

2.2 Universal quantifiers

2.2.1 Introduction

Pragmatic and conventionalist accounts differ in their predictions about the interpretation of sentences in which a scalar expression is embedded under a universal quantifier such as 'all' or 'every':

(26) All students read some of Chierchia's papers.

According to traditional pragmatic accounts, this sentence implies at most that, according to the speaker, not all of the students read all of Chierchia's papers. This interpretation can be computed as follows. (Here, '$\varphi[x]$' stands for 'All students read x of Chierchia's papers'.)

i. The speaker says '$\varphi[\text{some}]$'.
ii. He could have been more informative by saying '$\varphi[\text{all}]$'. Why didn't he?
iii. Presumably because he does not believe that it is true. $\neg\textsc{bel}_{\textsc{s}}\varphi[\text{all}]$
iv. The speaker is competent $\textsc{bel}_{\textsc{s}}\varphi[\text{all}] \vee \textsc{bel}_{\textsc{s}}\neg\varphi[\text{all}]$
v. It follows that the speaker believes the alternative is false. $\textsc{bel}_{\textsc{s}}\neg\varphi[\text{all}]$

Conventionalists predict the presence of a stronger interpretation according to which no student read all of Chierchia's papers. This interpretation can be explained most easily within a lexical framework: if ‘some’ is disambiguated as ‘some but not all’, the resulting interpretation entails that no student read all of Chierchia’s papers. Grammatical accounts according to which the 𝕆 operator is appended whenever this leads to a more informative sentence predict the same interpretation: the 𝕆 operator is appended to the embedded sentence ‘x read some of Chierchia’s papers’, which is then interpreted as ‘x read some but not all of Chierchia’s papers’. The whole sentence is subsequently interpreted as ‘All students read some but not all of Chierchia’s papers’, which entails that no student read all of them. To illustrate the different predictions, consider the following three situations. (Checkmarks indicate that student s read Chierchia’s paper c).

P1

	c_1	c_2	c_3	c_4
s_1	✓	✓	✓	✓
s_2	✓	✓	✓	✓
s_3	✓	✓	✓	✓
s_4	✓	✓	✓	✓

P2

	c_1	c_2	c_3	c_4
s_1	✓	✓	–	–
s_2	✓	✓	–	–
s_3	✓	✓	–	–
s_4	✓	✓	✓	✓

P3

	c_1	c_2	c_3	c_4
s_1	✓	✓	–	–
s_2	✓	✓	–	–
s_3	✓	✓	–	–
s_4	✓	✓	–	–

On its literal interpretation, with ‘some’ having a lower-bounded interpretation, (26) is true in all three situations. Pragmatic accounts assume that someone who utters this sentence conveys that not all students read all of Chierchia’s paper. This interpretation is only satisfied by situations P2 and P3. Conventionalist accounts propose that the sentence has a construal according to which none of the students read all of Chierchia’s papers. This interpretation is only satisfied by situation P3.

Theorists have different intuitions about which of these predictions is correct. Geurts & Pouscoulous (2009a) therefore undertook an experimental investigation, asking naive language users for their judgements. Their results suggest that listeners do not compute an embedded scalar inference, in support of a pragmatic account. In response, Clifton & Dube (2010) and Chemla & Spector (2011) provided experimental evidence that seems to support the opposite conclusion.

In the next section, I discuss the experimental record in greater detail. Afterwards, I argue that the results presented in the last two experiments are caused by typicality differences rather than scalar inferences, and that sentences like (26) do not license genuine embedded scalar inferences. I conclude with a note on the role of typicality in theoretical and experimental pragmatics.

2.2.2 The experimental record

Preliminary remarks

Geurts & Pouscoulous make use of a so-called *verification task* to test the interpretation of ‘some’ embedded under a universal quantifier. In verification tasks, participants have to provide truth judgements to sentences. Before discussing the details of the task Geurts & Pouscoulous used, it is important to understand what these truth judgements are about.

One might hypothesise that participants who provide truth judgements indicate whether or not the literal meaning of a sentence is true. It turns out that this hypothesis is mistaken. To illustrate, consider the following pair of sentences:

(27) a. The card is long.
b. The card is long or the number is even.

Theorists agree that the first sentence entails the second one. In other words, it is impossible that (27a) is true while (27b) is not. Nonetheless, participants are reluctant to judge the second sentence true in a situation in which they know the first one is true (Braine et al. 1984). A plausible explanation for this mismatch is that someone who utters (27b) communicates that one of the disjuncts must be true but also that he does not know which one. This ignorance inference cannot be reconciled with the information that the first disjunct is true.

So when giving truth judgements, participants are influenced by the *communicative content* of a sentence (Gibbs & Moise 1997, Gibbs 2002). The communicative content of a sentence includes both the literal meaning and the pragmatic inferences that are associated with an utterance of that sentence (van der Sandt 1991). Further evidence for this conclusion comes from the observation that it makes no difference whether participants are asked for truth judgements or for appropriateness judgements (Zondervan 2010).

The discrepancy between literal meaning and truth judgements is important when evaluating experimental results on the interpretation of scalar expressions. Consider the following sentence:

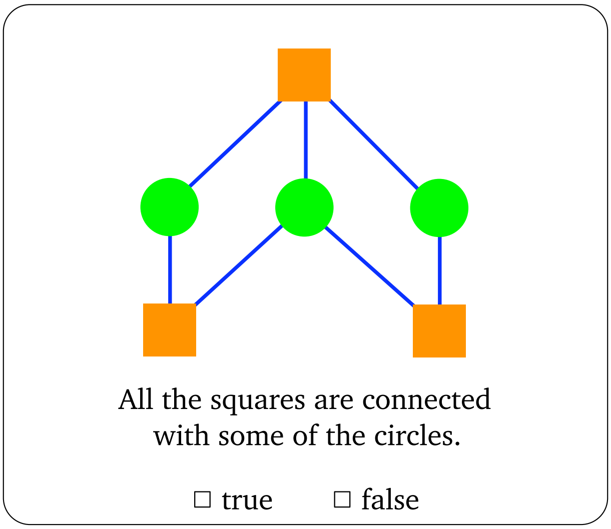
(28) Some elephants have trunks.

According to pragmatic accounts, the literal meaning of this sentence can be paraphrased as ‘At least some elephants have trunks’. Someone who utters this sentence might, in addition, implicate that not all elephants have trunks.

Participants who are presented with sentences like (28) judge them false around 60% of the time (Banga et al. 2009, Bott & Noveck 2004, Feeney et al. 2004, Noveck 2001, Pouscoulous et al. 2007, Zondervan 2010). Someone who assumes that truth judgements are purely based on the literal meaning of a sentence might construe this result as evidence that scalar inferences are part of the literal meaning of an utterance, thus contradicting the pragmatic account. The foregoing discussion indicates why this conclusion is mistaken: truth judgements are often informed by the communicative content of the sentence rather than just its literal meaning.

Geurts & Pouscoulous (2009a)

In order to determine if scalar expressions in the scope of universal quantifiers license embedded scalar inferences, Geurts & Pouscoulous presented participants with Figure 2.1. If an embedded scalar inference is computed, the target sentence implies that no square is connected with all of the circles. This interpretation is false in the corresponding picture, since the topmost square is connected with all of the circles. Therefore, if the target sentence licenses an embedded scalar inference, a substantial portion of the participants is expected to indicate that it is false in the corresponding picture, for the reasons discussed in the previous section.



All the squares are connected
with some of the circles.

□ true □ false

Figure 2.1: Experimental item used by Geurts & Pouscoulous (2009a).

As also noted in the previous section, similar verification tasks using sentences with ‘some’ in unembedded positions indicate that participants judge such sentences false around 60% of the time if the corresponding upper-bounded inference is false. In stark contrast to this result, Geurts & Pouscoulous found that not a single participant judged their target sentence false in the corresponding situation. This caused them to conclude that the target sentence normally does not license the inference that no square is connected to all of the circles. If this conclusion is correct, it invalidates one of the arguments against a pragmatic account of scalar inferences and challenges conventionalist accounts that consider a reading with an embedded scalar inference possible or even preferred.

In response, Clifton & Dube (2010) and Chemla & Spector (2011) have provided experimental evidence that seems to contradict Geurts & Pouscoulous’ findings. Both pairs of authors propose several adjustments to the verification task in order to facilitate the detection of embedded scalar inferences. One such adjustment concerns the role of contrast. In Geurts & Pouscoulous’ experiment, the target sentence occurred just once in the experiment and was presented with one target situation. In the experiments of Clifton & Dube and Chemla & Spector, by contrast, the target sentence occurred with several target situations. Moreover, in their experiments, participants were asked to indicate how well these situations were described by the sentence rather than whether or not the sentence was true. This introduced an element of contrast at the level of the target situations. Before addressing the effect of this contrast on the validity of the experimental task, I discuss the experiments by Clifton & Dube and Chemla & Spector in more detail.

Clifton & Dube (2010)

For their experiment, Clifton & Dube constructed sentences containing the scalar term ‘some’ in embedded and unembedded positions. Each sentence was presented with two target situations, and participants had to indicate which of these was best described by the sentence. In addition to the target situations, participants could choose the verbal options ‘both’ and ‘neither’. The inclusion of these verbal options is somewhat puzzling. Participants might choose ‘both’ if they cannot distinguish between the situations. But what about ‘neither’? The verbal response options seem to suggest that participants have to indicate whether the situations are correctly described by the sentence, rather than which situation is best. I come back to this interpretative ambiguity in section 2.2.8.

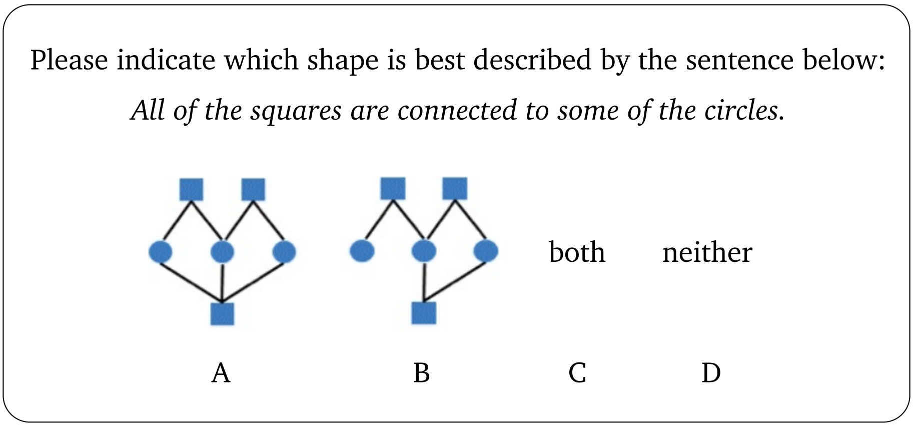
Clifton & Dube first ran this task using sentences in which ‘some’ occurred in an unembedded position:

(29) Some of the stars are in the box on the left.

As indicated, these sentences were presented with two target situations. In the first situation, the scalar inference was false; in the second situation, it was true. For example, in the first situation that was paired with (29), the box on the left contained all of the stars; in the second situation, it contained some but not all of them. When asked which of these situations was best described by the sentence, participants chose the situation in which the scalar inference was true most frequently (71%). The next most frequent choice was the verbal option ‘both’ (24%). The situation in which the scalar inference was false and the verbal option ‘neither’ were almost never chosen (3% and 2%).

Afterwards, Clifton & Dube ran the same task using sentences with ‘some’ embedded under a universal quantifier. The critical item came in two versions, one of which is illustrated in Figure 2.2. In this version, the first situation verified the scalar inference predicted by pragmatic accounts, according to which not all the squares are connected to all of the circles. The second situation, in addition, verified the embedded scalar inference predicted by conventionalist accounts.

In the second version, the first situation was the same as in the first version. In the second situation, however, all of the squares were connected to all of the circles. This situation thus falsified the scalar inference predicted by pragmatic accounts.



Please indicate which shape is best described by the sentence below:

All of the squares are connected to some of the circles.

Figure 2.2: Experimental item used by Clifton & Dube (2010).

Hence, there were three different pictures altogether: in each case, every square was connected to at least some of the circles, and in addition one of the following applied:

- All squares were connected to all of the circles. (P1)
- Some but not all squares were connected to all of the circles. (P2)
- No square was connected to all of the circles. (P3)

The results are provided in Table 2.1. In both versions of the critical trials, the verbal option 'both' was chosen most frequently (57% and 50%). In the first version, the P3 situation (= option B in Figure 2.2) was chosen in 39% of the cases, and the remaining options were almost never chosen (3% and 1%). In the second version, situation P2 (= option A in Figure 2.2) was the second most frequent choice (28%), followed closely by the verbal option 'neither' (17%). Situation P1 was chosen in 6% of the cases.

	P1	P2	P3	'both'	'neither'
Version 1	—	3	39	57	1
Version 2	6	28	—	50	17

Table 2.1: Percentages of choices for each option in Clifton & Dube's (2010) experiment. Data points claimed as evidence for embedded scalar inferences are shaded.

Clifton & Dube claim that these results prove that their participants computed an embedded scalar inference in a significant number of cases: they claim that this interpretation prevailed in 39% and 17% of the trials, respectively.

Chemla & Spector (2011)

Chemla & Spector conducted two experiments. In this section, we focus on their first experiment. Their second experiment will be discussed in section 2.3. In both experiments, participants were presented with a rating task. A trial consisted of a picture and a sentence, and participants had to indicate how "true" or "appropriate" the sentence was as a description of the picture by marking a value on a continuous scale. The critical sentence in Chemla & Spector's first experiment was:

(30) Every letter is connected to some of its circles.

This sentence was displayed with seven different kinds of situations, each of which consisted of six letters. Three of these situations are shown in Figure 2.3.

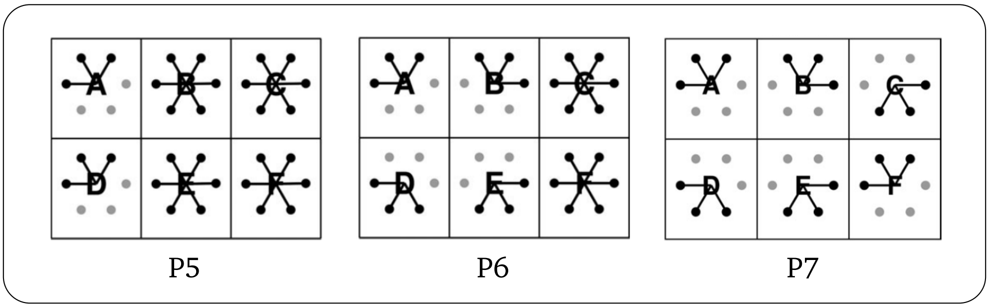


Figure 2.3: Three of the seven pictures used in Chemla & Spector's (2011) first experiment.

In each situation, a letter was connected either to none, some but not all, or all of its circles. I will refer to these as None, Mixed, and All cases, respectively. The distribution of the three kinds of letters in each situation, as well as the mean ratings participants gave to them, are provided in Table 2.2. As can be seen in the table, participants considered the target sentence less suitable in a situation with None cases than in a situation with All cases. In general, the more Mixed cases were present in the situation, the more suitable the sentence was judged. The degree of fit was the highest in the situation with only Mixed cases.

In their discussion of these results, Chemla & Spector focus their attention on the difference between the P5 and P6 conditions, on the one hand, and the P7 condition, on the other. Since P7 was the only picture that verified the target sentence with an embedded scalar inference, and its rating was significantly higher than those of the P5 and P6 items, Chemla & Spector conclude that their results support the existence of embedded scalar inferences.

Contrast

In both of these experiments, contrast plays an important role. In Clifton & Dube's experiment, participants had to choose between a limited range of options. The differences between the two versions of critical trials show that this had a significant

	P1	P2	P3	P4	P5	P6	P7
None cases	6	4	2	0	0	0	0
All cases	0	0	0	6	4	2	0
Mixed cases	0	2	4	0	2	4	6
Mean rating	0	12	24	44	63	73	99

Table 2.2: All pictures used in critical items in Chemla & Spector's (2011) first experiment, with their mean ratings (%).

effect. Although at first sight, this does not seem to hold for Chemla & Spector's experiment, two features of their design suggest that contrast played a role in their experiment, too. First, participants saw every picture up to four times. Second, Chemla & Spector presented their critical items in one continuous series, without using fillers to disguise the purpose of the experiment. These two features could have facilitated and even invited comparisons between items.

If that is what happened, how could contrast have affected Chemla & Spector's results? A natural hypothesis is that participants judged the target sentence less appropriate for a given picture only after having encountered a better instance. To test this hypothesis, we compared the mean ratings in the P5 and P6 conditions preceding any P7 items with the mean ratings for those conditions after having seen one or more P7 items. In line with the hypothesis, P5 and P6 items received higher ratings prior to P7, and the mean rating of the P6 items preceding P7 was statistically indistinguishable from that of the P7 items themselves ($W = 337.5$, p = n.s.), both being close to 100%.

	P5	P6
Trials following P7	63	73
Trials preceding P7	78	96
No contrast	93	100

Table 2.3: Mean ratings (%) for conditions P5 and P6 in Chemla & Spector's first experiment and the partial replication.

In order to corroborate this finding, we conducted a partial replication of Chemla & Spector's experiment, leaving out P7 entirely, and using P5 and P6 as the only critical situations. After three warm-up items, each of our twenty participants saw the target sentence with either P5 or P6. There was no difference at all between the P5 and P6 conditions in this experiment and the P7 condition in Chemla & Spector's (Table 2.3).

So, first appearances notwithstanding, Chemla & Spector's results are also influenced by the contrast between target situations. What consequences does this contrast have on the validity of the experimental paradigms used by Clifton & Dube and Chemla & Spector? I propose that it implies that these paradigms measure typicality differences instead of scalar inferences. Before providing evidence for this claim, I argue that Clifton & Dube's and Chemla & Spector's interpretation of their data in terms of embedded scalar inferences faces several problems.

2.2.3 Interpretation

A substantial number of participants in the experiments by Clifton & Dube and Chemla & Spector preferred sentences with 'some' in the scope of a universal quantifier to describe a situation that agreed with the embedded scalar inference over a situation that did not. Based on this finding, both pairs of authors conclude that these participants must have computed an embedded scalar inference.

But this conclusion raises the question why participants did not judge the same sentence false in a situation that directly and unambiguously contradicted the embedded scalar inference, as found by Geurts & Pouscoulous. After all, if 'some' occurs in an unembedded position, a substantial number of participants judge the sentence false in a situation in which the corresponding scalar inference is false. Additional assumptions have to be made about the effect of embedded scalar inferences on truth judgements in order to make sense of this finding.

One such explanation is proposed by Sauerland (2010). His argument takes its inspiration from the so-called *Principle of Charity* (Wilson 1959, Quine 1960, Davidson 1973). This principle can be seen as the listener's counterpart to Grice's maxim of Quality:

(31) *Principle of Charity*:
Try to interpret the speaker's utterance in such a way that it is true.

To understand Sauerland's argument, recall that Geurts & Pouscoulous' target sentence has at least three interpretations. The literal interpretation is that all squares are connected to at least some of the circles. Pragmatic accounts assume that someone who utters this sentence implies, in addition, that not all squares are connected to all of the circles. Conventionalist accounts predict a stronger interpretation according to which no square is connected to all of the circles.

According to Sauerland, the reason that participants judge the target sentence true in a situation in which its embedded scalar inference is false is that the Principle of Charity causes participants to adopt an interpretation whereby the target sentence is true, i.e., either the literal or the pragmatic interpretation.

The most pressing problem with this analysis is that it entails that scalar inferences should be systematically suppressed even in unembedded cases:

(32) Some elephants have trunks.

These sentences are true on their literal interpretation. Sauerland's analysis therefore predicts that participants adopt this interpretation and consequently judge such sentences true. This prediction is contradicted by the experimental record which consistently shows that participants judge these sentences false around 60% of the time, notwithstanding the presence of a true interpretation. So an explanation for Geurts & Pouscoulous' results is still wanting.

Chemla & Spector's interpretation faces an additional problem. Participants in their experiment considered (30) more suitable in a situation that agreed with the embedded scalar inference than in a situation that did not. According to Chemla & Spector, this suitability difference is caused by the computation of an embedded scalar inference. While this analysis accounts for the difference between the P6 and P7 situations, there were significant differences between all seven target situations. Many of these differences cannot be attributed to the computation of a scalar inference. For example, participants preferred (30) to describe the P6 over the P5 situation. Chemla & Spector concede that this difference is not caused by a scalar inference, arguing that some suitability differences are caused by scalar inferences, while others, like the difference between the P5 and P6 situation, are typicality effects. To justify this dual mechanism explanation, Chemla & Spector point at the different magnitudes of the suitability differences between the P5, P6, and P7 situations. The difference between the P6 and P7 situation (26%) is more than twice as big as the difference between the P5 and P6 situation (10%). This quantitative difference supposedly reflects the presence of two different causes.

This argument seems ad hoc: there is no a priori motivation to suppose that the magnitude of the suitability differences matters. Second and more importantly, some suitability differences are as big as the difference between the P6 and P7 situation but still do not correspond to a scalar inference. An example is the suitability difference between the P1 and P3 condition, which amounts to 24%. Chemla & Spector concede that this difference, despite its size, is not caused by a scalar inference but is a typicality effect. So their argument in favour of a dual mechanism explanation is wanting. This poses a serious problem for Chemla & Spector's interpretation, and calls for an alternative account that provides a unified explanation for the whole scale of suitability differences.

In what follows, I put forward an account that provides such a unified explanation in terms of typicality differences. Because typicality differences do not enter into the communicative content of an utterance (see section 2.2.5), participants in Geurts & Pouscoulous' verification task did not consider the target sentence false in a situation in which its embedded scalar inference was false. This explanation also accounts for the different results Clifton & Dube found for sentences with 'some' in unembedded and embedded positions. When 'some' occurred in an unembedded position, participants most often (71% of the cases) picked out the picture that verified the scalar inferences. For sentences with 'some' embedded under a universal quantifier, the option 'both' was the most popular choice (60% of

Member	Rank	Score	Member	Rank	Score
Robin	1	1.02	Duck	45	3.24
Sparrow	2	1.18	Peacock	46	3.31
Bluejay	3	1.29	Egret	47	3.39
Bluebird	4	1.31	Chicken	48	4.02
Canary	5	1.42	Turkey	49	4.09
Blackbird	6	1.43	Ostrich	50	4.12
Dove	7	1.46	Titmouse	51	4.35
Lark	8	1.47	Emu	52	4.38
Swallow	9	1.52	Penguin	53	4.53
Parakeet	10	1.53	Bat	54	6.15

Table 2.4: Sample of the typicality ratings Rosch (1975) found for different kinds of animals with respect to the category denoted by 'bird'.

the cases). This asymmetry will turn out to be compatible with the view that their results are influenced by typicality differences. Before delving into these matters, we first consider the notion of typicality.

2.2.4 What is typicality?

It is a well-established fact that not all members of a category are equally typical. Consider the category denoted by the word 'bird'. (In the following, I will denote categories with small caps.) Robins, ducks and penguins are all members of this category. But most people tend to agree that robins are more typical members than ducks, which in their turn are usually considered more typical than penguins.

Typicality differences are defined operationally on the basis of participants' direct judgements (Kahneman & Tversky 1972, Mervis & Rosch 1981). To illustrate, Rosch (1975) presented participants with a noun (e.g., 'bird') and a list of hyponyms (e.g., 'robin'). Participants had to indicate how typical the meaning of each of these hyponyms was on a seven-point scale, in which 1 denoted "highly typical" and 7 "highly atypical". This resulted in a robust typicality ordering (Figure 2.4).

Typicality structure has been demonstrated in a broad range of categories, including those denoted by adjectives (Lakoff 1973), locatives (Erreich & Valian 1979), and verbs (Fillmore 1975, Lakoff 1977). Categories that have been investigated include psychiatric classifications (Cantor et al. 1980), ad hoc categories (Barsalou 1983, 1987, 1991), perceptual categories (Rosch 1973), mathematical categories (Armstrong et al. 1983), computer programming categories (Adelson 1985), and human emotions (Fehr et al. 1982, Shaver et al. 1987).

Typicality differences have an effect on many dependent variables used in psychological research. To give a few examples, more typical members are learned earlier

(Anglin 1976, Mervis & Pani 1980, Rosch 1973, Rosch et al. 1976), recognised faster and more accurately as a member of the category (Armstrong et al. 1983, Rips et al. 1973, Rosch & Mervis 1975, Rosch et al. 1976, Sedivy et al. 1999), and produced earlier and more often when participants are asked to name members of the category (Barsalou 1983, Battig & Montague 1969, Van Overschelde et al. 2004).

There are obvious similarities between Rosch's rating task and Clifton & Dube's and Chemla & Spector's experiments. In the former, participants saw a category name with several instances and had to indicate how well these were described by the category name. In the latter, participants saw a sentence with several situations and had to indicate how well these were described by the sentence.

In fact, Chemla & Spector themselves propose that the results of their experiment are partly due to typicality differences. I propose to simplify this hypothesis: all of the evidence that Clifton & Dube and Chemla & Spector adduce for the presence of embedded scalar inferences is caused by typicality differences. In what follows, I provide substantive experimental evidence for this hypothesis. I show how it explains the whole scale of suitability differences found by Chemla & Spector (cf. Table 2.3), as well as the asymmetry between embedded and unembedded 'some' found by Clifton & Dube. But before addressing these issues, I discuss the role of typicality in communication.

2.2.5 Typicality and communication

As noted, Chemla & Spector propose that part of their results are caused by typicality effects. But they do not believe that this explanation obviates the need for postulating embedded scalar inferences:

> [If] the main factor explaining these results is the one hypothesized by the "typicality interpretation", what must the underlying metric be? More specifically, what kind of situations must be counted as "prototypical" instances of the sentence? As far as we can see, one should conclude that the best instances of the sentence among our various pictures are the ones used in the condition that receives the highest rating [i.e., P7]. We should thus be led to conclude that the best instances of the sentence are those that make the [embedded scalar inference] true. But it is hard to see how this could be so if the [interpretation with an embedded scalar inference] did not correspond to a salient reading of the sentence [...]. So the "typicality explanation", as far as we can see, would support our conclusion that the [embedded scalar inference] exists. (Chemla & Spector 2011, 381-382).

This argument can be paraphrased as follows: if a particular situation is the most typical instance of the category denoted by a sentence, then the sentence has a salient reading according to which it refers to that situation. In order to evaluate this argument, we need to establish what it means for an interpretation to be a salient reading of a sentence. One plausible criterion is that a salient reading can be part of the communicative content of a sentence. That is, someone who utters that sentence might intend to communicate that the salient reading is true and hence be held accountable if it is false.

Although it needs to be admitted that this definition is somewhat vague, it is precise enough to demonstrate that Chemla & Spector's argument is off the mark. Consider the following sentences:

(33) a. I just saw a bird in the garden.
b. I always eat vegetables with dinner.
c. The number of tulips in the vase is even.

It is clear that someone who utters (33a) does not intend to communicate that he just saw a robin in the garden and hence cannot be accused of being misleading if it turns out that he saw a pidgeon rather than a robin. The same conclusion goes for the other two sentences. Someone who utters (33b) does not intend to communicate that he always eats peas for dinner, that being the most typical instance of the category denoted by 'vegetable'. Perhaps the most striking illustration of the dissociation between typicality and communicative content involves sentences with mathematical expressions like 'even number'. Armstrong et al. (1983) found that these expressions, too, have typicality structure: two being the most typical even number. Nevertheless, it is clear that someone who utters (33c) does not communicate that there are two tulips in the vase.

In Chemla & Spector's words: there is no "salient reading" of (33c) which says that there are two tulips in the vase. Similarly, if among Chemla & Spector's situations, P7 is the best instance of the target sentence, then it does not follow that their target sentence has a salient reading according to which no letter is connected to all of its circles. So Chemla & Spector's argument is wanting.

In the following sections, I experimentally investigate the typicality structure of sentences with 'some' embedded under a universal quantifier. More precisely, I focus on the typicality structure of the categories denoted by the quantifiers 'every' and 'some'. My goal will be to show that these structures can explain the pattern of results found by Clifton & Dube and Chemla & Spector.

2.2.6 The typicality structure of EVERY

In the following discussion, I assume that the typicality structure of a category C is a fuzzy set (Goguen 1969, Lakoff 1973, Zadeh 1965, 1973). Specifically, the typicality structure of C is a function τ_{C} that associates with every element $d \in \mathbb{D}$ a number $\tau_{\text{C}}(d)$ in the left-open interval (0, 1].[3] $\tau_{\text{C}}(d)$ signifies the typicality of d in

[3] The range of typicality values is left-open, meaning that the value 0 is not an element in it. This complication is introduced for a technical reason: as will become apparent in the next section,

C. If C is of type $\langle a, t\rangle$ then all d are of type a. For example, if y is a robin and z is a parrot, then $\tau_{\text{BIRD}}(y) > \tau_{\text{BIRD}}(z)$. Sentences denote sets of situations. So 'It rains' refers to the category IT RAINS whose typicality structure $\tau_{\text{IT RAINS}}$ associates every situation $s \in \mathbb{S}$ with a value in the left-open interval (0, 1].

A classical idea in fuzzy predicate logic is that the typicality structure of a universally quantified sentence 'Every A B', equals the minimum of the typicality values of instances of A with respect to the category denoted by the predicate B:

(34) $\tau_{\text{EVERY A B}}(s) := min\{\tau_{\text{B}}(a_i)\}$, where $A = \{a_1, \ldots, a_n\}$

This definition is a straightforward generalisation of the interpretation of universally quantified sentences in predicate logic. To see whether it adequately approximates the typicality structure of a universally quantified sentence, I conducted a rating task in the style of Rosch (1975). Participants had to indicate how appropriate (35) was in eleven situations $s_0, \ldots, s_{10}$, each consisting of ten circles:

(35) Every circle is black

In every situation s_n, n circles were black and the remaining circles were white. If the minimum typicality definition is correct, situations s_0 through s_9 should receive a uniform rating that is significantly lower than the rating for s_{10}. Since participants see all eleven situations, the task involves a significant degree of contrast. In that respect it is similar to the experiments by Clifton & Dube and Chemla & Spector.

Experiment 1

Participants

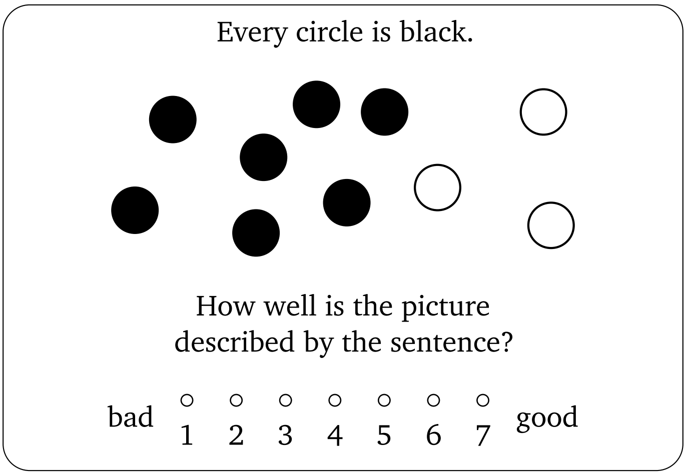
I posted surveys for 30 participants on Amazon's Mechanical Turk (mean age: 38; range: 21-61; 14 females).[4] Only workers with an IP address in the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. It turned out that all participants in this experiment were native speakers of English. None of them were excluded from the analysis.

determining the typicality of a situation with respect to a universally quantified sentence involves a division by the typicality values of the elements in that situation with respect to the predicate. This is obviously impossible if these typicality values sum to 0. Hampton (2007) argues that the range of typicality values is altogether open. This view is compatible with the proposed analysis.

[4] Mechanical Turk is a website where workers perform so-called 'Human Intelligence Tasks' (HITs) for financial compensation. It has been shown that the quality of data gathered through Mechanical Turk equals that of laboratory data (Buhrmester et al. 2011, Schnoebelen & Kuperman 2010, Sprouse 2011).

Materials

Each trial consisted of the target sentence (35) and a situation consisting of ten circles. There were altogether eleven situations $s_0, \ldots, s_{10}$. In every situation s_n, n circles were black and the remaining circles were white. Participants were instructed to indicate how well the sentence described each situation by marking a value on a seven-point scale. An example trial is shown in Figure 2.4. No fillers were included in the task. Ten randomised lists of items were prepared.



Every circle is black.

How well is the picture
described by the sentence?

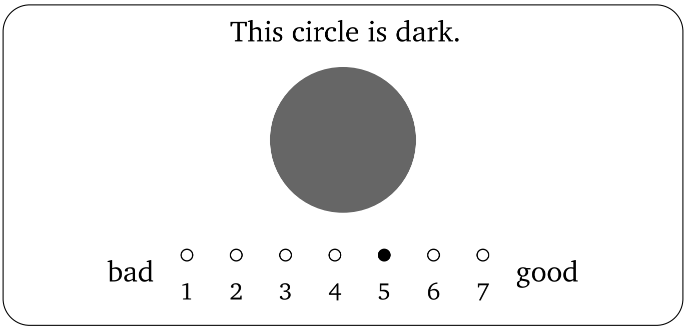
bad ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 good

Figure 2.4: Sample item used in Experiment 1

Procedure

In order to collect typicality judgements, participants were presented with the following instructions, which are based on the instructions used by Rosch (1975):

> This experiment is about how sentences are interpreted. Consider the sentence "This is a vehicle". Many people agree that this sentence is a better description of a car or a motorbike than of a sled or a tractor, even though they are strictly speaking all vehicles. Below is another example.



This circle is dark.

bad ○ 1 ○ 2 ○ 3 ○ 4 ● 5 ○ 6 ○ 7 good

> In my eyes, the picture is a reasonable instance of the sentence. I can imagine worse

> instances (for example a white circle) but I can also imagine better instances (for example a black circle). For that reason, I gave a rating that is in between the two extremes 1 and 7. However, the exact rating is a matter of taste and you might want to give a higher or lower rating. In this experiment, you will see one sentence with eleven pictures. For each picture, you have to indicate how well it is described by the sentence. It doesn't matter why you think that a sentence is a good or bad description of a particular picture. Just follow your intuition.

Results and discussion

The results of the EVERY experiment are summarised in Figure 2.5. The situation with ten black circles received the highest mean rating. The mean ratings of the other situations decreased with the number of white circles that they contained. Participants' responses were highly consistent (Cronbach's α = .982).

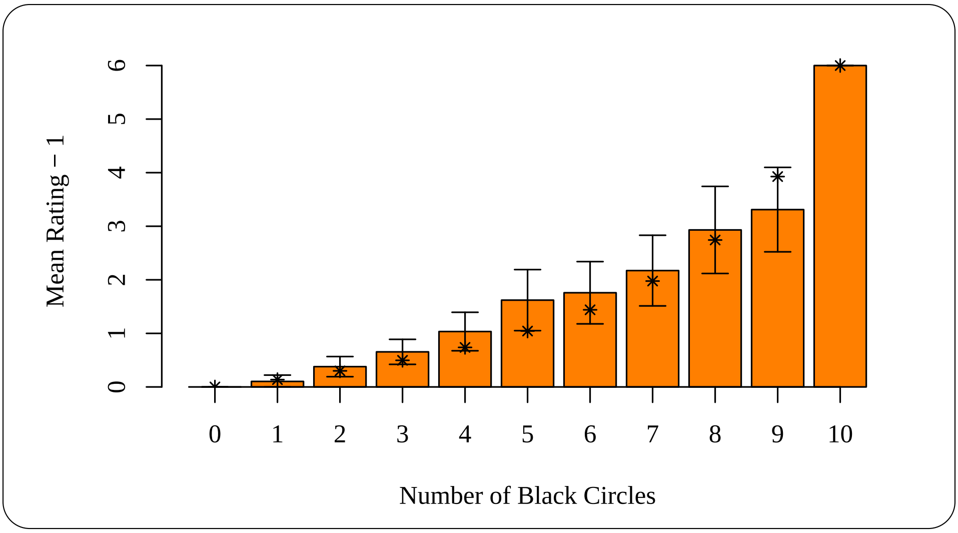


Figure 2.5: Mean typicality rating for the sentence 'Every circle is black' in situations with ten circles. The error bars represent 95% confidence intervals. The stars represent the scaled typicality ratings predicted by the definition in (37).

The minimum typicality definition in (34) is a poor fit to these data, since it predicts that every situation that contains a white circle should receive the same rating. This prediction is clearly not supported by the data: the mean rating of a situation depends on the typicality value of all of the individual circles with respect to the category IS BLACK, not just on the typicality value of the least typical instance. One way of modelling this would be to take the arithmetic mean of the typicality values:

(36) $\tau_{\text{EVERY A B}}(s) := \frac{\sum \tau_{\text{B}}(a_i)}{|A|}$, where $A = \{a_1, \ldots, a_n\}$

This is already a quite reasonable model of the pattern of ratings shown in Figure 2.5. The results, however, show that bad instances exert more influence on the mean rating of a situation than good instances: whereas the situation with one white circle received a significantly lower rating for (35) than a situation with only black circles, there was almost no difference between the situations with one black circle and none. This observation can be modelled by weighing the typicality values of the individual instances in such a way that lower typicality values carry more weight than higher ones, which can be done by taking the harmonic instead of the arithmetic mean:

(37) $\tau_{\text{EVERY A B}}(s) := \frac{|A|}{\sum \tau_{\text{B}}(a_i)^{-1}}$, where $A = \{a_1, \ldots, a_n\}$

According to this definition, the typicality of a situation depends on the typicality of every element in the domain with respect to the predicate, with bad instances exerting more influence on the typicality of the situation than good instances. Depending on the precise typicality values one assigns to the black and white circle with respect to the category IS BLACK, this definition predicts typicality values that correlate almost perfectly with the mean ratings found in the experiment. The correlation exceeds $r = .93$ for almost all possible values such that $\tau_{\text{IS BLACK}}(black) > \tau_{\text{IS BLACK}}(white)$. For example, if $\tau_{\text{IS BLACK}}(black) = .95$ and $\tau_{\text{IS BLACK}}(white) = .1$, $r = .97$, $p < .001$.

Let us return to sentences in which 'some' occurs in the scope of a universal quantifier:

(38) All the squares are connected to some of the circles.

Based on the foregoing discussion, we know that the typicality of a situation with respect to this sentence equals the harmonic mean typicality of the relevant squares with respect to the category IS CONNECTED TO SOME OF THE CIRCLES. In all of the experiments, a case (i.e. a square in the experiments by Geurts & Pouscoulous and Clifton & Dube, and a letter in the experiment by Chemla & Spector) was connected to none, some but not all, or all of the circles. How typical are these instances of the aforementioned category? This clearly depends on the typicality structure of SOME, to which I now turn.

2.2.7 The typicality structure of SOME

The typicality structure of 'some' has been investigated by a number of authors. In a series of experiments, Begg (1987) found that "the preferred meaning of *some* is less than half" (p. 62). Finer-grained results were found by Newstead et al. (1987), who asked participants to fill in the blanks in sentences like the following, where the total set size N was varied between items:

(39) If some of a group of N people are male, then _____ people are male.

For a total set size of 1,000 people, Newstead et al. found that participants estimated that on average 32% of the people are male. In order to corroborate these results in a more straightforward paradigm and in order to have finer-grained results that allow for a comparison with the results for 'every', we repeated Experiment 1 using the following target sentence:

(40) Some of the circles are black.

The remainder of the experiment was the same as Experiment 1.

Experiment 2

Participants

I posted surveys for 30 participants on Amazon's Mechanical Turk (mean age: 42; range: 18-70; 23 females). Only workers with an IP address in the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. One participant was excluded from the analysis because she was not a native speaker of English.

Materials and procedure

The materials and procedure of Experiment 2 were the same as for Experiment 1. But in this case, the target sentence was (40).

Results and discussion

The results for the SOME experiment are summarised in Figure 2.6. The situation with five black circles received the highest mean rating. The mean ratings for the other situations decreased with their distance from this prototype. Participants' responses were highly consistent (Cronbach's $\alpha = .981$).[5]

First of all, the results contradict Begg's conclusion that the preferred meaning of 'some' is "less than half", since the most typical situation was one in which half of the circles were black. One explanation for this difference is that Begg's experiments involved larger set sizes than the ten circles that were involved in this

[5] Degen & Tanenhaus (2011) found similar results using a comparable task. In their experiment, participants were asked how natural (i) was in situations in which the listener received zero, one, two, five, seven, eight, eleven, twelve, or thirteen out of a total of thirteen gumballs:

(i) You got some of the gumballs.

Degen & Tanenhaus' task differed from Experiment 2 in several respects: participants in their task did not see all possible situations, and encountered the critical situations multiple times. Furthermore, Degen & Tanenhaus included other quantifiers, like 'all' and 'none', and numerals, like 'two' and 'three', in their task. Despite these methodological difference, their results are highly similar to the data presented here.

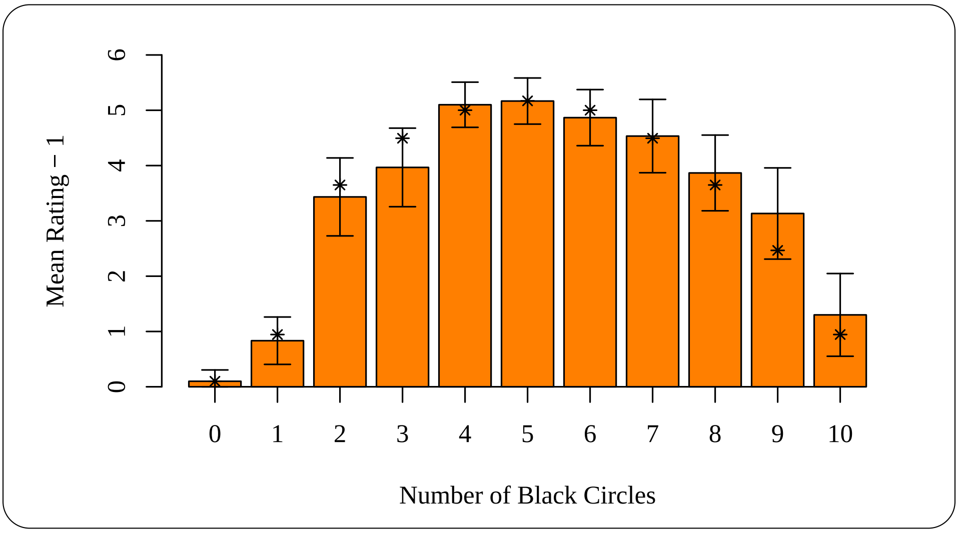


Figure 2.6: Mean typicality rating for the sentence 'Some of the circles are black' in situations with ten circles. The error bars represent 95% confidence intervals. The stars represent the scaled typicality ratings predicted by the definition in (41).

experiment. Newstead et al. (1987) found that the interpretation of 'some' was influenced by total set size: the larger the total set size the smaller the proportion that 'some' typically referred to. See section 4.2 for more discussion of this point.

Observe that the results cannot fully be explained in terms of scalar inferences. According to such an explanation, the situations with two to five black circles, as well as the situations with six to nine black circles, should have received a uniform rating. This was clearly not the case: there was a large amount of variation within these clusters. In addition, if participants interpreted 'some' as 'some but not most', the situations with six or seven black circles should have received a lower rating than the situations with four or five black circles. The former situations, however, received ratings that were statistically indistinguishable from those of the latter.

The rating for a situation decreased exponentially with its distance from the prototypical situation. In this experiment, the prototypical situation contained five black circles, but what a prototypical situation is varies across situations (Degen & Tanenhaus 2011, Newstead et al. 1987). The choice of prototype can even differ between participants, which would explain the slightly less than perfect rating for the situation with five black circles. In addition to distance from the prototype, the mean rating for a situation depended on the truth value of (39) in that situation: situations in which the target sentence was true received a substantially higher rating than situations in which it was false (i.e., the situations with zero or one black circle). The following definition captures these factors, where $v_{\text{some A B}}(s)$ denotes the truth value of 'Some A B' in s, and $dist(s, p)$ is a function of the distance between s and the prototypical situation p. In this experiment, the distance between two situations s_n and s_m equals the difference between n and m. The outcome has to be normalised to ensure that the typicality values occur in the interval (0, 1]:

(41) $\tau_{\text{SOME A B}}(s) := v_{\text{some A B}}(s) - dist(s, p)^2$

Depending on the relative importance of truth conditions compared to the effect of distance from the prototype, this model fits almost perfectly to the data. For example, assuming that the importance of truth conditions is three times as big as the effect of increasing the distance from the prototype by one, $r = .99, p < .001$.

The only motivation for the models of SOME and EVERY is that they provide a good fit to the observed ratings. The model for SOME is less fine-grained than the model for EVERY, which calculates the typicality value of a situation on the basis of the typicality values of the elements in the domain. Presumably, this can be done for SOME, too, but it would be inconsequential for my argument, which only requires the assumption that a situation in which some but not all A are B is a more typical instance of the category SOME A B than a situation in which all A are B, which in turn is more typical than a situation in which no A is B.

There is another, more fundamental difference between the two models: in the case of EVERY, the effect of increasing the distance from the prototype decreases with the distance from the prototype. In the case of SOME, the effect of increasing the distance from the prototype increases with the distance from the prototype. This is why in the definition for SOME the distance between a situation and the prototype is squared. It is possible that this difference is related to the experiment for EVERY being concerned mostly with situations in which the target sentence is false, and the experiment for SOME involving mostly situations in which it is true. For example, participants might have more pronounced typicality judgements when comparing genuine members of a category than when comparing non-members.

The typicality analysis of Experiment 2 does not preclude scalar inferences from having an effect alongside typicality. But, since typicality is defined operationally by means of participants' direct judgements, the conclusion remains that a Mixed case (i.e., a letter or a square that is connected to some but not all of the circles) is the most typical instance of the category IS CONNECTED TO SOME OF THE CIRCLES, followed by an All case (a letter or a square that is connected to all of the circles), and a None case (a letter or a square that is connected to none of the circles), in that order. In section 4.2, we explore the relationship between typicality and scalar inferences in some more depth.

2.2.8 Typicality structure in embedded scalars

Chemla & Spector

Having analysed the typicality structures of EVERY and SOME, I now return to the problem of embedded scalars. I first deal with Chemla & Spector's experiment, in which participants had to evaluate (42) in a wide range of situations (cf. Table 2.2).

(42) Every letter is connected to some of its circles.

Every situation consisted of six letters. A letter was connected either to none, some but not all, or all of its circles. As before, I refer to these as None, Mixed, and All cases, respectively. Based on the harmonic mean analysis of EVERY, it follows that the typicality of a situation s equals:

(43) $\tau_{\text{EVERY L C}}(s) = \dfrac{6}{\sum \tau_{\text{B}}(l_i)^{-1}}$, where $L = \{l_1, \ldots, l_6\}$

In this definition, EVERY L C is the category denoted by (42), and C is the category denoted by the predicate 'is connected to some of its circles'. Based on the typicality definition of SOME, it follows that:

(44) $\tau_{\text{C}}(\textit{Mixed}) > \tau_{\text{C}}(\textit{All}) > \tau_{\text{C}}(\textit{None})$

In order to determine the typicality values of the seven situations, there are still a number of free parameters that have to be determined: specifically, the typicality values of the three types of letters with respect to the category denoted by the predicate. It seems plausible that different participants converge on different typicality values for these cases. Therefore I conducted a Monte Carlo simulation. Monte Carlo investigations simulate actual conditions that contain some random element (e.g., Hammersly & Handscomb 1964).

I randomly generated 5,000 values for each of the three cases such that every triplet obeyed the constraint in (44). For each triplet that was generated, I calculated the typicality values for the seven situations based on the definition in (43). Ultimately, I derived the means of these values for comparison with the results of Chemla & Spector. The product-moment correlation between the mean typicality values in the Monte Carlo simulation and the mean suitability values found by Chemla & Spector was nearly perfect ($r = .995$, $p < .001$). This demonstrates that their results can be explained in terms of typicality differences between the seven situations.

To sum up: Chemla & Spector's data exhibit the kind of gradient pattern that, even by their own admission, calls for an explanation in terms of typicality. Nevertheless, they insist that a small part of this pattern proves the existence of embedded scalar inferences. I have argued that Chemla & Spector's argument cuts no ice: their findings can be fully accounted for by a typicality model that is entirely motivated on independent grounds, which obviates the need for positing embedded scalar inferences.

Clifton & Dube

The explanation is similar for Clifton & Dube's results. Participants in their experiment read sentences with 'some' in embedded or unembedded position. Every sentence was paired with two pictures. In the critical trials, the sentence with 'some' in unembedded position was paired with a picture that verified the scalar inference and a picture that falsified it. The sentence with embedded 'some' was paired either with a picture that verified the embedded scalar inference (i.e., with three Mixed cases) and a picture that verified only the scalar inference predicted by pragmatic accounts (i.e., with two Mixed cases and one All case), or with the latter picture and a picture that verified only the literal meaning of the sentence (i.e., with three All cases).

In the case of unembedded 'some', participants had a strong preference for the picture that verified the scalar inference. In the case of embedded 'some', participants preferred the verbal option 'both' in both conditions. Aside from this option, participants preferred the picture that verified the embedded scalar inference over the other two pictures, and the picture that verified the scalar inference predicted by pragmatic accounts over the picture that verified only the literal meaning. This preference ordering is predicted if participants are guided by typicality differences: Mixed cases are more typical of the category IS CONNECTED TO SOME OF THE CIRCLES than All cases, so a situation with only Mixed cases is more typical of the sentence than a situation with both Mixed and All cases, which in turn is more typical than a situation with only All cases.

However, this cannot be the whole story, since it does not explain why many participants chose the verbal option 'both'. I propose that the pattern of responses Clifton & Dube found is caused by an interpretative ambiguity in their task. Some participants took the instructions to heart and indicated which picture was best described by the sentence. These participants always picked the most typical picture. But other participants indicated whether the sentence was a correct description of the pictures. These participants chose the verbal option 'both' in the case of embedded 'some'. They also chose 'both' in the case of unembedded 'some' unless they computed a scalar inference, in which case they opted for the situation that agreed with the scalar inference. This explains why the preference for the picture that agreed with the scalar inference was more pronounced in the case of unembedded 'some' than the preference for the picture that agreed with the embedded scalar inference in the case of embedded 'some'.

This analysis makes two predictions about Clifton & Dube's task. First, that it is sensitive to typicality differences. Second, that sentences containing a category with a standard typicality structure will cause a response pattern that is similar to the response pattern for sentences with 'some' in embedded position:

(45) This is a bird.

In other words, when presented with (45) paired with pictures of a typical and an atypical instance of the category denoted by ‘bird’ (e.g., a robin and an ostrich), most participants should choose the verbal option ‘both’. A smaller but significant number of participants should opt for the typical instance. I tested these hypotheses in Experiment 3.

Experiment 3

Pretest

In order to select suitable categories with a standard typicality structure, I first established if participants considered the atypical category members that were to be included in the eventual experiment genuine category members. To this end, 10 Dutch participants (mean age: 22; range: 20-26; 5 females) filled in a questionnaire. These participants were not paid for their participation. One person was excluded from the analysis for making mistakes in two or more control items.

Each page of the questionnaire showed a sentence with a picture. Participants had to indicate whether the sentence was true or false as a description of the picture. In the critical condition, participants saw a sentence with an adjective or a noun that referred to a category with a standard typicality structure. Ten of these sentences were constructed: in five sentences, the category was denoted by an adjective, and in five by a noun. These sentences were paired with a picture of an atypical category member. A full list of the items that were used is provided in the appendix. Two examples are:

(46) a. This is a bird. *(paired with a picture of an ostrich)*
b. This animal is small. *(paired with a picture of a mouse)*

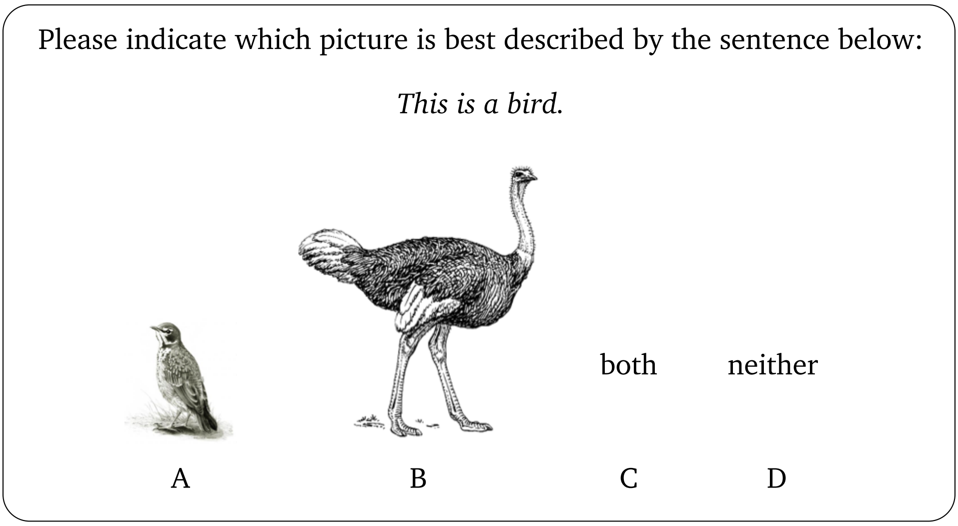
In 89% of the cases, participants indicated that the sentence was true as a description of the atypical category member. For no individual item were there more than two participants who considered the sentence false. I consider this rate sufficiently high to warrant the assumption that the atypical instances are considered genuine category members. So these items were included in the eventual experiment.

Participants

26 Dutch participants (mean age: 22; range: 18-26; 10 females) filled in the questionnaire. Participants were not paid for their participation. Five participants were excluded from the analysis because they made errors in two or more control items. This is a relatively high proportion of participants. It is possible that some of the participants were less motivated for lack of financial compensation.

Materials

Each page of the questionnaire showed a sentence with two pictures. Participants had to indicate which of the two pictures was best described by the sentence. In addition, they could choose the verbal options 'both' and 'neither'. The set-up of the questionnaire was thus the same as in Clifton & Dube's experiment. An example item is shown in Figure 2.7.



Please indicate which picture is best described by the sentence below:

This is a bird.

Figure 2.7: Sample item used in Experiment 3

Three kinds of sentences were constructed. First, the ten sentences with an adjective or a noun that referred to a category with a standard typicality structure. Second, five sentences with the scalar expression 'some' in unembedded position. Third, five sentences with 'some' embedded under a universal quantifier. These two conditions were also tested by Clifton & Dube, and were included to replicate their results in a different language and to be able to compare the results for these conditions directly to the results for the sentences denoting categories with a standard typicality structure. Example sentences from the embedded and unembedded 'some' conditions were:

(47) a. Every square is connected with some circles. *(Embedded 'some')*
b. Some stars are in the box on the left. *(Unembedded 'some')*

The sentences denoting categories with a standard typicality structure were paired with a picture of a typical instance and a picture of an atypical instance. I will refer to these options as O1 and O2. The sentences with 'some' in unembedded positions were paired with a picture that agreed with the scalar inference (O1) and a picture that did not (O2). Lastly, the sentences with 'some' embedded under a universal quantifier were paired with a picture that agreed with the embedded scalar inference (O1) and a picture that agreed only with the scalar inference predicted by pragmatic accounts (O2). The appendix gives an overview of the categories with a standard typicality structure and the corresponding typical and atypical category members.

Four lists were created, randomising the order of the items and pictures. Eight control items were included in the questionnaire. These control items were structurally similar to the target items, but had a single correct answer.

Procedure

The instructions went as follows:

> First of all, thank you for participating in this experiment. This questionnaire consists of multiple-choice questions like the following:

Please indicate which picture is best described by the sentence below:

This is a pet.

A

B

both

C

neither

D

> The question is the same for all multiple-choice questions. The sentence and the pictures differ. You simply have to indicate which picture best describes the sentence. To do so, circle one of the four possible answers. Note that the question is not whether you prefer a dog or a parrot, but which animal is best described by the sentence “This is a pet”.
>
> Fill in the items one by one and do not leaf forward or backward. There is no time limit, but do not think too long before giving an answer. Just follow your intuition.

Results and discussion

The results of the experiment are presented in Table 2.5. A multinomial regression analysis with the embedded ‘some’ condition as reference category shows that the distribution of answers in this condition did not significantly differ from the distribution of answers for categories with a standard typicality structure ($b = .157$, Wald $\chi^2(1) = 1.43$, $p = .232$), whereas it differed significantly from the distribution of answers for unembedded ‘some’ ($b = .787$, Wald $\chi^2(1) = 27.60$, $p < .001$).

The results in the lexical category condition show, first of all, that Clifton & Dube’s

	O1	O2	‘Both’	‘Neither’
Experiment 3				
Unembedded ‘some’	70	0	19	11
Embedded ‘some’	25	13	62	0
Lexical category	28	1	70	0
Clifton & Dube				
Unembedded ‘some’	71	3	24	2
Embedded ‘some’	39	3	57	1

Table 2.5: Percentage of choices for each of the four options in the three conditions of Experiment 3 alongside the results found by Clifton & Dube.

task is sensitive to typicality differences: although the verbal option ‘both’ was chosen most frequently, there was a pronounced preference for the typical over the atypical category member. In addition, the results indicate that there is indeed an interpretative ambiguity in Clifton & Dube’s task. It is clear that a robin is a better instance of the category BIRD than an ostrich. Participants who took the instructions to heart therefore chose option O1. Most participants, however, opted for ‘both’. These participants indicated whether the sentence was a correct description of the pictures, rather than which picture it best describes. As outlined in the previous section, this interpretative ambiguity caused the asymmetry between the results for embedded and unembedded ‘some’.[6]

2.2.9 Conclusion

In section 2.2.3, I discussed three observations that must be explained by any adequate account of the experimental data on the interpretation of sentences with a scalar expression in the scope of a universal quantifier:

i. Participants do not consider such sentences false in situations in which the embedded scalar inference is false (cf. Geurts & Pouscoulous).

[6] The results for the embedded and unembedded ‘some’ conditions mostly reflected the results found by Clifton & Dube. The only notable difference was the substantial number of participants who opted for the picture that falsified the embedded scalar inference in the case of embedded ‘some’. While this picture was chosen significantly less frequently than the picture that agreed with the embedded scalar inference ($F(1,20) = 7.7$, $p < .005$), it was chosen far more often than the 2% that Clifton & Dube found. Consequently, the preference for the picture that agreed with the embedded scalar inference was less pronounced in this experiment than in Clifton & Dube’s. A possible explanation for this difference is that I used ‘some’ instead of the partitive ‘some of the’. As noted by Degen & Tanenhaus (2011), the partitive more robustly triggers an upper-bounding inference than the simple form. In addition, I used the Dutch existential quantifier ‘enkele’ instead of ‘sommige’. There are several differences between these quantifiers (Banga et al. 2009, de Hoop & Kas 1989, de Jong 1983), one of which is that the upper-bounding inference associated with ‘sommige’ is stronger than the upper-bounding inference associated with ‘enkele’.

ii. Chemla & Spector found differences in suitability between situations that cannot be explained in terms of scalar inferences.

iii. Clifton & Dube found markedly different results for ‘some’ in embedded and unembedded positions.

A typicality explanation, as outlined in the previous sections, accounts for each of these observations. This obviates the need for stipulating embedded scalar inferences to account for the experimental data on ‘some’ embedded under a universal quantifier. More generally, it turns out that the experimental paradigms used by Clifton & Dube and Chemla & Spector are strongly influenced by typicality, even when it concerns the comprehension of complex quantified sentences.

The confounding effect of typicality further complicates the interpretation of sentences with a scalar term. At least three mechanisms are at work in these cases:

i. Pragmatic reasoning about the speaker’s intentions can lead to an ignorance inference according to which the speaker is uncertain about whether or not the stronger alternative is true. This is a conversational implicature.

ii. If a scalar expression is sufficiently emphasised by prosodic or contextual cues, its literal meaning can be narrowed to exclude its stronger scalemates.

iii. The present argument demonstrates that listeners have preferences about the meaning of an utterance that cannot be reduced to either of these mechanisms. For example, ‘Some A B’ is preferred to describe a situation in which between 40% and 50% of the A are B.

At the theoretical level, these factors can be distinguished as follows: only conversational implicatures and truth-conditional narrowing are intended by the speaker, thus determining the communicative content of an utterance. Of these two, only narrowing further affects the literal meaning of an utterance.

In section 2.5, we consider the distinction between these three meaning aspects in some more detail. But first we return to the issue of embedded scalar inferences. In the next section, I discuss the experimental data on the interpretation of ‘some’ embedded in the scope of non-monotone quantifiers.

2.3 Non-monotone environments

2.3.1 Introduction

In the previous section, we focused on sentences with a scalar term in the scope of a universal quantifier. Universal quantifiers are upward-entailing. This means the alternatives that are generated by replacing a scalar term with a stronger scalemate are more informative than the original sentence. The converse is true for downward-entailing quantifiers: replacing a scalar term in such an environment with a stronger scalemate leads to a less informative sentence. In the case of non-monotone quantifiers, the corresponding alternatives are neither more nor less informative. For example, it is possible that (48a) is true but (48b) is false and vice versa:

(48) a. There was only one key that fit some of the locks.
b. There was only one key that fit all of the locks.

If (48a) licenses an embedded scalar inference, it is interpreted as:

(49) There was only one key that fit some but not all of the locks.

The embedded scalar inference is predicted within lexical accounts if 'some' is disambiguated as 'some but not all'. Whether grammatical accounts also predict embedded scalar inferences in non-monotone environments depends on the constraints on the distribution of the $\mathbb{O}$ operator. Based on the constraint that the operator is appended if it leads to a more informative sentence, no embedded scalar inference is predicted, since (49) is neither more nor less informative than (48a). (That is, there are situations in which one sentence is false and the other one is not.) But Chemla & Spector (2011) propose that the $\mathbb{O}$ operator is appended whenever it does not lead to a less informative sentence. In that case, (48) is predicted to license an embedded scalar inference, which is generated by appending the $\mathbb{O}$ operator to the embedded sentence 'x fit some of the locks'.

According to a pragmatic account, embedded scalar inferences should only occur when the scalar expression is marked with contrastive stress:

(50) There was only one key that fit SOME of the locks.

Therefore, the question that needs to be answered is whether (49) is freely and systematically available as an interpretation of (48a). This question was put to the test experimentally by Geurts & Pouscoulous (2009a) and Chemla & Spector (2011), with markedly different results. In the next section, I describe their experiments. Afterwards, I argue that the results obtained by Chemla & Spector are caused by some of their participants reading the target sentence with contrastive stress.

2.3.2 The experimental record

Geurts & Pouscoulous (2009a)

Geurts & Pouscoulous (2009a) tested the interpretation of scalar expressions in non-monotone environments by means of a verification task. Participants had to indicate if (51a) and (51b) were true or false as a description of pictures P2 and P3, respectively (Figure 2.8):

(51) a. Exactly two squares are connected with some of the circles.

b. Exactly two circles are connected with some of the squares.

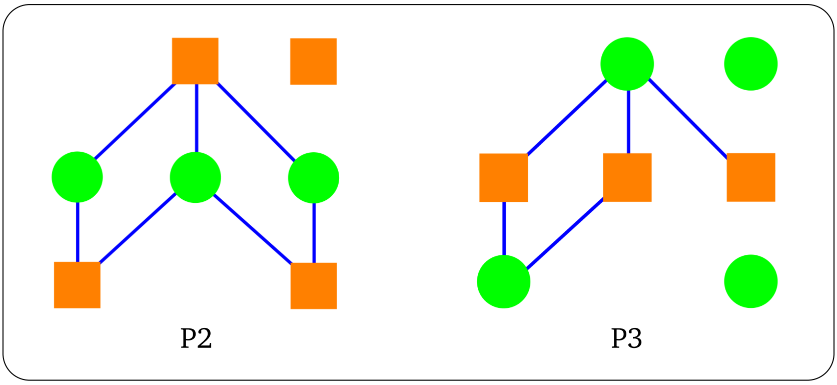


Figure 2.8: Pictures used in Geurts & Pouscoulous' (2009) experiment.

On its literal interpretation, the target sentence is false in situation P2 but true in situation P3. But if an embedded scalar inference is computed, the truth values are reversed: the target sentence is true in P2 but false in P3. Geurts & Pouscoulous found no evidence that an embedded scalar inference was computed: all of their participants indicated that the sentence was false in situation P2 and true in situation P3.

Chemla & Spector (2011)

In their experiment, Chemla & Spector used the following target sentence:

(52) Exactly one letter is connected to some of its circles.

This sentence was presented with the four pictures shown in Figure 2.9. In situation P0, the sentence is unambiguously false. In P1, it is literally true. In P2, it is true if an embedded scalar inference is computed. Lastly, in situation P3, the sentence is true both on its literal interpretation and if a *global scalar inference* is computed, according to which it is true that there is exactly one letter connected to some or all of its circles and that the letter in question is not connected to all of its circles.[7]

[7] According to Chemla & Spector, the computation of this global scalar inference requires reasoning with the alternative 'Exactly one letter is connected to all of its circles' which is not more informative than the speaker's utterance. Geurts (2010, section 8.4), however, offers a framework in which an alternative explanation can be accommodated that does not require this controversial assumption. According to Geurts, there can be anaphoric dependencies between the speaker's utterance and its alternatives. This permits the listener to construct an alternative that can be paraphrased as 'This letter is connected to all of its circles', where 'this letter' refers to the same letter as 'exactly one letter' in the speaker's utterance. This alternative is more informative than the speaker's utterance. His proposal can be formalised in terms of discourse representation theory (e.g., Kamp 1981, Kamp & Reyle 1993). The speaker's utterance leads to the following discourse model:

Observe that, in the experiment of Geurts & Pouscoulous, the target sentence was false in situation P3 if an embedded scalar inference was computed, since one of the circles was connected to all of the squares. In Chemla & Spector's version, due to their choice of numeric expression, the sentence with an embedded scalar inference was true in situation P3.

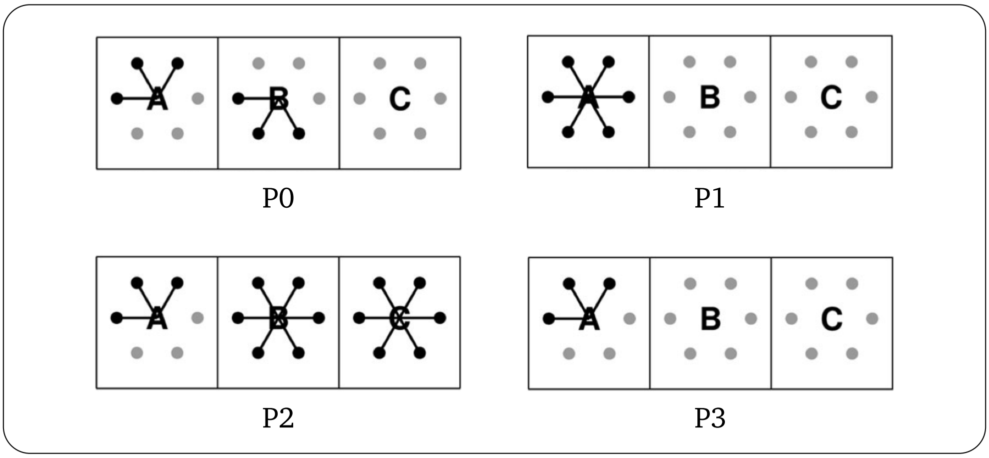


Figure 2.9: Pictures used in Chemla & Spector's (2011) second experiment.

Chemla & Spector asked participants to indicate on a continuous scale how well the situations were described by the sentence. Participants indicated that the sentence was the best description of the P3 situation (98%), which is in line with the results of Geurts & Pouscoulous. However, Chemla & Spector found that participants also considered the sentence quite suitable in the P2 situation (73%). The ratings for the P0 (7%) and P1 (37%) situations were substantially lower. These results led them to conclude that sometimes the 𝕆 operator is appended to sentences even if this does not lead to to a more informative sentence (p. 366).

	BEL$_S$:	X
(i)		|X| = 1 letter(X) X is connected to some of the circles
	?BEL$_S$:	X is connected to all of the circles

Here, '?' indicates that the listener asks himself if the speaker believes the corresponding discourse condition. Presumably the listener concludes that he does not. Thereupon the question mark is replaced with a negation operator. If, furthermore, the speaker is taken to be competent, the belief operator takes scope over the negation. The critical point is that the discourse referent 'X' in the alternative is the same discourse referent that was introduced by the speaker's utterance. So it is possible to compute the global scalar inference in a pragmatic account without suspending the assumption that alternatives have to be more informative than the speaker's utterance. See Geurts (2010, section 8.4) for further discussion.

2.3.3 Interpretation

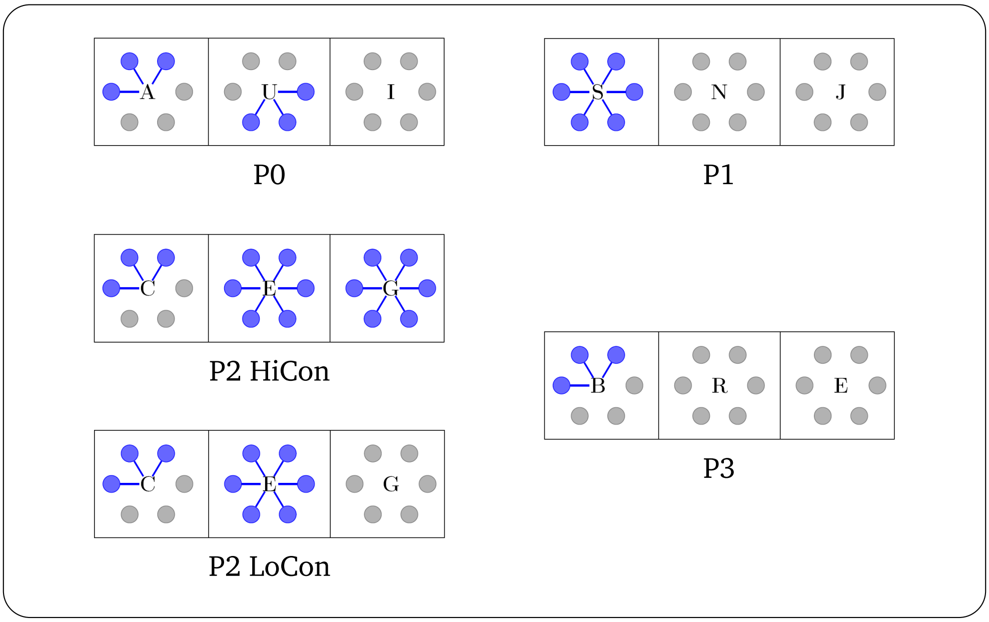
If Chemla & Spector are right and there is a general preference for computing scalar inferences, why did participants in Geurts & Pouscoulous' task unanimously consider the target sentence false in a situation in which the embedded scalar inference was true? In the case of 'some' embedded under a universal quantifier, Sauerland (2010) attempted to explain the results found by Geurts & Pouscoulous by assuming that participants tend to judge a sentence true as long as there is an interpretation according to which it is true (see the discussion in section 2.2.3). But this analysis is unavailable and even contradicted in this case, since participants in Geurts & Pouscoulous' experiment judged the target sentence false in the P2 situation despite the hypothesised presence of an interpretation according to which the sentence was true in this situation.

On the other hand, if Geurts & Pouscoulous are right and 'some' does not license embedded scalar inferences in the scope of non-monotone quantifiers, how is it possible that participants in Chemla & Spector's experiment judged the target sentence quite suitable in a situation that verified the embedded scalar inference? I propose that these apparently contradictory findings can be reconciled by assuming that participants in Chemla & Spector's experiment interpreted 'some' contrastively.

There are a number of factors that, taken together, caused 'some' to receive a contrastive interpretation in Chemla & Spector's experiment. First, participants had to evaluate the target sentence 52 times, and did not encounter any filler items that might have disguised the purpose of the experiment. Second, within the P2 situation presented in Figure 2.9, there is a salient contrast between the leftmost letter and the other two. Given that the predicate is unequivocally verified by the leftmost letter, participants were invited to interpret that letter as the one the speaker had in mind, and construe the predicate in such a way that it was falsified by the remaining letters. Put otherwise, the visual constrast in the P2 situation provoked a contrastive interpretation of the predicate. If this analysis is correct, Chemla & Spector's data have no bearing on the debate between pragmatic and conventionalist accounts of scalar inferences, since both agree that embedded scalar inferences occur if a scalar expression receives a contrastive interpretation.

This analysis implies that the rating for the P2 situation can be lowered by reducing the visual contrast within that situation. In order to test this prediction, we replicated Chemla & Spector's experiment with materials that were nearly identical to theirs, but where low-contrast and high-contrast versions of the P2 situation were presented to different groups of participants. The critical difference between the high-contrast and low-contrast versions was that in the former but not the latter case, there was a salient visual contrast between one letter, which unequivocally verified the predicate of the target sentence, and the remaining letters. All of the items are shown in Figure 2.10.

Questionnaires for 54 participants were posted on Amazon's Mechanical Turk.



P0

P1

P2 HiCon

P3

P2 LoCon

Figure 2.10: Experimental items used in the in the high/low contrast study.

All participants were presented with the four situations in 2.10. For half of the participants, (52) was paired with the low-constrast (LoCon) version of the P2 situation, for the other half, the same sentence was presented with the high-contrast (HiCon) version (i.e., Chemla & Spector's situation P2). The other three items were identical for both groups of participants. The order of presentation was randomised across nine lists. As in Chemla & Spector's experiment, there were no fillers in the experiment, so there were four items in both conditions. Participants were instructed to indicate, on a seven-point scale, how well each of the pictures was described by sentence (52).

Three participants were excluded from the analysis because they were not native speakers of English. The average ratings produced by the remaining participants are given in Table 2.6. As expected, in the high-contrast condition, the P2 situation received a rating that was significantly higher than the P0 situation ($t(25) = 4.0$, $p < .001$). In the low-contrast condition, on the other hand, the rating for the P2 situation did not differ significantly from the P0 situation in which the sentence was unambiguously false. These findings confirm the hypothesis that the high rating Chemla & Spector found for P2 was caused by the visual contrast within that picture.[8]

[8] Interestingly, Chemla & Spector also tested both HiCon and LoCon pictures in their experiment, but failed to find a difference between them: both received a comparatively high rating of 73%. A likely explanation for this discrepancy is that, once participants had encountered a HiCon item in Chemla & Spector's experiment, the scalar inference triggered by this item had knock-on effects on subsequent items. This would also explain why, in our experiment, the difference between the P1 and P3 items was significant for the HiCon version ($t(25) = 4.1$, $p < .001$), but only marginally so

	LoCon	HiCon
P0	30	32
P1	74	49
P2	37	64
P3	84	80

Table 2.6: Average ratings (%) in the high/low contrast study

In the previous section, we provided evidence that scalar expressions embedded under universal quantifiers do not license upper-bounding inferences. In this section, we have argued that the same conclusion holds for scalar expressions in the scope of non-monotone quantifiers. Taken together, these findings are problematic for all versions of conventionalism that predict embedded scalar inferences in one or both of these cases.

In the next section, I extend the analysis to free choice inferences. As observed in section 1.2.2, some authors explain these as a kind of quantity inference. So if free choice inferences survive in the scope of universal quantifiers, this might provide support for a conventionalist view on quantity inferences.

2.4 Universal free choice

2.4.1 Introduction

Free choice inferences occur whenever ‘or’ is embedded under an operator that existentially quantifies over individuals or over possible worlds that are deontically, epistemically, or dynamically accessible (e.g., Chemla 2009, Fox 2007, Franke 2009, Geurts 2005, 2010, Kamp 1973, Klinedinst 2007, Zimmermann 2000):

(53) Johan may take a ham or cheese sandwich. *Deontic*
$\rightsquigarrow$ Johan may take a ham sandwich.
$\rightsquigarrow$ Johan may take a cheese sandwich.

(54) Some players received a yellow or red card during the match. *Existential*
$\rightsquigarrow$ Some players received a yellow card during the match.
$\rightsquigarrow$ Some players received a red card during the match.

(55) According to the doctor, Piet might have the flu or the measles. *Epistemic*
$\rightsquigarrow$ According to the doctor, Piet might have the flu.
$\rightsquigarrow$ According to the doctor, Piet might have the measles.

in the LoCon condition ($t(24) = 2.0, p = .062$). In Chemla & Spector’s study, such effects may have been reinforced by the extreme repetitiveness of the experiment, in which every item was presented four times, which is to say that every participant saw the critical sentence no less than 52 times.

(56) Karin can write a haiku or a limerick. *Dynamic*
⇝ Karin can write a haiku.
⇝ Karin can write a limerick.

As explained in section 1.2.2, these unembedded, or *singular*, free choice inferences can be explained within both pragmatic and grammatical accounts of quantity inferences. But only grammatical accounts predict free choice inferences to occur in embedded positions:

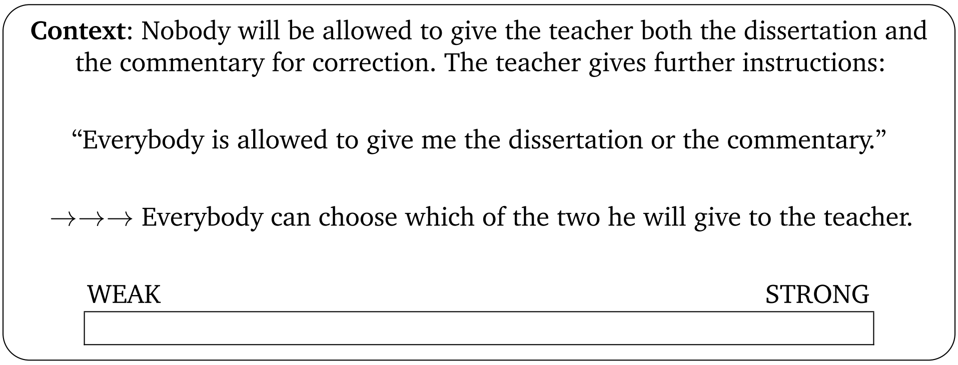
(57) Every visitor may take a ham or cheese sandwich.
?⇝ Every visitor may take a ham sandwich.
?⇝ Every visitor may take a cheese sandwich.

Chemla (2009) conducted an experiment to test the robustness of such embedded, or *universal*, free choice inferences. He found that these inferences were as robust as their unembedded counterparts. In response, Geurts & Pouscoulous (2009b) attempted to incorporate Chemla's findings within a pragmatic framework. In the next section, we discuss Chemla's study and Geurts & Pouscoulous' response.

2.4.2 The experimental record

Chemla (2009)

Chemla (2009) tested the robustness of universal free choice inferences by asking participants to indicate on a continuous scale how strongly they would infer singular and universal free choice inferences. An example item from his experiment is provided in Figure 2.11. Chemla found that participants considered universal free choice inferences as robust as their singular counterparts, both receiving ratings of around 80%. If free choice inferences are quantity inferences, the robustness of embedded free choice inferences presents a challenge to pragmatic accounts.



Context: Nobody will be allowed to give the teacher both the dissertation and the commentary for correction. The teacher gives further instructions:

“Everybody is allowed to give me the dissertation or the commentary.”

→→→ Everybody can choose which of the two he will give to the teacher.

WEAK STRONG

Figure 2.11: Experimental item from Chemla (2009).

Geurts & Pouscoulous (2009b)

In response to to Chemla's experiment, Geurts & Pouscoulous (2009b) argued that universal free choice inferences are restricted to permission-giving sentences. They cite the examples below to show that the inferences appear to be less robust in other kinds of sentences:

(58) Every day of the week, some of the guests order scrambled eggs or an omelet.
?$\rightsquigarrow$ Every day of the week, some of the guests order scrambled eggs.
?$\rightsquigarrow$ Every day of the week, some of the guests order an omelet.

(59) If things had turned out differently, everyone in my department could have been a banker or a lawyer.
?$\rightsquigarrow$ Everyone in my department could have been a banker.
?$\rightsquigarrow$ Everyone in my department could have been a lawyer.

(60) Every student of mine can write a haiku or play the Moonlight Sonata.
?$\rightsquigarrow$ Every student of mine can write a haiku.
?$\rightsquigarrow$ Every student of mine can play the Moonlight Sonata.

If these intuitions are correct, the problem for pragmatic accounts might not be as daunting as it seemed. Of course, it remains to give an explanation for why universal free choice inferences are licensed in permission-giving sentences. Geurts & Pouscoulous illustrate their explanation with the following example. Suppose Inspector Ambrose addresses constables Bacon and Champion:

(61) You$_{[\text{PLU}]}$ may have a daiquiri.

On its most accessible interpretation, this utterance implies that both Bacon and Champion are allowed to have a daiquiri. According to Geurts & Pouscoulous, this interpretation can be explained by assuming that Ambrose uses (61) to perform two speech acts instead of one. It is as if he told both of his addressees:

(62) You$_{[\text{SING}]}$ may have a daiquiri.

In the same vein, the teacher in Chemla's scenario in Figure 2.11 might be understood as if he told each of his students:

(63) You$_{[\text{SING}]}$ are allowed to give me the dissertation or the commentary.

If this is correct, the problem of explaining universal free choice inferences boils down to the problem of explaining singular free choice inferences, which has been addressed by several pragmatic accounts (Franke 2011, Geurts 2010, Kratzer & Shimoyama 2002, Schulz 2005).

So Geurts & Pouscoulous provide an intuitively plausible explanation that salvages the pragmatic account in light of the robustness of universal free choice inferences.

But is their explanation correct? There are at least three arguments that might be adduced against their proposal:

i. Geurts & Pouscoulous' account hinges on their intuitions about complex sentences. Since intuitions about such sentences are likely to be heavily theory-laden, naive participants should be probed for their judgements.

ii. Only a small number of example sentences are provided to argue that universal free choice inferences are restricted to permission-giving contexts. These examples might be idiosyncratic in some way. It is unclear if Geurts & Pouscoulous' intuitions are vindicated across a wider variety of examples.

iii. One reason for the fragility of universal free choice inferences in sentences that cannot be used to grant permission might be that the corresponding singular free choice inferences are less robust as well. If so, the fragility of the universal free choice inferences is due to this difference instead being caused by the presence of an embedding operator.

In order to determine if these concerns are valid, I conducted an inference task similar to Chemla's, testing the robustness of free choice inferences in case 'or' was embedded under a deontic, epistemic, or dynamic modal operator, or under an existential quantifier. If Geurts & Pouscoulous' explanation is correct, only free choice inferences in sentences that can be used to grant permission should survive embedding under a universal quantifier.

Two versions of this task were created. In the baseline version, participants encountered only singular free choice sentences. In the universal version, the same sentences were embedded under a universal quantifier. The inclusion of a baseline version was necessary to factor in a possible explanation for Geurts & Pouscoulous' intuitions: the reason that universal free choice inferences seem to be less robust in sentences that cannot be used to grant permission could be that free choice inferences in general are less robust in such contexts. Therefore I investigated the effect of embedding rather than just the robustness of universal free choice inferences. In the next section, I discuss the details of the task.

The experiment

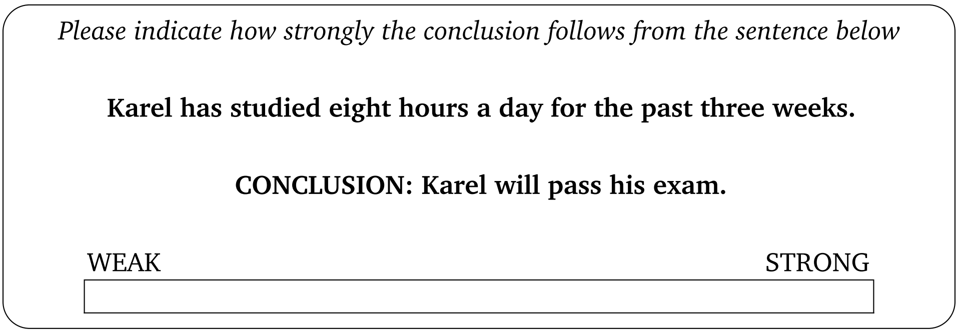
Participants

54 undergraduates in Philosophy filled in the questionnaire before a lecture on argumentation theory. 30 of them made the universal version, and the remaining 24 the baseline version. Participants were not paid for their participation. No participants were excluded from the analysis. The questionnaire took between 10 and 15 minutes to complete.

Materials and procedure

Each page of the questionnaire showed a sentence followed by a conclusion. An example trial is provided in Figure 2.12. Participants were instructed to indicate how strongly the conclusion followed from the sentence. The full instructions were based on those used by Chemla (2009) and went as follows:

> First of all, thank you for your participation in this survey. The survey consists of questions like the following:



Please indicate how strongly the conclusion follows from the sentence below

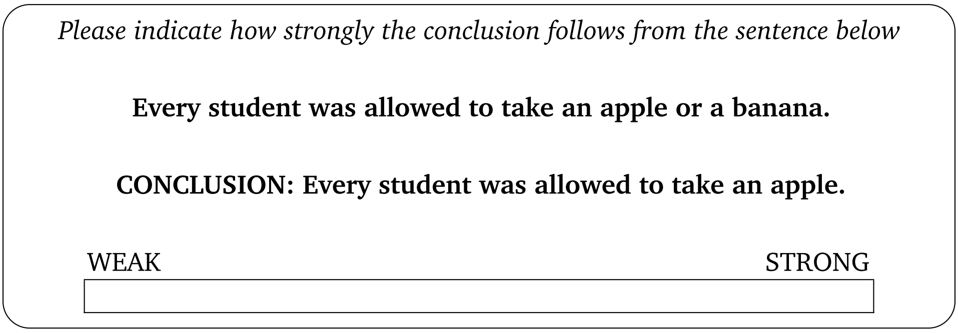
Karel has studied eight hours a day for the past three weeks.

CONCLUSION: Karel will pass his exam.

WEAK STRONG

> The question is the same for all items. The sentence and the conclusion differ per item. It is your task to indicate how strongly the conclusion follows from the sentence. To this end, you can draw a line on the bar. The further to the right the line is drawn, the stronger the conclusion follows from the sentence.
>
> Fill in the items one by one and do not leaf forward or backward. There is no time limit, but do not think too long before giving an answer. Just follow your first intuition.

Participants could set a line anywhere on a bar. The further to the right the line was set, the more strongly the conclusion was felt to follow. The dependent variable was the proportion of bar to the left of the line. The set-up of the task was essentially the same as the inference task used by Chemla (2009).



Please indicate how strongly the conclusion follows from the sentence below

Every student was allowed to take an apple or a banana.

CONCLUSION: Every student was allowed to take an apple.

WEAK STRONG

Figure 2.12: Experimental item used to test the robustness of universal free choice inferences.

The questionnaire consisted of 35 items in seven conditions. In the four target conditions, ‘or’ was embedded under a deontic, epistemic, or dynamic modal operator, or under an existential quantifier.[9] Three filler conditions were included. The first filler condition consisted of sentences with 'and' instead of 'or'. The second filler condition consisted of sentences with 'or' embedded in the antecedent of a conditional. The final filler condition mirrored the target conditions but involved a conclusion that was not supported by the sentence. Examples of the three filler conditions from the universal version are:

(64) Every guest took a sausage and a slice of bacon from the barbecue.
Conclusion: Every guest took a sausage from the barbecue.

(65) If everyone votes for or against the motion, a second vote won't be necessary.
Conclusion: If everyone votes for the motion, a second vote won't be necessary.

(66) Every fan was allowed to interview Dennis Bergkamp or Patrick Kluivert.
Conclusion: Every fan was allowed to interview Edgar Davids.

In every condition but the last, the conclusion was the same as the target sentence except that the disjunction or conjunction was replaced by one of the disjuncts or conjuncts. Every condition consisted of five items. The order of the items and the disjunct or conjunct that was presented in the conclusion was randomised across multiple lists. A list of the target sentences is given in the appendix.

Note that the conclusion that was offered to participants differed from the one Chemla included in his inference task. Rather than one of the disjuncts, Chemla asked participants, for example, whether it follows from the target sentence in Figure 2.12 that every guest may choose between taking an apple and taking a pear. I avoided this choice of conclusion because it does not readily extend to other free choice licensing environments. The results for the deontic condition did not differ from those found by Chemla, so this difference appears to be immaterial.

Results

The results are summarised in Figure 2.13. In the baseline version, the mean strength of the free choice inferences was the highest in the deontic condition (80%). Free choice inferences were slightly less robust in the epistemic condition (75%), followed by the existential (51%) and dynamic (47%) conditions. I conducted an analysis of variance with condition (deontic, dynamic, epistemic, or existential) as fixed factor and item as a random factor nested under condition. There was a significant main effect of condition ($F(3, 16) = 18.7$, $p < .001$, partial $\eta^2 = .78$) and a marginally significant effect of item ($F(16, 380) = 1.6$, $p = .066$, partial $\eta^2 = .06$). Planned comparisons between conditions using Tukey's

[9] The Dutch verb 'mogen' was used to express deontic modality. Both epistemic and dynamic modality were expressed by 'kunnen'. In the epistemic case, every sentence began with 'According to ...' so that it was clear which modality was intended. The existential quantifier used in the Existential condition was 'enkele'. The universal quantifier was varied between 'alle' and 'iedere', which correspond to English 'all' and 'every'.

procedure showed that all of the means were significantly different ($p < .001$), except for the means in the deontic and epistemic ($p = .542$), and in the existential and dynamic conditions ($p = .749$).

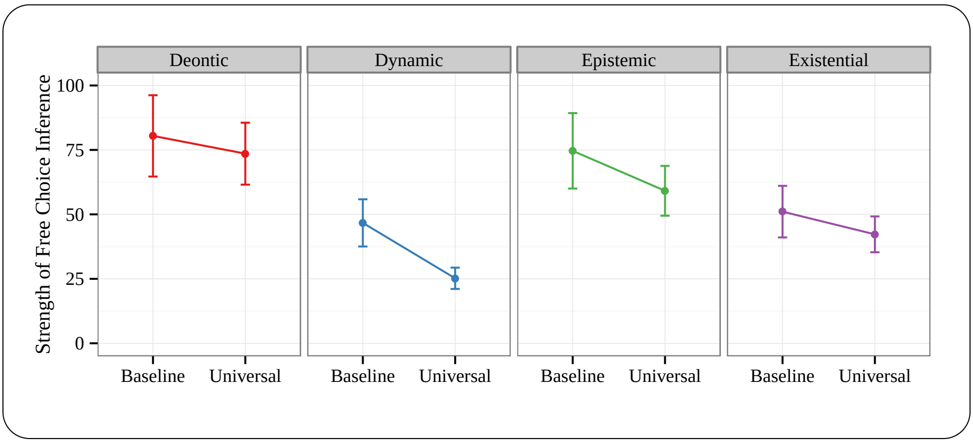


Figure 2.13: Mean strength of singular and universal free choice inferences in different contexts. Errors bars represent 95% confidence intervals.

The results in the universal version displayed a similar pattern: the mean strength of the free choice inferences was the highest in the deontic condition (74%). Free choice inferences were less strong in the epistemic condition (59%), followed by the existential condition (42%). The mean strength of universal free choice inferences was the lowest in the dynamic condition (25%). Just as for the baseline version, I conducted an analysis of variance with condition as fixed factor and item as a random factor nested under condition. There were significant main effects of condition ($F(3, 16) = 14.2$, $p < .001$, partial $\eta^2 = .73$) and item ($F(16, 553) = 4.6$, $p < .001$, partial $\eta^2 = .12$). Planned comparisons based on Tukey's procedure showed that all of the means in the four conditions were significantly different (all p's $< .002$).

To test the effect of embedding on the robustness of free choice inferences, I conducted a three-way analysis of variance with version (baseline or universal) and condition as fixed factors and item as a random factor nested under condition. There were significant main effects of version ($F(1, 16) = 30.2$, $p < .001$, partial $\eta^2 = .65$), condition ($F(3, 16) = 20.2$, $p < .001$, partial $\eta^2 = .79$), and item ($F(16, 16) = 3.0$, $p = .017$, partial $\eta^2 = .75$). The interactions between version and condition ($F(3, 16) = 1.9$, $p = .160$), and version and item ($F(16, 933) = 1.4$, $p = .128$) did not reach significance. The absence of a significant interaction between version and item demonstrates that the effect of embedding did not depend on the particular items involved.

For a more detailed picture, I conducted a two-way analysis of variance within each condition with version as a fixed factor and item as a random factor. The interaction between version and item was significant in the epistemic condition

($F(4, 234) = 3.3$, $p = .013$, partial $\eta^2 = .05$) but not in the deontic, dynamic, or existential conditions (all p's $> .4$). On closer inspection, the interaction in the epistemic condition was almost entirely caused by one item:

(67) According to the professor, [this / every] research question can be answered by means of a survey or an experiment.

The singular free choice inferences licensed by (67) were highly robust (91%), in stark contrast to its universal counterparts (30%). Given the minimal size of the effect, however, I do not think that this interaction has any theoretical consequences.

For comparison, it is instructive to consider the three filler conditions. The mean strength of conjunctive inferences was 87% in the baseline version and 89% in the universal version; that of disjunctive inferences from conditionals 77% in both versions; and that of unwarranted disjunctive inferences 13% in the baseline version and 9% in the universal version. The mean strength of all of the target inferences occurred inbetween the extremes of conjunctive inferences and unwarranted disjunctive inferences.

2.4.3 General discussion

Geurts & Pouscoulous hypothesised that universal free choice inferences are restricted to permission-giving sentences. In line with this hypothesis, universal free choice inferences were the most robust in the deontic condition. But a glance at Figure 2.13 suffices to see that universal free choice inferences are not absent in at least the epistemic and existential conditions. Only in the dynamic condition might it be argued that free choice inferences actually disappear under embedding. In line with these observations, the interaction between condition and version fails to reach significance. So the effect of embedding is the same for all target conditions.

Further evidence against Geurts & Pouscoulous' explanation comes from the particular items used in the deontic condition. The modal verbs in the target sentences in this condition were either in the present or in the past tense:

(68) Iedere bezoeker mag een cocktail of een glas limonade nemen.
Every visitor may.PRES a cocktail or a glass lemonade take
'Every visitor may take a cocktail or a glass of lemonade.'

(69) Iedere student mocht een appel of een peer nemen.
Every student may.PAST a apple or a pear take
'Every student was allowed to take an apple or a pear.'

While the former sentences can be used to give permission, the latter can only be used to report permission. Nonetheless, the effect of embedding on the robustness of free choice inferences licensed by permission-giving sentences (82% in the baseline version against 72% in the embedded version) was not significantly different from the effect of embedding on permission reports (78% against 77%), as shown by a two-way analysis of variance with version and tense (past or present) as fixed factors ($F(1, 240) = 1.3, p = .259$). This provides strong evidence against the hypothesis that universal free choice inferences can be explained by means of distributive speech acts.

An interesting observation is that the robustness of both singular and universal free choice inferences largely depends on the kind of operator that is involved. Free choice inferences are highly robust when 'or' is embedded under a deontic or epistemic modal operator. The inferences are much less robust under a dynamic modal operator or an existential quantifier. This variability is not expected on any account of free choice inferences and stands in need of clarification.

One explanation might be that sentences involving deontic or epistemic modalities require a greater deal of speaker competence than sentences involving dynamic modality or existential quantification. Properties and abilities can usually be observed, permission and knowledge cannot. Partial knowledge about permission and knowledge is thus less likely than partial knowledge about properties and abilities. Since a reading without free choice inferences expresses partial knowledge, this might have caused more participants to derive free choice inferences for sentences involving deontic or epistemic modality.

The results of the experiment pose a serious problem for pragmatic accounts of quantity inferences that predict universal free choice inferences to be restricted to permission-giving contexts. However, Franke (2011) suggests that universal free choice inferences in other contexts can be accounted for within a pragmatic framework as well. If its free choice inferences are not licensed, a sentence like (70) is associated with a situation in which some of the guests are allowed to take coffee while the others are allowed to take tea:

(70) Every guest may take coffee or tea.

According to Franke, if this situation is sufficiently improbable, an interpretation without free choice inferences might be excluded on a priori grounds. This leads the listener to adopt the correct interpretation according to which every guest is allowed to choose between taking coffee and taking tea.

Against this proposal, it must be noted that other quantity inferences such as scalar inferences seem to be insensitive to plausibility considerations. We have already discussed the following example from Geurts (2010):

(71) Cleo threw all her marbles in the swimming pool. Some of them sank to the bottom.

Even though it is highly unlikely that some but not all of the marbles sank to the bottom, this is exactly what is implied by the sentence. If Franke is right, then, universal free choice inferences are sensitive to world knowledge considerations in a way that other quantity inferences are not. Franke's account stands in need of explaining this asymmetry: under what circumstances can plausibility considerations affect the derivation of quantity inferences?

In personal correspondence, Franke has suggested that there is an important difference between cases like (71), where world knowledge is unable to mitigate the force of the scalar inference, and cases like (70), where it is hypothesised that a particular situation is ruled out on the basis of world knowledge considerations. In the former case, ruling out the implausible situation would clash with the assumption that the speaker is rational: if he believed that all of the marbles sank to the bottom, he could have said so by using the same statement with 'all' instead of 'some'. But in the case of (70), there is no alternative sentence of equal complexity available that unambiguously conveys the universal free choice inferences. This contrast might explain the difference between these cases: world knowledge can only affect the computation of scalar inferences if the resultant interpretation does not clash with the assumption of rationality.

On the other hand, universal free choice inference are also problematic for grammatical accounts. In section 2.2, I argued that scalar inferences are not licensed in the scope of a universal quantifier. So in that respect, they differ from free choice inferences, which do survive embedding under a universal quantifier. Grammatical accounts have to explain this asymmetry: under what circumstances is the $\mathbb{O}$ operator appended to sentences embedded under a universal quantifier?

In any case, before evaluating the ramifications of universal free choice inferences on the various accounts of quantity inferences, a more pressing matter is to find out whether free choice inferences are actually a variety of quantity inferences in the first place. The arguments that have been adduced in favour of this position are anything but conclusive. One such argument is that free choice inferences mirror scalar inferences in their propensity to disappear in downward-entailing environments. That is, someone who utters (72) does not merely convey that you are not allowed to choose between taking an apple or a pear, but makes the stronger claim that you are not allowed to take either one:

(72) You may not take an apple or a pear.

This observation indeed poses a problem for theories that locate the source of free choice inferences in the semantics of 'or' (Geurts 2005, Zimmermann 2000), but other theories are able to account for it without assuming that free choice inferences are quantity inferences (e.g., Barker 2010).

In fact, there are a number of observations that speak against the view that free choice inferences are quantity inferences. Perhaps the most pressing problem is that free choice inferences occur even if the disjunction takes wide scope over the modal operator (cf. Jennings 1994):

(73) Mr. X might be in Victoria or he might be in Brixton.

This sentence implies that Mr. X might be in Victoria and that he might be in Brixton. This observation is problematic for all accounts that explain free choice inferences as a variety of quantity inferences, since one of the hypothesised alternatives to (73) is ‘Mr. X might be in Victoria and he might be in Brixton.’ Inferring that the speaker does not believe that this alternative is true is tantamount to rejecting the free choice inferences. Theories that explain free choice inferences as a kind of quantity inferences have to stipulate ancillary assumptions to account for these cases: for example, Geurts (2010, 107) suggests that ‘or’ has “a non-Boolean flavor” in such contexts. It is clear that this falls short of providing an explanatorily adequate account of free choice in sentences in which the disjunction takes wide scope.

Are free choice inferences a variety of quantity inferences? There are at least two possible ways to answer this question. The first is to investigate whether free choice inferences are licensed in other embedding environments as well:

(74) Johan hopes that he may take a ham or cheese sandwich.
?⇝ Johan hopes that he may take a ham sandwich?
?⇝ Johan hopes that he may take a cheese sandwich.

(75) Johan believes that he may take a ham or cheese sandwich.
?⇝ Johan believes he may take a ham sandwich.
?⇝ Johan believes he may take a cheese sandwich.

If so, this would speak against the view that free choice inferences are a kind of quantity inferences, since these do not occur systematically and freely in embedded environments.

A second route is based on the observation that some quantity inferences are associated with a processing cost. If free choice inferences are a kind of quantity inferences, one might expect that the processing of these inferences is costly, too. In the next chapter, I discuss and test this hypothesis.

Universal free choice inferences present a challenge for both pragmatic and grammatical accounts of quantity inferences. Pragmatic accounts have to assume that some possible interpretations are ruled based on considerations of world knowledge. Grammatical accounts have to provide an explanation for why free choice inferences but not scalar inferences survive embedding under a universal quantifier. A more fundamental issue, however, is to establish if free choice inferences are a kind of quantity inferences in the first place. One method of determining this will be discussed in the next chapter.

2.5 Conclusion

Various theories have been developed to explain the interpretation of scalar expressions. While all of these theories agree that both pragmatic reasoning and truth-conditional narrowing influence the interpretation of scalar expressions, the division of labour between these mechanisms has been debated at length. On the one hand, pragmatic accounts assume that truth-conditional narrowing is restricted to situations in which the scalar expression is marked by prosodic or contextual cues. On the other hand, conventionalist accounts propose that truth-conditional narrowing occurs across the board. These theories can be distinguished on the basis of the interpretation of scalar expressions in embedded environments.

Testing these predictions, however, has proven difficult because of the different meaning aspects that influence the interpretation of scalar expressions. In addition to pragmatic reasoning and truth-conditional narrowing, the interpretation of these expressions is affected by typicality differences. For example, listeners prefer a sentence of the form 'Some A are B' to describe a situation in which 32% of the A are B. In section 2.2 and 2.3, I have argued that the extant evidence for embedded scalar inferences can be explained in terms of truth-conditional narrowing and typicality differences, thus favouring pragmatic accounts of scalar inferences over their conventionalist competitors.

What is the relationship between pragmatic reasoning, truth-conditional narrowing, and typicality differences? In the previous sections, I have propagated the view that typicality differences, unlike pragmatic inferences and truth-conditional narrowing, are not part of the communicative content of a sentence. But this is not the only possible perspective. An alternative account is based on the assumption that communicative content is a fluid concept. According to this view, someone who utters 'Some A are B' is committed to each of the following claims:

i. At least one of the A is B.
ii. Not all of the A are B.
iii. Around a third of the A are B.

But the strength to the speaker's commitment to these claims differs. Whereas the commitment to *i* is undefeasible, the commitment to *iii* can be waived without much effort. Claim *ii* seems to fall inbetween these extremes.

These differences in commitment strength can be accounted for by considering the salience of the alternatives. In the case of truth-conditional narrowing, the alternative is often given in the linguistic context, whereas it remains implicit in the case of pragmatic reasoning. Alternatives also have an effect in shaping typicality structure: the best instances of a category are those that are maximally distinct from instances of categories denotes by alternatives. For example, bats and whales are poor instances of the category denoted by 'mammal' because they have much in common with instances of the categories denoted by the alternatives 'bird' and 'fish'. However, the effect of these contrasting categories has been shown to be quite limited (Frake 1969, Malt & Johnson 1992, Markman & Wisniewski 1997, Rosch & Mervis 1975, Rosch 1978, Verbeemen et al. 2001).

So alternatives are increasingly more salient in the case of typicality differences, pragmatic reasoning, and truth-conditional narrowing. It seems plausible that the likelihood that the speaker considered and excluded an alternative is an increasing function of its salience. Perhaps this feature can explain the differences in commitment strength between the three claims. These considerations suggest that it might be possible to explain each of the three meaning aspects on the basis of one underlying mechanism. In section 4.2 and 5.2.1, we consider the relationship between these three meaning aspects in some more detail.

From an empirical perspective, the foregoing discussion emphasised the importance of considering construct validity when conducting experiments into language understanding. Construct validity refers to the desideratum that an experimental task measures what it aims to measure. The experimental tasks by Clifton & Dube (2010) and Chemla & Spector (2011) do not conform to this principle because they do not distinguish between the three meaning aspects discussed. In section 5.2.2, I briefly return to this point.

While the current experimental data fails to provide evidence for the systematic presence of embedded scalar inferences, this does not mean that the matter is now settled. Future experiments might provide new insights. See, for example, Tian et al. (2012) who use an act-out task to test for exhaustivity inferences in different embedded environments. As it stands, however, the experimental record on embedded scalar inferences favours a pragmatic account.

Unlike scalar inferences, free choice inferences do survive embedding under a universal quantifier. This finding is problematic for both pragmatic and conventionalist accounts. In order to accommodate universal free choice inferences, pragmatic accounts have to assume that considerations of world knowledge can rule out particular interpretations. Conventionalist accounts, on the other hand, have to provide a principled explanation for the difference in the effect of embedding on scalar inferences and free choice inferences. Why do free choice inferences but not scalar inferences survive embedding under a universal quantifier? This asymmetry warrants a closer look at the assumption that free choice inferences are a kind of quantity inferences. In sections 3.3 and 5.2.3, I return to this issue.

Aside from the interpretation of embedded scalar expressions, pragmatic and conventionalist accounts also make different predictions about the processing profile of scalar inferences. I will discuss and test these predictions in the next chapter. In addition, I take a closer look at the construct validity of an experimental paradigm that is often used in experiments about the processing of scalar inferences. This investigation also offers an insight into the question whether free choice inferences are a kind of quantity inferences.

3 Processing data

3.1 Introduction

In the previous chapter, we have seen that pragmatic and conventionalist theories of quantity inferences make different predictions about the interpretation of sentences with scalar expressions in embedded positions. On the one hand, conventionalist theories predict scalar inferences to occur "systematically and freely in arbitrarily embedded positions" (Chierchia et al. 2012, 2297) because these theories suppose that scalar inferences are an aspect of the literal meaning of an utterance. According to pragmatic theories, on the other hand, scalar inferences are computed after the literal meaning of an utterance has been established. Scalar inferences are therefore predicted to be unable to interact with other aspects of the literal meaning.

More generally, conventionalist accounts assume that scalar inferences are part of the literal meaning of an utterance, whereas pragmatic accounts assume that they are computed on the basis of the previously established literal meaning. One of the corollaries of this difference is that, according to conventionalist accounts, scalar inferences are computed in tandem with the literal meaning of a sentence. Chierchia (2004, 60) is particularly lucid about this point:

> Now, since the plan is to deal with each implicature as soon as possible, I will define α^{ALT} [= the set of alternatives] in such a way that it yields the alternatives induced solely by the last scalar element in the tree (i.e. the highest or topmost one). The rationale for this is that scalar terms below the topmost (if present) will have already been taken care of (by the terms below the topmost). This is a locality constraint driven by the guiding idea of our attempt (namely, that implicatures are processed locally in the order in which their triggers appear).

In the next section, I present an experiment in which I tested this prediction. This experiment is based on the observation that the computation of scalar inferences leads to a delay in verification times (Bott & Noveck 2004). For a proper evaluation of the findings of this experiment, it will be necessary to determine exactly which aspect of the computation of scalar inferences causes the delay in verification times. In order to address this question, I tested three other kinds of quantity inferences in the same experiment. These three quantity inferences differ from scalar inferences in several respects, allowing us to gauge which aspect of the computation process is responsible for the delay in verification times. The results of this part of the experiment will be discussed in section 3.3.

3.2 Locating the processing cost of underinformativeness

3.2.1 The processing cost of underinformativeness

Following up on work by Noveck (2001) and Noveck & Posada (2003), Bott & Noveck (2004) investigated how long it takes for participants to determine the truth value of sentences like the following:

(1) a. Some parrots are birds.
    b. Some dogs are mammals.

Such sentences are true if 'some' receives its literal interpretation as 'at least some', but they are false if interpreted with an upper bound. I will follow tradition in referring to 'true' answers as *literal* and to 'false' answers as *pragmatic*. These are purely descriptive terms: it might turn out that the mechanism that underwrites scalar inferences is not pragmatic in nature after all.

Participants in Bott & Noveck's third experiment read these sentences in a forced reading task in which each word was flashed on the screen for 200 milliseconds, and could provide their answer immediately after the final word was flashed on the screen. Bott & Noveck measured the time between the onset of the final word and the moment one of the response buttons was pressed. Many participants were ambivalent about the truth of sentences like (1), varying their answer between structurally similar trials. Comparing the reaction times of these ambivalent participants, Bott & Noveck found that it took them significantly longer to answer 'false' than it took them to answer 'true'. This difference in reaction times was absent in a control condition, in which the sentence was unambiguously true or false:

(2) a. Some birds are parrots.
    b. Some dogs are insects.

In this condition, 'false' answers did not take significantly longer than 'true' answers. Bott & Noveck attribute the difference in reaction times between literal and pragmatic answers to the time it takes to derive a scalar inference.

Breheny et al. (2006, experiment 2 and 3) corroborated Bott & Noveck's finding in a self-paced reading experiment, in which participants read discourses like:

(3) a. John was taking a university course and working at the same time. For the exams, he had to study from short and comprehensive sources. Depending on the course, he decided to read the class notes or the summary.
b. John heard that the textbook for Geophysics was very advanced. Nobody understood it properly. He heard that if he wanted to pass the course he should read the class notes or the summary.

In (3a) but not in (3b), the scalar inference associated with 'or' is supported by the preceding discourse. It therefore seems natural to suppose that the likelihood of a scalar inference is greater in the former case than in the latter. So if scalar inferences are time-consuming, it is predicted that reading times for 'or' are longer in (3a) than in (3b). Breheny et al.'s findings confirm this hypothesis.

More recently, Grodner et al. (2010) reported on results from an eye-tracking study that seem to undermine the conclusion that scalar inferences are time-consuming. Their participants watched a display while listening to sentences like:

(4) Click on the girl who has some of the balls.

The display contained, among several other characters, a girl with some but not all of the balls and a girl with all of the balloons. Grodner et al. found that, upon hearing 'some', participants immediately fixated on the girl with some but not all of the balls. While their design prevented them from directly comparing the fixation time for an circumbounded interpretation of 'some' to the fixation time for a literal interpretation, the speed of the fixation time corresponded to the time it takes to interpret any lexical item, and therefore suggests that scalar inferences are computed immediately upon hearing the scalar term. Nevertheless, several considerations speak against the view that Grodner et al.'s results disprove the hypothesis that scalar inferences are time-consuming.

A first consideration is that, in a similar eye-tracking experiment, Huang & Snedeker (2009) found that participants fixate on the target character only after having heard the entire sentence, in contradistinction to Grodner et al.'s results. According to Degen & Tanenhaus (2011), the critical difference between these studies was that Huang & Snedeker also included trials with numeric expressions like 'exactly two', whereas Grodner et al. only used the quantifiers 'none', 'some', and 'all'. Degen & Tanenhaus argue that eye movements in these experiments are modulated by considerations of typicality: listeners immediately fixate on the most typical referent of 'some'. The typicality of a referent with respect to 'some' depends on whether there are alternative labels that have previously been mentioned, like 'exactly two' in Huang & Snedeker's experiment, and whether the cardinality of the referent falls into the subitising range (1-4 items). In this range, the cardinality of a referent can be established with particular ease and confidence, and the referent should therefore be singled out by means of a numeric expressions like ‘exactly two’. Participants in Grodner’s experiment immediately fixated on the target character because of the absence of alternative labels.

Degen & Tanenhaus demonstrate that, even if there are alternative labels, participants immediately fixate on the target character provided that the cardinality of the referent of ‘some’ falls outside the subtitising range. Importantly, this explanation does not entail that participants computed a scalar inference upon hearing ‘some’. Just like in Clifton & Dube’s and Chemla & Spector’s experiments (section 2.2), the eye-tracking paradigm measures something different from genuine scalar inferences.

Even if it were assumed that Grodner et al.’s results disprove the hypothesis that scalar inferences are time-consuming, it remains to be seen how the results found by Bott & Noveck and Breheny et al. can be accounted for. Grodner et al. tentatively suggest that the difference in reaction times between literal and pragmatic answers, as found by Bott & Noveck, might be caused by a circumbounded interpretation being more complex than a logical interpretation (cf. Geurts 2010, 90 for a similar explanation). This proposal has some intuitive force: a circumbounded interpretation requires representing both a minimum and a maximum, whereas a logical interpretation only requires the representation of a minimum. But Grodner et al.’s proposal predicts that pragmatic answers to sentences like (1) should pattern with ‘false’ answers to sentences like (5), for which the evaluation also requires representing both a minimum and a maximum, and that both of them take significantly longer than logical answers.

(5) a. Only some parrots are birds.
b. Only some dogs are mammals.

In contrast to this prediction, Bott et al. (2012) found that ‘false’ answers to sentences like (5) pattern with logical answers to sentences like (1), and that both are significantly faster than pragmatic answers.

Geurts et al. (2010) provide further evidence that contradict Grodner et al.’s explanation. They investigated how long participants took to verify sentences such as the following:

(6) a. There are exactly three As.
b. There are at least three As.
c. There are more than three As.

If we follow Grodner et al.’s reasoning, ‘exactly three’ is more complex than the other quantifiers, since it requires representing both a minimum and a maximum, whereas ‘at least three’ and ‘more than three’ only involve a minimum. Nevertheless, participants were significantly faster and more accurate when verifying sentences like (6a) than (6b) or (6c). This result indicates that circumbounded interpretations are not always more complex than lower-bounded interpretations.

In the absence of an alternative explanation for the difference in reaction times found by Bott & Noveck, I assume, for the moment, that the computation of scalar inferences is time-consuming. In the next section, I explain that pragmatic and conventionalist accounts make divergent predictions about when the difference in reaction times found by Bott & Noveck is expected to occur. Afterwards, I put these predictions to the test in a sentence-picture verification task.

3.2.2 When are scalar inferences computed?

According to pragmatic accounts, scalar inferences are a kind of conversational implicature. If this is correct, the difference in response times between literal and pragmatic answers is predicted to occur after the parsing stage. All listeners initially arrive at a literal interpretation of the sentence. Some of them take this interpretation as input to further pragmatic reasoning, which leads to a circumbounded construal of 'some'. Conventionalist theories, on the other hand, assume that listeners who compute a scalar inference immediately arrive at a circumbounded interpretation. Assuming that the difference in response times for literal and pragmatic answers is caused by the computation of a scalar inference, it therefore seems natural to suppose that, according to these theories, upper-bounded answers lead to longer parsing times instead. We put these predictions to the test in an experiment.

Each trial in our experiment started out with a sentence. Participants read this sentence and pressed a button once they had understood it, whereupon it disappeared and was replaced with a picture after a 250 millisecond interval. Participants then indicated if the sentence was true or false as a description of the picture. Both reading times and decision times were measured. We assume that reading times provide a reliable approximation of parsing times (cf. Aaronson & Scarborough 1976, Geurts et al. 2010). More precisely, we assume that participants finish parsing the sentence before they enter the decision stage. In support of this assumption, the reading stage was separated from the decision stage by a 250 millisecond interval, which makes it unlikely that that participants were still in the process of parsing the target sentence while in the decision stage. Wrap-up effects generally take around 150 milliseconds at the end of a paragraph boundary, and around half of that at the end of a sentence (Just & Carpenter 1980). We thus do not consider it particularly likely that parsing efforts leaked into the decision stage. In the results section, we provide further evidence that this assumption is warranted.

Based on this assumption, we predict longer reading times for pragmatic answers than for literal answers if scalar inferences are computed during the parsing stage, and longer decision times if they are derived afterwards. Another possibility is that the increased response times occur both during and after the parsing stage, in which case we predict longer reading and decision times. In the next section, we present the details of our experiment.

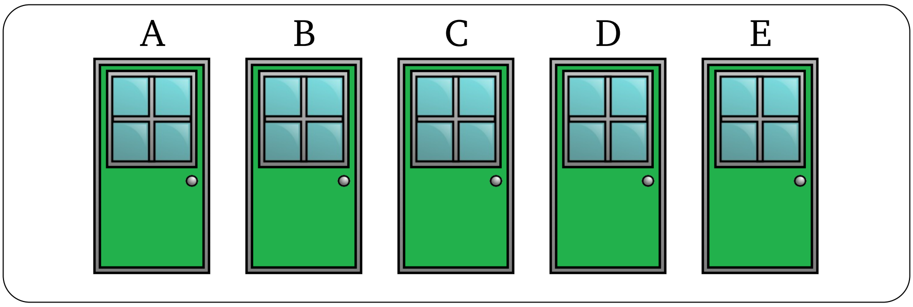


Figure 3.1: Target picture used in the experiment. The target sentence 'You may open some of the doors' is true on its literal interpretation but false if an upper-bounded inference is derived.

The experiment (I)

Participants

Seventy-four students from the Psychology department at the Katholieke Universiteit Leuven, all native speakers of Dutch, were paid for participating in this experiment. The participants' ages ranged between 18 and 28. Their mean age was 20. Eighteen of the participants were male.

Materials

All of the sentences were of the form 'You may open…'. The corresponding pictures consisted of five doors, labeled A through E. Each of these doors was either green or red, indicating whether it was allowed to open it. Some but not all of the doors had windows in them. The target trial consisted of (7), followed by a situation in which all of the doors were green, as depicted in Figure 1.

(7) Je mag sommige deuren openen
You may some doors open
'You may open some of the doors.'

Although, in principle, this sentence is syntactically ambiguous, only the reading where the existential quantifier takes scope over the modal operator is naturally available in Dutch. Its English translation has an interpretation according to which the listener is allowed to freely choose which doors she wants to open. This reading is unavailable in the Dutch original, which only allows for an interpretation according to which there is a determinate set of doors that is allowed to be opened.

The target trial occurred five times in the experiment, which consisted of 80 trials in total. Sentence (7) was used an additional six times in control situations in which it was uncontroversially true or false. All other trials involved different quantifiers in both positive and negative environments, disjunctions, and conditional statements. The results of some of these conditions will be discussed in section 3.3. Examples are given in (8) in the situation depicted in Figure 3.1.

(8) a. Je mag minstens drie deuren openen
You may at least three doors open
'You may open at least three doors.'
b. Je mag precies vier deuren openen
You may exactly four doors open
'You may open exactly four doors.'

Participants agreed almost unanimously (in 92% and 97% of the cases) that these sentences were true and false, respectively. Such fillers were included to ensure that participants understood the significance of the door types. The order of the trials was pseudo-randomised across four lists, making sure that target or control trials never occurred consecutively.

Procedure

The experiment was run on a desktop computer using the E-Prime software package. On each trial, the target sentence was displayed first. Participants were instructed to press the space bar as soon as they had read and understood the sentence. Thereupon, the sentence disappeared and was replaced by a picture after a 250 millisecond delay. Participants had to decide as quickly as possible whether the sentence was true or false as a description of the depicted situation, and had to register their decision by pressing one of two keys. Reading times were recorded from sentence onset to the point at which the space bar was pressed. Decision times were recorded from situation onset to the point at which the 'yes' or 'no' key was pressed. Participants were familiarised with the paradigm by means of six practice trials.

Results

Eight participants were excluded from the analysis because they answered less than 70% of the fillers correctly. We furthermore removed trials for which either the reading or decision time differed more than three standard deviations from the corresponding mean, which resulted in the removal of 3% of the data. The target sentence was judged true in 52% of the target trials. Thirty-four participants responded with a single type of response, either all literal (16 participants) or pragmatic (18 participants). The average reading and decision times for sentences containing the scalar term 'some' are summarised in Figure 3.2.

Although the reading times might seem especially fast (on average 1551 milliseconds), note that all of the sentences in the experiment were structurally the same, varying only in the polarity and the quantifier involved. In this respect, our experiment differs, for example, from Bott & Noveck's. In their experiment, sentences also contained unpredictable content words like 'dogs' and 'mammals'. As further evidence for our assumption that participants read and understood the sentence before proceeding to the decision stage, the variation in reading times for all trials,

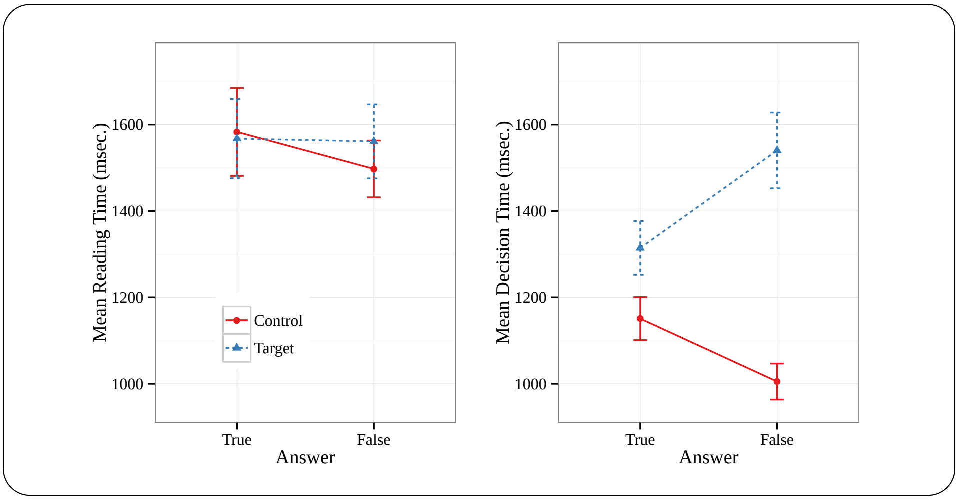


Figure 3.2: Mean reading and decision times for the sentence 'You may open some of the doors' in target and control conditions. Error bars represent standard errors.

where the standard deviation totalled 1883 milliseconds, was almost twice as large as the variation in decision times, where the standard deviation was 1012 milliseconds. Differences in complexity between sentences mainly caused differences in reading times, as would be expected if participants parse the sentence during the reading stage.

All of the following analyses were conducted on the basis of reaction times that were logarithmised to reduce the positive skewness of their distribution. In order to establish for which conditions there was a significant difference in reaction times, we first analysed the reaction times of ambivalent participants, who gave both literal and pragmatic answers to target trials. An important advantage of this analysis is that overall differences in reaction times between participants are balanced. For instance, participants who provide primarily literal responses might be overall faster than participants who provide primarily upper-bounded responses. Such potential idiosyncracies are balanced in this analysis, which only takes into account differences within participants. We analysed our results by fitting a series of linear mixed models predicting reaction times, with condition (control or target) and response (true or false) as independent factors, and participants as random factor (Barr et al. 2013). We conducted our analysis in R, a programming language and environment for statistical computing (R Development Core Team 2006), using the `nlme` package (Pinheiro et al. 2013). The results of this analysis are summarised in Table 3.1.

We first determined if we replicated Bott & Noveck's finding that pragmatic answers take longer than literal answers by analysing the summed reading and decision times. There was a significant interaction between condition and response. Closer investigation into this interaction showed that 'false' answers took significantly

Model term	***β***	**SE**	***t*(675)**	***p***
Reading times				
Condition ∗ Response	-0.03	0.06	-0.59	.555
Condition	0.03	0.04	0.85	.395
Response	-0.03	0.04	-0.94	.345
Decision times				
Condition ∗ Response	-0.19	0.07	-2.72	.007
Condition	0.30	0.05	6.23	.000
Response	0.09	0.04	2.01	.045
Total times				
Condition ∗ Response	-0.22	0.09	-2.32	.002
Condition	0.33	0.07	5.07	.000
Response	0.05	0.06	0.89	.374

Table 3.1: Parameters of the linear mixed models predicting logarithmised reaction times in the reading stage, decision stage, and entire trial, on the basis of condition (control or target) and response (true or false).

longer than 'true' answers in the target condition (β = -0.20, *SE* = 0.08, $t(283)$ = -2.42, p = .016), while there was no significant difference between these responses in the control condition (β = 0.05, *SE* = 0.06, $t(319)$ = 0.07, p = .429). Given that we replicated Bott & Noveck's finding, we further investigated whether this difference in reaction times was caused by differences in reading times, decision times, or both.

Analysing the reading times, we did not find a significant interaction between condition and response, and no main effects of either condition or response. For the decision times, there was a significant interaction between condition and response. Closer investigation showed that 'false' answers took significantly longer than 'true' answers in the target condition (β = -0.13, *SE* = 0.06, $t(283)$ = -2.07, p = .039), while 'true' answers took significantly longer in the control condition (β = 0.09, *SE* = 0.04, $t(319)$ = 2.16, p = .032). So the difference in response times between literal and upper-bounded answers occurred during the decision stage, and not during the reading stage. An explanation for the difference in decision times in the control condition is that the 'false' situation was relatively easier to recognise, since it depicted only red doors, whereas the 'true' situation depicted a combination of red and green doors.

We corroborated the results of our within participants analysis in a between participants analysis. To this end, we categorised participants as literal or pragmatic responders, depending on whether they provided more literal or pragmatic answers. For each literal responder, the average reaction time for their logical answers was calculated, and correspondingly for the pragmatic responders. These averages were compared by means of Welch's t-tests. In line with our previous findings, both the summed reading and decision times ($t(71) = 1.68$, one-sided $p = .048$), and the decision times ($t(66) = 2.10$, one-sided $p = .020$) were significantly faster for literal responders than for pragmatic responders, while there was no such difference in reading times ($t(72) = 0.41$, two-sided $p = .685$). These results strengthen the conclusion that, overall, pragmatic answers take longer than literal answers, and that this difference occurs in the decision stage rather than the reading stage.

3.2.3 General discussion

Bott & Noveck (2004) found that it takes participants longer to judge sentences like (9) false than it takes to judge them true:

(9) a. Some parrots are birds.
    b. Some dogs are mammals.

In our experiment, we investigated whether this difference in response times occurs during the parsing of the sentence or afterwards. To that end, we separated reading times from decision times, assuming that participants finish parsing the sentence before they enter the decision stage, which, in addition, was temporarily separated by a 250 millisecond interval. It was found that the computation of scalar inferences leads to longer decision times but not to longer reading times.

Can this finding be incorporated within the three theories of scalar inferences we discussed? It fits in neatly with a pragmatic account, according to which the derivation of scalar inferences involves reasoning about the speaker's intentions on the basis of the compositional meaning of the sentence. In this framework, participants first arrive at a literal interpretation, which, if a scalar inference is computed, is subsequently used as input to further pragmatic reasoning. This reasoning process, which occurs after the sentence has been parsed, is responsible for the increase in decision times.

Note that, while our finding can be naturally explained within a pragmatic account, this does not mean that the account would have been disproved had we found that pragmatic answers lead to longer reading times instead. In such a scenario, it might have been argued that participants already engage in pragmatic reasoning before proceeding to the decision stage. This assumption, however, would have meant an additional complication of the theory that might be falsified by further experiments. For example, it would entail that reading times for the control conditions should lie somewhere in between the reading times for literal and pragmatic answers, since in half of these trials, too, participants should have computed a scalar inference. The finding that pragmatic answers lead to longer decision times than literal answers can be explained more naturally within a pragmatic account of scalar inferences.

The opposite is true for both conventionalist theories; incorporating a post-parsing latency within a lexical or grammatical account is not trivial. We discuss three potential explanations, only the last of which provides an adequate account of our data. Recall that, according to both lexical and grammatical theories, a sentence such as (10) is ambiguous between a literal and a pragmatic interpretation:

(10) You may open some of the doors.

A first explanation for the difference in decision times between literal and pragmatic answers is that participants hold both possible interpretations in mind until they reach the decision stage, at which point they make a choice between them. The difference in decision times is caused by the circumbounded interpretation being more difficult to retrieve than the literal interpretation. Alternatively, but similarly, participants arrive at the literal interpretation by default, and then, at least in some cases, reinterpret the sentence in the decision stage to compute the upper bound.

An important problem with this first line of explanation is that it contradicts the Principle of Charity which states that hearers normally try to interpret an ambiguous sentence in such a way that it is true (section 2.2.3). In the first case, participants have both a lower-bounded and a circumbounded interpretation in mind, and then choose the one that makes the sentence false instead of the one that makes it true. The situation is even more striking in the second case, where participants initially compute an interpretation that makes the sentence true, and then reanalyse the sentence to retrieve an interpretation that makes it false. To salvage this first line of explanation, it would be necessary to assume that the choice between both interpretations, in the former case, or the reanalysis of the sentence, in the latter, is somehow impervious to the principle that listeners prefer true interpretations over false ones. If that were the case, the difference in response times between literal and pragmatic answers might also be attributed to the perceived tension between the aforementioned preference for truth and the decision to indicate that the sentence is false. However, given that all of these explanations are at odds with such a widely accepted principle, we do not consider them particularly attractive, especially if a less controversial explanation is available.

A second possible explanation is based on the observation that, at least in some situations, sentences like (10) will only imply that the speaker does not know whether you may open all of the doors. Lexical and grammatical theories agree that this ignorance inference cannot be accounted for by means of an ambiguity, and requires pragmatic reasoning about the speaker's intentions (see the discussion in section 2.1.1). It seems plausible that this kind of reasoning occurs after the compositional meaning of the sentence has been calculated. Perhaps, then, the difference in decision times is due to the computation of this ignorance inference, which caused participants to judge the sentence false.

One issue with this explanation is that the computation of an ignorance inference is not sufficient for concluding that the sentence is false in a situation in which you may open all of the doors. After all, this situation does not disprove the conclusion that the speaker does not know whether you may open all of the doors, unless further assumptions about his competence are made. But if we assume that it is possible to arrive at a strong enough inference to judge (10) false by means of pragmatic reasoning about the speaker's intentions, lexical and grammatical theories would have to incorporate a pragmatic mechanism for deriving scalar inferences. While it is possible to adopt a hybric account according to which scalar inferences can be computed both by semantic and pragmatic means, doing so would change the landscape of theories and the debate between pragmatic and conventionalist accounts of scalar inferences.

A final explanation, which we consider the most plausible, is that decision times for pragmatic answers are longer than for literal answers because determining the truth of the circumbounded interpretation is somehow more difficult than evaluating a literal interpretation. As argued before, this difficulty cannot be ascribed to circumbounded interpretations requiring the representation of both a maximum and a minimum, since pragmatic answers to sentences like (9) are significantly slower than response times for sentences like (11), which also involve the representation of a maximum and a minimum:

(11) a. Only some parrots are birds.
b. Only some dogs are mammals.

But there is an important difference between these sentences: in the case of (11), the upper bound is made explicit, whereas it remains implicit in the case of (9). Explicit information is known to be more available than implicit information, which can be seen, for instance, by observing that only an explicit upper bound can be referred to with 'because'. (See section 1.2.1 for some other differences.) For that reason, (12b) is pragmatically felicitous, while (12a) is not (Horn 2004):

(12) a. Because you ate some of your spinach, you don't get dessert.
b. Because you ate only some of your spinach, you don't get dessert.

Given this difference in availability, it does not seem unreasonable to suppose that an implicit upper bound is more difficult to reason with than an explicit upper bound, which might explain the increased decision times for pragmatic answers. A lexical or grammatical account can thus assume that the difference in decision times is caused by the increased difficulty of reasoning with implicit information. In the next section, we evaluate this explanation in a follow-up analysis based on the results for three other kinds of quantity inferences.

3.3 Processing quantity inferences

3.3.1 A closer look at the verification paradigm

In the previous section, we have seen that, in situations in which all of the doors were allowed to be openened, participants took longer to judge the following sentence false than they took to judge it true:

(13) You may open some of the doors.

What causes this difference in response times between literal and pragmatic answers? It seems inevitable that it is somehow related to the process of computing the scalar inference, since there was no difference in response times in the control conditions. However, it is unclear which step in the calculation process caused the delay in response times. There are at least three possibilities that suggest themselves:

i. *Reasoning with implicit information is time-consuming.*
   This explanation was proposed in the previous section. Scalar inferences constitute implicit information. Such information is less available or prominent than explicit information, and might therefore be more difficult to reason with. Pragmatic responses involve reasoning with implicit information while literal responses do not. This might explain the difference in response times.
ii. *Reasoning with alternatives is time-consuming.*
   Computing the scalar inference that is associated with (1a) requires considering the statement 'You may open all of the doors'. According to this explanation, considering alternative statements is time-consuming.
iii. *Constructing alternatives is time-consuming.*
   This explanation was suggested by Chemla & Bott (2014). The computation of scalar inferences is unlike that of other quantity inferences in that it requires constructing alternatives by substituting elements in the speaker's utterance with expressions from the lexicon. If this explanation is correct, the retrieval of expressions from the lexicon is time-consuming.

In order to decide between these competing explanations, it is necessary to look at the processing times of other kinds of quantity inferences, such as exhaustivity inferences, conditional perfection, and free choice inferences (see section 1.2.2 for discussion). Relevant examples are given below:

(14) I had strawberries for breakfast. *Exhaustivity inference*
$\rightsquigarrow$ I only had strawberries for breakfast.

(15) If you mow the lawn, I will give you \$5. *Conditional perfection*
$\rightsquigarrow$ I will not give you \$5 unconditionally.

(16) You may have coffee or tea. *Free choice inferences*
$\rightsquigarrow$ You may choose between having coffee and tea.

Let us see how these inferences are predicted to behave given the three hypotheses outlined above. First, consider the view that the processing cost of quantity inferences depends on whether the verification procedure involves reasoning with implicit information. To test whether a quantity inference is implicit, we determine how felicitous it is to refer to it with the expression 'because' compared to the felicity of 'because' when the quantity inference is made explicit. Exhaustivity inferences and conditional perfection pattern with scalar inferences in that reference by means of 'because' results in an infelicitous discourse. That is, according to our intuitions, the (b) sentences in the following minimal pairs are more felicitous than the corresponding (a) sentences:

(17) a. Because I had strawberries for breakfast, I'm still hungry.
b. Because I only had strawberries for breakfast, I'm still hungry.

(18) a. Because you'll give me $5 if I mow the lawn, I'm calling the police.
b. Because you'll only give me $5 if I mow the lawn, I'm calling the police.

In contrast with scalar inferences, exhaustivity inferences, and conditional perfection, free choice inferences constitute explicit information. So in the following pair of sentences, neither is more felicitous than the other one.

(19) a. Because I may open door A or B, I'm opening door A.
b. Because I may choose between opening door A and B, I'm opening door A.

In summary, if the processing cost of quantity inferences depends on whether it involves reasoning with implicit information, only free choice inferences are not predicted to lead to a significant delay in response times.

According to the second hypothesis, the processing cost of quantity inferences depends on whether their computation involves reasoning with alternatives. Since all of the quantity inferences under discussion involve reasoning with alternatives, this hypothesis predicts a significant delay in response times across the board.

Lastly, the third hypothesis states that whether the derivation of a quantity inference takes time depends on whether the construction of the alternatives requires substituting elements in the speaker's utterance with expressions from the lexicon. This is the case for scalar inferences but not for any of the other three kinds of quantity inferences. Therefore, if this hypothesis is correct, only the computation of scalar inferences should be associated with a processing cost.

Table 3.2 summarises the characteristics of each of the four kinds of quantity inferences we included in our analysis. Our investigation builds on that of Chemla & Bott (2014), who compared scalar inferences and free choice inferences. In the next section, we discuss their task and results.

	Implicit	Alternatives	Substitution
Scalar inference	+	+	+
Exhaustivity inference	+	+	–
Conditional perfection	+	+	–
Free choice inferences	–	+	–

Table 3.2: Characteristics of the four kinds of quantity inferences under investigation.

3.3.2 The experimental record

Chemla & Bott investigated whether the computation of free choice inferences is time-consuming by means of a sentence verification task. Since it is difficult to construct sentences that are introspectively true on their literal interpretation but false if free choice inferences are computed, participants were first familiarised with the following vignette:

> The planet faces imminent destruction. A group of zoologists and engineers are allowed to save one object each. Given their expertise, zoologists are only allowed to save living beings, while engineers are only allowed to save artifacts.

Afterwards, participants read sentences in a forced reading experiment similar to Bott & Noveck's, and had to determine their truth values based on the aforementioned vignette. The target condition consisted of sentences like:

(20) a. Beverly-the-zoologist is allowed to save a hammer or a lion.
b. Essie-the-engineer is allowed to save a kangaroo or a fork.

These sentences are true on their literal interpretation but false if free choice inferences are computed, since zoologists are not allowed to save artifacts, and engineers are not allowed to save living beings.

Like in Bott & Noveck's experiment, many participants were ambivalent about the truth of sentences like (20), varying their answer between structurally similar trials. Chemla & Bott analysed the response times of these ambivalent participants, but they did not find a significant difference between literal and pragmatic answers. They also included a control condition involving sentences that were unambiguously true or false:

(21) a. Federico-the-engineer is allowed to save a hammer or a fork.
b. Martina-the-engineer is allowed to save a lion or a kangaroo.

Comparing these two conditions, Chemla & Bott found an interaction between condition (target or control) and answer (true or false). In the control condition, 'false' answers were significantly slower compared to 'true' answers than in the target condition. So computing free choice inferences reduced the increase in response times that was present in the control condition. In this respect, free choice inferences differ from scalar inferences, for which it was found that 'false' answers took longer than 'true' answers in the target but not in the control condition.

What do these results tell us about the three hypotheses outlined above? The finding that free choice inferences are not associated with a processing cost seems to contradict the hypothesis that the processing cost of scalar inferences is caused by reasoning with alternatives, since the derivation of free choice inferences also involves alternatives. But this conclusion depends on the contested assumption that free choice inferences are a kind of quantity inferences. In other words, Chemla & Bott's results are consistent with the hypotheses that the processing cost of quantity inferences depends on their informational status or on whether the construction of alternatives involves substitution (i.e., hypotheses *i* and *iii* in the list above) but also with the hypothesis that the processing cost of quantity inferences is caused by having to reason with alternatives if it turns out that free choice inferences are not a kind of quantity inferences after all.

In order to obtain more decisive evidence about the three explanations for the processing cost of scalar inferences, we tested four kinds of quantity inferences in the task that was introduced in the previous section. The sentences we used to test these four inferences are provided in Table 3.3. In the target condition, these sentences were followed by a situation in which the corresponding quantity inference was false; in the control condition, by a situation in which they were unambiguously true or false. Table 3.3 provides example situations from both conditions for each kind of quantity inference.

In the target sentences for free choice inferences, we varied whether the noun 'door' was repeated in the second disjunct. The target sentence for testing exhaustivity inferences was chosen the basis of a pretest. We intuited that participants would be reluctant to judge sentences like 'You may open door A' false in a situation in which door A was not the only door that was allowed to be opened. Therefore we presented 25 participants, who were drafted on Amazon's Mechanical Turk, with the sentences in (22). Each sentence was followed by a situation in which its exhaustivity inference was false. Participants were asked to indicate if the sentences were true or false in these situations. The experiment also included 8 filler items, for which the truth value of the sentence was unambiguous.

(22) a. You may open doors B, C and D.
b. You may not open door B or C.
c. You may not open door D.
d. You may not open any door that has a window.

The percentages of 'false' responses for the sentences in (22) were 4, 10, 20, and 26, respectively. For our current purposes, it was important that participants provide a substantial number of literal and pragmatic responses. Therefore we decided to include (22d) in the experiment, since that sentence yielded the highest percentage of 'false' responses.

Scalar inference

You may open some of the doors

⇝You may not open all of them

	Control					Target				
	A	B	C	D	E	A	B	C	D	E
True	🟩	🟥	⊞🟩	⊞🟩	⊞🟩	🟩	🟩	⊞🟩	⊞🟩	⊞🟩
	A	B	C	D	E					
False	🟥	🟥	⊞🟥	⊞🟥	⊞🟥					

Free choice inference

You may open door A or (door) B

⇝You may open either door

	Control					Target				
	A	B	C	D	E	A	B	C	D	E
True	🟩	🟩	⊞🟥	⊞🟥	⊞🟥	🟩	🟥	⊞🟥	⊞🟥	⊞🟥
	A	B	C	D	E					
False	🟥	🟥	⊞🟥	⊞🟥	⊞🟥					

Conditional perfection

You may open every door if it has a window

⇝ You may not open every door

	Control					Target				
	A	B	C	D	E	A	B	C	D	E
True	🟥	🟥	⊞🟩	⊞🟩	⊞🟩	🟩	🟩	⊞🟩	⊞🟩	⊞🟩
	A	B	C	D	E					
False	🟥	🟥	⊞🟥	⊞🟩	⊞🟩					

Exhaustivity inference

You may not open any door with a window

⇝You may open every door without a window

	Control					Target				
	A	B	C	D	E	A	B	C	D	E
True	🟩	🟩	⊞🟥	⊞🟥	⊞🟥	🟩	🟥	⊞🟥	⊞🟥	⊞🟥
	A	B	C	D	E					
False	🟩	🟩	⊞🟩	⊞🟥	⊞🟥					

Table 3.3: Target and control situations for the four kinds of quantity inference tested in the experiment.

In the next section, we report the results of the experiment. Since this was the same experiment as the one described in section 3.2, we omit discussion of participants, materials, and procedure. The interested reader is referred to that section.

The experiment (II)

Results

For scalar inferences (52% literal responses), free choice inferences (53%), and conditional perfection inferences (53%), literal responses were roughly as frequent as pragmatic responses. In line with our pretest, participants in the case of exhaustivity inferences had a distinct preference for literal responses: 81% of the responses were literal. Some participants consistently gave literal or pragmatic responses. For scalar inferences (17 literal responders versus 18 pragmatic responders), free choice inferences (20/17), and conditional perfection inferences (19/15), there were roughly as many literal responders as there were pragmatic responders. For exhaustivity inferences there were 27 literal responders and no pragmatic responders. Even in this condition, however, the large majority of the participants (42 out of 69) were ambivalent about the truth of the target sentence, varying their answer between literal and pragmatic responses.

All of the following analyses were conducted on the basis of reaction times that were logarithmised to reduce the positive skewness of their distribution. In order to establish for which conditions there was a significant difference, we analysed the reaction times of ambivalent participants, who gave both literal and pragmatic answers to target trials. The analyses involving reading and decision times are discussed separately.

Reading times

For each kind of inference, we constructed a mixed model predicting reading times. Condition (control or target), response (true or false), their interaction, and trial number were included as independent factors. Trial number was included to account for learning effects: it seems plausible to suppose that participants require increasingly less time reading the sentences during the experiment. In all of the subsequent analyses, random slopes were included for participants and items (cf. Barr, Levy, Scheepers, & Tily 2013).

The effect of condition was not significant for scalar inferences (β = 0.04, *SE* = 0.05, $t(662) < 1$), conditional perfection inferences (β = -0.04, *SE* = 0.05, $t(645) < 1$), or exhaustivity inferences (β = -0.01, *SE* = 0.08, $t(647) < 1$). There was, however, a significant effect of condition for free choice inferences (β = -0.14, *SE* = 0.05, $t(139)$ = -2.93, $p < .001$): reading times were significantly shorter in the target condition than in the control condition. On closer inspection, this effect is caused by the skewed distribution of short ('You may open door A or B') and long

('You may open door A or door B') target sentences: 72% of the sentences were long in the target condition, as opposed to 55% in the control condition. A mixed model with sentence length and trial number as independent factors indicates that reading times for long statements were significantly longer than for short statements (β = -0.20, *SE* = 0.04, $t(491)$ = -5.71, $p < .001$).

Response did not have a significant effect for scalar inferences (β = -0.01, *SE* = 0.04, $t(662) < 1$), free choice inferences (β = -0.06, *SE* = 0.05, $t(638)$ = -1.33, p = .183), conditional perfection inferences (β = -0.03, *SE* = 0.05, $t(645) < 1$), or exhaustivity inferences (β = -0.05, *SE* = 0.05, $t(647) < 1$). The interaction between condition and response was also not significant for scalar inferences (β = -0.05, *SE* = 0.07, $t(662) < 1$), free choice inferences (β = 0.11, *SE* = 0.07, $t(638)$ = 1.54, p = .125), conditional perfection inferences (β = 0.02, *SE* = 0.08, $t(645) < 1$), or exhaustivity inferences (β = 0.04, *SE* = 0.09, $t < 1$). Trial number had a significant effect for scalar inferences (β = -0.00, *SE* = 0.00, $t(662)$ = -4.48, p = .000), free choice inferences (β = -0.01, *SE* = 0.00, $t(638)$ = -7.14, p = .000), and conditional perfection inferences (β = -0.00, *SE* = 0.00, $t(645)$ = -3.73, p = .000), but not for exhaustivity inferences (β = -0.00, *SE* = 0.00, $t(647)$ = -1.51, p = .130).

Decision times

The average decision times for each kind of quantity inference are summarised in Figure 3.3. For each kind of inference, we constructed a mixed model predicting decision times. Condition (control or target), response (true or false), their interaction, and trial number were included as independent factors. As already observed in the previous section, there was a significant interaction between condition and response in the case of scalar inferences (β = -0.19, *SE* = 0.07, $t(662)$ = -2.72, p = .007). A mixed model with decision times as dependent variable, and response as independent variable shows that in the target condition, 'false' answers took significantly longer than 'true' answers (β = -0.14, *SE* = 0.07, $t(275)$ = -2.08, p = .038), whereas the opposite was true in the control condition (β = 0.11, *SE* = 0.04, $t(320)$ = 2.87, p = .004). For conditional perfection inferences (β = -0.01, *SE* = 0.06, $t(645) < 1$), and exhaustivity inferences (β = -0.01, *SE* = 0.08, $t(647) < 1$) the interaction between condition and response was not significant, and it was marginally so for free choice inferences (β = 0.11, *SE* = 0.06, $t(638)$ = 1.68, p = .094). In this condition, however, the interaction went in the opposite direction as was the case for scalar inferences: 'false' answers took longer than 'true' answers in the control condition, and vice versa in the target condition. A mixed model with decision times as dependent factor, and response as independent factor shows that the difference was significant in the control condition (β = -0.11, *SE* = 0.04, $t(296)$ = -2.99, p = .003), but not in the target condition (β = 0.02, *SE* = 0.06, $t(275) < 1$).

Furthermore, there was a main effect of trial number for each kind of inference:

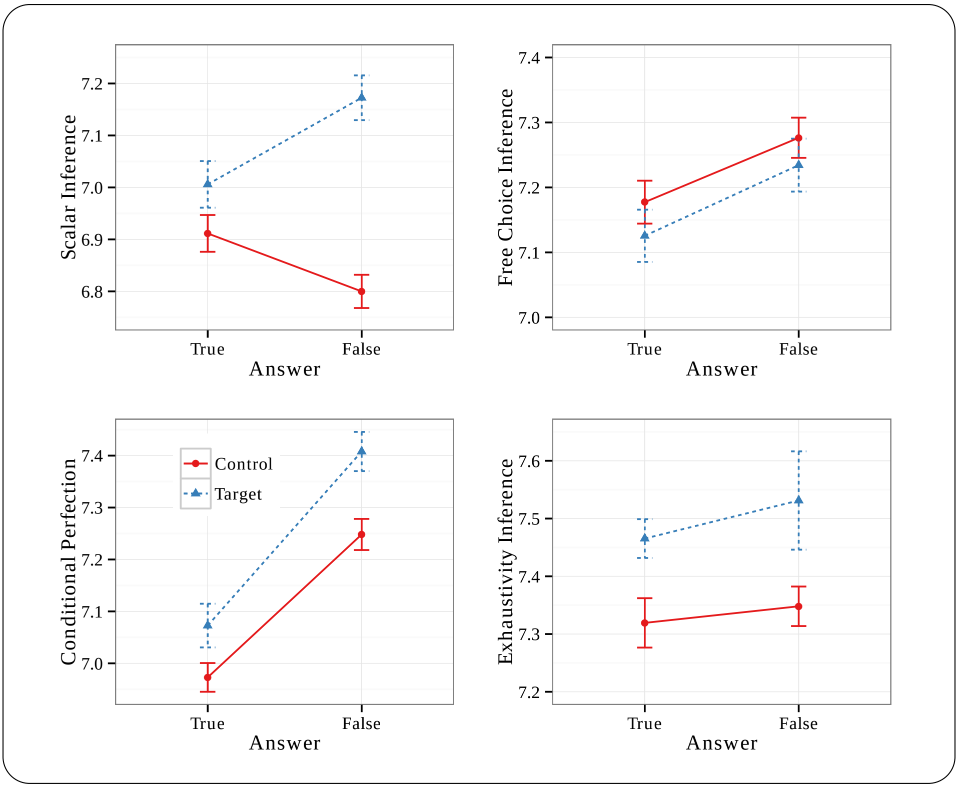


Figure 3.3: Mean logarithmised decision times for the four kinds of quantity inference. Error bars represent standard errors.

scalar inferences (β = -0.00, *SE* = 0.00, $t(662)$ = -4.41, p = .000), free choice inferences (β = -0.00, *SE* = 0.00, $t(638)$ = -3.49, p = .000), conditional perfection inferences (β = -0.00, *SE* = 0.00, $t(645)$ = -2.81, p = .005), and exhaustivity inferences (β = -0.00, *SE* = 0.00, $t(647)$ = -3.32, p = .001). Condition had a significant effect for scalar inferences (β = 0.33, *SE* = 0.05, $t(662)$ = 6.81, p = .000), conditional perfection inferences (β = 0.12, *SE* = 0.05, $t(645)$ = 2.66, p = .008), and exhaustivity inferences (β = 0.18, *SE* = 0.07, $t(647)$ = 2.52, p = .012), but not for free choice inferences (β = -0.07, *SE* = 0.04, $t(638)$ = -1.53, p = .126). Lastly, response had a significant effect for scalar inferences (β = 0.10, *SE* = 0.04, $t(662)$ = 2.25, p = .025), free choice inferences (β = -0.09, *SE* = 0.04, $t(638)$ = -2.15, p = .031), conditional perfection (β = -0.27, *SE* = 0.04, $t(645)$ = 6.44, p = .000), but not for exhaustivity inferences (β = -0.02, *SE* = 0.05, $t(647) < 1$).

To determine if the interaction between condition and response was the same for different kinds of inferences, we made pairwise comparisons. For each pair of inferences, we constructed a mixed model predicting decision times, with inference type, condition, response, and their interactions as independent factors. As in our previous analyses, we also included trial number as an independent factor. There were significant three-way interactions between inference type, condition, and response for scalar inferences and free choice inferences (β = -0.30, *SE* = 0.09, $t(1369)$ = -3.22, p = .001), scalar inferences and conditional perfection inferences (β = -0.24, *SE* = 0.09, $t(1376)$ = -2.61, p = .009), and scalar inferences and exhaustivity inferences (β = -0.25, *SE* = 0.11, $t(1378)$ = -2.33, p = .020). The three-way interactions for free choice inferences and conditional perfection inferences (β = 0.10, *SE* = 0.09, $t(1352)$ = 1.13, p = .261), free choice inferences and exhaustivity inferences (β = 0.12, *SE* = 0.10, $t(1354)$ = 1.21, p = .227), and conditional perfection inferences and exhaustivity inferences (β = 0.02, *SE* = 0.10, $t(1361) < 1$) were not significant. So the specific interaction between condition and response that Bott & Noveck found for scalar inferences is absent for all of the other kinds of quantity inference that were investigated.

3.3.3 General discussion

Unlike scalar inferences, the processing of exhaustivity inferences, conditional perfection, and free choice inferences is not accompanied by an increase in decision times. This finding replicates and extends previous results by Bott & Noveck (2004) and Chemla & Bott (2014).

In the introduction to this section, we outlined three possible explanations for the finding that the computation of scalar inferences leads to a delay in response times. These can be succinctly summarised as follows:

i. Reasoning with implicit information is time-consuming.
ii. Reasoning with alternatives is time-consuming.

iii. Constructing alternatives by substituting elements from the speaker's utterance with expressions from the lexicon is time-consuming.

The unique status of scalar inferences argues against the first hypothesis, according to which the increase in decision times for scalar inferences is caused by the process of reasoning with implicit information. Like scalar inferences, conditional perfection and free choice inferences are also implicit pieces of information. But neither of these inferences was associated with an increase in decision times. The results also contradict the hypothesis that the increase in decision times for scalar inferences is related to the presence of alternatives in the derivation procedure. After all, alternatives feature in the derivation of all four kinds of quantity inferences whereas only scalar inferences led to a delay in response times.

The results of this experiment are compatible with the third hypothesis, which states that the increase in decision times for scalar inferences is caused by the need to consult the lexicon for constructing alternatives. In this respect, scalar inferences differ from the other kinds of quantity inferences included in the experiment. So the critical difference between scalar inferences and the three other kinds of quantity inferences centers on the nature of the required alternatives.

What about the assumption that free choice inferences are a kind of quantity inferences? Since it is unclear what the predictions are if this assumption is false, the results of this experiment do not resolve this question. Nonetheless, we observe that the behaviour of free choice inferences is in line with the assumption that they are a kind of quantity inferences. The results can therefore be adduced as indirect evidence in favour of this assumption.

Returning to the previous section, the falsification of the hypothesis that reasoning with implicit information is time-consuming reopens the challenge for conventionalist accounts to provide an explanation for the post-parsing processing cost of scalar inferences. If scalar inferences are part of the literal meaning of an utterance, how is it possible that their computation occurs after having parsed the corresponding sentence? We find it difficult to see how one could solve this conundrum even in principle.

3.4 Conclusion

According to conventionalist accounts, the circumbounded interpretation of scalar expressions is part of their literal meaning. One of the corollaries of this assumption seems to be that the processing cost that is associated with the computation of scalar inferences is incurred during the parsing of the sentence and not afterwards. We tested this prediction in a straightforward sentence-picture verification task, and discovered that the processing cost occurs after the sentence has been parsed. This result seems surprising from a conventionalist perspective: if scalar inferences are computed in tandem with the literal meaning of a sentence, how is it possible that we see the psychological traces of this computation process after the literal meaning has been determined?

However, in order to evaluate this finding, it is important to determine what causes the increase in verification times for scalar inferences. Therefore we included three other kinds of quantity inferences in the experiment: exhaustivity inferences, conditional perfection, and free choice inferences. These other quantity inferences differed from scalar inferences along two dimensions: their status as implicit or explicit information and the structural nature of the required alteratives. None of these three quantity inferences were associated with a processing cost.

So scalar inferences are unique in leading to an increase in verification times. Scalar inferences are also unique in that their computation involves alternatives that have to be generated by substituting scalar expressions for their scalemates. We have therefore proposed that this feature explains the unique behaviour of scalar inferences in the verification paradigm.

In order to provide further evidence for this hypothesis, more kinds of quantity inference would have to be tested. Unfortunately, there do not seem to be other kinds of quantity inference that pattern with scalar inferences in that the generation of alternatives requires drawing on elements from the lexicon. Therefore an alternative test would be to extend the verification paradigm to a more diverse range of scalar expressions. As will be emphasised in section 4.3, most of the experimental work on scalar inferences has focused on just two scalar expressions: ‘some’ and ‘or’. If structural properties of the alternatives are responsible for the special behaviour of scalar inferences, similar findings are predicted for other scalar expressions like ‘warm’ implying ‘not hot’ and ‘cheap’ implying ‘not free’. In section 4.3, we will show that at least in some respects different scalar expressions display markedly different behaviour in experimental tasks.

In stark contrast to the perhaps undue focus on one or two scalar expressions, researchers have employed a wide range of dependent variables to investigate scalar inferences. To give just a small sample, computing a scalar inference has been operationalised as…

i. looking at a target that satisfied the scalar inference.
ii. answering affirmatively when asked if one would compute the scalar inference.
iii. indicating that the sentence is false if its scalar inference is false.
iv. judging a situation where the scalar inference is true more appropriate than a situation in which it is false.

In the previous chapter, we have seen that these measures are influenced by different meaning aspects. The current chapter has shown that even if a variable measures scalar inferences, it is not obvious which aspect of the computation process it measures. A closer look at the provenance of some of the dependent variables used in experimental research on scalar inferences is therefore warranted.

4

New issues

4.1 Introduction

In the previous two chapters, we have focused on the debate between pragmatic and conventionalist accounts of quantity inference; first by investigating the interpretation of scalar expressions in embedded contexts, and afterwards by analysing the processing cost of different kinds of quantity inferences. In this chapter, I will not be concerned with this debate per se and focus instead on two issues that are not incorporated in any of the current theories of quantity inferences.

The first issue involves the role of typicality and its relationship with literal meaning and pragmatic inferences. In chapter 2, we have seen that each of these three meaning aspects influences the interpretation of scalar expressions. But in that chapter, we left open the question of how these meaning aspects are related. This open question will be addressed in section 4.2, in which we investigate how these three meaning aspects influence the interpretation of quantifiers. The second issue focuses on a tacit assumption that seems to underlie much of the literature on scalar inferences, namely that all scalar expressions behave alike in experimental tasks. In section 4.3, we provide evidence that discredits this *uniformity assumption* and discuss a number of explanations for the differences between scalar expressions.

4.2 Truth and typicality in the interpretation of quantifiers

4.2.1 The interpretation of quantifiers

The interpretation of quantifiers has been investigated from a range of perspectives. The standard view in natural language semantics is that quantifiers denote relations between sets (e.g., Barwise & Cooper 1981, Keenan & Stavi 1986, Montague 1973). For example, ‘All A are B’ is true if the set of A is a subset of the set of B, ‘Some A are B’ is true if the intersection between the set of A and the set of B is nonempty, and ‘Most A are B’ is true if the number of A that are B is greater than the number of A that are not B.

These set-theoretic definitions assign binary truth values: quantified sentences are either true or false in a situation. No finer-grained differences between situations are therefore expected. But, as we have seen in section 2.2, quantified statements often convey finer-grained information than what is encoded in their set-theoretic definitions. For example, ‘Some of the circles are black’ is a better description of a situation in which five out of ten circles are black than of a situation in which nine of them are. To provide another example, Newstead et al. (1987) asked participants to fill in the blanks in sentences like the following, in which the quantifier Q and the total set size N were varied between items:

(1) If Q of a group of N people are male, then _____ people are male.

In addition, Newstead et al. asked participants what they would expect to be the minimum and maximum number of people that satisfied the predicate given the truth of the antecedent. For statements where the total set size was 60, Newstead et al. found the following mean estimates for the previously mentioned quantifiers ‘all’, ‘some’, and ‘most’:

	Min.	**Mean**	**Max.**
All	100	100	100
Some	17	33	45
Most	66	83	90

Table 4.1: Mean estimates (%) for the quantified sentence ‘Q people are male’ in situations in which there are in total 60 people.

The results for ‘all’ are in line with its set-theoretic definition: ‘All people are male’ implies that there are no females. The response ranges for ‘some’ and ‘most’, however, are much smaller than suggested by their set-theoretic definitions. For example, even though, according to its set-theoretic definition, ‘Some people are male’ is true whenever more than one person is male, participants infer from this statement that between 17% and 45% of the people are male. Further research has shown that the precise estimates are influenced by extralinguistic factors, such as:

i. *Total set size*. Newstead et al. (1987) found for some quantifiers that the estimated number of people that satisfies the predicate depends on the total set size. An example is ‘Some people are male’: if there are just twelve people, participants estimate that 37% of them are male, whereas in a situation with ten thousand people their estimation drops to about 27%.
ii. *World knowledge*. Estimates for statements of the form ‘Q A are B’ depend on the estimated likelihood that As are B: they will be higher for statements like ‘Q people find Miss Sweden attractive’ than for statements like ‘Q earthquakes

occurred in California in 1951', since the likelihood of someone finding Miss Sweden attractive is greater than that of an earthquake happening in California (e.g., Moxey & Sanford 1993, Pepper & Prytulak 1974).

iii. *Audience design*. Yildirim et al. (2014) provide evidence that listeners tailor their interpretation of quantified statements to the idiosyncracies of the speaker. If a speaker consistently refers to situations in which half of the A are B with 'Some A are B' instead of 'Many A are B', listeners take this information into account in their estimates.

iv. *Alternatives*. Chase (1969) found that estimates for quantified statements depend on the alternative expressions that feature in the experiment. He asked participants to rate the likelihood of an event on a five-point scale. In one condition, these events were described by means of high-frequency quantifiers (e.g., 'very often', 'usually'); in the other condition, by means of low-frequency quantifiers (e.g., 'seldom', 'occasionally'). In many cases, Chase found that the mean likelihood ratings in these conditions were statistically indistinguishable.

These findings can be modelled in various ways. Some authors have proposed that quantifiers denote probability distributions over situations. In other words, 'Q A are B' denotes a function from situations to numerical values that sum to one (e.g., Yildirim et al. 2014). But there are a number of issues with this proposal. First, it is not immediately obvious what the numerical values represent. Suppose that the function assigns a value p to a particular situation. One interpretation is that this means that the listener believes that the likelihood of this situation is p. Another interpretation is that the likelihood that a listener believes this situation is the one the speaker had in mind is p. Yet another interpretation is that the listener believes that the speaker believes that the likelihood of that situation is p. For our current purposes, the differences between these proposals are immaterial but it is an issue that stands in need of further analysis.

A more pressing problem with the probabilistic account pertains to the interpretation of universal quantifiers. In section 2.2, we provided evidence that appropriateness ratings for 'Every circle is black' steadily increase with the number of black circles in the situation, which means that all situations except for the one without black circles received a positive rating. It seems implausible, however, to conclude from these findings that, for example, the probability of a situation in which three out of ten circles are black given the utterance 'Every circle is black' is anything other than zero.

In order to avoid these issues, we will model the finer-grained interpretation of quantified statements by assuming that these refer to functions from situations to typicality values (Rosch 1975). This is a more general approach than the probabilistic account because it does not require that the numerical values sum to one. The numerical values represent the "typicality" of a situation with respect to the quantified statement. In section 2.2, we explain the notion of typicality in some more detail. Typicality values can be converted to probability values by dividing them by the sum of the typicality values, which will only be possible if information about the total set size is available.

What is the relationship between typicality structure and the set-theoretic truth definitions proposed by natural language semanticists? Are these meaning aspects disparate, or are they reflections of one underlying dimension? Are set-theoretic definitions of quantifiers still needed in light of the findings from psychological experiments? In section 4.2.3, we address these questions on the basis of the results of two experiments that will be discussed in section 4.2.2. This investigation follows the lead of McCloskey & Glucksberg (1978), who inquired into the interpretation of nouns. In the next section, we consider their arguments in some detail.

4.2.2 McCloskey & Glucksberg (1978)

Nouns like 'bird' and 'furniture' refer to categories. According to the classical view, which was first propagated by Socrates in Plato's *Statesman* and later popularised by Aristotle, categories are sets of objects that fulfill a list of necessary and sufficient conditions. For example, 'bird' refers to the set of objects that are warm-blooded and egg-laying vertebrates with feathers and wings. According to this account, all objects are either birds or nonbirds.

Psychological research, however, suggests that listeners often make finer-grained distinctions between objects than the binary distinction imposed by the classical definition of a category. For example, Rosch (1975) found that participants consider sparrows to be more typical birds than penguins or chickens. Some authors have argued that these typicality judgements indicate that category membership itself is a matter of degree: sparrows are birds to a greater degree than are penguins or chickens (e.g., Lakoff 1973). Other authors, however, have criticised this view (e.g., Kamp & Partee 1995).

What is the relationship between typicality judgements and category membership? Do listeners have access to classical definitions for categories denoted by nouns like 'bird' and 'furniture'? To address these questions, McCloskey & Glucksberg (1978) probed participants for typicality judgements and category membership judgements for a range of categories and objects.

Table 4.2 provides a sample of the results for the categories denoted by 'bird' and 'furniture'. In both cases, the average typicality values line up along a continuum between the two extremes. In the case of 'furniture', the percentages of positive responses in the category membership task also form such a continuum. In the case of 'bird', by contrast, almost all percentages of positive responses are close to the extremes of 0 and 100. Participants were thus more concurrent in their judgements of category membership for 'bird' than for 'furniture'.

Do participants have access to a well-defined category for these nouns, or are their category membership judgements fully determined by typicality differences? In

Object	τ	ς	Object	τ	ς
Bird			*Furniture*		
Robin	10.00	100	Chair	9.95	100
Eagle	9.58	100	Table	9.83	100
Partridge	8.42	100	Bed	9.58	98
Goose	8.29	97	Rug	6.25	48
Condor	8.23	100	Lampshade	5.70	63
Buzzard	8.08	98	Sewing machine	5.32	11
Turkey	7.92	100	Refrigerator	5.07	18
Chicken	7.75	95	Waste basket	4.70	31
Loon	7.43	100	Bookends	4.53	43
Ostrich	7.25	97	Ironing board	4.32	16
Penguin	6.96	92	Pillow	4.12	31
Bat	3.63	17	Electric fan	3.78	13
Flying squirrel	2.63	5	Ashtray	3.45	21
Vampire	2.29	13	Door	2.87	10
Bee	2.04	3	Ceiling	2.03	0
Locust	1.83	9	Fence	1.87	0

Table 4.2: Sample of the results of the typicality and category membership tasks for the categories denoted by 'bird' (left column) and 'furniture' (right column) as found by McCloskey & Glucksberg (1978). τ: Average typicality on a 10-point scale; ς: Percentage of positive responses in the category membership task.

order to answer this question, we constructed two models predicting proportions of positive responses in the category membership task: one based on a classical definition and one based on typicality judgements. The classical model for 'bird' assigned 1 to all biological birds and 0 to all other objects. In the case of 'furniture', there was no straightforward criterion for distinguishing category members from nonmembers. Therefore a cutoff point in the typicality ratings was used: all objects that scored higher than 5.5 were assigned 1 and all other objects 0. The typicality model was formed by the normalised typicality ratings. For both nouns, we compared the absolute differences between the predicted and attested proportions of positive responses in the category membership task by means of Welch t-tests. In the case of 'bird', the classical model provided a better fit than the typicality model (mean differences of .06 and .17, $t(54)$ = -3.94, $p < .001$), whereas the converse was the case for 'furniture' (mean differences of .19 and .09, $t(34) = 2.48$, $p = .02$).

The results for 'bird' are thus in accordance with the classical account of categorisation. For this noun, judgements of category membership were relatively crisp. This suggest that participants have access to a well-defined category of birds. The results for 'furniture' are in accordance with the typicality account, since judgements of category membership were better approximated by typicality judgements than by any classical definition. These observations can be formalised as follows. Here, $\varsigma_{\mathrm{A}}(x)$ is the proportion of participants who indicate that x is an instance of the category denoted by ‘A’, and $\tau_{\text{A}}(x)$ is the normalised mean typicality rating for x in the category denoted by ‘A’. These are values in the interval (0, 1]. $x \in \textsc{a}$ means that x is a member of the category denoted by ‘A’ according to its classical definition. This equals a value in the set $\{0, 1\}$.

$$\begin{aligned} \varsigma_{\textsc{bird}}(x) &= x \in \textsc{bird} \\ \varsigma_{\textsc{furniture}}(x) &= \tau_{\textsc{furniture}}(x) \end{aligned}$$

A further question that stands in need of an explanation is what determines the typicality judgements McCloskey & Glucksberg found. In the case of ‘bird’, category membership plays a prominent role in the typicality judgements as well: the difference in mean typicality rating between the least typical birds (i.e., 6.96 for penguins) and the most typical nonbirds (i.e., 4.96 for pterodactyls) is much bigger than the difference between any other pair of neighbours on the ς-scale. No such effect is visible in the case of ‘furniture’. In addition, typicality judgements are often explained in terms of distance from the prototype (e.g., Rosch & Mervis 1975). A prototype is an object that is especially representative of a category because it satisfies most or all of the characteristics that are standardly associated with that category. For example, a prototypical bird might be an animal that is capable of flight, relatively small, and not too exotic. These observations can be formalised as follows. Here, $dist(x, p)$ is a measure of the distance between x and the prototype p. This equals a value in the interval [0, 1]. The resultant typicality values occur in the interval (-1, 1] and should therefore be normalised to the (0, 1] interval.

$$\begin{aligned} \tau_{\textsc{bird}}(x) &= x \in \textsc{bird} - dist(x, p) \\ \tau_{\textsc{furniture}}(x) &= dist(x, p) \end{aligned}$$

In order to address the questions we posed at the end of the previous section, we conducted two experiments analogous to McCloskey & Glucksberg’s to determine and model the relationship between set-theoretic definitions and typicality structure in the interpretation of quantifiers. To that end, we gathered and analysed truth value judgements and typicality judgements for quantified statements. Since sentences refer to situations instead of individuals, we used pictures of situations instead of words referring to individuals. An example of a trial is shown in Figure 4.1. The quantifiers that were included in the experiments are listed in Table 4.3.

In the first experiment, participants had to indicate on a seven-point scale how well the situation was described by the statement. The design of this experiment was the same as that of the experiments reported in section 2.2.6 and 2.2.7. In the second experiment, they had to indicate whether the statement was true or false in the depicted situation. One of our goals was to investigate if truth value judgements are better approximated by set-theoretic definitions or by typicality judgements. The set-theoretic definitions we used to this end are also listed in Table 4.3.

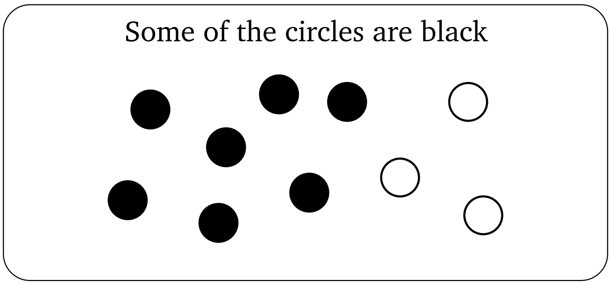


Figure 4.1: Sample item used in the experiments.

Quantifier	**Definition**	**Quantifier**	**Definition**
All	$A \subseteq B$	Most	$\|A \cap B\| > \|A \cap \bar{B}\|$
Every	$A \subseteq B$	None	$A \cap B = \varnothing$
Few	$\|A \cap B\| < \delta$	Some	$A \cap B \neq \varnothing$
Many	$\|A \cap B\| > \delta$	Not all	$A \not\subseteq B$
More than half	$\|A \cap B\| > \|A \cap \bar{B}\|$	Not many	$\|A \cap B\| \leq \delta$

Table 4.3: Quantifiers used in the experiments and the corresponding set-theoretic definitions.

These definitions are the standard ones from the literature. ‘Few’ and ‘many’ are vague quantifiers. This implies that their meaning depends on the context. ‘Few A are B’ expresses that the number of A that are B is surprisingly low. This can be formalised by means of a contextually determined threshold value δ below which the real number of A that are B is supposed to fall. ‘Many A are B’ conversely implies that the number of A that are B is surprisingly high, and should therefore exceed some contextually determined threshold value δ. It turns out that in our experiment the value that most participants assigned to δ was 5 for both quantifiers.

There is one probable complication that warrants some further discussion. We have already seen that truth value judgements are influenced by scalar inferences. To illustrate, the truth conditions of ‘Some A are B’ are compatible with all situations in which one or more A is B. Nonetheless, someone who utters this sentence will often exclude some of these situations because of the derivation of scalar inferences. This particular utterance carries at least three possible inferences. First, it might implicate that it is not the case that only one A is B. This inference is triggered by the plural marking on the subject and verb. Some authors have argued that this plurality inference is pragmatic in nature (e.g., Spector 2006). Second, the utterance might implicate that it is not the case that all of the A are B. This is a scalar inference based on the lexical scale ⟨some, all⟩. Third, it might implicate that it is not the case that most of the A are B. This scalar inference is based on the lexical scale ⟨some, most⟩. Zevakhina (2012) provides evidence that the ‘not all’ inference is more robust than the ‘not most’ inference.

Although it has been shown that some participants judge sentences false in situations in which their scalar inferences are false, this does not necessarily mean that these participants consider a sentence like ‘Some A are B’ equally bad in a situation in which all of the A are B as they do in a situation in which none of them are, and in which the set-theoretic truth conditions of the sentence are thus violated. In the experiment reported in section 2.2.7, we observed that participants considered the sentence ‘Some of the circles are black’ more appropriate in a situation with only black circles than in a situation with only white circles.

Most of the quantifiers in our investigation licensed scalar inferences. The quantifiers ‘many’, ‘more than half’, and ‘most’ license the inference that not all of the circles are black; the quantifiers ‘few’, ‘not all’, and ‘not many’ license the inference that at least one of the circles is black. Note that ‘some’ is exceptional in that it carries three potential scalar inferences, whereas all of the other quantifiers have just one possible inference. It seems plausible to suppose that these scalar inferences will have an effect on the results of both experiments. We will discuss this issue in more detail in the Results section. First, however, we report the details of the two experiments.

The experiments

Participants

We posted surveys for 340 participants on Amazon’s Mechanical Turk. Only workers with an IP address in the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. 120 participants provided truth value judgements (mean age: 34; range: 18-61; 68 females). All of these participants were native speakers of English. 220 participants provided typicality judgements (mean age: 37; range: 18-70; 135 females). 30 participants provided typicality judgements for the quantifiers ‘some’ and ‘every’. The results of these participants have already been discussed in sections 2.2.6 and 2.2.7. 20 participants provided typicality judgements for all of the other quantifiers. 5 participants provided typicality judgements for more than one quantifier. 11 of the 220 participants in the typicality task were excluded from the analysis because they were not native speakers of English.

Materials

Sentences were of the following form:

(2) Q {circle is / of these circles are} black.

Q was instantiated by the quantifiers in Table 4.3. The corresponding pictures consisted of ten circles which were either black or white. The distribution of black and white circles was manipulated, thus creating eleven situation s_0, …, s_{10}. In a situation s_n, n of the circles were black and the remaining $10 - n$ circles were white. An example trial is shown in Figure 4.1.

Surveys in the truth value judgement task consisted of twenty trials. Each of the ten quantifiers was instantiated twice with two different pictures. The pictures for one quantifier always differed in at least three circles. Because there was an uneven number of situations, one of the them, s_5, occurred twice as often as the other situations: 40 instead of 20 times. The order of the items was randomised for each participant, making sure that the same quantifier never occurred consecutively. The truth value judgement task differed from McCloskey & Glucksberg's, who asked participants to categorise all possible instantiations. We avoided this procedure because it might lead to contrastive readings of quantifiers. Intuitively, if 'some' receives a contrastive reading, which usually manifests itself by means of prosodic stress, it excludes 'all' by entailment. We wanted to avoid this potential confound as much as possible.

In the typicality experiment, participants were presented with one quantifier in all eleven situations. The order of the items was randomised for each participant.

Procedure

To collect truth value judgements, participants were presented with the following instructions:

> In the following survey, we will show you pairs of pictures and sentences. In each case, we ask you to decide whether or not the sentence gives a correct description of the picture. If you feel that the sentence is true, check "True". If not, check "False".
>
> We are interested in your spontaneous judgments, so please don't think too long about your answers.

The instructions used in the typicality task are provided in full in the Procedure section of 2.2.6.

Results

Figure 4.2 provides the normalised mean typicality judgements and the proportions of positive responses in the truth value judgement task. The figure suggests that, in general, typicality judgements were less pronounced than proportions of positive responses in the truth value judgement task. Furthermore, the average typicality values were more evenly distributed across the space of possible answers, whereas the proportions of positive responses in the truth value judgement task clustered around the extremes of 0 and 1. This suggestion is confirmed by a comparison of the variances: the variance in normalised typicality ratings was significantly bigger than the variance in the proportions of positive responses in the truth value judgement task ($F(109, 109) = 1.9$, $p < .001$). Both of these observations are captured by the density plot in Figure 4.3: the modes of the average typicality values are closer to the center than the modes of the proportions of positive responses in the truth value judgement task, and there are more values in the middle region of the space of possible answers in the average typicality values than in the proportions of positive answers in the truth value judgement task.

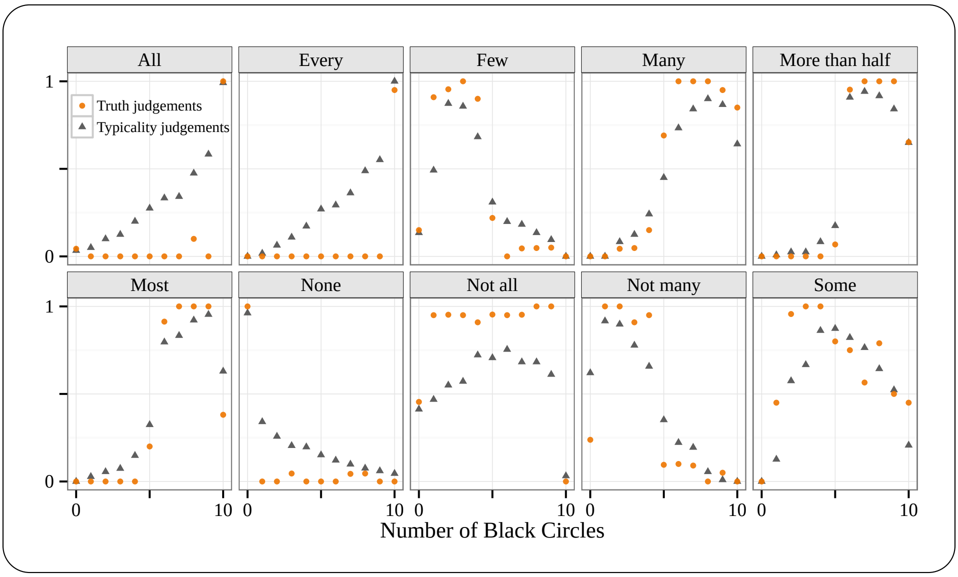


Figure 4.2: Normalised typicality judgements and proportions of positive responses in the truth value judgement task for ten quantifiers.

One anomalous observation is the proportion of positive responses for ‘some’ in s_8 (M = .79). This proportion is unexpectedly higher than in situations with seven (M = .56) or nine (M = .50) black circles. The difference, however, is not statistically significant in either case. It is presumably caused by the between-participants design of the truth value judgement task: s_8 was judged by different participants than s_7 or s_9. Apparently s_8 was judged by more charitable participants than the other two situations.

Which situations are prototypical of the quantifiers that were investigated? There are at least two ways of answering this question. The first is to take the situations with the highest mean typicality judgements. The second is to take the situations that received the highest typicality judgements from the largest number of participants. For almost all of the quantifiers, these methods lead to the same prototypes. The sole exception is ‘not all’. For this quantifier, the highest mean typicality judgement was for s_6, whereas s_9 was assigned the highest typicality rating by the most participants. This discrepancy reflects a high degree of disagreement between participants in the typicality task for this quantifier. Some participants gave the highest rating to s_0, some to s_9, and some to situations inbetween these extremes. Note that this lack of agreement is not visible in the results of the truth value judgement task.

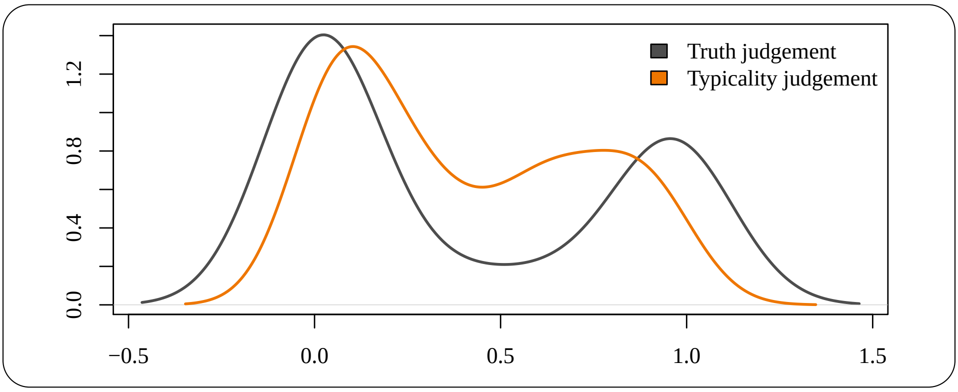


Figure 4.3: Density plot of the normalised typicality ratings and proportions of positive responses in the truth value judgement task.

As hypothesised, there was a strong effect of scalar inferences on both truth value and typicality judgements. We compared the results for situations in which a scalar inference was violated with the results for the nearest situation where this was not the case. The mean difference between these situations was higher in the truth value judgement task ($M = .46$) than in the typicality task ($M = .28$). This difference was marginally significant ($t(7) = 1.87$, one-sided $p = .05$). Focusing on the results of the truth value judgement task, the difference was higher for negative quantifiers ($M = .69$) than for positive quantifiers ($M = .32$, $t(6) = 2.46$, one-sided $p = .03$). There was no analogous effect of monotonicity on the effect of scalar inferences in the average typicality judgements.

What factor underlies truth value judgements in this experiment? Based on the discussion in section 4.2.2, two possible answers suggest themselves: truth value judgements are determined either by typicality judgements or by set-theoretic definitions. In order to decide between these two possible answers, we constructed three models predicting proportions of positive responses in the truth value judgement task. The first model used the set-theoretic truth definitions given in Table 4.3 as predictor variable. The second and third model were based on the normalised mean typicality ratings. The second model mapped these typicality ratings straight onto proportions of positive responses in the truth value judgement task. The intuition that underlies this model is that the typicality of an object reflects the likelihood that the sentence is considered true in that situation. The third model dichotomised the typicality judgements on the basis of a cutoff point. We calculated that the optimal cutoff point was 4.4. Values below the cutoff point were mapped to 0 and values above it to 1.

According to the third model, listeners make truth value judgements by dichotomising their typicality judgements. Unlike the typicality model, it is not obvious what the rationale behind the dichotomised model is. Why do listeners dichotomise typicality judgements on the basis of an apparently arbitrary cut-off point? One possible answer is that this cut-off point reflects whether the sentence is true according to its set-theoretic definition. But in that case the model draws upon set-theoretic definitions just like the first model. If this model is to be a competitor to the set-theoretic model, it thus stands in need of a more principled motivation.

The mean distances between these three models and the results from the truth value judgement task are given in Table 4.4. For the set-theoretic model, the mean distances range from .00 for 'every' to .25 for 'some'. For the typicality model, the mean distances range from .04 for 'more than half' to .27 for 'not all'. For the dichotomised typicality model, the mean distances range from .00 for 'every' to .24 for 'not all'.

Quantifier	Sem	Typ	Dich	Quantifier	Sem	Typ	Dich
All	.01	.22	.10	Most	.08	.15	.13
Every	.00	.22	.00	None	.01	.13	.01
Few	.13	.13	.14	Some	.25	.18	.23
Many	.06	.12	.10	Not all	.08	.27	.24
More than half	.04	.04	.04	Not many	.11	.18	.11

Table 4.4: Mean difference between the proportions of positive responses in the truth value judgement task and the set-theoretic definitions in Table 4.3 (= ***Sem***), the normalised typicality judgements (= ***Typ***), and the dichotomised typicality judgements (= ***Dich***)

We compared the distances between these models and the proportions of positive responses in the truth value judgement task. The mean distance was significantly greater for the typicality model ($M = .16$) than for the set-theoretic model ($M = .08$, $t(109) = 3.77$, $p < .001$). It was also significantly greater for the typicality model than for the dichotomised typicality model ($M = .11$, $t(109) = 2.42$, $p = .002$). The difference in mean distances between the set-theoretic and dichotomised typicality models was marginally significant ($t(109) = 1.52$, one-sided $p = .07$) in favour of the set-theoretic model.

A substantial part of the lack of fit in all of the models is caused by the confounding effect of quantity inferences. Therefore we also used the models to predict a restricted set of the data points that were not affected by quantity inferences. This improved all of the models. We once again compared the distances between these limited models and the proportions of positive responses in the truth value judgement task. The mean distance was significantly greater for the typicality model ($M = .15$) than for the set-theoretic model ($M = .03$, $t(96) = 8.27$, $p < .001$) and for the dichotomised typicality model ($M = .06$, $t(96) = 5.01$, $p < .001$). The mean distance for the dichtomised typicality model was significantly greater than for the set-theoretic model ($t(96) = 1.94$, one-sided $p = .03$).

Proportions of positive responses in the truth-value judgement task are thus better approximated by set-theoretic definitions or dichotomised typicality values than by simple typicality values. Moreover, there is some evidence that the set-theoretic model is more appropriate than the dichotomised model: it is a marginally better predictor of the proportions of positive responses including data points that are influenced by quantity inferences and a significantly better predictor of the proportions of positive responses excluding those data points. In addition, the dichotomised model lacks a principled explanation for the use of a seemingly arbitrary cut-off point in the mean typicality ratings. Moreover, the sole plausible motivation seems to invoke set-theoretic truth conditions.

What factor underlies typicality judgements in this experiment? Based on the discussion in section 4.2.2, two possible answers suggest themselves: typicality judgements are determined either by set-theoretic definitions and distance from the prototype or by distance from the prototype alone. In order to decide between these possible answers, we constructed two models predicting mean typicality judgements. The first model included the set-theoretic definitions from Table 4.3 and distance from the prototype as predictor variables, whereas the second model included distance from the prototype alone. Before these models can be operationalised, however, a number of parameters have to be set. First, what are the prototypes associated with quantified sentences? Second, how to operationalise distance from the prototype? Third, what is the relative importance of set-theoretic definitions and distance from the prototype in the first model. We discuss these issues in turn.

What are the prototypes associated with quantified sentences? As noted before, quantifiers differ in how unambiguous and salient the prototype is: for quantifiers like ‘every’ and ‘none’, all participants converged on the same prototype, for quantifiers like ‘some’ and ‘most’, there was a consensus but not a unanimous one, and for ‘not all’ there was a large amount of disagreement among participants. The choice of prototype seems to be determined by at least two factors: set-theoretic truth conditions and competing quantifiers. To start with the first factor, prototypes are always situations in which the sentence is true according to its set-theoretic truth definition. For some quantifiers, however, this still leaves a number of situations to choose from. In that case, the specific choice might be affected by competing quantifiers: a prototypical situation is one that is maximally distinct from the prototypical situations of competing quantifiers. This criterion can explain the different choices of prototypes for ‘not all’. Participants who assume that ‘not all’ competes with ‘all’ will consider s_0 as the prototype because that situation is maximally distinct from the prototype for ‘all’. Participants who also took into consideration ‘none’ as a competitor will opt for a situation with around five black circles. Lastly, participants who also considered ‘not many’ might have converged on a prototype that lies somewhere on the upper end of the scale.

This discussion leaves open some further questions. What determines the choice of competing quantifiers? Which quantifiers are in principle available as competitors? How do the alternatives that determine the choice of prototype relate to the alternatives that are involved in the computation of scalar inferences? While we believe these questions are interesting and might warrant further analysis, we will simply stipulate that the prototypes are the situations that received the highest mean typicality ratings.

A second question is how to operationalise distance from the prototype. This question has a straightforward answer: the distance between a prototypical situation s_n containing n black circles and another situation s_i equals the absolute difference between n and i.

The final question involves the relative importance of set-theoretic truth conditions and distance from the prototype in the model that included both of these factors. Because we do not have specific expectations about these parameters, we simulated them by means of Monte Carlo methods. To this end, we assigned 5,000 random values to both parameters. For each pair of values, we calculated the predicted typicality values and the correlation between these predicted values and the attested typicality values. In the optimal situation, the effect of set-theoretic truth conditions was more than seven times as large as the effect of increasing the distance from the prototype by one. We therefore weighed the two factors accordingly in the first model. As observed by Michael Franke, this analysis is biased in favour of parametrised models. So the important observation is that set-theoretic truth conditions are added as a pronounced and positive factor.

Figure 4.4 provides a visual overview of the goodness of fit of both models. The correlation between the typicality values predicted by the first model containing both set-theoretic definitions and distance from the prototype, and the attested typicality values was $r = .94$. The correlation between the typicality values predicted by the second model consisting of distance from the prototype alone and the attested typicality values was $r = .83$. We compared the two models on the basis of the absolute differences between predicted and attested typicality values. The mean difference was significantly higher for the second model that included only distance from the prototype ($M = .21$) than for the first model that also included set-theoretic truth conditions ($M = .12$, $t(109) = 6.53$, $p < .001$). As before, the fit of both models is relatively poor for situations that are excluded by pragmatic means. Excluding those data points from consideration leads to correlations of $r = .96$ for the first model and $r = .87$ for the second one.

What do these results tell us about the questions we posed at the end of the introduction? What is the relationship between set-theoretic truth conditions and typicality judgements? Do listeners have access to set-theoretic definitions of quantified sentences? In the following section, we discuss these questions on the basis of the foregoing results.

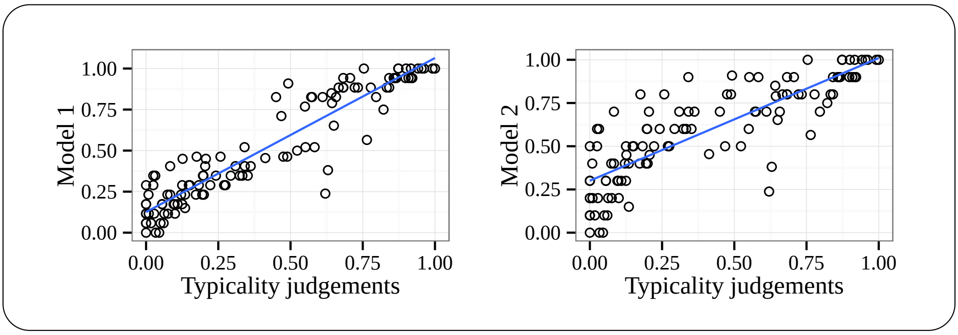


Figure 4.4: Scatterplot with the typicality values predicted by the two models and the typicality judgements found in the experiment.

4.2.3 General discussion

The interpretation of quantified statements is often finer-grained than what is encoded in their set-theoretic definitions. For example, according to its set-theoretic truth conditions, the statement 'Some A are not B' is true in all situations in which not all of the A are B. But when participants are asked how many A are B given that this statement is true, Newstead et al. (1987) found that their estimates are much more precise and range between 55% and 89% of the total number of A. These findings can be modelled in different ways. Some authors have proposed that quantifiers denote probability distributions over situations. However, this proposal seems implausible in light of the results for the universal quantifiers: even though 'Every A is B' receives a positive rating in situations in which a subset of the A are B, it seems that the probability of such a situation obtaining is zero simpliciter.

In order to avoid this issue, we have modelled the finer-grained interpretation of quantifiers by assuming that these denote typicality functions. In that case, is it still necessary to suppose that listeners have access to set-theoretic definitions of quantifiers, or do their typicality structures provide sufficient information? The results of our experiments provide a number of arguments to substantiate the role of set-theoretic definitions in the interpretation of quantifiers. First, we constructed three models to predict proportions of positive responses in the truth value judgement task. One model consisted of set-theoretic definitions; the second model consisted of typicality judgements; and the third model dichotomised these typicality judgements based on a cutoff point. We found that the absolute deviations of the first model were significantly smaller than those of the second model, and marginally smaller than those of the third model.

These findings provide further evidence that the typicality of a situation should not be equated to the likelihood that the corresponding statement is judged true. But the evidence against the view that truth value judgements are dichotomised typicality judgements was less convincing. Nevertheless, even if we assume that truth value judgements are dichotomised typicality judgements, set-theoretic definitions still feature in the explanation of the typicality judgements themselves.

We constructed two models to predict typicality judgements. The first model consisted of set-theoretic definitions and distance from the prototype, where the former factor was given seven times as much weight as the latter. The second model consisted of distance from the prototype alone. We found that absolute deviations from the first model were significantly smaller than those from the second model. This finding also indicates that listeners have access to set-theoretic definitions of quantifiers.

A related argument in favour of the view that set-theoretic definitions feature in the interpretation of quantified statements is the rarity of disagreement between participants about whether a quantified statement was true or false. More precisely, aside from situations in which pragmatic inferences played a role, there were four proportions of positive responses in the range between 0.1 and 0.9. Three of these involved the s_5 situation for the proportional quantifiers 'most', 'many', and 'few', and one involved the s_4 situation for 'many'. These three quantifiers all involve some kind of threshold value. So a possible explanation for these anomalous observations is that participants might have disagreed about the exact value of this threshold. By contrast, there was a substantial amount of disagreement in the typicality judgements for all quantifiers. If truth value judgements were calculated on the basis of typicality judgements, we would have expected a similar amount of disagreement in both of these measures.

Furthermore, it might be argued that set-theoretic definitions are necessary to account for the interpretation of embedded quantifiers. For example, it seems inevitable to suppose that the interpretation of 'not all' is a function of the interpretation of 'all'. This relationship is apparent in the truth value judgements but not in the typicality judgements: the correlation between the proportions of positive responses in the truth value judgements task for 'all' and 'not all' is -.88 and, if the effect of the scalar inference for 'not all' is taken into account, -.98. This strength of association is absent in the typicality judgements: the correlation between typicality judgements for 'all' and 'not all' is -.47. (Note, however, that the typicality structure of 'many' does provide a reliable fit to the typicality structure of 'not many'.)

We have thus provided a number of arguments in favour of the view that listeners have access to set-theoretic interpretations of quantifiers. One apparent exception to this rule was 'some'. The set-theoretic definition for this quantifier was a poor fit to the actual truth value judgements. Truth value judgements were equivocal for all situations other than those with zero black circles, where it was judged false by all participants, and those with two, three, or four black circles, where it was judged true by all or almost all participants. The exceptional position of 'some' is striking given the number of experiments that are conducted on the basis of this quantifier. In the next section, we provide evidence that 'some' is also special among other scalar expressions.

4.3 Scalar diversity

4.3.1 Lexical scales

Scalar expressions are associated with lexical scales whose members are ordered in terms of informativeness. For instance, 'some' evokes the scale ⟨some, all⟩, where 'all' is more informative than 'some'. There is no uncontroversial definition of what makes a pair of expressions into a lexical scale. However, it is widely assumed that lexical scales contain expressions that are ordered in terms of informativeness and lexicalised to the same degree (e.g., Horn 1972, Gazdar 1979, Atlas & Levinson 1981). We will confine our attention to scales that meet these conditions. This means that we will not be concerned with ranked orderings or ad hoc scales (e.g., Hirschberg 1991, Levinson 2000). The lexical scales in Table 4.5 (= Table 1.1) all count as lexical scales according to this traditional definition.

Category	Examples	
Adjectives	⟨intelligent, brilliant⟩	⟨difficult, impossible⟩
Adverbs	⟨sometimes, always⟩	⟨possibly, necessarily⟩
Connectives	⟨or, and⟩	
Determiners	⟨some, all⟩	⟨few, none⟩
Nouns	⟨mammal, dog⟩	⟨vehicle, car⟩
Verbs	⟨might, must⟩	⟨like, love⟩

Table 4.5: Sample scales for various grammatical categories.

The debate about scalar inferences has, for the most part, centered on the question how these inferences come about. A fair number of experiments have been done to compare the predictions of various theories. One striking feature of these experiments is that, for the most part, they are confined to just two scalar expressions, namely 'some' and 'or'. To illustrate, Table 4.6 provides an overview of the scalar expressions that have been used in a representative sample of the research on the development and processing of scalar inferences. A comparison with Table 4.5 makes it clear that several classes of scalar expressions, most notably nouns, adjectives and adverbs, have been consistently overlooked. Even within the classes that have been investigated, the variety of scalar expressions is limited.

Apparently, the tacit assumption underlying these experiments is that the scalar expressions in Table 4.5, and especially 'some' and 'or', are representative for the entire family of scalar expressions. Until recently, the *uniformity assumption*, as we will call it, had not been questioned, but it was put to the test by Doran et al.

Scale	Sources	
⟨some, all⟩	Noveck (2001)	Noveck & Posada (2003)
	Papafragou & Musolino (2003)	Bott & Noveck (2004)
	Feeney et al. (2004)	Guasti et al. (2005)
	Breheny et al. (2006)	De Neys & Schaeken (2007)
	Pouscoulous et al. (2007)	Banga et al. (2009)
	Geurts & Pouscoulous (2009a)	Huang & Snedeker (2009)
	Clifton & Dube (2010)	Grodner et al. (2010)
	Barner et al. (2011)	Chemla & Spector (2011)
	Bott et al. (2012)	Geurts & van Tiel (2013)
	van Tiel (2014)	Degen & Tanenhaus (2014)
⟨or, and⟩	Noveck et al. (2002)	Storto & Tanenhaus (2005)
	Breheny et al. (2006)	Chevallier et al. (2008)
	Pijnacker et al. (2009)	Zondervan (2010)
	Chemla & Spector (2011)	
⟨might, must⟩	Noveck (2001)	
⟨start, finish⟩	Papafragou & Musolino (2003)	

Table 4.6: Scalar expressions used in a representative sample of experiments on the interpretation, development, and processing of scalar inferences.

(2009), following up on a study by the same group (Larson et al. 2009). Doran et al.'s findings suggest that there is significant variability between the rates at which scalar terms of different lexical categories give rise to upper-bounded inferences. However, as we will argue in the following, Doran et al.'s experimental design precludes a straightforward interpretation of their data. We therefore undertook a study based on a simpler design, which also provided us with finer-grained results than previous studies. Furthermore, we investigated several candidate explanations for the variability we observed.

4.3.2 Extant evidence for diversity

Before Doran et al. put the uniformity assumption to the test, a number of experimental findings had already cast doubt on the view that all scalar expressions behave alike. For example, Noveck (2001) found that children and adults were more likely to interpret 'might' with an upper bound than 'some'. The experiments in which these scalar expressions were tested, however, differed along a number of dimensions, thus precluding a straightforward comparison.

More direct evidence against the uniformity assumption comes from the interpretation of the existential quantifier in Dutch and French. This quantifier can be instantiated as 'enkele' or 'sommige' in Dutch, and as 'quelques' or 'certains' in French. Banga et al. (2009) found that 'sommige' licenses an upper-bounding inference more often than 'enkele'. Pouscoulous et al. (2007) found the same result for 'quelques' when compared to 'certains'. Moreover, a comparison between these studies shows that Dutch 'sommige' and 'enkele' were substantially more likely to be interpreted with an upper bound than their French counterpart 'certains'. These findings indicate that the likelihood of a scalar inference varies both within and between languages.

A similar conclusion can be distilled from Geurts's (2010, 98-99) survey of ten experiments employing the verification paradigm. In these experiments, participants had to decide whether target sentences were true or false in states of affairs where the scalar inference was false. For example, Bott & Noveck's (2004, experiment 3) participants rejected statements like those in (3) 59% of the time:

(3) a. Some parrots are birds.
b. Some dogs are mammals.

The main point transpiring from Geurts's survey is that, across the collated experiments, the mean rate of scalar inferences for 'or' was clearly lower than for 'some': 35% against 57%. This suggests rather strongly that scalar inference rates are higher for 'some' than for 'or'.

These preliminary observations aside, Doran et al. (2009, 2012) were the first to test the uniformity assumption in an integrated experimental design. In both of their studies, participants were presented with stories like the following:

(4) Irene: How much cake did Gus eat at his sister's birthday party?
Sam: He ate most of it.
Fact: By himself, Gus ate his sister's entire birthday cake.

(5) Irene: How would you say Alex is doing financially?
Sam: He's comfortable.
Fact: Alex just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.

Participants had to decide whether Sam's answers were true or false. The premiss was that if Sam's statement was deemed to be false, then participants must have derived a scalar inference.

One further manipulation introduced in Doran et al.'s first paper was that, in addition to the condition illustrated in (4) and (5), there were two other conditions: one in which Irene's question contained a scalar term that was stronger than the one used by Sam in his answer, as in (6a) and (7a), and one in which Irene's question, in effect, offered Sam three scalar expressions to choose from, as in (6b) and (7b):

(6) a. Did Gus eat all of his sister's birthday cake?
b. Did Gus eat some, most, or all of his sister's birthday cake?

(7) a. Would you say Alex is financially wealthy?

b. Would you say that Alex is poor, comfortable, or wealthy?

In the following, we will use the terms *neutral* and *(one- or two-way) contrastive* to label these conditions: (4) and (5) count as neutral, (6a) and (7a) are one-way contrastive, and (6b) and (7b) are two-way contrastive.

Doran et al.'s first main finding was that, whereas quantified statements were rejected 32% of the time, for sentences with adjectives, the rejection rate was only 17%. Scalar inferences were thus about twice as frequent for quantifiers as for adjectives. Secondly, Doran et al. found that only adjectival items were affected by the difference between the neutral and contrastive conditions: within the adjectival category, the two-way contrastive items elicited significantly more 'false' responses than the neutral and the one-way contrastive ones; otherwise, the neutral/contrastive distinction was inert.

Although Doran et al.'s findings provide convincing evidence against the uniformity assumption, there are a number of reasons for going over the same ground with a different experimental design and a finer-grained analysis. Firstly, Doran et al. adopted a rather coarse-grained categorisation of experimental items, grouping together quantifying expressions with measure phrases and modal adverbs, for example. The fact that they found a dichotomous distinction between quantifying and adjectival expressions may have been due to this, and it is quite possible that a finer-grained analysis would have produced results that speak against such a dichotomy. Such a finer-grained analysis is also a prerequisite for determining what factors underlie the variable rates of scalar inferences.

Secondly, Doran et al.'s experiment employed a verification task for gauging the frequency of scalar inferences, but it is unique in that it presented the relevant facts by way of verbal description. A potential problem with this approach is that it is difficult to standardise the descriptions of the relevant facts. To illustrate, compare the fact descriptions in (4) and (5). A number of differences stand out. First, the fact description for 'comfortable' is more verbose than for 'most', which makes Sam's response seem almost like an ironic understatement in the case of 'comfortable'. Second, the fact description for 'most' contains the scalar expression 'entire' which is a possible scalemate of 'most'. This may have rendered the lexical scale for 'most' more available than for 'comfortable'. Such differences may have contributed to the results that Doran et al. found. We therefore repeated Doran et al.'s experiment using a different paradigm and a finer-grained analysis, and then considered a number of potential explanations for the observed variability.

Instead of Doran et al.'s verification task, we decided to adopt an inference task, which has been widely used in the psychology of reasoning, and has occasionally been used in experimental studies on scalar inference (Chemla 2009, Geurts & Pouscoulous 2009a). It has been shown that the inference paradigm yields higher rates of scalar inferences than the verification paradigm, but since we were primarily interested in relative frequencies of scalar inferences, that was no cause for concern.

Experiment 1

Participants

We posted surveys for 25 participants on Amazon's Mechanical Turk (mean age: 35; range: 21–63; 28 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question.

Materials and procedure

Figure 4.5 shows an example of a critical item (the full list of materials is given in the appendix). In each trial, a character named John or Mary made a statement containing a scalar expression, which always occurred in predicate position, and participants had to decide whether or not this implied that, according to the speaker, the statement would have been false if that expression had been replaced with a stronger scale member. The statements were kept as bland as possible, so that participants would not be guided by expectations based on their world knowledge. This was done mainly by using pronouns instead of complex noun phrases, but also by using generic predicates like 'go inside' and 'do that'. (Experiment 2, which is reported in the next section, replicated the current experiment with more informative sentences.) Pronouns were never congruent with the speaker's gender in order to prevent them from being interpreted as referring to the speaker.

John says:

> *She is intelligent.*

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

□ Yes □ No

Figure 4.5: Sample item used in Experiment 1.

Materials comprised a selection of scales consisting of quantifiers (2 scales), adverbs (1), auxiliary verbs (2), main verbs (6), and adjectives (32). A complete list is given in Table 4.7. Our selection of scalar expressions was guided in part by examples discussed in the literature (e.g., Horn 1972, Hirschberg 1991, Doran et al. 2009). However, adjectival scales, which were used in 70% of the experimental items, were selected by searching the internet and several corpora (the British National

Corpus, the Corpus of Contemporary American English, and the Open American National Corpus) for constructions of the form 'X if not Y', 'X or even Y', and 'not just X but Y', which yielded a large number of candidate scales. In the final selection, we made sure to include scales whose weaker term occurred more frequently than the stronger term, based on word counts in the Corpus of Contemporary American English (Davies 2008), and scales for which the opposite was true; we did this because we wanted to test the hypothesis that relative frequency has an effect on the rate at which a scalar inference is derived (Section 4.3.4).

Randomised lists were created for each participant, varying the order of the items. Seven control items were included, which involved statements that either entailed (e.g., an inference from 'wide' to 'not narrow') or were completely unrelated to (e.g., an inference from 'sleepy' to 'not rich') the critical inference (see appendix).[1]

Results and discussion

One participant was excluded from the analysis for making mistakes in three of the control items. Four out of a total of 1250 answers were missing. Control items were answered correctly on 94% of the trials. The results for the target trials are shown in Figure 4.6. It is evident from this graph that there was considerable variation among critical items, with positive responses ranging along a continuum from 4% (for seven adjective scales) to 100% (for ⟨cheap, free⟩ and ⟨sometimes, always⟩). The results of our first experiment thus disprove the uniformity assumption: different scalar expressions yield widely different rates of scalar inferences.

In this experiment, we used materials that were as neutral as possible, which was done mainly by using pronouns instead of complex noun phrases, but also by using generic predicates. One potential drawback of this approach is that it may have had a disorienting effect, leaving participants to wonder who or what these pronouns referred to, which, in its turn, may have affected out findings. Though it is difficult to see how this confusion could be responsible for the contrasts between scales, it will be instructive to gauge the robustness of the results by replicating Experiment 1 with less neutral materials and comparing the results.

[1] In a pilot experiment we gauged whether the number of control items had an effect on the results of the inference task. We presented 50 participants (mean age: 35; range: 18–67; 30 females) on Mechanical Turk with 10 of the target items included in Experiment 1 alongside 32 control items. In 16 of the control trials, the target inference was clearly valid; in the remaining 16 controls, it was clearly not valid. The results of this pilot experiment correlated almost perfectly with the results from Experiment 1 ($r = .97$, $t(8) = 11.66$, $p < .01$). Apparently, the number of control items does not have a substantial effect on the contrasts between scales.

Scale	SI		Cloze		Cat	Freq	LSA	Dist	Bnd
	+N	–N	+N	–N					
⟨cheap, free⟩	100	93	0	0	O	-0.66	.19	5.52	+B
⟨sometimes, always⟩	100	86	80	90	O	-1.05	.60	5.70	+B
⟨some, all⟩	96	89	67	87	C	-0.12	.79	5.83	+B
⟨possible, certain⟩	92	93	55	31	O	0.10	.42	5.65	+B
⟨may, will⟩	87	89	83	80	C	0.68	.51	5.41	+B
⟨difficult, impossible⟩	79	96	13	10	O	0.46	.60	6.22	+B
⟨rare, extinct⟩	79	79	40	34	O	1.05	.29	5.83	+B
⟨may, have to⟩	75	71	83	80	C	-1.22	.64	5.26	+B
⟨warm, hot⟩	75	64	70	38	O	-0.28	.51	5.00	–B
⟨few, none⟩	75	54	20	30	C	0.75	.47	5.35	+B
⟨low, depleted⟩	71	79	23	60	O	2.29	.16	4.87	+B
⟨hard, unsolvable⟩	71	71	10	10	O	2.87	.08	5.26	+B
⟨allowed, obligatory⟩	67	82	20	47	O	-0.85	.02	5.35	+B
⟨scarce, unavailable⟩	62	57	40	17	O	0.29	.18	4.78	+B
⟨try, succeed⟩	62	39	37	57	O	1.23	.35	5.82	+B
⟨palatable, delicious⟩	58	61	67	47	O	-0.89	.32	5.52	–B
⟨memorable, unforgettable⟩	50	54	23	60	O	0.56	.29	4.83	+B
⟨like, love⟩	50	25	80	57	O	0.23	.37	5.74	–B
⟨good, perfect⟩	46	39	60	23	O	1.00	.42	6.09	+B
⟨good, excellent⟩	37	32	60	57	O	1.34	.46	5.48	–B
⟨cool, cold⟩	33	46	23	40	O	-0.21	.61	4.30	–B
⟨hungry, starving⟩	33	25	63	40	O	0.71	.52	5.74	–B
⟨adequate, good⟩	29	32	33	57	O	-1.52	.27	3.52	–B
⟨unsettling, horrific⟩	29	25	37	37	O	-0.48	NA	5.65	–B
⟨dislike, loathe⟩	29	18	93	90	O	0.46	.16	5.87	–B
⟨believe, know⟩	21	61	67	67	O	-0.70	.46	5.04	+B
⟨start, finish⟩	21	21	43	50	O	0.70	.40	4.95	+B
⟨participate, win⟩	21	18	7	37	O	-0.62	.21	6.35	+B
⟨wary, scared⟩	21	14	40	37	O	-0.48	.06	4.39	–B
⟨old, ancient⟩	17	36	50	33	O	1.08	.24	5.39	–B
⟨big, enormous⟩	17	21	83	37	O	1.13	.21	5.43	–B
⟨snug, tight⟩	12	21	87	87	O	-1.05	.30	2.86	–B
⟨attractive, stunning⟩	8	21	53	72	O	0.37	.07	5.78	–B
⟨special, unique⟩	8	14	50	30	O	0.54	.32	3.48	+B
⟨pretty, beautiful⟩	8	11	73	50	O	-0.46	.41	5.04	–B
⟨intelligent, brilliant⟩	8	7	17	3	O	-0.12	.27	4.74	–B
⟨funny, hilarious⟩	4	29	50	33	O	1.17	.07	5.04	–B
⟨dark, black⟩	4	29	30	27	O	-0.49	.40	4.04	+B
⟨small, tiny⟩	4	25	80	27	O	0.80	.54	4.22	–B
⟨ugly, hideous⟩	4	18	37	31	O	0.86	.48	5.27	–B
⟨silly, ridiculous⟩	4	14	77	40	O	0.01	.43	4.17	–B
⟨tired, exhausted⟩	4	14	57	41	O	0.92	.45	5.13	–B
⟨content, happy⟩	4	4	87	50	O	-0.85	.13	4.52	–B

Table 4.7: List of scales used in the experiments reported in this section. Legend: ***SI*** = percentages of participants who derived a scalar inference; ***Cloze*** = percentages of participants who mentioned a stronger scalar term in the modified cloze task (Exp. 3, lenient analysis); *+N* = neutral condition (Exp. 1); *–N* = non-neutral condition (Exp. 2); ***Lex*** = lexical class (*O* = open, *C* = closed); ***Freq*** = logarithm of the ratio between the frequency of the weaker scalar term and the frequency of the stronger scalar term; ***LSA*** = semantic relatedness based on latent semantic analysis; ***Dist*** = mean perceived semantic distance (Exp. 4); ***Bnd*** = boundedness (*+B* = bounded, *–B* = non-bounded).

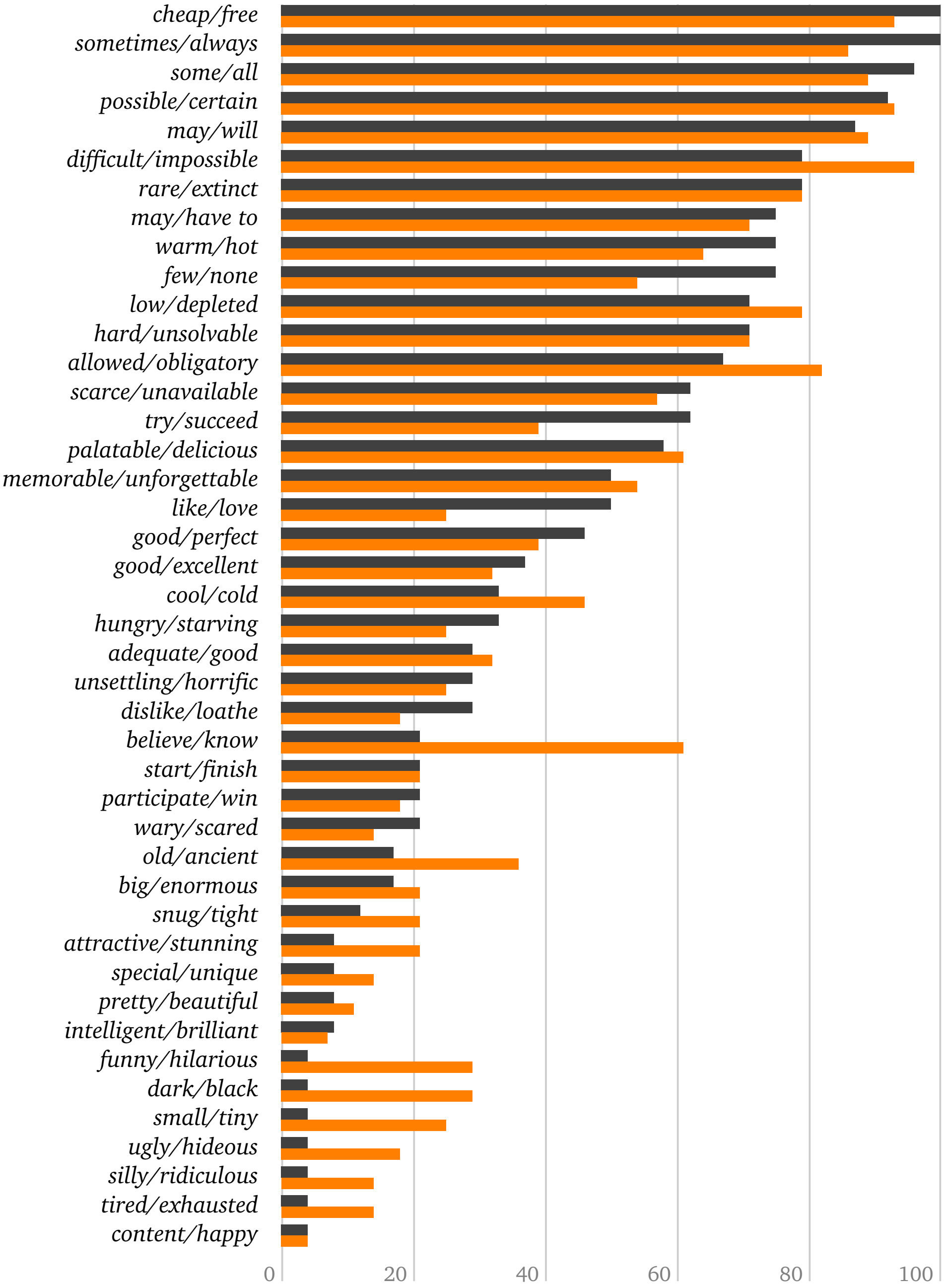


Figure 4.6: Percentages of positive responses in Experiment 1 (neutral content, dark grey) and Experiment 2 (non-neutral content, orange). The acceptance rates for entailments and unfounded inferences were 92% and 6%.

Experiment 2

Participants

We posted surveys for 30 participants on Amazon's Mechanical Turk (mean age: 32; range: 21–62; 14 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. One participant was excluded from the analysis because she was not a native speaker of English. None of the participants in Experiment 2 had already participated in Experiment 1.

Materials and procedure

We tested the same scales as in Experiment 1, using the same procedure. But in this case, the statements made by John and Mary contained more specific predicates and full noun phrases rather than pronouns. These statements were created on the basis of the following pretest. Ten participants (mean age: 35; range: 21–60; 6 females), all of them U.S. residents and native speakers of English, were drafted through Amazon's Mechanical Turk. Participants saw sentences containing a gap, like the following:

(8) a. The ________ is attractive but she isn't stunning.
    b. He is sometimes ________ but not always.

Statements always contained both the weaker and the stronger scalar term because we wanted to avoid confusion about the meaning of the weaker scalar term. Otherwise, scalar terms like 'low' and 'hard', for instance, might have received an interpretation on which they are incompatible with 'depleted' and 'unsolvable', respectively. Participants were instructed to indicate how the blanks could be filled in so as to yield a natural-sounding sentence, and had to provide three completions for every statement.

Out of all the completions suggested by the participants in the pretest, we selected three per scale, applying two constraints. First, we sought to ensure sufficient variation for each scalar expression. To illustrate, in the case of (8a), we chose 'nurse', rather than 'singer', in addition to 'model' and 'actress'. Second, whenever possible, we selected two relatively frequent and one relatively infrequent completion for each scale; if the variation of suggested completions was too great to apply this criterion, a random selection was made. Thus we constructed three statements for every scale. An example trial is given in Figure 4.7. Every statement was encountered by 10 participants (i.e. 1 in 3). Lastly, we included seven control items per list, in which the statement either entailed or was unrelated to the critical inference. The target and control statements are listed in the appendix.

John says:

This student is intelligent.

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

□ Yes □ No

Figure 4.7: Sample item used in Experiment 2.

Results and discussion

One participant was excluded from the analysis for making mistakes in four control items. Four out of a total of 1500 answers were missing. Figure 4.6 shows the mean acceptance rates for each scale.

Paired chi-square tests showed that only two scales yielded different rates of scalar inferences in the two experiments, namely ⟨believe, know⟩, where the rate of positive responses increased from 20% to 60% ($\chi^2(1) = 7.42$, $p = .01$), and ⟨funny, hilarious⟩, where the rate of positive responses went from 4% to 30% ($\chi^2(1) = 4.05$, $p = .04$). Accordingly, the product-moment correlation between the proportions of positive answers for corresponding items in the two experiments was quite high ($r = .91$, $t(41) = 13.98$, $p < .01$). Overall, the rates of positive responses (42% versus 44%) did not differ significantly across the two experiments ($\chi^2(1) = 0.85$, $p = .37$). Paired chi-square tests showed that there was no pair of statements for any scale that yielded significantly different rates of positive answers (though it should be noted that there were at most ten observations per statement).

Adding more content to the materials had a relatively small effect on the overall results, and did not affect the general conclusions we drew from the results of Experiment 1. This finding suggests that the general pattern of responses is robust to changes in the sentential context. Given our own data and Doran et al.'s, we can safely say that the uniformity assumption is false: the rates at which scalar expressions yield upper-bounding inferences fluctuate wildly.

Before moving on, we first consider a potential methodological issue with the inference task. Consider the example trial in Figure 4.7. This trial asks participants if, according to the speaker, the student is 'not brilliant'. It has been observed that negated expressions sometimes cause an inference to the antonym. In other words, 'not brilliant' sometimes conveys a mitigated sense of dumbness (e.g., Fraenkel & Schul 2008, Horn 1989, Krifka 2007). Perhaps, then, the variable rates of scalar inferences that we observed in Experiments 1 and 2 are affected by the likelihood with which the negated scalemate licensed an inference to the antonym. According to this explanation, inferences to the antonym should occur more often with, for example, ‘not exhausted’ and ‘not tight’ than with ‘not free’ and ‘not hot’.

There are, however, a number of reasons to assume that inferences to the antonym did not confound the general pattern of results. Firstly, the effect of inferences to the antonym might be preempted by the content of the speaker’s statement. For example, participants might avoid interpreting ‘not brilliant’ as rather dumb because John just stated that she is intelligent. The question is much less trivial if the negated adjective receives its literal interpretation. Secondly, inferences to the antonym are especially robust if the negated expression contains a negative element itself (e.g., Horn 1989, Krifka 2007). We tested a number of such expressions: ‘impossible’, ‘none’, ‘unsolvable’, ‘unavailable’, and ‘unforgettable’. However, all of these expressions generated scalar inferences in more than 50% of the cases. Thirdly, Doran et al. (2009, 2012) compared scalar inference rates for quantifying expressions and gradable adjectives in a verification task. This paradigm does not involve negated expressions and is therefore not susceptible to the problem of inferences to the antonym. The relative proportions of scalar inferences for quantifying expressions and gradable adjectives in Doran et al.’s task (32% versus 17% negative responses) were the same as for scalar expressions from closed and open grammatical categories in Experiments 1 and 2 (76% versus 40% positive responses).

We conclude that the results of Experiments 1 and 2 provide a reliable indication of the likelihood with which different lexical scales license upper-bounding inferences. The variable rates of scalar inferences suggest that lexical scales differ in one or more aspects that are relevant for the computation of scalar inferences. In what follows, we discuss two such aspects: availability and distinctness. Afterwards, we measure the contribution of these factors to the rates of scalar inferences by operationalising them in a number of ways.

4.3.3 Explaining diversity

In order to compute a scalar inference, one has to assume that the speaker considered using a stronger scalemate of the scalar expression he used in his utterance. Otherwise it would be mistaken to infer from the speaker’s utterance that he believes the stronger scalar expression is inappropriate. So perhaps the variable rates of scalar inferences are caused by differences in the *availability* of lexical scales.

Doran et al. (2009) provide some evidence to suggest that lexical scales are indeed available to different degrees. As discussed in the previous section, participants in their experiment were presented with stories in which Irene asked a question. In the neutral condition, Irene’s question did not contain any scalar expressions; in the one-way contrastive condition, it mentioned a scalar expression that was stronger than the one used in Sam's answer; in the two-way contrastive condition, Irene's answer offered Sam three scalar expressions to choose from:

(9) a. How much cake did Gus eat at his sister's birthday party?
b. Did Gus eat all of his sister's birthday cake?
c. Did Gus eat some, most, or all of his sister's birthday cake.

It seems plausible that mentioning the scalemates of the scalar expression in Sam's answer makes the corresponding lexical scale more available and thus increases the likelihood of a scalar inference. In line with this prediction, Doran et al. observed higher rates of scalar inferences for adjectival scales in the two-way contrastive condition compared to the neutral and one-way contrastive conditions. No such effect, however, was found for quantificational scales. These observations can be construed as implying that quantificational scales are by default more available than adjectival scales. Explicit mentioning therefore has an effect on the rates of scalar inferences for adjectival but not quantificational scales.

Even if the lexical scale is available, a scalar inference can be preempted if the speaker used the weaker scalar term for a reason other than his believing that the utterance with the stronger scalar term is false. One such alternative reason is that the speaker is uncertain which scalar expression is appropriate. The likelihood that such a situation obtains will depend inter alia on the *distinctness* of the scale members, i.e., how easy it is to perceive the distinction between them. To illustrate, consider the scalar expressions 'some' and 'intelligent'. Intuitively, it is easier to establish if someone solved some or all of the problems than if that person is intelligent or brilliant. This difference in distinctness might explain why upper-bounding inferences were more prevalent for 'some' than for 'intelligent'. More generally, perhaps the variable rates of scalar inferences can be attributed to differences in the distinctness of the scalar expressions on a scale.

In order to determine to what extent availability and distinctness can account for the variable rates of scalar inferences, we operationalised these notions in a number of ways. As measures of availability, we considered strength of association, grammatical class, word frequencies, and semantic relatedness. As measures of distinctness, we considered semantic distance and boundedness. In the following sections, we discuss these factors in greater detail.

4.3.4 Availability

Association strength

The most straightforward measure of the availability of a lexical scale is the strength of association between the scalar expression used in the speaker's utterance and its stronger scalemate. The greater the association strength, the more likely it is that the speaker considered using the stronger scale member. So perhaps the

She is intelligent.

She is __________
She is __________
She is __________

Figure 4.8: Sample item used in Experiment 3 (–N condition).

differential rates of scalar inferences can be explained in terms of differences in association strengths. To illustrate, consider the scalar expressions ‘warm’ and ‘big’. The reason that scalar inferences were more prevalent for ‘warm’ than for ‘big’ might be that the strength of association between ‘warm’ and ‘hot’ is much greater than between ‘big’ and ‘enormous’. Thus we arrive at the following hypothesis:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is an increasing function of the strength of association between α and β.

In order to test this hypothesis, we need to measure the strength of association between two scalar expressions. To this end, we conducted a modified cloze task. A standard cloze task, like the one we used to obtain materials for Experiment 2, consists of sentences or text fragments with certain words removed, where participants are asked to replace the missing words. We modified this design by underlining instead of removing words. Participants were asked to list three alternatives to a given sentence $\varphi[\alpha]$ by replacing the underlined scalar term α with whatever expression they saw fit. We assumed that the stronger the association between α and β, the more likely it would be that participants replaced α with β.

Experiment 3

Participants

We posted surveys for 60 participants on Amazon’s Mechanical Turk (mean age: 36; range: 21–57; 21 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. All participants were native speakers of English. Two of the participants had already participated in Experiment 1 or 2. We included these participants in the analysis we discuss below. Excluding them would not change the statistical significance of any of the *p*-values we report.

Materials and procedure

Figure 4.8 shows an example of a critical item. Each trial consisted of a sentence with a scalar term that was underlined. Participants were instructed to indicate which words could have occurred instead of the underlined word. Half of the participants saw the neutral statements used in Experiment 1; the other half saw the non-neutral statements from Experiment 2. We constructed two minimally different sets of instructions. One version is given below:[2]

> In the following you will see 43 sentences. In every sentence, one word will be highlighted, like this:
>
> She is angry.
>
> Which words could have occurred instead of the highlighted one? Some of the alternatives that may come to mind are *beautiful*, *happy*, *married*, and so on. We ask you to tell us the first three alternative words that occur to you when you read these sentences. We are interested in your spontaneous responses, so don't think too long about it.

In the second version, the first sample alternative (here 'beautiful') was replaced with a scalar term that was stronger than the highlighted expression (namely 'furious'). We did this to control for the possibility that mentioning or not mentioning a stronger expression in the instructions might have an effect on the responses. More precisely, participants might be more likely to provide stronger scalemates if a stronger scalemate had been mentioned in the instructions. A different list was constructed for each of the participants, varying the order of the trials.

Results and discussion

Seven out of a total of 2550 answers were missing. We annotated our results in two different ways. For each trial, we first coded if the participant mentioned the stronger scalar term we used in the inference tasks. However, this measure might be too strict because, first, there were scalar terms for which participants often mentioned a synonymous or nearly synonymous stronger scalar term, like 'every' instead of 'all'. Second, it might be the case that participants compute a scalar inference if they consider a stronger scalar term different from the one in the question. For instance, a participant who associates 'possible' with 'probable', and computes a scalar inference on the basis of the scale ⟨possible, probable⟩, thereby also infers that it is not certain, even though she did not consider that particular alternative. Therefore we also determined for each trial whether any stronger scalar term was mentioned. In this measure, we did not include scalar expressions that were stronger than scalar expression in the question, such as 'perfect' for the

[2] Note that the neutral version included only 41 statements, the reason being that the statements for ⟨good, excellent⟩ and ⟨good, perfect⟩, on the one hand, and ⟨may, have to⟩ and ⟨may, will⟩, on the other, were identical in this version of the task. In the analysis reported below, we paired the results for these statements with the results on the inference task for ⟨good, excellent⟩ and ⟨may, have to⟩, respectively. Changing this pairing did not have any effect on the results.

⟨adequate, good⟩ scale and ‘freezing’ for the ⟨cool, cold⟩ scale. After all, someone who infers from (10a) that, according to the speaker, it is not perfect does not necessarily infer that it is not good. Similarly for (10b): someone who infers that it is not freezing does not necessarily infer that it is not cold.

(10) a. It is adequate.
b. That is cool.

The results of our analyses are summarised in Table 4.7. We start with the strict coding scheme. We first conducted a loglinear analysis to test whether the probability that the stronger scalar term used in the inference task was mentioned was affected by (a) whether or not the target sentences were neutral (+N vs. –N) and (b) whether or not a stronger scalar expression was mentioned in the instructions (+S vs. –S). A summary of the effects of these factors is given in Table 4.8. Overall, the stronger scalar term was mentioned in 25% of the trials. It was mentioned significantly more often with neutral statements (27%) than with non-neutral ones (22%, $G^2(1) = 11.53, p < .001$). However, this effect interacted with the form of the instructions ($G^2(2) = 14.22, p = .001$): it was only significant if the instructions did not contain a stronger scalar term ($G^2(1) = 12.28, p < .001$). The stronger scalar term was also mentioned significantly more often when the instructions contained a stronger scalar term (27%) than when they did not (22%, $G^2(1) = 7.22, p < .01$), and again there was an interaction with the neutral/non-neutral factor ($G^2(2) = 9.91, p < .01$): the effect reached significance for non-neutral statements only ($G^2(1) = 9.12, p < .005$).

	–N	+N
+S	25	29
–S	18	26

Strict coding

	–N	+N
+S	47	51
–S	40	46

Lenient coding

Table 4.8: Percentages of responses in Experiment 3 which mentioned either the same scalar term we used in our inference tasks *(Strict coding)* or any stronger scalar term *(Lenient coding)*. Instructions either contained a stronger scalar term *(+S)* or not *(–S)*, and sentences were neutral *(+N)* or not *(–N)*.

A possible explanation for why stronger scalar terms were mentioned more often in the neutral condition is that in this condition, the scalar term was more or less the only thing to go on, whereas in the non-neutral condition, associations were constrained by the sentential context as well. To illustrate, compare the following sentences:

(11) a. That house is old.
b. It is old.

Whereas in the case of (11a) participants might mention properties they associate with houses or old houses, (11b) is much less constraining. Mentioning a stronger scalar term in the instructions dampened this effect.

With the lenient coding scheme, we found a very similar pattern. A stronger scalar term was mentioned in 46% of the trials. It was mentioned significantly more often with neutral than non-neutral sentences (49% vs. 44%, $G^2(1) = 6.41, p < .025$). As with the strict coding scheme, this effect interacted with the form of the instructions ($G^2(2) = 6.87, p < .05$): it only reached significance if the instructions did not contain a stronger scalar term ($G^2(1) = 5.01, p < .025$). Stronger scalar terms were mentioned significantly more often if the instructions contained a stronger scalar term than when they did not (49% vs. 43%, $G^2(1) = 9.57, p < .01$). There was an interaction with the neutral/non-neutral factor: the effect was only significant with non-neutral statements ($G^2(1) = 6.98, p < .01$).

Let us now examine the association hypothesis in light of the foregoing results. In order to determine which factors are significant predictors of the rates of scalar inferences in Experiments 1 and 2, we used the `lme4` package (Bates & Maechler 2009) to construct a binomial mixed model with the responses in the inference tasks as dependent variable, and the measures with which we operationalised the notions of availability and distinctness as independent factors, and participants and items as random factors (Barr et al. 2013). The parameters of the mixed model are provided in section 4.2.3.

The proportion of participants in Experiment 3 who mentioned a stronger scalemate was not a significant predictor of the rates of scalar inferences in the corresponding inference task ($\beta = 0.16$, $SE = 0.31$, $Z < 1$). The same conclusion holds for the strict analysis in which we counted the proportion of participants who mentioned the exact stronger scalemate that was used in the inference task ($\beta = 0.11$, $SE = 0.31$, $Z < 1$). Note that, for both measures, the direction of the effect even goes in the opposite direction of what is predicted by the association hypothesis.

Therefore, whether or not a scalar inference is computed does not seem to depend on association strength, as operationalised in the modified cloze task. To illustrate, in the case of ‘snug’, nearly all participants in Experiment 3 mentioned ‘tight’ as an alternative, but in Experiments 1 and 2 the average rate of the scalar inference was only 16%; similar observations hold for ⟨pretty, beautiful⟩ and ⟨dislike, loathe⟩. On the other hand, there was a substantial group of scales that yielded high rates of scalar inferences, but for which stronger scalar terms were rarely mentioned in Experiment 3, clear examples being ⟨cheap, free⟩, ⟨hard, unsolvable⟩ and ⟨difficult, impossible⟩. In sum, the findings of this experiment argue against the hypothesis that rates of scalar inferences are determined by the strength of the connections between weaker and stronger scalar terms.

It might be objected that the modified cloze task is a poor measure of association strength because participants who computed a scalar inference based on the target sentence might therefore not have mentioned a stronger scalar term. According to this explanation, participants were guided in part by the inferences that could be made on the basis of the target sentence. However, this prediction is incorrect, since antonyms were among the most frequently given answers: participants mentioned an antonym in 35% of the items. Apparently, participants were not constrained by the information conveyed by the target sentence. We thus conclude that strengths of association do not have an effect on the rates of scalar inferences.

A more pressing issue is that the cloze task does not provide an absolute measure of the strength of association between two expressions. Even if the association strength between a scalar expression α and its stronger scalemate β is high, this might not be visible in the results of the cloze task because there are at least three expressions with which it is even more strongly associated. Conversely, even though the association strength between α and its stronger scalemate β is low, this might not be visible in the results of the cloze task because there are no other expressions with which it is more strongly associated. In order to address this concern, we discuss and test three other measures of availability.

Grammatical class

A first alternative measure of availability involves the distinction between open and closed grammatical classes. The domain of closed grammatical classes, like quantifiers and auxiliary verbs, is much smaller than that of open grammatical classes, like adjectives, adverbs, and main verbs. In consequence, the search space of alternatives is much smaller for closed grammatical classes than for open ones. So it seems plausible to suppose that lexical scales are more available when their elements are from a closed grammatical class than from an open one. The following hypothesis captures this explanation:

> The availability of a lexical scale $\langle\alpha, \beta\rangle$ is greater if α and β are from a closed grammatical class.

To test this hypothesis, we subdivided the scalar expressions into open and closed grammatical classes, as can be seen in Table 4.7. Although the average rate of scalar inferences was higher for scales from closed (76%) than open (40%) grammatical classes, the distinction between them did not have a significant effect on the rates of scalar inferences (β = -0.47, *SE* = 0.47, *Z* = -1.00, *p* = .32). One factor contributing to this nonsignificant result is that, in our experimental items, all closed class scales were also bounded scales (but not the other way around). We discuss the distinction between bounded and non-bounded scales in section 4.3.5.

Word frequencies

A third measure of availability is based on word frequencies. To see how these could have an effect, we compare the scales ⟨warm, hot⟩ and ⟨big, enormous⟩, which gave rise to scalar inferences 65% and 19% of the time, respectively. It might be that this discrepancy was caused by the fact that, whereas 'hot' is a quite common word that should be readily available to the speaker in a context in which she uttered 'warm', 'enormous' is rare relative to 'big', which might explain why the speaker did not use it even if, strictly speaking, it was more appropriate than 'big'. This explanation can be generalised and made more precise as follows:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is an increasing function of the frequency of β relative to that of α.

In order to test this hypothesis, we extracted the frequencies of all scalar expressions in our materials from the Corpus of Contemporary American English (Davies 2008). For each scale, we divided the frequency of the stronger scalar term by the frequency of the weaker one, and logarithmised the outcome to reduce the skewness of the resulting distribution. The results of this analysis are given in Table 4.7. The logarithmised ratio of the frequencies of the scalemates did not have a significant effect on the rates of scalar inferences that we found in Experiments 1 and 2 (β = -0.15, SE = 0.21, $Z < 1$).

An alternative possibility is that it is not relative frequency, but rather the bare frequency of the stronger alternative that determines the likelihood with which a scalar inference is derived. The idea would be that, even if 'horrific' is more frequent than 'unsettling', a speaker who uses 'unsettling' might not have considered 'horrific' simply because it is a rare word. To test this hypothesis, we carried out an analysis similar to the one reported in the last paragraph, but this time using logarithmised frequencies of the stronger scalar terms as predictor variable. Again, the frequencies did not have a significant effect on the results of Experiments 1 and 2 (β = -0.14, SE = 0.24, $Z < 1$).

To sum up: it appears that neither the relative frequency of the scalar expressions nor the absolute frequency of the stronger term had a significant effect on whether or not a scalar inference is computed. We conclude, therefore, that frequency does not have a major effect on the distribution of scalar inferences.

Semantic relatedness

As a final test for the hypothesis that the variable rates of scalar inferences are caused by differences in the availability of the corresponding scale, we consider semantic relatedness. Words that are semantically related tend to occur in similar linguistic environments. To illustrate, 'warm' and 'hot' often occur with words like 'food', 'climate', 'water', and 'sand', whereas 'warm' and 'stunning' do not have such shared collocations. It has been demonstrated that words that tend to occur in the same environments also prime each other in word recognition tasks (Landauer et al. 1998). It seems plausible to suppose, then, that semantic relatedness provides a good measure of availability:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is an increasing function of the semantic relatedness of α and β.

A common measure of semantic relatedness is latent semantic analysis (Landauer & Dumais 1997). LSA constructs a matrix with words from a corpus as rows and columns. A row consists of binary values that represent whether the words in question occur in the same sentence; so words that co-occur in a sentence have a 1 in the same column. Words that are semantically related are expected to occur relatively often with the same words and thus have a lot of 1s in the same columns. Based on this matrix, LSA computes a value in the interval [0, 1] that denotes the semantic relatedness of different words. For example, the LSA value for 'warm' and 'hot' is .51 compared to .02 for 'warm' and 'stunning'. Note that these LSA values do not indicate how often the scalemates co-occur but rather how often they co-occur with the same words.

On the basis of Landauer et al.'s (1998) LSA implementation, we obtained relatedness values for each pair of scalar terms through pairwise, term-to-term comparisons with "general reading up to first year of college" as topic space. These relatedness values, which are provided in Table 4.7, were used as an estimator of the results of Experiments 1 and 2. LSA values were not a significant predictor of the rates of scalar inferences ($\beta = 0.01$, $SE = 0.01$, $Z < 1$). We thus conclude that semantic relatedness has no effect on the rates of scalar inferences that we observed in Experiments 1 and 2.

Conclusion

In order to compute a scalar inference, one has to assume that the speaker considered the corresponding lexical scale. Otherwise it would be mistaken to attribute his choice for a weaker scalar expression to the belief that the stronger scale member is inappropriate. Based on this observation, we hypothesised that the differential rates of scalar inferences in Experiments 1 and 2 were caused by differences in availability. In the previous sections, we have operationalised the notion of availability in various ways. But none of these measures made a significant contribution to the rates of scalar inferences. Availability thus seems to have a marginal role at best in shaping the results of Experiments 1 and 2.

It might be countered that the absence of a significant contribution of availability has a methodological cause. In our inference tasks, the question participants had to answer contained a scale member that was stronger than the one used in the target statement. One might suppose that this feature caused all lexical scales to be rendered available, thereby obviating the effect of intrinsic measures of availability like the ones we tested in the previous sections.

A number of observations speak against this explanation. First and foremost, recall that Doran et al. (2009) made a comparison between neutral, one-way contrastive, and two-way contrastive items. In the neutral condition, Irene's question did not contain scale members; in the one-way contrastive condition, it contained one scale member that was stronger than the one used in Sam's answer; and in the two-way contrastive condition, Irene, in effect, provided Sam with three scale members to choose from. The items in our inference tasks most closely resemble the items in Doran et al.'s one-way contrastive condition, since both involve a question that contains a scale member stronger than the one used in the target statement. Nevertheless, Doran et al. found no difference between the neutral and one-way contrastive items. This result provides strong evidence that mentioning one stronger scale member does not affect the availability of the lexical scale.

In addition, even if the question in the inference task made the lexical scale available to the participants, it does not follow that, according to these participants, it was also available to the speaker. After all, the question that mentions the stronger scalar expression was not presented to the speaker. In this respect, our inference tasks differ from Doran et al.'s one-way contrastive condition, in which the question that contains the stronger scalar expression is presented to the speaker character. So if mentioning a stronger scalar term affects the availability of lexical scales, this effect should be more pronounced in Doran et al.'s task than in our inference tasks. The lack of an effect in Doran et al.'s task thus makes it unlikely that such an effect should have occurred in our inference tasks.

We conclude that availability plays a marginal role in fixing the likelihood of a scalar inference. In the next section, we discuss a second possible factor: distinctness. If a scalar inference is computed, it has to be assumed that the speaker is able to determine which scalar expression is most appropriate. So if distinguishing between scalar expressions is difficult, it might be less likely that a scalar inference is derived. In the next section, we discuss two measures to operationalise the notion of distinctness: semantic distance and boundedness.

4.3.5 Distinctness

Semantic distance

It seems likely that expressions that are semantically distant are easier to distinguish than expressions that are semantically close. The notion of semantic distance was inspired by an observation by Horn (1972, 90). Consider:

(12) a. Many of the senators voted against the bill.
b. Most of the senators voted against the bill.
c. All of the senators voted against the bill.

It seems to us that, normally speaking, an utterance of (12a) would be more likely to implicate the negation of (12c) than the negation of (12b). The reason for this divergence is that the difference in semantic strength between (12a) and (12c) is greater than that between (12a) and (12b). This intuition was partially vindicated by Zevakhina (2012), who found that participants considered the semantic distance between 'some' and 'all' greater than between 'some' and 'most'; although she did not find such a pattern for adjectival scales like ⟨warm, hot, sweltering⟩. The idea underlying the following hypothesis is that the highly variable rates at which scalar inferences are drawn might be explained in terms of the semantic distance between the weaker and the stronger term:

> Given a lexical scale $\langle\alpha, \beta\rangle$, the distinctness of α and β is an increasing function of the semantic distance between these expressions.

Obviously, this hypothesis presupposes that it makes sense to compare pairs of expressions from different scales, and thus requires an absolute measure of semantic distance. Assuming that there is such a thing and that speakers have reliable intuitions about it (and neither assumption seems entirely unreasonable to us), the distance hypothesis leads us to expect that speakers' intuitions about semantic distance should at least be a partial predictor of the likelihood of a scalar inference. Therefore, we conducted an experiment in which participants were asked, for all scales $\langle\alpha, \beta\rangle$ used in Experiments 1 and 2, how much stronger $\varphi[\beta]$ is relative to $\varphi[\alpha]$, and compared the results to the findings of those experiments.

(Note that the notion of semantic distance is not interdependent with the notion of semantic relatedness. It is possible for two expressions to be related but distant or unrelated but close. For example, 'warm' and 'cold' are related but distant.)

Experiment 4

Participants

We posted surveys for 25 participants on Amazon's Mechanical Turk (mean age: 33; range: 20–62; 15 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. One participant was excluded from the analysis because she was not a native speaker of English. Two participants had also participated in Experiment 1 or 2. We included these participants in the analysis. Excluding them would not change the statistical significance of any of the *p*-values we report.

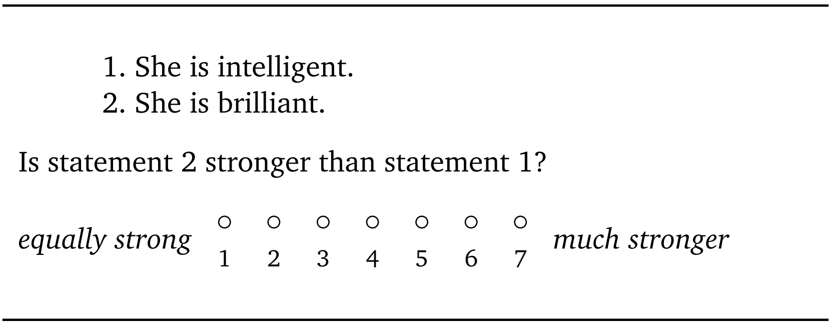
Materials and procedure

An example trial is given in Figure 4.9. Participants were instructed to indicate whether and, if so, to what extent a statement with the higher-ranked scalar term was stronger than the same statement with the lower-ranked scalar term, by selecting a value on a seven-point scale. The instructions went as follows:

> Consider the following claims:
>
> 1. This is okay.

> 2. This is fantastic.
>
> Clearly, claim 2 is stronger than claim 1.
> Now compare the following claims:
>
> 3. This is fantastic.
> 4. This is marvelous.
>
> Here, neither claim seems much stronger than the other, if they differ in strength at all. In this questionnaire, we will show you a number of sentence pairs like the ones above. In each case, we ask you to indicate on a 7-point scale how much stronger the second claim is, where 1 means that the two claims are equally strong, and 7 means that the second claim is much stronger than the first one.

For this test, the neutral statements of Experiment 1 were used. Different lists of items were constructed for all participants, varying the order of the trials. Seven control items were included, each of which involved two statements which were synonymous or nearly so. These control statements involved the following pairs of words: ‘enormous’/‘immense’, ‘fantastic’/‘sensational’, ‘gifted’/‘talented’, ‘obvious/‘clear’, ‘unbearable’/‘intolerable’, ‘unexpected’/‘unforeseen’, and ‘unpleasant’/‘disagreeable’.



1. She is intelligent.
2. She is brilliant.

Is statement 2 stronger than statement 1?

equally strong ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 *much stronger*

Figure 4.9: Sample item used in Experiment 4.

Results and discussion

Eight out of a total of 1250 answers were missing. One participant was excluded from the analysis because her mean rating for the control items exceeded two standard deviations from the grand mean for these items. The results of the experiment are presented in Table 4.7. The mean distance for the synonymous control items was 2.81. The 95% confidence interval around this mean was 2.53–3.09. There was only one lexical scale whose mean distance fell within that confidence interval: ⟨snug, tight⟩. This finding suggests that, except for this outlier, participants were able to perceive a difference in strength between scalemates.

The mean ratings on the distance task made a significant contribution to the rates of scalar inferences ($\beta = 0.65$, $SE = 0.27$, $Z = 2.36$, $p = .02$). This finding confirms the prediction made by the distance hypothesis. In the conclusion to this section, we discuss the variance explained by this and other factors.

Boundedness

A second measure of distinctness is more structural in nature. We have seen that rates of scalar inferences differ even within scalar expressions of the same grammatical class. For example, the percentages of positive responses for adjectival scales range from 4% for ⟨content, happy⟩ to 95% for ⟨cheap, free⟩. However, there is an important difference between these two scales: in the case of 'cheap', but not in the case of 'content', the stronger scale member denotes an end point on the dimension over which the scalar terms quantify (Kennedy & McNally 2005, Rotstein & Winter 2004). We will refer to scales with such a terminal expression as *bounded*, as opposed to *non-bounded* scales like ⟨content, happy⟩. Note that boundedness depends on the semantics of the stronger scalar expression alone.

Scalar expressions on bounded scales can be distinguished on formal grounds alone: one scalar term denotes an interval and the other one an end point. By contrast, distinguishing scalar expressions on non-bounded scales requires inspecting the reach of the intervals denotes by both non-terminal expressions. It might therefore be hypothesised that scalar expressions on bounded scales are easier to distinguish than on non-bounded scales:

> Given a lexical scale $\langle\alpha, \beta\rangle$, the distinctness of α and β is greater if β is a terminal expression.

To test this hypothesis, we subdivided the lexical scales from Experiments 1 and 2 according to whether the stronger scalar expression denotes an end point, as can be seen in Table 4.7. It turned out that this classification subsumed the classification into open and closed grammatical classes. That is, all scalar expressions from closed grammatical classes occurred on bounded scales but not vice versa. This is not necessarily so: scales like ⟨some, most⟩ and ⟨sometimes, often⟩ are open even though they consist of elements from a closed grammatical class.

It was found that bounded scales indeed licensed higher rates of scalar inferences than non-bounded scales (62% versus 25%). Boundedness made a significant contribution to the rates of scalar inferences in Experiments 1 and 2 (β = -1.87, SE = 0.40, Z = -4.72, $p < .01$). The likelihood of a scalar inference is predicted in part by the distinction between bounded and non-bounded lexical scales. In the following section, we discuss a measure of the variance explained by boundedness.

Conclusion

If the distinction between scalar expressions is unclear, the speaker might choose to use a weaker expression because he is uncertain about whether the stronger expression is appropriate. Based on this observation, we hypothesised that the general pattern of results in Experiments 1 and 2 is shaped by the distinctness of the scale members. In the previous sections, we operationalised the notion of distinctness by means of semantic distance and boundedness. Both of these measures turned out to have a significant effect on the rates of scalar inferences: the likelihood of a scalar inference increased with the semantic distance between scalar expressions, and scales with a terminal expression caused significantly higher rates of scalar inferences than scales without a terminal expression. We conclude that the likelihood of an upper-bounding inference is partly predicted by distinctness.

4.3.6 General discussion

In recent years, neither the experimental nor the theoretical literature on scalar inferences has shown much concern for the diversity of scalar expressions, and by and large has confined its attention to less than a handful of items, notably ‘some’ and ‘or’. Presumably, the tacit assumption has been that these are representative of the whole family of scalar terms. That assumption turns out to be mistaken: following up on studies by Doran et al. (2009, 2012), we have shown that the rates at which scalar expressions give rise to upper-bounding inferences could hardly be more diverse, and that the ⟨some, all⟩ scale, which has been the workhorse of recent research on scalar inferences, is an extreme case (Experiments 1 and 2).

This was our main finding, but a large part of the foregoing discussion addressed the question of how the observed diversity can be accounted for. We considered two factors that might help to explain the variable rates of scalar inferences: availability and distinctness. Availability refers to how likely it is, according to the hearer, that the speaker considered stronger scalemates. Distinctness refers to how likely it is, according to the hearer, that the speaker considers the distinction between the weaker and the stronger scalar expression substantial enough that it is reasonable to assume that he should have used the latter if possible. In a series of analyses, we operationalised these factors in various ways.

We introduced two measures of distinctness, both of which made a significant contribution to the rates of scalar inferences:

i. *Semantic distance*
The difference in strength between $\varphi[\alpha]$ (e.g. ‘It is warm’) and $\varphi[\beta]$ (e.g. ‘It is hot’) showed a positive correlation with the likelihood that $\varphi[\alpha]$ would trigger the inference that $\neg\varphi[\beta]$.

ii. *Boundedness*
Scalar expressions that inhabit a bounded scale, on which the stronger scalar term refers to an end point, were more likely to give rise to scalar inferences than their non-bounded counterparts. While bounded scales predominate in the upper half of the distribution in Figure 4.6, the lower half is populated

mainly by non-bounded scales. However, there is no strict dichotomy: inference rates were high for some non-bounded scales, too, and low for some of the bounded scales.

In constrast to these two measures of distinctness, none of our four measures of availability had a significant effect on the variable rates of scalar inferences:

i. *Association strength*
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$ might have correlated with the association strength between α and β (relative to the sentence frame $\varphi[\;]$) or with the association strength between α and any other stronger scalemate of α's. However, we didn't find evidence for either hypothesis.

ii. *Grammatical class*
In their study, Doran et al. contrasted quantificational scales with adjectival scales. We included a similar subdivision between scalar expressions from open and closed grammatical classes. This distinction did not have an effect on the rates of scalar inferences.

iii. *Word frequencies*
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$, where β is a stronger scalemate of α, might be correlated with the frequency of β. We tested two versions of this idea, measuring β's frequency either in absolute terms or relative to α's frequency, but neither version was supported by the data.

iv. *Semantic relatedness*
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$ might depend on how often α and β occur in similar linguistic environments. We determined the relatedness between expressions by means of latent semantic analysis (Landauer & Dumais 1997), but the outcome did not predict the rates of scalar inferences observed in Experiments 1 and 2.

In order to gauge how much variance was explained by each of the foregoing factors, we employed the measure of explained variance introduced by Nakagawa & Schielzeth (2012). The full mixed model, which included participants and items as random factors, and association strength, grammatical class, relative word frequencies, semantic relatedness, semantic distance, and boundedness as fixed factors, explained 52% of the variance in the results of Experiments 1 and 2 (Table 4.9). 22% of this variance was explained by the fixed factors and the remaining 30% by differences between items and participants. As for the independent factors, we found that none of our measures of availability explained more than 1% of the results. Distinctness turned out to be a more substantial factor, with semantic distance explaining 3% and boundedness explaining 10% of the results. Note that these percentages do not sum to 22%, because some of the variance explained by a particular factor may be explained by another factor if the first factor is omitted from the model. For example, grammatical class explained a substantial part of the variance explained by boundedness in models where the latter factor was omitted.

Parameter	β	**SE**	**Z**	p	R^2
(Intercept)	-2.80	1.73	-1.62	.104	–
Association strength	0.16	0.31	0.51	.611	.000
Grammatical class	-0.38	0.74	-0.52	.606	.001
Relative frequency	-0.15	0.21	-0.74	.461	.003
Semantic relatedness	0.01	0.01	0.93	.355	.006
Semantic distance	0.65	0.27	2.36	.018	.027
Boundedness	-1.87	0.40	-4.72	.000	.108

Table 4.9: Parameters of a mixed model with the results from Experiments 1 and 2 as dependent variable, the strengths of association based on the lenient coding scheme (Experiment 3), open or closed lexical class, the logarithms of the ratio between the frequencies of scalemates, the semantic relatedness between scalemates, averages of the perceived semantic distance between scalemates (Experiment 4), and boundedness as independent variables, and participants and items as random variables.

To summarise, the full model explained roughly half of the observed variance; one fifth of the variance could be accounted for by factors we manipulated in our experiments, and half of that was due to boundedness. What could explain the remaining variance? One candidate factor that is often mentioned in the literature is that the likelihood of a scalar inference is determined by the question under discussion (e.g., van Kuppevelt 1996, van Rooij & Schulz 2004, Zondervan 2010). On this view, a scalar expression will only give rise to an upper-bounding inference if it is part of the focus of an utterance. That is to say, B's answer in (13), but not in (14), should imply that she did not eat all of the cookies.

(13) A: How many cookies did you eat?
B: I ate [some]$_F$ of the cookies.

(14) A: Who ate some of the cookies?
B: [I]$_F$ ate some of the cookies.

Since in our experiments no questions were asked, a possible explanation for the differential ratings of sentences with, e.g., 'warm' and 'big' is that participants tended to contextualise these sentences in different ways, with 'warm' having a preference for a focus interpretation and 'big' having a preference a non-focus interpretation.

However, there are rather compelling reasons to doubt that this explanation is on the right track. In our experiments, scalar adjectives always occurred in predicate position, which is widely agreed to be focused by default (Ward & Birner 2004, 154). Furthermore, in Experiment 1, grammatical subjects were always pronominal, and pronouns rarely receive focus (*ibid.*, 158). To illustrate, it is obvious that, in the following examples, the adjectives are highly likely to be focused:

(15) a. It is cheap.
b. It is small.

But whereas (15a) triggered scalar inferences in all cases, (15b) did so only 4% of the time. Although we cannot rule out the possibility that focus contributed to the rates of scalar inferences in Experiments 1 and 2, these observations suggest that focus is unlikely to be an important factor.

Which brings us back to our initial question: how to explain the remaining variance in the data of Experiments 1 and 2? In the foregoing, we have looked at all the candidate factors we could think of. Almost none of these factors explained a substantial portion of the observed variance; the exception was boundedness, and even its contribution was a mere 10%. In the absence of more successful candidates, we are forced to conclude that a major part of the observed variance was unsystematic. In Experiments 1 and 2, participants had to decide whether they would draw a scalar inference $\neg\varphi[\beta]$ from an utterance $\varphi[\alpha]$ that, save for the speaker's name, was not overtly contextualised. Making this decision requires an estimate of the likelihood that the speaker considered $\varphi[\beta]$ at least as relevant as $\varphi[\alpha]$. Our findings suggest that these estimates were by and large impervious to differences in word frequencies and various abstract semantic factors.

Perhaps it is not too surprising that this should be so. It is a well-established fact that speakers and hearers are alert to all manner of statistical patterns in language use (e.g., Seidenberg 1997), and therefore we might conjecture that language users keep track of the frequencies with which scalar expressions give rise to upper-bounded interpretations. If that is what underlies the remaining variance in Experiments 1 and 2, there is no reason to suppose that, e.g., the fact that sentences with 'silly' and 'tired' received the same rates of scalar inferences cannot be idiosyncratic. This may seem a defeatist conclusion, but it is not, for in essence it is just to say that scalar inferences are strongly context-dependent.

It must be stressed that this line of reasoning is predicated on the absence of better explanations for our data, and is therefore highly tentative. However, if it is on the right track, it invites speculation about the processing of scalar expressions along the following lines. In the psychological literature, it is generally assumed that upper-bounded interpretations of scalars must be either defaults or due to nonce inferences (e.g., Bott & Noveck 2004, Breheny et al. 2006). But if it is true that, in our experiments, participants based their judgments on statistical patterns in their previous experience with scalar expressions, another view suggests itself. For it may be the case that, inside and outside the lab, hearers rely both on statistical regularities and on honest-to-Grice implicatures, employing the former to help them gauge the prior likelihood that an alternative expression will be relevant to the speaker, and the latter to derive their scalar inferences.

Even if an alternative is readily available, the speaker need not consider it sufficiently relevant to take it into account in his utterances. The concept of relevance is notoriously slippery, and it may not always be clear to the hearer whether or not a given alternative counts as sufficiently relevant or not. Whenever such quandaries arise, past experience may be brought to bear on the issue. If this picture is correct, the reason why young children are more cautious than adults in drawing scalar inferences (e.g., Noveck 2001, Guasti et al. 2005, Pouscoulous et al. 2007) may be due, at least in part, to their more limited exposure to scalar expressions.

5

Conclusion

5.1 Summary

The information that is conveyed by an utterance comes in a variety of kinds and flavours. Information that is intimately connected to the words in the utterance is called semantic, whereas information that depends on the context in which the utterance is made is called pragmatic. I have focused on one type of information whose position in the semantics/pragmatics spectrum has been debated at length: quantity inferences. Various theories about the provenance of quantity inferences have been developed. Pragmatic accounts assume that quantity inferences are a kind of conversational implicatures, whereas conventionalist accounts take them to be semantic in nature.

In order to decide between these competing accounts, I first focused on the interpretation of scalar expressions in embedded environments. Scalar expressions in unembedded environments often license upper-bounding inferences. These scalar inferences are usually explained as a kind of quantity inferences. If scalar inferences persist when scalar expressions occur in embedded environments, this would suggest that they are semantic in nature because conversational implicatures cannot be calculated on the basis of embedded sentences. Examining the extant experimental record, I concluded that what evidence has been adduced for the existence of embedded scalar inferences is actually caused by two other meaning aspects of scalar expressions: typicality structure and truth-conditional narrowing (sections 2.2 and 2.3).

These findings imply that any adequate account of linguistic communication has to accommodate at least these three meaning aspects: typicality, pragmatic inferences, and literal meaning. In section 4.2, I investigated the relationship between these three meaning aspects. It turns out that typicality is influenced by quantity inferences and literal meaning, and literal meaning in its turn is influenced by quantity inferences. Rather being disparate, then, typicality, pragmatic inferences, and literal meaning are three distinct but tightly interwoven meaning aspects.

Unlike scalar inferences, free choice inferences, another kind of quantity inferences, do persist when the licensing sentence is embedded under a universal quantifier (section 2.4). In order to account for these universal free choice inferences, both pragmatic and conventionalist accounts have to make a number of stipulations.

The conventionalist assumption that scalar inferences are semantic in nature also has implications for their predictions about the processing of these inferences. One such prediction is that the processing cost associated with the computation of scalar inferences is incurred during sentence parsing. In section 3.2, I tested and falsified this prediction in a sentence-picture verification task. Afterwards, in section 3.3, I took a closer look at the validity of the experimental task in question by including three other kinds of quantity inference. None of these caused an increase in verification times. A possible explanation for their unique psychological profile is that only the computation of scalar inferences requires constructing alternatives by substituting expressions in the speaker's utterance by other expressions from the lexicon. In other words, the processing cost associated with quantity inferences depends on structural characteristics of the alternatives.

A blind spot in most of the current research on scalar inferences is the excessive focus on a small subset of the full range of scalar expressions. Almost all of the experimental research has been concerned with the scalar expressions 'some' and 'or'. The tacit assumption that underlies this research is that 'some' and 'or' are representative for the whole range of scalar expressions. In section 4.3, I challenged this uniformity assumption: the rates at which scalar expressions license upper-bounded inferences could hardly be more diverse. These variable rates of scalar inferences can be attributed in part to the distinctness of the scalemates. But a large part of the variation seems idiosyncratic and dependent on statistical regularities about the usage of these expressions.

These explorations have also raised a number of new issues that could not be investigated in the scope of this thesis. In the next section, I outline these as possible avenues for further research.

5.2 Further research

5.2.1 Typicality and interpretation

An important conclusion of chapter 2 is that the interpretation of scalar expressions is affected by typicality differences. For example, listeners tend to assume that 'Some A are B' describes a situation in which around 33% of A are B.

I discussed two views about the relationship between truth-conditional narrowing, pragmatic inferencing, and typicality differences. According to the first, these three meaning aspects are theoretically distinct. Only truth-conditional narrowing and pragmatic inferencing affect the communicative content of an utterance. Of these, only truth-conditional narrowing further affects the literal meaning. The second view assumes that these three meaning aspects are reflections of one underlying mechanism of alternative exclusion. The differences between these three meaning aspects are the result of differences in the salience of the alternatives. The alternatives are most salient in the case of truth-conditional narrowing and least salient in the case of typicality differences.

A logical next step is to determine which of these accounts is most adequate. There are several possible methods to decide this issue. A first method is to investigate if the differences between these meaning aspects are fundamental or rather a matter of degree. A second one is to operationalise the notion of communicative content to determine if it is possible that typicality differences enter into the communicative content of an utterance. We leave this issue for future analysis.

5.2.2 A shopping list of response variables

Research on scalar inferences has employed a variety of measures to obtain an indication that a scalar inference was computed. The following is an overview of some of the measurements that have been employed to this end:

- Proportion of participants indicating that the sentence is false in a situation in which the scalar inference is false (e.g. Noveck 2001).
- Proportion of participants who indicate that they would conclude the scalar inference (e.g., Chemla 2009, Geurts & Pouscoulous 2009a).
- Difference between the mean rating for situations in which the scalar inference is true and in which it is false (Chemla & Spector 2011, Katsos & Bishop 2010).
- Proportion of participants who fixate on a target that satisfies the scalar inference (e.g., Huang & Snedeker 2009, Grodner et al. 2010).
- Proportion of participants who change a situation in which the scalar inference is false into one in which it is true (Pouscoulous et al. 2007).
- Proportion of participants who consider it impossible that a situation in which the scalar inference is false obtains (e.g., Zondervan 2010).
- Reaction time difference between 'true' and 'false' answers verifying a sentence in situations in which the scalar inference is false (e.g., Chemla & Bott 2014).
- Proportion of participants who prefer a situation in which the scalar inference is true over one in which it is false (Clifton & Dube 2010).

We have also seen that the interpretation of scalar expressions is influenced by at least three factors: pragmatic inferences, truth-conditional narrowing, and typicality differences. I have argued that at least some of the measurements mentioned above are affected by typicality differences. In further research, it should be established what each of these response variables is measuring.

5.2.3 Disjunction

The main issues for both pragmatic and conventionalist accounts of scalar inferences are posed by disjunctive statements. First, consider the issue of free choice inferences. The following sentences are problematic for all current accounts that explain free choice inferences as a kind of quantity inference:

(1) a. Everyone can have an apple or a pear.
 b. You can have an apple or you can have a pear.

Someone who says (1a) implies that everyone can have an apple and that everyone can have a pear. In order to explain these inferences within a pragmatic account, it has to be assumed that certain interpretations can be ruled out for reasons of plausibility. A grammatical account has to postulate that free choice inferences but not scalar inferences persist when embedded under a universal quantifier. Neither solution seems particularly attractive.

Furthermore, an explanation for sentences like (1b) is not forthcoming on either account. This sentence implies the free choice inferences that you can have an apple and that you can have a pear. However, these inferences cannot be computed because the conjunction of permission sentences is of one of its alternatives. Geurts (2010, 107) remarks that ‘or’ has “a non-Boolean flavor” in these cases, but it is clear that this falls short of providing an explanatorily adequate account of free choice in sentences in which the disjunction takes wide scope.

A more adequate account of free choice inferences is therefore called for. In order to determine if an explanation in terms of quantity inferences is feasible, it has to be established if free choice inferences survive other types of embedding:

(2) Johan hopes that he may take a ham or cheese sandwich.
 ?⇝ Johan hopes that he may take a ham sandwich?
 ?⇝ Johan hopes that he may take a cheese sandwich.

(3) Johan believes that he may take a ham or cheese sandwich.
 ?⇝ Johan believes he may take a ham sandwich.
 ?⇝ Johan believes he may take a cheese sandwich.

If free choice inferences are robust in these environments, this would speak against the view that they are a kind of quantity inferences.

A descriptively adequate account of free choice inferences should ideally also explain the processing finding that the computation of free choice inferences is not time-consuming (see section 3.3), Hurford's constraint, and the behaviour of 'or' in sentences like the following:

(4) a. I must go now, or I'll be late.
    b. Has Beulah arrived? Or is that Myrtle's voice?

The development of such an account will presumably require much further theoretical and empirical work.

5.2.4 Towards an integrated pragmatic account

A final issue is to develop an integrated pragmatic account of scalar inferences. Recent developments have necessitated at least two changes to traditional pragmatic theories of scalar inferences (e.g., Gazdar 1979, Soames 1982). The first involves the notion of alternatives. It is generally agreed that someone who says (5) implies that the letter in question is not connected to all of its circles:

(5) Exactly one letter is connected to some of its circles.

There are two ways of explaining this inference within a pragmatic framework. The first is to allow for alternatives that are logically independent from the original utterance to enter into the reasoning process. The second is to assume that referential expressions in alternatives can refer anaphorically to entities that are introduced by the original utterance. The second option is discussed by Geurts (2010) who couches his explanation in terms of discourse representation theory (Kamp 1981). According to this proposal, alternatives are sentences that are interpreted in the context as it is after the interpretation of the original utterance.

The full consequences of the second analysis have yet to be investigated. For example, as discussed in section 2.1.2, this accounts makes particular predictions about the interpretation of sentences with multiple scalar expressions. In addition, Geurts' analysis also predicts that scalar inferences can occur in presupposed material. For example, someone who says (6) is predicted to convey that Fred lost some but not all of the cookies on the previous occasion. These predictions require experimental vindication.

(6) Fred lost some of the cookies again.

A second issue is to make explicit the role of probabilistic cues in the decision to compute a scalar inferences. In section 4.3, we showed that scalar expressions give rise to variable rates of scalar inferences. While some of this variation can be attributed to structural cues, a considerable part of it seems to depend on idiosyncratic statistical regularities. The role of such regularities in the derivation of scalar inferences is another aspect that needs to be worked out in more detail.

Bibliography

Aaronson, D., & Scarborough, H. S. (1976). Performance theories for sentence coding: some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(1), 56–70.

Adelson, B. (1985). Comparing natural and abstract categories: a case study from computer science. *Cognitive Science*, *9*(4), 417–430.

Aher, M. (2012). Free choice in deontic inquisitive semantics (DIS). In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz, & M. Westera (Eds.) *Logic, Language and Meaning*, vol. 7218 of *Lecture Notes in Computer Science*, (pp. 22–31). Berlin: Springer.

Anglin, J. (1976). *Word, object, and conceptual development*. New York: Norton.

Ariel, M. (2004). Most. *Language*, *80*(4), 658–706.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Asher, N. (2012). Implicatures and discourse structure. *Lingua*, *132*, 13–28.

Atlas, J., & Levinson, S. (1981). *It*-clefts, informativeness, and logical form. In P. Cole (Ed.) *Radical pragmatics*, (pp. 1–61). New York: Academic Press.

Bach, K. (1994). Conversational impliciture. *Mind & Language*, *9*(2), 124–162.

Banga, A., Heutinck, I., Berends, S. M., & Hendriks, P. (2009). Some implicatures reveal semantic differences. In B. Botma, & J. van Kampen (Eds.) *Linguistics in the Netherlands 2009*, (pp. 1–13). Amsterdam: John Benjamins.

Barker, C. (2010). Free choice permission as resource-sensitive reasoning. *Semantics and Pragmatics*, *3*(10), 1–38.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition, 11*(3), 211–227.

Barsalou, L. W. (1987). The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.) *Concepts and conceptual development: ecological and intellectual factors in categorization,* (pp. 101–140). New York: Cambridge University Press.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.) *The psychology of learning and motivation, vol. 27,* (pp. 1–64). San Diego: Academic Press.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-32. URL: `http://CRAN.R-project.org/package=lme4`.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories: a replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80*(3), 1–46.

Begg, I. (1987). Some. *Canadian Journal of Psychology, 41*(1), 62–73.

Blome-Tillmann, M. (2008). Conversational implicature and the cancellability test. *Analysis, 68*(2), 156–160.

Borge, S. (2009). Conversational implicatures and cancellability. *Acta Analytica, 24*(2), 149–154.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*(1), 123–142.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language, 51*(3), 437–457.

Braine, M. D. S., Reiser, B. J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In G. H. Bower (Ed.) *The psychology of learning and motivation,* (pp. 317–371). New York: Academic Press.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An online investigation into the role of context in generating pragmatic inferences. *Cognition, 100*(3), 434–463.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5.

Cantor, N., Smith, E. E., French, R. D., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology*, *89*(2), 181–193.

Chase, C. I. (1969). Often is where you find it. *American Psychologist*, *24*(11), 1043.

Chemla, E. (2009). Universal implicatures and free choice effects: experimental data. *Semantics and Pragmatics*, *2*(2), 1–33.

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: Disjunctions and free choice. *Cognition*, *130*(3), 380 – 396.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded implicatures. *Journal of Semantics*, *28*(3), 359–400.

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology*, *61*(11), 1741–1760.

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

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

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.) *An international handbook of natural language meaning*, (pp. 2297–2332). Berlin: Mouton de Gruyter.

Clifton, C., & Dube, C. (2010). Embedded implicatures observed: a comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics*, *3*(7), 1–13.

Colonna Dahlman, R. (2013). Conversational implicatures are still cancellable. *Acta Analytica*, *28*(3), 321–327.

Davidson, D. (1973). Radical interpretation. *Dialectica*, *27*(3-4), 314–328.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990 – present. URL: `http://corpus.byu.edu/coca/`.

de Hoop, H., & Kas, M. (1989). Sommige betekenisaspecten van enkele kwantoren, oftewel: enkele betekenisaspecten van sommige kwantoren. *Interdisciplinair Tijdschrift voor Taal & Tekstwetenschap*, *9*(1), 31–49.

de Jong, F. (1983). *Sommige* niet, andere wel: de verklaring van een raadselachtig verschil. *GLOT*, *6*(4), 229–246.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.) *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, (pp. 3299–3304). Austin, TX: Cognitive Science Society.

Degen, J., & Tanenhaus, M. K. (2014). Processing scalar implicature: a constraint-based approach. To appear in: *Cognitive Science*.

Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, *1*(2), 1–38.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel paradigm for distinguishing between what is said and what is implicated. *Language*, *88*(1), 124–154.

Erreich, A., & Valian, V. (1979). Children's internal organization of locative categories. *Child Development*, *50*(4), 1071–1077.

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

Fehr, B., Russell, J. A., & Ward, L. M. (1982). Prototypicality of emotions: a reaction time study. *Bulletin of the Psychonomic Society*, *20*(5), 253–254.

Fillmore, C. D. (1975). An alternative to checklist theories of meaning. *Proceedings of the 1st Annual Meeting of the Berkeley Linguistics Society*, *1*, 123–131.

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). Houndmills: Palgrave Macmillan.

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives. *Intercultural Pragmatics*, *5*(4), 517–540.

Frake, C. O. (1969). The ethnographic study of cognitive systems. In S. Tyler (Ed.) *Cognitive anthropology*, (pp. 28–41). New York: Holt, Rinehart and Winston.

Franke, M. (2009). *Signal to act: game theory in pragmatics*. Ph.D. thesis, University of Amsterdam.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*(1), 1–82.

Gazdar, G. (1979). *Pragmatics: implicature, presupposition, and logical form*. New York: Academic Press.

Geis, M. L., & Zwicky, A. M. (1971). On invited inferences. *Linguistic Inquiry*, *2*(4), 61–66.

Geurts, B. (1999). *Presuppositions and pronouns*. Oxford: Elsevier.

Geurts, B. (2005). Entertaining alternatives: disjunctions as modals. *Natural Language Semantics*, *13*(4), 383–410.

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

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: logic, acquisition, and processing. *Language and Cognitive Processes*, *25*(1), 130–148.

Geurts, B., & Pouscoulous, N. (2009a). Embedded implicatures?!? *Semantics and Pragmatics*, *2*(4), 1–34.

Geurts, B., & Pouscoulous, N. (2009b). Free choice for all: a response to Emmanuel Chemla. *Semantics and Pragmatics*, *2*(5), 1–10.

Geurts, B., & van Tiel, B. (2013). Embedded scalars. *Semantics and Pragmatics*, *6*(9), 1–37.

Gibbs, R. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, *34*(4), 457–486.

Gibbs, R., & Moise, J. (1997). Pragmatics and understanding what is said. *Cognition*, *62*(1), 51–74.

Goguen, J. A. (1969). The logic of inexact concepts. *Synthese*, *19*(3-4), 325–373.

Greenhall, O. (2008). Against Chierchia's computational account of scalar implicatures. *Proceedings of the Aristotelian Society*, *108*(3), 373–384.

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

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

Grice, H. P. (1978). Some further notes on logic and conversation. In P. Cole (Ed.) *Syntax and semantics, volume 9: Pragmatics,* (pp. 113–127). New York: Academic Press.

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.

Groenendijk, J., & Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*. Ph.D. thesis, University of Amsterdam.

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

Hammersly, J., & Handscomb, R. (1964). *Monte Carlo methods*. New York: John Wiley.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*(3), 355–383.

Hirschberg, J. (1991). *A theory of scalar implicature*. New York: Garland Press.

Hobbs, J., Stickel, M., Appelt, D., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*(1-2), 69–142.

Hobbs, J. R. (2004). Abduction in natural language understanding. In L. R. Horn, & G. Ward (Eds.) *The handbook of pragmatics*, (pp. 724–741). Oxford: Blackwell.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles. Distributed by Indiana University Linguistics Club.

Horn, L. R. (1989). *A natural history of negation*. Chicago: Chicago University Press.

Horn, L. R. (2004). Implicature. In L. R. Horn, & G. Ward (Eds.) *The handbook of pragmatics*, (pp. 2–28). Blackwell.

Horn, L. R. (2006). The border wars: a neo-Gricean perspective. In K. von Heusinger, & K. Turner (Eds.) *Where semantics meets pragmatics*, (pp. 21–48). Berlin: Mouton de Gruyter.

Horn, L. R. (2009). WJ-40: Implicature, truth, and meaning. *International Review of Pragmatics*, *1*(1), 3–34.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

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

Ippolito, M. (2010). Embedded implicatures? Remarks on the debate between globalist and localist theories. *Semantics and Pragmatics*, *3*(5), 1–15.

Jaszczolt, K. (2009). Cancellability and the primary/secondary meaning distinction. *Intercultural Pragmatics*, *6*(3), 259–289.

Jayez, J., & van Tiel, B. (2012). Only ‘only’? an experimental window on exclusiveness. In M. Aloni, V. Kimmelman, F. Roelofsen, G. Sassoon, K. Schulz, &

M. Westera (Eds.) *Logic, Language and Meaning*, vol. 7218 of *Lecture Notes in Computer Science*, (pp. 391–400). Berlin: Springer.

Jennings, R. E. (1994). The 'or' of free choice permission. *Topoi*, *13*(1), 3–10.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kamp, H. (1973). Free choice permission. *Proceedings of the Aristotelian Society*, *74*, 57–74.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, & M. Stokhof (Eds.) *Formal methods in the study of language*, (pp. 277–322). Amsterdam: Mathematical Centre Tracts 135.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191.

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Dordrecht: Kluwer.

Katsos, N., & Bishop, D. V. M. (2010). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, *20*(1), 67–81.

Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, *30*(6), 669–690.

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, *9*(3), 253–326.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381.

Klinedinst, N. (2007). Plurals, possibilities, and conjunctive disjunction. *UCL Working Papers in Linguistics*, *19*, 261–284.

Kratzer, A., & Shimoyama, J. (2002). Indeterminate pronouns: the view from Japanese. In Y. Otsu (Ed.) *The Proceedings of the Third Tokyo Conference on Psycholinguistics*, (pp. 1–25). Tokyo: Hituzi Syobo.

Krifka, M. (2007). Negated antonyms: creating and filling the gap. In U. Sauerland, & P. Stateva (Eds.) *Presupposition and implicature in compositional semantics*, (pp. 163–177). Houndmills: Palgrave Macmillan.

Lakoff, G. (1973). Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, *2*(4), 458–508.

Lakoff, G. (1977). On linguistic gestalts. *Chicago Linguistics Society Papers*, *13*, 236–287.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284. URL: `http://lsa.colorado.edu/`.

Landman, F. (1998). Plurals and maximalization. In S. Rothstein (Ed.) *Events and grammar*, (pp. 237–271). Basingstoke: Kluwer.

Larson, M., Doran, R., McNabb, Y., Baker, R., Berends, M., Djalili, A., & Ward, G. (2009). Distinguishing the said from the implicated using a novel experimental paradigm. In U. Sauerland, & K. Yatsushiro (Eds.) *Semantics and pragmatics: from experiment to theory*, (pp. 74–93). Berlin: Palgrave MacMillan.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Lipton, P. (2004). *Inference to the best explanation* (2nd edition). London: Routledge.

Magri, G. (2011). Another argument for embedded scalar implicatures based on oddness in downward entailing environments. *Semantics and Pragmatics*, *4*(6), 1–51.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language*, *31*(2), 195–217.

Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: the differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 54–70.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

Merin, A. (1992). Permission sentences stand in the way of Boolean and other lattice-theoretic semantices. *Journal of Semantics*, *9*(2), 95–162.

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology*, *12*(4), 496–522.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. M. E. Moravcsik, & P. Suppes (Eds.) *Approaches to natural language*, (pp. 221–242). Dordrecht: Reidel.

Moxey, L. M., & Sanford, A. J. (1993). Prior expectations and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology*, *5*(1), 73–91.

Nakagawa, S., & Schielzeth, H. (2012). A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, *18*(3), 178–182.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Noveck, I. A., Chierchia, G., Chevaux, F., Guelminger, R., & Sylvestre, E. (2002). Linguistic-pragmatic factors in interpreting disjunction. *Thinking and Reasoning*, *8*(4), 297–326.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain and Language*, *85*(2), 203–210.

Noveck, I. A., & Sperber, D. (2007). The why and how of experimental pragmatics: the case of 'scalar inferences'. In N. Burton-Roberts (Ed.) *Advances in pragmatics*, (pp. 184–212). Basingstoke: Palgrave.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *78*(3), 253–282.

Pepper, S., & Prytulak, L. S. (1974). Sometimes frequently means seldom: context effects in the interpretation of quantitative expressions. *Journal of Research in Personality*, *8*(1), 95–101.

Pijnacker, J., Hagoort, P., van Buitelaar, J., Teunisse, J.-P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*(4), 607–618.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2013). *nlme: linear and nonlinear mixed effects models*. R package version 3.1-108.

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.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

R Development Core Team (2006). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Recanati, F. (2003). Embedded implicatures. *Philosophical Perspectives*, *17*(1), 299–332.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*(1), 1–20.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*(3), 328–350.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology, 104*(3), 192–233.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.) *Cognition and categorization,* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.

Rotstein, C., & Winter, Y. (2004). Total adjectives vs. partial adjectives: scale structure and higher-order modification. *Natural Language Semantics, 12*(3), 259–288.

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics, 23*(4), 361–382.

Sadock, J. M. (1991). On testing for conversational implicature. In S. Davis (Ed.) *Pragmatics: a reader,* (pp. 365–376). Oxford: Oxford University Press.

Sauerland, U. (2004a). On embedded implicatures. *Journal of Cognitive Science, 5*, 107–137.

Sauerland, U. (2004b). Scalar implicatures in complex sentences. *Linguistics and Philosophy, 27*(3), 367–391.

Sauerland, U. (2010). Embedded implicatures and experimental constraints: a reply to Geurts & Pouscoulous and Chemla. *Semantics and Pragmatics, 3*(2), 1–13.

Sauerland, U. (2012). The computation of scalar implicatures: pragmatic, lexical or grammatical? *Language and Linguistics Compass, 6*(1), 36–49.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43*(4), 441–464.

Schulz, K. (2005). A pragmatic solution to the paradox of free choice permission. *Synthese, 147*(2), 343–377.

Sedivy, J. C., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71*(2), 109–147.

Seidenberg, M. S. (1997). Language acquisition and use: learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603.

Shaver, P., Schwarz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061–1086.

Simons, M. (2010). A Gricean view on intrusive implicatures. In K. Petrus (Ed.) *Meaning and analysis: new essays on H. Paul Grice*, (pp. 138–169). London: Palgrave Macmillan.

Simons, M. (2014). Local pragmatics and structured contents. *Philosophical Studies*, *168*(1), 21–33.

Soames, S. (1982). How presuppositions are inherited: a solution to the projection problem. *Linguistic Inquiry*, *13*(3), 483–545.

Spector, B. (2003). Scalar implicatures: exhaustivity and Gricean reasoning. In B. ten Cate (Ed.) *Proceedings of the eighth ESSLLI student session*, (pp. 277–288). Vienna.

Spector, B. (2006). *Aspects de la pragmatique des opérateurs logiques*. Ph.D. thesis, Université de Paris VII.

Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, *7*(5), 273–294.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Storto, G., & Tanenhaus, M. K. (2005). Are scalar implicatures computed online? In E. Maier, C. Bary, & J. Huitink (Eds.) *Proceedings of Sinn und Bedeutung 9*, (pp. 431–445). Nijmegen: Nijmegen Centre for Semantics.

Thomason, R. H. (1990). Accommodation, meaning, and implicature: interdisciplinary foundations for pragmatics. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.) *Intentions in communication*, (pp. 325–363). Cambridge, MA: MIT Press.

Tian, Y., Breheny, R., & van Tiel, B. (2012). Embedded implicatures: do they exist? Poster presented at *AMLaP* 2012, Riva del Garde, Italy.

van der Sandt, R. A. (1991). Denial. In *Papers from the parasession on negation*, (pp. 331–344). Chicago: Chicago Linguistic Society.

van Kuppevelt, J. (1996). Inferring from topics: scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy*, *19*(4), 393–443.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: an updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335.

van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information, 13*(4), 491–519.

van Rooij, R., & Schulz, K. (2006). Pragmatic meaning and non-monotonic reasoning: the case of exhaustive interpretation. *Linguistics and Philosophy*, *26*(2), 205–250.

van Tiel, B. (2012). Universal free choice? In A. Aguilar Guevara, A. Chernilovskaya, & R. Nouwen (Eds.) *Proceedings of Sinn und Bedeutung 16: Volume 2*, (pp. 627–638). MIT Working Papers in Linguistics.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177.

van Tiel, B., & Geurts, B. (2014). Truth and typicality in the interpretation of quantifiers. In U. Etxeberria, A. Fălăuş, A. Irurtzun, & B. Leferman (Eds.) *Proceedings of Sinn und Bedeutung 18*, (pp. 451–468). Bayonne and Vitoria-Gasteiz.

Verbeemen, T., Vanoverberghe, V., Storms, G., & Ruts, W. (2001). The role of contrast categories in natural language concepts. *Journal of Memory and Language*, *44*(4), 618–643.

Ward, G., & Birner, B. (2004). Information structure and non-canonical syntax. In L. R. Horn, & G. Ward (Eds.) *Handbook of pragmatics*, (pp. 153–174). Malden, MA: Blackwell.

Weiner, M. (2006). Are all conversational implicatures cancellable? *Analysis*, *66*(2), 127–130.

Wilson, D., & Carston, R. (2007). A unitary approach to lexical pragmatics: relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.) *Pragmatics*, (pp. 230–259). London: Palgrave Macmillan.

Wilson, N. L. (1959). Substances without substrata. *The Review of Metaphysics*, *12*(4), 521–539.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2014). Predicting pragmatic reasoning in language games. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, (pp. 3835–3840). Austin, TX: Cognitive Science Society.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*(1), 28–44.

Zevakhina, N. (2012). Strength and similarity of scalar alternatives. In

A. Aguilar Guevara, A. Chernilovskaya, & R. Nouwen (Eds.) *Proceedings of Sinn und Bedeutung 16: Volume 2*, (pp. 647–658). MIT Working Papers in Linguistics.

Zimmermann, T. E. (2000). Free choice disjunction and epistemic possibility. *Natural Language Semantics*, *8*(4), 255–290.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Ph.D. thesis, Utrecht University.

Appendices

Appendix A: Universal quantifiers

Experiment 3: List of lexical categories

Category	**Typical**	**Atypical**
	Noun	
BIRD	Robin	Ostrich
FURNITURE	Chair	Stool
WEAPON	Gun	Catapult
VEHICLE	Car	Tank
INSTRUMENT	Guitar	Bagpipes
	Adjective	
BIG	Elephant	Hippopotamus
SLOW	Snail	Turtle
SMALL	Ant	Mouse
LONG AGO	Caveman	Knight
RICH	€8.000.000	€500.000

Table 5.1: Lexical categories used in Experiment 3 and the corresponding typical and atypical instantiations that were depicted.

Appendix B: Universal free choice

Target sentences

Deontic

(1) a. Jan was allowed to take an apple or a banana.
    b. Every student was allowed to take an apple or a banana.

(2) a. Gerard was allowed to take a bottle of water or lemonade.
    b. Every sportsman was allowed to take a bottle of water or lemonade.

(3) a. Richard may take a card or roll the dice again.
b. Every player may take a card or roll the dice again.

(4) a. Ria may take a cocktail or a glass of lemonade
b. Every visitor may take a cocktail or a glass of lemonade.

(5) a. Johan may take a ham or cheese sandwich.
b. Every visitor may take a ham or cheese sandwich.

Dynamic

(1) a. Thomas can speak French or Italian.
b. Every employee can speak French or Italian.

(2) a. Hans can play the guitar or the drums.
b. Every band member can play the guitar or the drums.

(3) a. Karel can read Greek or Latin.
b. Every student can read Greek or Latin.

(4) a. Karin can write a haiku or a limerick.
b. Every pupil can write a haiku or a limerick.

(5) a. Marcel can navigate a tank or an airplane.
b. Every soldier can navigate a tank or an airplane.

Epistemic

(1) a. According to the doctor, Piet might have the flu or the measles.
b. According to the doctor, every pupil might have the flu or the measles.

(2) a. According to the gardener, the seed might be a plant or a tree next year.
b. According to the gardener, every seed might be a plant or a tree next year.

(3) a. According to the secret service, the warehouse might be a drugs laboratory or a weed plantation.
b. According to the secret service, every warehouse might be a drugs laboratory or a weed plantation.

(4) a. According to the professor, this research question might be answered by means of a survey or an experiment.
b. According to the professor, every research question might be answered by means of a survey or an experiment.

(5) a. According to the commander, the city might be a target for bomb squadrons or war ships.
b. According to the commander, every city might be a target for bomb squadrons or war ships.

Existential

(1) a. Some people took a dog or a cat out of the animal home today.
b. Every day, some people take a dog or a cat out of the animal home.

(2) a. Some children had the flu or the measles.
b. Every week, some children have the flu or the measles.

(3) a. Some guests ordered an omelette or a sandwich this morning.
b. Every morning, some guests order an omelette or a sandwich.

(4) a. The computer had some problems with the software or the hardware.
b. Every computer had some problems with the software or the hardware.

(5) a. Some players received a yellow or red card during the match.
b. Every match, some players received a yellow or red card.

Results subdivided by target sentence

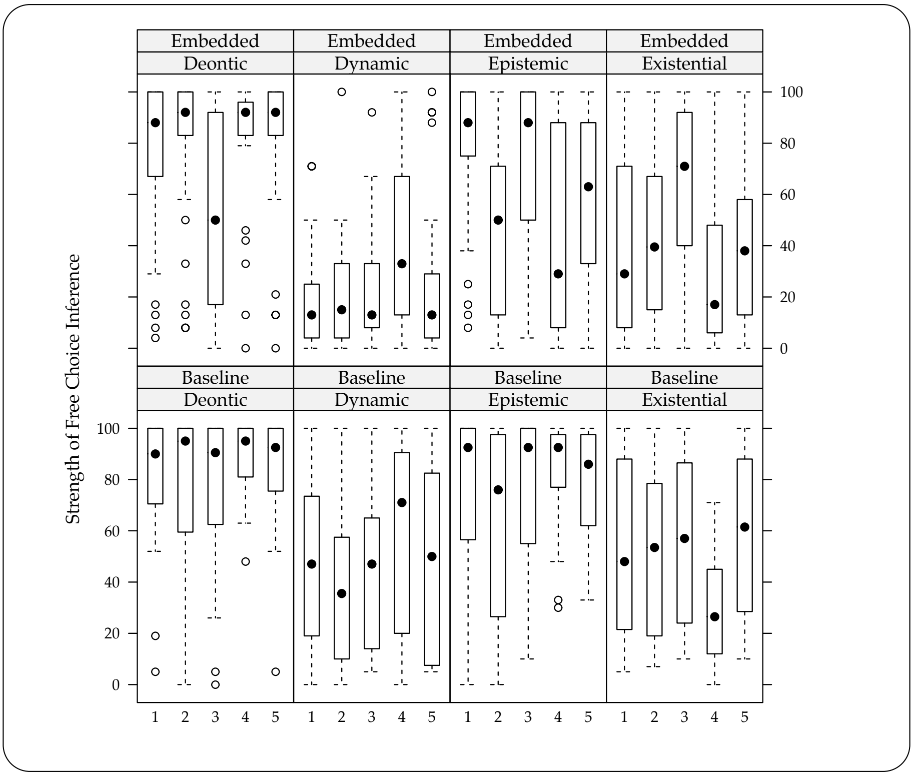


Figure 5.1: Strength of the free choice inference when 'or' is embedded under a deontic, dynamic or epistemic modal, or under an existential quantifier, further subdivided by item.

Appendix C: Scalar diversity

Notation: ‘It || The food (5) | salary (1) | solution (1) is adequate’ means that in Experiment 1 the target sentence was ‘It is adequate’, while in Experiment 2 the target sentences were ‘The food is adequate’,‘The salary is adequate’, and ‘The solution is adequate’, and that ‘food’, ‘salary’, and ‘solution’ were mentioned 5, 1, and 1 times, respectively, in the pretest where 10 participants were prompted for completions to the sentence ‘The _________ is adequate but it is not good’ (see the Materials section for Experiment 2).

Target sentences

• *adequate/good:* It || The food (5) | salary (1) | solution (1) is adequate. • *allowed/obligatory:* It || Copying (2) | Drinking (4) | Talking (2) is allowed. • *attractive/stunning:* She || That ring (1) | This model (7) | The singer (2) is attractive. • *believe/know:* She believes it. The student (1) believes it will work out (1). The mother (3) believes it will happen (1). The teacher (6) believes it is true (1). • *big/enormous:* It || That elephant (4) | The house (1) | That tree (1) is big. • *cheap/free:* It || The water (2) | electricity (1) | food (5) is cheap. • *content/happy:* She || This child (3) | The homemaker (1) | The musician (1) is content. • *cool/cold:* That || The air (1) | weather (4) | room (1) is cool. • *dark/ black:* That || That fabric (1) | The sky (3) | The shirt (1) is dark. • *difficult/impossible:* It || The task (6) | journey (1) | problem (3) is difficult. • *dislike/loathe:* He dislikes it. The boy (1) dislikes broccoli (1). The teacher (2) dislikes fighting (1). The doctor (3) dislikes coffee (1). • *few/none:* He saw few of them. The biologist (1) saw few of the birds (2). The cop (1) saw few of the children (1). The observer (1) saw few of the stars (1). • *funny/hilarious:* It || This joke (3) | The play (1) | This movie (7) is funny. • *good/excellent:* It || The food (2) | That movie (2) | This sandwich (1) is good. • *good/perfect:* It || The layout (1) | This solution (1) | That answer (1) is good. • *hard/unsolvable:* It || That problem (6) | The issue (3) | The puzzle (5) is hard. • *hungry/starving:* He || The boy (5) | dog (3) | elephant (1) is hungry. • *intelligent/ brilliant:* She || The assistant (1) | That professor (2) | This student (3) is intelligent. • *like/love:* She likes it. The princess (2) likes dancing (1). The actress (1) likes the movie (1). The manager (1) likes spaghetti (1). • *low/depleted:* It || The energy (2) | This battery (1) | The gas (5) is low. • *may/have to:* He may do it. The child (2) may eat an apple (1). The boy (3) may watch television (0). The dog (2) may sleep on the bed (1). • *may/will:* He may do it. This lawyer (1) may appear in person (0). The teacher (3) may come (2). The student (1) may pass (0). • *memorable/unforgettable:* It || This party (2) | The view (1) | This movie (3) is memorable. • *old/ancient:* It || That house (2) | mirror (1) | table (1) is old. • *palatable/delicious:* It || The food (3) | That wine (2) | The dessert (1) is palatable. • *participate/win:* She || The freshman (1) | runner (2) | skier (1) participated. • *possible/certain:* It || Happiness (1) | Failing (2) | Success (2) is possible. • *pretty/beautiful:* She || This model (5) | That lady (1) | The girl (4) is pretty. • *rare/extinct:* It || That plant (3) | This bird (2) | This fish (1) is rare. • *scarce/unavailable:* It || This recording (1) | resource (4) | mineral (2) is scarce. • *silly/ridiculous:* It || That song (3) | joke (6) | question (1) is silly. • *small/tiny:* It || The room (1) | The car (1) | This fish (2) is small. • *snug/tight:* It || The shirt (4) | That dress (2) | This glove (1) is snug. • *some/all:* He saw some of them. The bartender (1) saw some of the cars (2). The nurse (1) saw some of the signs (1). The mathematician (1) saw some of the issues (1). • *sometimes/always:* He is sometimes inside. The assistant (1) is sometimes angry (3). The director (1) is sometimes late (2). The doctor (2) is sometimes irritable (1). • *special/unique:* It || That dress (1) | That painting (1) | This necklace (1) is special. • *start/finish:* She || The athlete (1) | dancer (2) | runner (2) started. • *tired/exhausted:* He

|| The quarterback (1) | runner (1) | worker (3) is tired. *try/succeed:* He || The candidate (1) | athlete (1) | scientist (1) tried. • *ugly/hideous:* It || The wallpaper (2) | That sweater (1) | That painting (3) is ugly. • *unsettling/horrific:* It || The movie (6) | This picture (1) | The news (2) is unsettling. • *warm/hot:* That || The weather (5) | sand (1) | soup (3) is warm. • *wary/scared:* He || The dog (3) | victim (1) | rabbit (1) is wary.

Control sentences

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

Samenvatting

Inleiding

Door dingen te zeggen kunnen we informatie overbrengen. Maar niet alle overgebrachte informatie is in alle opzichten gelijk. Neem bijvoorbeeld de volgende dialoog:

A: Was het examen moeilijk?
B: De meeste studenten zijn gezakt.

Spreker B brengt met zijn antwoord onder meer de volgende informatie over:

i. Meer dan de helft van de studenten zijn gezakt.
ii. Niet alle studenten zijn gezakt.
iii. Het examen is moeilijk.

Dergelijke informatie-eenheden heten *proposities*. Propositie *i* wordt opgeroepen door de letterlijke betekenis van de woorden in B's uiting en is daarmee *semantisch* van aard. Propositie *iii* hangt juist in sterke mate af van de context waarin B's uiting plaatsvindt. Met andere woorden, iemand die zegt dat de meeste studenten zijn gezakt suggereert daarmee niet noodzakelijk dat het examen moeilijk is. Informatie die afhankelijk is van eigenschappen van de context wordt *pragmatisch* genoemd.

Hoe zit het met propositie *ii*? Dit is een zogenaamde *kwantiteitsinferentie*. Kwantiteitsinferenties ontstaan als een spreker minder informatief is dan hij had kunnen zijn. In dit geval had de spreker meer informatie kunnen geven door te zeggen: 'Alle studenten zijn gezakt'. Op basis van de observatie dat de spreker kiest voor een minder informatieve uiting concludeert de hoorder dat het informatievere alternatief volgens de spreker onwaar is. In dit geval concludeert de hoorder dus dat niet alle studenten zijn gezakt.

In deze dissertatie richt ik me op de vraag of kwantiteitsinferenties semantisch of pragmatisch van aard zijn. Semantische theorieën gaan ervanuit dat kwantiteitsinferenties een onderdeel zijn van de letterlijke betekenis van een uiting. Volgens deze theorieën is de betekenis van 'de meeste' te parafraseren als 'de meeste maar niet alle'. Pragmatische theorieën stellen daarentegen dat hoorders kwantiteitsinferenties afleiden door te redeneren over waarom de spreker niet zo informatief was als hij had kunnen zijn.

‘Ingebedde implicaturen’

Semantische en pragmatische theorieën maken op een aantal punten verschillende voorspellingen. In hoofdstuk 2 onderzoek ik de interpretatie van schaaltermen, zoals ‘de meeste’, in verschillende soorten van inbedding:

> Alle studenten had de meeste antwoorden goed.
> Twee studenten hadden de meeste antwoorden goed.
> Je haalt een voldoende als je de meeste antwoorden goed hebt.

Volgens semantische theorieën betekent ‘de meeste’ in elk van deze gevallen ‘de meeste maar niet alle’. In paragrafen 2.2 en 2.3 beargumenteer ik dat het experimentele bewijs voor dergelijke ingebedde kwantiteitsinferenties niet steekhoudend is, omdat de gebruikte experimentele taken geen kwantiteitsinferenties maar andere betekenisaspecten meten. Een van deze betekenisaspecten is *typicaliteit*. Het blijkt dat hoorders uitgesproken preferenties hebben over de interpretatie van zinnen als ‘Sommige A zijn B’. Zo vinden hoorders deze zin een betere beschrijving van een situatie waarin 40% van de A B zijn dan een situatie waarin 10% van de A B zijn, ondanks dat de zin strikt genomen waar is in beide situaties.

In paragraaf 2.4 onderzoek ik het effect van inbedding op een ander soort kwantiteitsinferenties: *free choice inferenties*. Als ik iemand vertel dat hij koffie of thee mag, dan zal hij daaruit concluderen dat hij koffie mag en dat hij thee mag. Deze free choice inferenties vormen een probleem voor het wijdverbreide idee dat ‘of’ equivalent is aan logische disjunctie. Veel recente theorieën lossen dit probleem op door free choice inferenties te verklaren als een soort van kwantiteitsinferenties. In tegenstelling tot schaalinferenties blijken ingebedde free choice inferenties niet wezenlijk te verschillen van oningebedde varianten. Deze asymmetrie dwingt zowel semantische als pragmatisch theorieën tot complicerende aannames.

De verwerking van kwantiteitsinferenties

In hoofdstuk 3 benader ik het debat tussen semantische en pragmatische theorieën over kwantiteitsinferenties vanuit het perspectief van psychologische verwerking. Uit eerder onderzoek is gebleken dat het afleiden van schaalinferenties tijd kost. Uit dit onderzoek blijkt echter niet wanneer deze vertraging in reactietijden optreedt. Semantische theorieën voorspellen dat deze vertraging optreedt tijdens het lezen van een zin. Pragmatische theorieën voorspellen dat deze vertraging juist na afloop van de zinsinterpretatie optreedt. Aan de hand van een verificatietaak laat ik zien dat de tweede voorspelling correct is: de vertraging in reactietijden in het geval van schaalinferenties wordt pas zichtbaar nadat proefpersonen een zin hebben gelezen.

Om deze bevinding op waarde te schatten is het nodig om erachter te komen welk onderdeel van de afleiding van schaalinferenties verantwoordelijk is voor de vertraagde reactietijden. Daartoe onderzoek ik in paragraaf 3.3 de psychologische verwerking van drie andere soorten kwantiteitsinferenties. Het blijkt dat alleen de afleiding van schaalinferenties een vertraging in reactietijden veroorzaakt. Schaalinferenties zijn tevens uniek omdat hoorders de benodigde alternatieven bouwen door termen in de uiting te vervangen door andere woorden uit het lexicon. Het lijkt er dus op dat deze factor verantwoordelijk is voor de vertraagde reactietijden in het geval van schaalinferenties.

Samengenomen bieden deze twee hoofdstukken bewijs voor een pragmatische theorie van kwantiteitsinferenties. Hoofdstuk 2 laat zien dat er een belangrijk verschil is in de interpretatie van schaaltermen in ingebedde en oningebedde omgevingen. Hoofdstuk 3 bewijst dat hoorders pas na zinsinterpretatie kwantiteitsinferenties afleiden. Beide bevindingen impliceren dat kwantiteitsinferenties pragmatisch van aard zijn.

Nieuwe kwesties

In hoofdstuk 4 behandel ik twee onderwerpen die buiten het bereik van de huidige theorievorming over kwantiteitsinferenties vallen. In paragraaf 4.2 onderzoek ik de relatie tussen typicaliteit en andere betekenisaspecten. In tegenstelling tot suggesties in de literatuur is het niet mogelijk op basis van typicaliteitsverschillen de interpretatie van gekwantificeerde zinnen te verklaren. Desondanks is er een nauw verband tussen typicaliteit en andere betekenisaspecten. In paragraaf 4.3 onderzoek ik de stilzwijgende aanname in de literatuur dat alle schaaltermen in zekere zin hetzelfde zijn. Op basis van een experiment laat ik zien dat deze uniformiteitsaanname onhoudbaar is. Bovendien bespreek ik een aantal factoren die de verschillen tussen schaaltermen kunnen verklaren. Daaruit blijkt dat de verschillen tussen schaalinferenties tot op zekere hoogte bepaald worden door de afstand tussen schaaluitdrukkingen. Aangezien de mate waarin ‘alle’ sterker is dan ‘sommige’ een stuk groter is dan de mate waarin bijvoorbeeld ‘afzichtelijk’ sterker is dan ‘lelijk’, is de kans dat ‘sommige’ geïnterpreteerd wordt als ‘sommige maar niet alle’ groter dan dat ‘lelijk’ geïnterpreteerd wordt als ‘lelijk maar niet afzichtelijk’.

Curriculum vitae

Bob van Tiel was born in Eindhoven, the Netherlands, on May 19, 1986. In 2004, he obtained his high school diploma from Gymnasium Beekvliet, Sint-Michielsgestel. Afterwards he went to the Radboud University Nijmegen to study Dutch Language and Literature, graduating in 2009 with a Master's thesis about the development of the linking element in Dutch nominal compounds. From 2006 onwards, he combined his studies in Dutch Language and Literature with a degree in Philosophy, which he obtained in 2010 with a Master's thesis about generalised quantifiers. In 2010, he also started his Ph.D. at the Philosophy department in the NWO-funded project "Quantity matters: Building a theory of Q-implicature", which he finished in 2014. He currently works as a research associate on the SFB project "Alignment in communication" at Bielefeld University.