Embedded scalars and typicality

Bob van Tiel

Radboud University Nijmegen

Abstract

In recent years, the interpretation of scalar terms in embedded environments has been investigated extensively. Some experimentalists have been concerned with sentences like (1), in which a scalar term is embedded under a universal quantifier. The controversy involves the question whether 'some' in these sentences is interpreted as 'some but not all', thus leading to the embedded upper-bounded inference that no square is connected to all of the circles.

(1) All the squares are connected with some of the circles.

Geurts & Pouscoulous (2009) conducted a verification task that seemed to prove that the inference is not licensed. In response, Clifton & Dube (2010) and Chemla & Spector (2011) gathered evidence suggesting the opposite conclusion. By experimentally investigating the typicality structure of complex quantified sentences like (1) above, I show how the results of the latter two experiments can be explained as typicality effects. Typicality plays a significant role in the comprehension of quantified sentences, thus complicating the interpretation of data from the kinds of experiments that have been employed to test between theories of upper-bounded inferences.

1. Introduction

A speaker who says (1) usually implies that John did not eat all of the apples. The scalar term 'some' receives an upper-bounded construal (UBC)

and thus comes to exclude 'all'.

(1) John ate some of the apples.

Traditionally, these UBCs have been explained as conversational implicatures (e.g., Horn 1972). Someone who hears (1) may reason as follows: the speaker could have made a stronger statement by saying that John ate all of the apples. Why didn't he do so? Presumably because he does not believe that John ate all of the apples. Assuming that the speaker knows whether or not John ate all of the apples, it follows that he believes that John did not eat all of the apples. Alternatively, it has been argued that sentences with a scalar term are ambiguous between a reading with and without the UBC. This ambiguity resides either in the lexical meaning of scalar terms (e.g., Levinson 2000) or in the optional presence of a syntactic exhaustivity operator (e.g., Chierchia 2006, Fox 2007).

I refer to the aforementioned theories as *pragmatic* and *conventionalist*, respectively. These two accounts make divergent predictions about the interpretation of scalar terms in embedded contexts. Some of these predictions have been subjected to experimental analysis. Part of this research has been concerned with scalar terms embedded under universal quantifiers. For example, Geurts & Pouscoulous (2009) experimentally investigated the interpretation of sentences like (2).

(2) All the squares are connected with some of the circles.

Most versions of the pragmatic account predict that (2) implies at most that not all the squares are connected to all of the circles. I refer to pragmatic accounts that make this prediction as *standard* pragmatic accounts (e.g., Horn 2006, Geurts 2010). The predicted inference is derived by negating the stronger statement 'All the squares are connected to all of the circles'. According to conventionalist accounts, as well as *nonstandard* pragmatic accounts (e.g., van Rooij 2002, Spector 2003), (2) may entail that no square is connected to all of the circles. To arrive at this embedded UBC by conventionalist means, the exhaustivity operator is attached to the embedded sentence before the universal quantifier is appended.[1]

Geurts & Pouscoulous provide evidence suggesting that listeners do not derive the embedded UBC of (2). In response, Clifton & Dube (2010) and Chemla & Spector (2011) present experimental data that seem to support the opposite claim. In the next section, I give a concise overview of these

1. These embedded UBCs are sometimes referred to as *embedded implicatures*. I avoid this term because conversational implicatures by definition cannot be embedded (cf. Cohen 1971, Walker 1975, but see Simons 2010 for a dissenting point of view).

three studies. I argue that the results presented in the latter two experiments are caused by typicality differences rather than UBCs. To this end, I experimentally investigate the typicality structure of complex quantified sentences like (2). I conclude with a note on the role of typicality in theoretical and experimental pragmatics.

Chemla & Spector have provided another argument against both standard and nonstandard pragmatic accounts, based on the interpretation of scalar terms in non-monotonic environments. These data cannot straightforwardly be explained by means of typicality. Therefore they fall outside the purview of the present article. My aim is thus not to vindicate a pragmatic account of UBCs but to dispel one of the counterarguments launched against its standard version by investigating the effect of typicality on the interpretation of complex quantified sentences. However, Geurts & van Tiel (2012) argue that the interpretation of scalar terms in non-monotonic environments can be accommodated within a standard pragmatic framework as well.

2. *Experiments*

2.1. Geurts & Pouscoulous

The discussion about sentences like (3) has centered on the following question: does this sentence convey the embedded UBC that no square is connected to all of the circles? To answer this question, Geurts & Pouscoulous (2009) presented participants with this sentence in a situation where the embedded UBC was false, as depicted in Figure 1.[2]

(3) All the squares are connected with some of the circles.

Since the topmost square is connected to all of the circles, this situation falsifies the embedded UBC. Participants in Geurts & Pouscoulous' experiment were asked to indicate if the sentence was true or false in this situation. If the embedded UBC is part of the communicative content of (3), it is expected that at least some of the participants that compute it will consider the sentence false.

This way of detecting UBCs has been validated in numerous experiments with sentences containing a scalar term in an unembedded position. Take, for example, (4).

2. Figure 1 and 2, as well as the figures in the appendix, originally appeared in color. To see these pictures in color, the reader is referred to the electronic version of this paper.

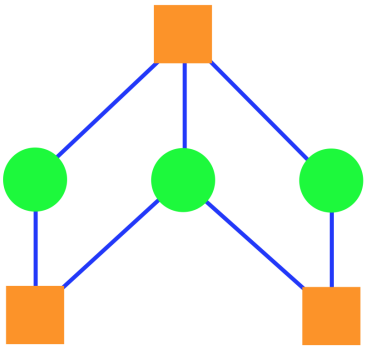


Figure 1: Experimental item used in the verification task by Geurts & Pouscoulous.

(4) Some elephants have trunks.

This sentence usually conveys the UBC that not all elephants have trunks, which is manifestly false. So if the UBC is part of the communicative content of (4), at least some of the participants that compute it should judge the sentence to be false. Indeed, it is found quite consistently that roughly 60% of the participants consider sentences like (4) false (e.g., Noveck 2001, Bott & Noveck 2004, Feeney et al. 2004, Pouscoulous et al. 2007, Banga et al. 2009, Zondervan 2010). This suggests that at least 60% of them compute the UBC associated with (4).

In stark contrast to these findings, Geurts & Pouscoulous found that not a single participant considered (3) false in a situation that falsified the embedded UBC. This led them to conclude that participants do not compute the embedded UBC of (3). If this conclusion is correct, it invalidates an important argument against a standard pragmatic account of UBCs.

However, the suitability of Geurts & Pouscoulous' experimental task has been put into question by Clifton & Dube (2010) and Chemla & Spector (2011). Both pairs of authors propose several adjustments to the verification task to facilitate the detection of embedded UBCs. One such adjustment concerns the role of contrast. In Geurts & Pouscoulous' experiment, any form of contrast between different situations was avoided. The target sentence (3) occurred only once in the experiment and was presented with only one target situation. In addition, the critical item was hidden among an ample amount of filler items. In the experiments by Clifton & Dube and Chemla & Spector, participants saw a sentence like (3) with several target situations, instead of just one. Participants had to indicate how well these situations were described by the sentence, instead of 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 give a more detailed overview of the experiments by Clifton & Dube and Chemla & Spector.

2.2. Clifton & Dube

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. Participants had to indicate which of these situations was best described by the sentence. Additionally, they could choose the options 'both' and 'neither'. The inclusion of these two options is somewhat puzzling. Participants might choose 'both' if they cannot distinguish between the situations. But what about 'neither'? The response options intuitively suggest that participants have to indicate whether the situations are adequately described by the sentence, rather than which situation is best. I will come back to this interpretative ambiguity in section 9.

Clifton & Dube first ran this task using sentences with 'some' in an unembedded position, like (5). As indicated, these sentences were presented with two target situations. In the first situation, the box on the left contained all of the stars. In the second situation, the box on the left contained some but not all of the stars. Only the latter situation agreed with the UBC associated with (5).

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

When asked which of these situations was best described by the sentence, participants chose the situation where the UBC was true most frequently (71%). The next most popular choice was the option 'both' (24%). The situation falsifying the UBC and the option 'neither' were practically never chosen (3% and 2%).

Next, Clifton & Dube ran the same task using sentences with 'some' embedded under a universal quantifier, such as (6).

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

Again, these sentences were presented with two target situations. In the critical condition, the first situation contained one square connected to all of the circles and two squares connected to some but not all of the circles (this picture was similar to the one in Figure 1). In the second situation, three squares were connected to some but not all of the circles. These two situations correspond to options A and B in Figure 2. The embedded UBC

is true only in the latter situation.

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

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

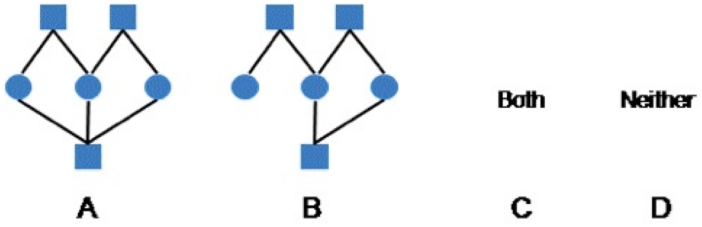


Figure 2: Target item used in the experiment by Clifton & Dube.

In this condition, participants picked out the option 'both' most frequently (60%). The target situation that verified the embedded UBC was chosen in 38% of the cases. Again, the other two options were practically never chosen (2% and 0%). According to Clifton & Dube, these results show that 38% of the participants computed the embedded UBC.

2.3. Chemla & Spector

Chemla & Spector conducted two experiments. Their second experiment investigated the interpretation of scalar terms in non-monotonic contexts. As noted in the introduction, the results of that experiment fall outside the scope of my argument. In their first experiment, Chemla & Spector presented (7) in a wide variety of situations.

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

On each target trial, participants saw (7) with one of these situations. Participants had to indicate how 'true', 'appropriate', or 'suitable' the sentence was in the situation at hand. To this end, they could set a line anywhere on a continuous bar. The further to the right the line was set, the more suitable the sentence was judged to be.

Every situation that was paired with (7) consisted of six letters. 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. By varying the distribution of these cases, Chemla & Spector constructed seven situations. Four of these are depicted in Figure 3. The distribution of different letters in each of the situations is given in Table 1. The table also shows how suitable participants judged the sentence on average in the respective situations.

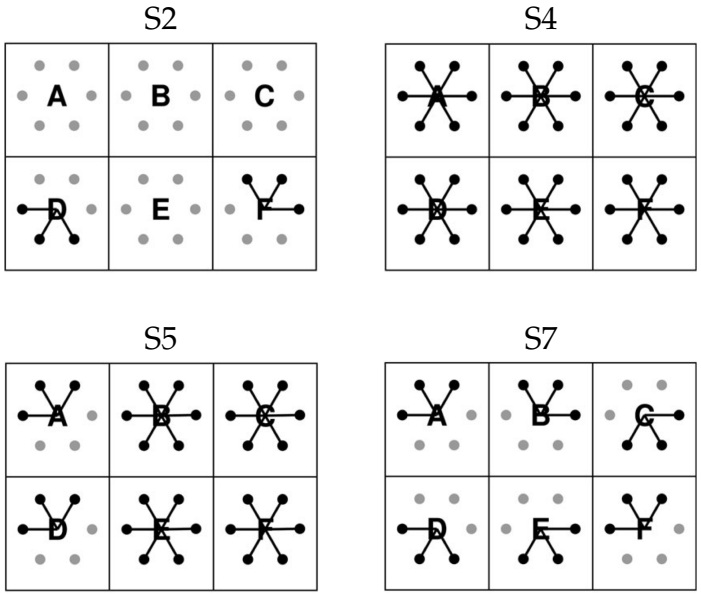


Figure 3: Four of the seven target situations used in Chemla & Spector's first experiment.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
None	6	4	2	0	0	0	0
Mixed	0	2	4	0	2	4	6
All	0	0	0	6	4	2	0
μ	0	12	24	44	63	73	99

Table 1: Mean suitability ratings for the target situations in Chemla & Spector's first experiment.

As can be seen in the table, participants considered the sentence less suitable in a situation with None cases than in a situation with All cases. In general, the more Mixed cases were added to the situation, the more suitable the sentence was judged to be. The degree of fit was the highest in a situation with six Mixed cases. This was the only situation that verified the embedded UBC. Chemla & Spector took this preference to show that their participants computed the embedded UBC.

Like Clifton & Dube, then, Chemla & Spector found that if a scalar term is embedded under a universal quantifier, participants prefer the sentence to describe a target situation that agrees with the embedded UBC over a situation that does not. In the following sections, I discuss the cause of this preference.

3. Interpretation

A substantial number of participants in Clifton & Dube's and Chemla & Spector's experiments preferred sentences like (8) to describe a situation that agreed with the embedded UBC over a situation that did not. This led both pairs of authors to conclude that these participants apparently computed the embedded UBC.

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

This conclusion immediately raises the question why participants did not consider the sentence false in a situation that directly and unambiguously contradicted the embedded UBC, as demonstrated by Geurts & Pouscoulous. After all, if 'some' occurs in an unembedded context, a substantial number of participants consider a sentence with a false UBC to be altogether false. Additional assumptions about the truth-conditional status of embedded UBCs have to be made in order to make sense of Geurts & Pouscoulous' results (cf. Sauerland 2010, whose explanation is discussed and criticized in Geurts & van Tiel 2012).

Chemla & Spector's interpretation faces an additional difficulty. Participants in their experiment considered (8) on average more suitable in a situation that agreed with the embedded UBC than in a situation that did not. The explanation offered by Chemla & Spector is that this suitability difference is caused by the computation of an embedded UBC. This diagnosis is problematic for the following reason: Chemla & Spector concentrate their attention on the difference between the S6 and S7 condition, but there are significant differences between all seven target situations. Many of these suitability differences cannot be attributed to the computation of a UBC. For

instance, participants preferred (8) to describe the S6 over the S5 situation. This difference is manifestly not caused by a UBC, as Chemla & Spector themselves concede. According to them, some suitability differences are caused by a UBC and some differences are caused by another factor. Chemla & Spector propose that the latter differences are typicality effects. To justify this dual mechanism explanation, they point at the different magnitudes of the suitability differences between the S5, S6 and S7 situations. The difference between the S6 and S7 situation (26%) is more than twice as big as the difference between the S5 and S6 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 S6 and S7 situation but still do not correspond to a UBC. An example is the suitability difference between the S1 and S3 condition, which amounts to 24%. Chemla & Spector concede that this difference, despite its size, is not caused by a UBC but is a typicality effect. So their argument in favour of a dual mechanism explanation is wanting. This poses a serious problem for the UBC 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 4.2 below), participants in Geurts & Pouscoulous' verification task did not consider a sentence false in a situation where its embedded UBC was false. My 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 context, participants most often (71% of the cases) picked out the picture where the UBC was true. For sentences with 'some' embedded under a universal quantifier, the option 'both' was the most popular choice (60% of the cases). This asymmetry will turn out to be compatible with the view that their results are influenced by typicality differences.

4. Typicality

4.1. What is typicality?

It is a well-established fact that not all members of a category are equally typical (or representative). Take, for instance, the category BIRD. Robins,

ducks and penguins are all members of this category. But most people tend to agree that a robin is more typical of the category BIRD than a duck, which in its turn is usually considered more typical than a penguin.

Typicality differences are defined operationally on the basis of participants' direct judgements (Kahneman & Tversky 1972, Mervis & Rosch 1981). For example, 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 7-point Likert scale. This resulted in robust typicality orderings.

Typicality structure has been demonstrated in a broad range of categories. This encompasses categories denoted by nouns (Hampton & Gardiner 1983), adjectives (Lakoff 1973), locatives (Erreich & Valian 1979), quantifiers (Begg 1987), 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 but a few examples, more typical members are learned earlier (Rosch 1973, Anglin 1976, Rosch et al. 1976, Mervis & Pani 1980), recognized faster and more accurately as a member of the category (Rips et al. 1973, Rosch & Mervis 1975, Rosch et al. 1976, Armstrong et al. 1983, Sedivy et al. 1999), and produced earlier and more often when participants are asked to name members of the category (Battig & Montague 1969, Barsalou 1983, 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 see a category name with several instances and have to indicate how well these are described by the category name. In the latter, participants see a sentence with several situations and have to indicate how well these are described by the sentence. In fact, Chemla & Spector themselves propose that their results are partly due to typicality differences. I propose to simplify this hypothesis: all results found by Clifton & Dube and Chemla & Spector are 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 1), as well as the asymmetry between embedded and unembedded 'some' noted by Clifton & Dube. But before addressing these issues, I discuss the role of typicality in communication.

4.2. Typicality and communication

Even though a robin is a more typical instance of the category BIRD than a duck, both are unequivocally birds. There is thus an important distinction between typicality and category membership (cf. Fuhrmann 1991, Kamp & Partee 1995, Osherson & Smith 1997). Even though the notions are related (cf. McCloskey & Glucksberg 1978, Hampton 2007), there is a broad consensus that they are not identical (for extensive discussion, see Margolis & Laurence 1999).

The distinction between typicality and category membership is critical for understanding the role of typicality in communication. A speaker who uses a category name denotes a category member but not necessarily a typical one. That is, (9a) conveys that Fred sees a member of the category BIRD but not necessarily a robin or a sparrow. By analogy, (9b) does not imply that Sue always eats peas at dinner, that being the most typical instance of the category VEGETABLE.

(9) a. Fred sees a bird.
b. Sue always eats vegetables at dinner.

This is not to deny that typicality differences affect interpretation. For instance, it is well established that participants first think of typical cases when asked to list members of a category. But this effect is not intended by the speaker, as the examples in (9) show. By contrast, UBCs are intended by the speaker, from both the pragmatic and conventionalist perspective. In the former case, UBCs are conversational implicatures. In the latter case, UBCs are part of the propositional content.

In the following, 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 'some' and 'every'.

5. The typicality structure of EVERY

In the following discussion, I assume that the typicality structure of a category C is a fuzzy set (e.g., Zadeh 1965, 1973, Goguen 1969, Lakoff 1973). In particular, the typicality structure of C is a function $\rho_C : D \to (0, 1]$ that associates with every element $d \in D$ a number $\rho_C(d)$ in the left-open interval $(0, 1]$.[3] $\rho_C(d)$ signifies the typicality of d in C. If C is of type $\langle a, t \rangle$

3. The range of typicality values is left-open, which entails that the value 0 is not an element in the range of typicality values. This complication is introduced for a technical reason. As

then all $d \in D$ are of type a. To give an example, if y is a robin and z is an ostrich, then $\rho_{\text{BIRD}}(y) > \rho_{\text{BIRD}}(z)$.

What is the typicality structure of sentences like (10)? I take sentences to denote sets of situations. That is, 'It rains' denotes the category $\lambda S(it\ rains\ in\ S)$. In the following, I omit the lambdas and simply refer to the denotation of a sentence by using small caps.

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

A classical idea in fuzzy predicate logic is that $\rho_{\text{EVERY A B}}(S)$ equals the minimum of typicality values of instances of A with respect to the predicate B (e.g., Zadeh 1965, Priest 2001).

(11) $\rho_{\text{EVERY A B}}(S) = min\{\rho_B(a)|a \in A\}$

This definition is a straightforward generalization of the standard interpretation of universally quantified sentences. 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).

6. Experiment 1

6.1. Participants

30 participants were recruited on Amazon's Mechanical Turk. All participants were paid for their participation. Only U.S. residents were eligible for participation. Participants were asked to submit their native language but payment was not contingent on their response to this question. Only native speakers of English were included in the analysis. One response was excluded from analysis because the same participant had already filled in the experiment once.

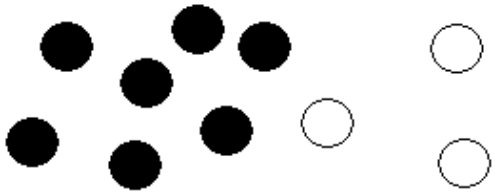
6.2. Materials and design

On each trial, participants saw sentence (12) with a picture (cf. Figure 4). Participants were instructed to indicate how well the sentence described the picture, by marking a value on a 7-point Likert scale. The full instructions are given in the appendix. No fillers were included in the task.

---

will become apparent in the next section, 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, which is obviously impossible if these typicality values sum to 0.

Every circle is black.



How well is the picture described by the sentence?

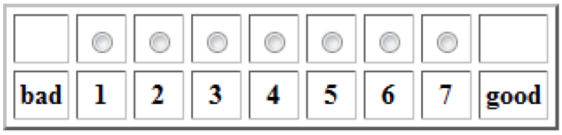


Figure 4: Experimental item from the rating experiment for EVERY.

(12) Every circle is black.

The target sentence was presented in situations consisting of ten circles. There were eleven situations altogether. These situations had anywhere between zero and ten black circles. Every participant encountered every situation once. Ten randomized lists of items were prepared.

6.3. Results and discussion

The results of the EVERY experiment are summarized in Figure 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 they contained. The typicality scale for EVERY was highly reliable (Cronbach's $\alpha = 0.94$).

The minimum typicality definition in (11) fits these data poorly, 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 for a situation depends on the typicality value of all of the individual instances with respect to the category BLACK CIRCLE, not just on the value of the least typical instance. One way of modelling this would be to take the mean of the typicality values. However, the data show that bad instances (i.e., white circles) exert more influence on the mean rating of a situation than good instances (i.e., black circles), since adding one white circle to a situation with only black circles leads to a significant drop in the mean rating, while adding one black circle to a situation with only white

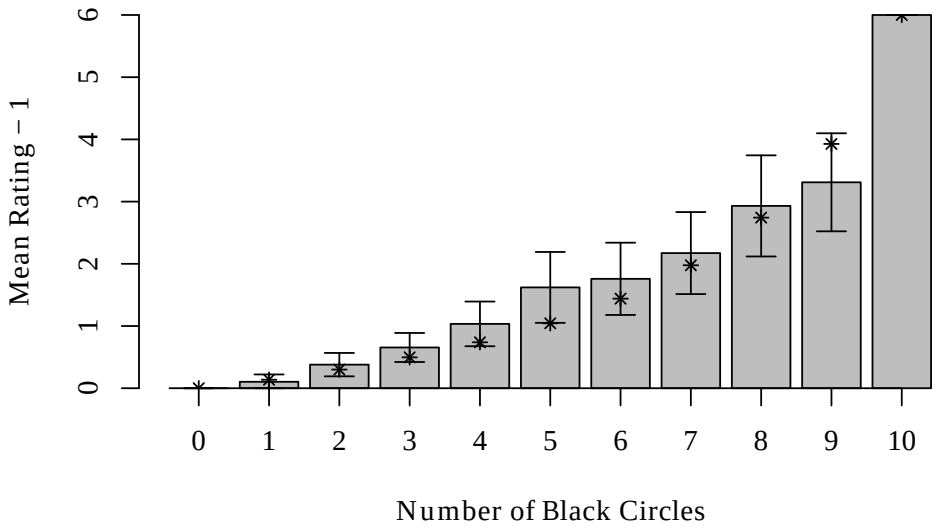


Figure 5: Mean typicality rating for the sentence 'Every circle is black' in situations with ten circles. The error bars represent 95% confidence intervals. The stars represent the scaled typicality ratings predicted by the definition in (13).

circles has barely any effect. 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 typicality values. This can be done, for instance, by weighing the typicality values by their reciprocal, which yields the harmonic mean of the typicality values.

(13) $\rho_{\text{EVERY A B}}(S) = \dfrac{n}{\Sigma\rho_B(a_i)^{-1}}$, where $A = \{a_i, \ldots, a_n\}$

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 circles with respect to the category BLACK CIRCLE, 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 $\rho_C(black) > \rho_C(white)$. For instance, if $\rho_C(black) = .95$ and $\rho_C(white) = .1$, $r = .97$, $p < .001$. Therefore I adopt this definition in the following.

Let us now return to sentences like (14).

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

The typicality of a situation with respect to (14) equals the harmonic mean typicality of the relevant cases with respect to the category CONNECTED TO SOME OF ITS CIRCLES. In the experiments by Clifton & Dube and Chemla & Spector, a case (i.e., a square or a letter) was connected either 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.

7. *The typicality structure of* SOME

The typicality structure of SOME has been investigated experimentally by Begg (1987). He found that "the preferred meaning of *some* is 'less than half'" (p. 62). Similar findings were obtained by Borges & Sawyers (1974) and Newstead et al. (1987). These findings make theoretical sense. It is well known that the typicality of an object in a category correlates negatively with the degree to which that object represents contrasting categories (cf. Frake 1969, Rosch & Mervis 1975, Rosch 1978, Cantor et al. 1980, Niedenthal & Cantor 1984). Take the category MAMMAL, which contrasts with the categories BIRD and FISH. A bat is a poor instance of the category MAMMAL because it has many properties in common with the category BIRD. Similarly, a whale is a poor instance because it shares many properties with the category FISH.

In a small-scale experiment, I investigated what the contrasting categories for SOME are. Twenty-two Dutch participants were asked to read the following instructions, which are the standard way of determining contrasting categories (e.g., Rosch & Mervis 1975, Malt & Johnson 1992, Markman & Wisniewski 1997, Verbeemen et al. 2001).

> A friend of yours returns from a tennis tournament. During the tournament, he played several matches. He asks you to guess how many matches he won. The word below [= 'some'] is your first answer. This answer turns out to be wrong. Think about what a good second guess would be and write down that answer.

The contrasting categories for SOME turned out to be NONE, ALL (both mentioned seven times) and MOST (mentioned three times). The results of the multidimensional scaling undertaken by Routh (1994) point in the same direction. A set containing less than half of the total number of elements is unlike the sets denoted by these contrasting categories. This supports Begg's conclusion that the most representative instance of SOME is a set containing less than half of the total number of elements.

Nonetheless, it could be argued that Begg's results are caused by the derivation of a UBC. 'Some' forms a Horn scale with 'most' and 'all'. So participants might have interpreted 'some' as 'some but not most' or 'some but not all'. This could be responsible for Begg's experimental results. To address this concern, it was necessary to obtain more fine-grained data. To this end, I conducted another rating task modelled after Rosch (1975). Participants had to rate (15) in situations with ten circles, anywhere between zero and ten of them being black.

(15) Some of the circles are black.

What results are expected if participants base their results only on whether a UBC is violated? In the scenario of the aforementioned experiment, the following clusters are expected to arise (where each n refers to the corresponding situation with n black circles):

(16) $\{2,3,4,5\} > \{6,7,8,9\} > \{10\} > \{0,1\}$

I assume that the sentence is unambiguously false in a situation with only one black circle because of the plural morpheme in 'circles'. Importantly, the within-cluster variation should be smaller than the between-cluster variation. If there is a large amount of within-cluster variation, possibly paired with a small degree of between-cluster variation, the UBC interpretation loses much of its appeal. In the next section, I report the results of a rating task conducted to test these predictions.

8. Experiment 2

8.1. Participants

30 participants were recruited on Amazon's Mechanical Turk. All participants were paid for their participation. The selection procedure was the same as for Experiment 1. No participants were excluded from the analysis.

8.2. Materials and design

The materials and design of Experiment 2 were the same as for Experiment 1. But in this case, the target sentence was (17).

(17) Some of the circles are black.

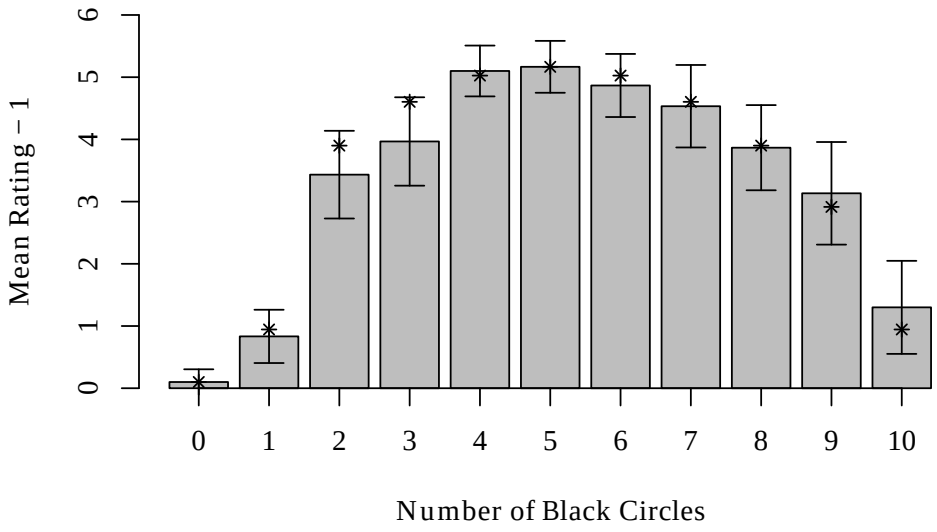


Figure 6: Mean typicality rating for the sentence 'Some of the circles are black' in situations with ten circles. The error bars represent 95% confidence intervals. The stars represent the scaled typicality ratings predicted by the definition in (19).

8.3. Results and discussion

The results for the SOME experiment are summarized in Figure 6. The situation with five black circles received the highest mean rating. The mean ratings of the other situations decreased with their distance from this prototype. The typicality scale for SOME was highly reliable (Cronbach's $\alpha = 0.8$).

Similar results were found by Degen & Tanenhaus (2011), who asked how natural (18) was in situations where the hearer received zero, one, two, five, seven, eight, eleven, twelve, or thirteen out of a total of thirteen gumballs.

(18) You got some of the gumballs.

Degen & Tanenhaus' task differed from the present experiment 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 differences, their results are in accordance with the data presented here.

The results that were found cannot fully be explained in terms of UBCs.

On such an interpretation, 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, as 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. However, the former situations received a rating that was statistically indistinguishable from the ratings for the latter.

The mean rating for a situation decreased exponentially with its distance from the prototypical situation. In our case, the prototypical situation contained five black circles, but what a prototypical situation is may vary across contexts, depending on factors like total set size (Newstead et al. 1987) and subitizability (Degen & Tanenhaus 2011). The choice of prototype may even differ between participants, which could 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 (17) in that situation. Situations where the sentence was true received a substantially higher rating than situations where it was false (i.e., the situations with zero or one black circle). The following definition captures these factors, where $v_S(\text{'some A B'})$ denotes the truth value of 'some A B' in S, and Z is a normalizing factor to ensure that the typicality values occur in the interval $(0, 1]$ (e.g., Lesot et al. 2005).

$$\rho_{\text{SOME A B}}(S) = 1 - \frac{1}{Z}(dist(S - P)^2 + v_S(\text{'some A B'})) \tag{19}$$

Depending on what value is associated with the distance between adjacent situations, this model fits almost perfectly to the data (for instance, assuming that the distance between any two adjacent situations equals 1, $r = .95, p < .001$).

The models for SOME and EVERY are only proposed because they provide a good fit to the mean ratings found in the experiments. The model for SOME is less fine-grained than the one for EVERY, which calculated 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 where some but not all of the circles are black is more typical of (17) than a situation where all of the circles are black, which in turn is more typical than a situation where none of the circles are black. There is another, more fundamental distinction 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 (19) the distance between a situation and the prototype is squared. This difference might be due to the experiment for EVERY being concerned mostly with situations where the target sentence is false and the experiment for SOME involving mostly situations where it is true. For instance, participants might have more pronounced typicality judgements when comparing genuine members of a category than when comparing non-members.

The typicality analysis outlined in the foregoing does not preclude UBCs from having an effect alongside typicality. But, as typicality is defined operationally by means of participants' direct judgements, the conclusion remains that a Mixed case (i.e., a letter or square that is connected to some but not all of the circles) is the most typical instance of CONNECTED TO SOME OF THE CIRCLES, followed by an All case (a letter or square connected to all of the circles) and a None case (a letter or square connected to none of its circles), in that order.

9. Typicality structure in embedded scalars

Having analyzed the typicality structures of SOME and EVERY, I now turn to the problem of embedded scalars. I first deal with the experiment of Chemla & Spector. These authors presented participants with (20) in a wide range of situations (cf. Table 1).

(20) Every letter is connected with some of its circles.

Each 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 with $A = \{l_1, \ldots, l_6\}$ equals:

(21) $$\rho_{\text{EVERY A B}}(S) = \frac{6}{\Sigma \rho_B(l_i)^{-1}}$$

Based on the typicality definition of SOME, it follows that:

(22) $$\rho_B(Mixed) > \rho_B(All) > \rho_B(None)$$

I ran a Monte Carlo simulation to gauge the typicality values of the seven situations based on this constraint on the typicality values of the three cases. In general, Monte Carlo investigations simulate actual conditions

that contain some random element (e.g., Hammersly & Handscomb 1964). Using R's programming interface (R Development Core Team 2006), I randomly generated 5,000 values for each of the three cases such that every triplet obeyed the constraint in (22). For each triplet that was generated, I calculated the typicality values for the seven situations based on the definition in (21). Ultimately, I derived the mean 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 ($r = 0.995, p < .001$). This demonstrates that Chemla & Spector's results can almost entirely be explained as typicality effects.

A typicality explanation accounts for several aspects of Chemla & Spector's results which are not accounted for by an explanation in terms of UBCs. As discussed before, Chemla & Spector found suitability differences that uncontroversially do not correspond to a UBC. In particular, sentence (20) was judged suitable to a significantly different degree in the S1, S2 and S3 situations, as well as in the S5 and S6 situations. For this reason, they had to adopt a dual mechanism explanation. These differences come out as entirely natural on a typicality explanation. Importantly, these differences are governed by the same mechanism that determines the difference between, for instance, the S6 and S7 situation. Typicality thus provides a uniform explanation for the whole scale of suitability differences found by Chemla & Spector.

What about Clifton & Dube's results? Participants in their experiment read sentences containing 'some' in embedded and unembedded positions. Each sentence was paired with two situations. In the embedded 'some' condition, one situation contained three Mixed cases, thus agreeing with the embedded UBC, while the other situation contained two Mixed cases and one All case. In the unembedded 'some' condition, one situation contained a Mixed case, thus agreeing with the corresponding UBC, while the other situation contained an All case. Participants were instructed to choose which situation was best described by the sentence. Alternatively, they could choose the options 'both' and 'neither'. In the next section, I provide experimental evidence that the inclusion of these two options led to an interpretative ambiguity in Clifton & Dube's task. On the one hand, participants were instructed to indicate which situation was best described by the sentence. On the other hand, the presence of the options 'both' and 'neither' suggested that participants had to indicate whether the situations were adequately described by the sentence. Consequently, participants could adopt either a suitability response pattern or an adequacy response pattern.

When 'some' occurred in an unembedded context, participants most often picked out the situation where the UBC was true. For sentences with 'some' embedded under a universal quantifier, the option 'both' was the most popular choice. This asymmetry is caused by the aforementioned interpretative ambiguity in Clifton & Dube's task. Participants who adopted the suitability response pattern chose the situation that agreed with the corresponding UBC in both conditions. After all, since Mixed cases are more typical than All cases, it follows that a Mixed case, or a situation with only Mixed cases, is more typical than an All case, or a situation with Mixed as well as All cases. By contrast, participants who adopted the adequacy response pattern always chose 'both' in the embedded 'some' condition. These participants also chose 'both' in the unembedded 'some' condition *unless* they computed the corresponding UBC. In that case they opted for the situation that agreed with the UBC. After all, for these participants the UBC was part of the communicative content of the sentence. This led to an elevated preference for the situation that agreed with the UBC in the unembedded 'some' condition compared to the embedded 'some' condition.

This account makes two clear predictions about Clifton & Dube's task. First, that it is sensitive to differential typicality. Second, that the same response pattern arises in the case of sentences containing a category name with a standard typicality structure. That is, when presented with (23) alongside a typical (e.g., a robin) and atypical (e.g., an ostrich) instance of the category BIRD, most participants should answer 'both'. A smaller but significant number of participants should opt for the typical instance.

(23) This is a bird.

I tested these hypotheses in an experiment.

10. Experiment 3

10.1. Pretest

In order to select suitable lexical categories, a pretest was conducted. For this pretest, 10 Dutch participants (5 males and 5 females, average age 22) were asked to fill 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 clear typicality structure. Ten such sentences were constructed. In five sentences, the category was denoted by an adjective, in five sentences by a noun. These sentences were paired with a picture of an atypical category member. A list of the items used is provided in the appendix. Two example sentences are given in (24). These ten target sentences were supplemented with 18 control items.

(24) a. This is a bird.
b. This animal is small.

In 89% of the cases, participants indicated that the sentence was true as a description of the atypical category member. For no individual item, there were 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 members of the respective categories. So these items were included in the eventual experiment.

10.2. Participants

26 Dutch participants (12 males and 14 females, average age 22) were asked to fill 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, possibly due to inattentiveness.

10.3. Materials and design

Each page of the questionnaire showed a sentence with two pictures (cf. Figure 7). Participants had to indicate which of the two pictures was best described by the sentence. Additionally, they could choose the options 'both' and 'neither'. The set-up of the experiment was thus the same as in Clifton & Dube's experiment. The full instructions are provided in the appendix.

Three kinds of sentences were constructed. First, the aforementioned ten sentences with an adjective or a noun that referred to a category with a clear typicality structure. Second, five sentences with the scalar term 'some' in an unembedded position. Third, five sentences with 'some' embedded under a universal quantifier. The latter two conditions correspond to the sentences used in Clifton & Dube's task, and were included to replicate their results in a different language, and to be able to compare the results for these two conditions directly to the results for the first condition. Example sentences from the embedded and unembedded 'some' conditions are given in (25).

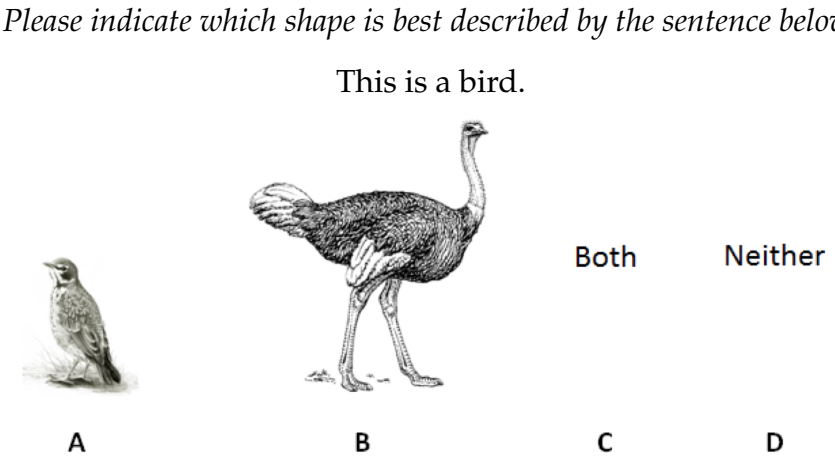


Figure 7: Experimental item from the lexical category condition.

(25) a. Every square is connected with some circles.
b. Some of the stars are in the box to the left.

The sentences with 'some' in an unembedded position were presented with a picture where the UBC was true and one where it was false. The sentences with 'some' embedded under a universal quantifier were presented with a picture where the embedded UBC was true and one where only the UBC predicted by standard pragmatic accounts was true. Lastly, the sentences with a lexical category were presented with a picture of a typical category member and a picture of an atypical category member. The appendix provides an overview of the lexical categories and their corresponding category members. The former picture in each condition is referred to as 'option A' and the latter as 'option B'. Four lists were created, randomizing 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. An example is provided in the appendix.

10.4. Results and discussion

The results of the experiment are presented in Table 2. A multinomial regression analysis with the embedded 'some' condition as reference category showed that the distribution of answers for this condition did not differ significantly from the distribution of answers for lexical categories ($b = .157$, Wald $\chi^2(1) = 1.43, p = .232$) whereas it differed significantly from the distribution of answers for unembedded 'some' ($b = .787$, Wald

$\chi^2(1) = 27.6, p < .001$).

Condition	Option A	Option B	'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'	38	2	60	0

Table 2: Percentage of choices for each of the four options in the three conditions of Experiment 3 alongside the results found by Clifton & Dube.

First of all, these results show that Clifton & Dube's task is sensitive to typicality differences. Second and more importantly, the distribution of answers for the lexical category condition shows that there is indeed an interpretative ambiguity in Clifton & Dube's task. Quite clearly, a robin is a better instance of the category BIRD than an ostrich. Participants who adopted the suitability response pattern therefore chose option A. However, most participants opted for 'both'. These participants followed an adequacy response pattern, indicating whether the pictures were adequately described by the sentence. As outlined in the previous section, this interpretative ambiguity caused the asymmetry between the results for embedded and unembedded 'some'.[4]

11. Conclusion

In section 3, I discussed three observations that must be explained by any adequate account of the experimental data on the interpretation of scalar

4. 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 UBC. While this picture was chosen significantly less frequently than the picture that agreed with the UBC ($F(1, 20) = 7.7, p < .005$), it was chosen far more often than the 2% found by Clifton & Dube. Consequently, the preference for the picture that agreed with the embedded UBC was less pronounced in this experiment than in Clifton & Dube's. A possible explanation is that I used 'some' instead of the partitive 'some of the'. As noted by Banga et al. (2009) and Degen and Tanenhaus (2011), use of the partitive leads to a slightly elevated rate of UBC computation.

terms embedded under a universal quantifier:

i. Participants do not consider the target sentence false in a situation where the embedded UBC is false.
ii. Chemla & Spector found differences in appropriateness between situations that manifestly do not correspond to a UBC.
iii. Clifton & Dube found markedly different results for 'some' in embedded and in unembedded positions.

A typicality explanation, as outlined in the previous sections, accounts for each of these observations. This indicates that there is an ambiguity in the interpretation of results from the experiments used by Clifton & Dube and Chemla & Spector, which can be explained either in terms of typicality or in terms of UBCs. More generally, it turns out that results from their experimental paradigms are strongly influenced by typicality, even when it concerns the comprehension of complex quantified sentences.

The confounding effect of typicality complicates the interpretation of sentences with a scalar expression, like (26).

(26) John ate some of the apples.

At least three mechanisms may be at work in these cases:

i. As outlined in the introduction, a UBC may be derived by means of pragmatic reasoning. In the case of (26), the speaker could have made the stronger statement that John ate all of the apples. Why didn't he do so? Presumably because he does not believe that John ate all of the apples. This UBC does not enter into the truth-conditional content of an utterance, since it is defeasible.
ii. If 'some' is sufficiently emphasized, by prosodic or contextual cues, a UBC can enter into the truth-conditional content of an utterance. In that case, 'some' means 'some but not all'. Geurts (2009) calls this process *truth-conditional narrowing*. Conventionalists and pragmaticists seem to agree that both narrowing and pragmatic reasoning play a role in sentence processing (e.g., Horn 2006, Fox 2007, Geurts 2010).
iii. The present paper demonstrates that the typicality structure of scalar terms affects the evaluation of sentences like (26) in different tasks. Typicality differences thus influence the interpretation not only of predicates like 'bird' and 'red', but also of quantifiers like 'some' and 'every'. While the typicality structure in many cases is not part of the intended meaning of an utterance, it affects a wide range of dependent variables used in psycholinguistic research (see section 4). In particular, a sen-

tence is considered more suitable in typical than in atypical situations. Some of these typicality differences are similar to, but distinct from, UBCs.

At the theoretical level, these three factors may be distinguished as follows: only conversational implicatures and truth-conditional narrowing are intended by the speaker, thus determining the communicative content of an utterance (cf. Armstrong et al. 1983, Kamp & Partee 1995). Of these two, only narrowing further affects the truth conditions of an utterance (cf. Geurts 2010). But these theoretical distinctions might not straightforwardly correspond to listeners' actual perception of what the intended meaning of an utterance is. Further research is thus required to decide whether the proposed theoretical boundaries between these meaning aspects are psychologically real. If so, an important task for experimental pragmatics is to distinguish between these (and possibly even more) factors empirically by employing suitable experimental paradigms.

Author's address

Bob van Tiel
Department of Philosophy
Radboud University Nijmegen
P.O. Box 9103
6500 HD Nijmegen
The Netherlands
E-mail: bobvantiel@gmail.com

Acknowledgements

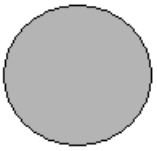
Thanks to Emmanuel Chemla, Chuck Clifton, Chad Dube, Michael Franke, Bart Geurts, Anneke Neijt, Benjamin Spector, Sammie Tarenskeen, Natalia Zevakhina, and the editor and reviewers of Journal of Semantics for their comments on the argument developed in this paper. This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO), which is gratefully acknowledged.

Appendix

Experiment 1–2: Instructions

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.

EXAMPLE: This circle is dark.



How well is the picture described by the sentence?

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 ten 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. Good luck.

Experiment 3: Instructions

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

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

This is a pet.

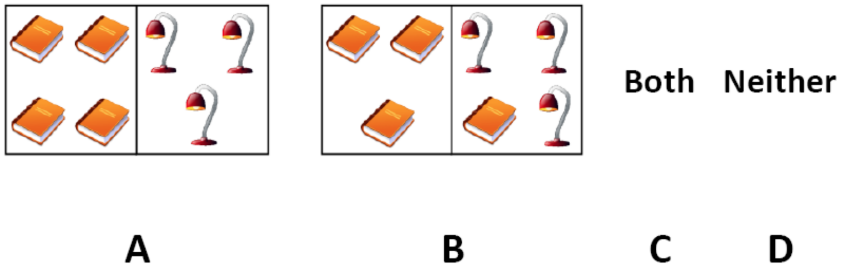
The question is the same for all multiple-choice questions. The sentence and the pictures differ. You simply have to indicate which picture best describes the sentence. To do so, circle one of the four possible answers. Note that the question is not whether you prefer a parrot or a dog, but which animal you think 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 first intuition.

Experiment 3: Example of a control item

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

There are exactly three books in the box on the left.



Experiment 3: List of lexical categories

Sentence	**Typical**	**Atypical**
	Noun	
'This is a bird'	Robin	Ostrich
'This is a piece of furniture'	Chair	Stool
'This is a weapon'	Gun	Catapult
'This is a vehicle'	Car	Tank
'This is a music instrument'	Guitar	Bagpipes
	Adjective	
'This animal is big'	Elephant	Hippopotamus
'This animal is slow'	Snail	Turtle
'This animal is small'	Ant	Mouse
'This person lived a long time ago'	Caveman	Knight
'The person having this sum is rich'	€8.000.000	€500.000

References

Adelson, B. (1985). Comparing natural and abstract categories: a case study from computer science. *Cognitive Science*, *9*, 417–430.

Anglin, J. (1976). *Word, object, and conceptual development*. New York: Norton.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308.

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.

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, *11*, 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.

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.

Borges, M., & Sawyers, B. (1974). Common verbal quantifiers: usage and interpretation. *Journal of Experimental Psychology*, *102*, 335–338.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language*, *51*, 437–457.

Cantor, N., Smith, E. E., French, R. D., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology*, *89*(2), 181–193.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded implicatures. *Journal of Semantics*, *28*(3), 359–400.

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

Clifton, C., & Dube, C. (2010). Embedded implicatures observed: a comment on Geurts and Pouscoulous (2009). *Semantics & Pragmatics*, *3*(7), 1–13.

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

Degen, J., & 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.

Erreich, A., & Valian, V. (1979). Children's internal organization of locative categories. *Child Development*, *50*, 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*, 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.

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.

Fuhrmann, G. (1991). Note on the integration of prototype theory and fuzzy-set theory. *Synthese*, *86*, 1–27.

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., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics & Pragmatics*, *2*(4), 1–34.

Geurts, B., & van Tiel, B. (2012). Scalar expressions under embedding. To appear in: *Semantics & Pragmatics*.

Goguen, J. A. (1969). The logic of inexact concepts. *Synthese*, *19*, 325–373.

Hammersly, J., & Handscomb, R. (1964). *Monte Carlo methods*. New York: John Wiley.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, *31*, 355–383.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: a correlational analysis of normative data. *British Journal of Psychology*, *74*, 491–516.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, UCLA. Distributed by Indiana University Linguistics Club.

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.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*, 129–191.

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.

Lesot, M.-J., Mouillet, L., & Bouchon-Meunier, B. (2005). Fuzzy prototypes based on typicality degrees. In B. Reusch (Ed.) *Computational Intelligence, Theory and Applications*, vol. 33 of *Advances in Soft Computing*, (pp. 125–138). Berlin: Springer.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language*, *31*, 195–217.

Margolis, E., & Laurence, S. (1999). Concepts and cognitive science. In E. Margolis, & S. Laurence (Eds.) *Concepts: core readings*, (pp. 3–81). Cambridge, MA: MIT Press.

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*, 54–70.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Mervis, C. L., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology*, *12*, 496–522.

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.

Niedenthal, P. M., & Cantor, N. (1984). Making use of social prototypes: from fuzzy concepts to firm decisions. *Fuzzy Sets and Systems*, *14*, 5–27.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*, 165–188.

Osherson, D., & Smith, E. E. (1997). On typicality and vagueness. *Cognition*, *64*, 189–206.

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.

Priest, G. (2001). *Introduction to non-classical logic*. Cambridge: Cambridge University Press.

R Development Core Team (2006). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

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*, 1–20.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 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*, 573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Routh, D. A. (1994). On representations of quantifiers. *Journal of Semantics*, *11*, 199–214.

Sauerland, U. (2010). Embedded implicatures and experimental constraints: a reply to Geurts & Pouscoulous and Chemla. *Semantics & Pragmatics*, *3*(2), 1–13.

Sedivy, J. C., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–147.

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*, 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.

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.

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*, 289–335.

van Rooij, R. (2002). Relevance only. In J. Bos, M. Foster, & C. Matheson (Eds.) *Proceedings of the sixth workshop on the semantics and pragmatics of discourse (EDILOG 2002)*, (pp. 155–160).

Verbeemen, T., Vanoverberghe, V., Storms, G., & Ruts, W. (2001). The role of contrast categories in natural language concepts. *Journal of Memory and Language*, *44*, 618–643.

Walker, R. C. (1975). Conversational implicatures. In S. W. Blackburn (Ed.) *Meaning, reference and necessity: new studies in semantics*, (pp. 133–181). Cambridge: Cambridge University Press.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 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.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Ph.D. thesis, Utrecht University.