Summary of pilot studies on questions and scalar items

Alexandre Cremers

Klinedinst and Rothschild (2011) suggested that the strong and intermediate readings of sentences containing embedded questions are obtained with the same EXH-operator as in grammatical theories of scalar implicatures (Fox, 2007; Chierchia et al., 2008). More precisely, in their theory the intermediate exhaustive reading corresponds to global (matrix) exhaustification, while the strong exhaustive reading corresponds to local (embedded) exhaustification.

The main idea of this series of experiments was to test a natural hypothesis: if embedded questions and scalar items involve the same strengthening mechanism, we may expect correlations between the rates of strong readings people access for the two types of sentences. For instance, people who easily derive local implicatures with scalar items like 'some' may be more likely to derive strongly exhaustive readings for embedded questions.

1 Experiment 1: 'surprise' vs. 'know'

1.1 Goal

In upward entailing contexts, local exhaustification gives rise to the strongest readings, which entail the readings with global exhaustification. The latter in turn entail the readings without any exhaustification. Therefore, we may observe a correlation simply due to the fact that some participants are charitable and tend to give more *True* responses across the board, while some others are very demanding and tend to respond *False* as soon as a sentence is not clearly true.

To avoid this confound, we decided to use both upward and downward entailing contexts. Concretely, this was achieved with the embedding verbs 'know' and 'surprise', respectively.

1.2 Material

The experiment consisted of a truth-value judgment task. On each trial, participants were presented with a picture which contained a card made of 4 colored squares (red, green or blue) and showed how a character, Mary, remembered the card. Participants were asked to judge a sentence such as "Mary knows which squares are red". In some cases, Mary was unsure about some of the 4 squares and this was represented with a grey square and a question mark (see Figure 1).

The sentences were built by combining one of the two embedding environments in (1) with one of the five complements in (2):

(1) **Know:** Mary knows...

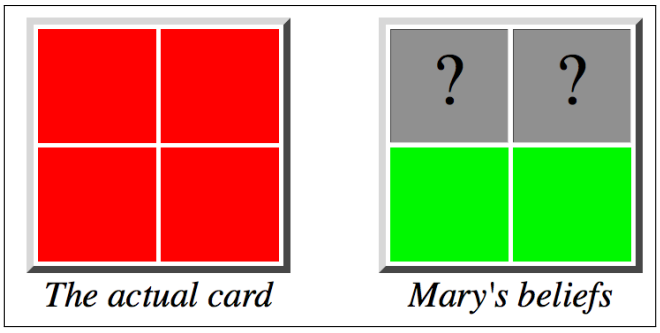

Figure 1: Example picture from Experiment 1

(2)
Surp:	It will surprise Mary...
Wh:	...which of the four squares are [blue].
WhN:	...which of the four squares are not [blue]
S:	...that some of the four squares are [blue]
A:	...that all of the four squares are [blue]
AN:	...that not all of the four squares are [blue]

Most of the resulting sentences were ambiguous between 2 or 3 different readings. For simplicity, we will describe the readings as *literal, global* and *local* with respect to the place of exhaustification as assumed in Klinedinst and Rothschild (2011).[1]

(3) **[Know-Wh]:** "Mary knows which squares are blue"
   a. Literal (=WE): For each blue square, she knows it is blue.
   b. Global (=IE): For each blue square, she knows it is blue and she does not have false beliefs about non-blue squares.
   c. Local (=SE): For each square she knows whether it is blue or not.

(4) **[Surp-Wh]:** "It will surprise Mary which squares are blue"
   a. Literal (=WE): She expects at least some of the actual blue squares not to be blue.
   b. Local (=SE): She expects some of the actual blue squares not to be blue or some non-blue squares to be blue.

(5) **[Know-S]:** "Mary knows that some of the squares are blue"
   a. Literal: [All squares are blue] and Mary knows that some or even all of them are.

[1]Note that in the case of scalar items, it is not obvious which readings are derived and what the presuppositions of the sentences are. These cases have rarely been studied and Spector and Sudo (2014) show that prosody may play a role as well. We follow the simple following rule: for the local exhaustification, there is no interaction. For the global exhaustification, we follow the suggestions of Spector and Sudo (2014) and assume a default prosody. For the literal case, we try to satisfy the presuppositions predicted for the global case. Furthermore, 'know' may give rise to anti-presupposition due to the competition with 'believe' (Percus, 2006; Sauerland, 2008).

b. Global: [All squares are blue] and Mary only knows that some of them are.

c. Local: [Some but not all squares are blue] and Mary knows exactly that.

(6) **[Know-AN]:** "Mary knows that not all of the squares are blue"

a. Literal: [Some or all squares are non-blue] and Mary knows that some or even all of them are.

b. Local: [Some but not all squares are blue] and Mary knows exactly that.

(7) **[Surp-S]:** "It will surprise Mary that some of the squares are blue"

a. Literal: [Some or all squares are blue] and Mary expects no square to be blue.

b. Local: [Some but not all squares are blue] and Mary expects that all of them are or that none of them are.

(8) **[Surp-A]:** "It will surprise Mary that all of the squares are blue"

a. Literal: [All squares are blue] and Mary expects some or all squares to be non-blue.

b. Global: [All squares are blue] and Mary expects only some squares to be blue.

(9) **[Surp-AN]:** "It will surprise Mary that not all of the squares are blue"

a. Literal: [Some or all squares are non-blue] and Mary expects all squares to be blue.

b. Local: [Some but not all squares are blue] and Mary expects all or none of them to be.

For each sentence, we built one unambiguously true and one unambiguously false control. In addition to these, we tried to build situations that would allow us to distinguish between each possible reading. Table 1 presents the list of combinations we included in the experiment.

1.3 Participants

43 participants (19 female, 24 male, all native speaker of English) were recruited on Amazon's Mechanical Turk. 8 of them were removed because their error rate on control sentences was above 30%. Other participants made at most 12.5%.

1.4 Results

1.4.1 Controls

A logit mixed model on responses to controls showed that sentence with 'surprise' gave rise to slightly more errors than 'know', that sentences with 'all' gave rise to less errors and sentences with 'not all' slightly more than average. However, all differences were under 1% and performance was excellent in all conditions (above 95%).

1.4.2 Targets

Figure 2 presents the detailed results from this experiment. Without going into details, we would like to highlight the following points:

Environment	Complement	*literal*	*global*	*local*
Mary knows...	that some squares are blue	1	0	#
		1	1*	0
		1	1	0
	that not all squares are blue	1	0	0
		1	1	0
	which squares are blue	1	0	0
		1	1	0
	which squares are not blue	1	0	0
		1	1	0
I will surprise Mary...	that some squares are blue	0/#	0/#	1
	that all squares are blue	1	0	0
	that not all squares are blue	0	0	1
	which squares are blue	0	0	1
	which squares are not blue	0	0	1

Table 1: List of situations tested for each experimental sentence in addition to the true and false controls. '#' indicates that the corresponding reading was a presupposition failure.

- Scalar item mostly gave rise to literal readings (without implicatures).
- 'not all' gave rise to more local implicatures than 'some' (about 25% vs less than 10%).
- Both 'which' and 'which..not' gave rise to surprisingly high rates of *True* responses in the '100' condition where only the weakly exhaustive reading is meant to be true.
- 'which..not' also gave rise to a high rate of strongly exhaustive answers, both with 'surprise' and 'know' (*True* responses to the '110' condition of 'know', *False* responses to the '001' condition of 'surprise'). 'which' did so with 'surprise' as well.

We further measured all possible correlations between participants 'True' responses to all conditions. Once we corrected for multiple comparisons, 9 out of the 528 correlations turned out significant (corrected $p' < .05$, i.e. $p < 9.5 \times 10^{-5}$).

- 4 of these 9 correlations were anti-correlations between one true and one false control.
- Interestingly, the 3 correlations between symmetric 'which' and 'which..not' ambiguous conditions were among the 9 (namely: '100' and '110' conditions for 'know', '001' condition for 'surprise').
- The remaining 2 correlations were positive correlations between the '110-know' conditions of 'which' and 'some' on the one hand and 'which..not' and 'not all' on the other hand.

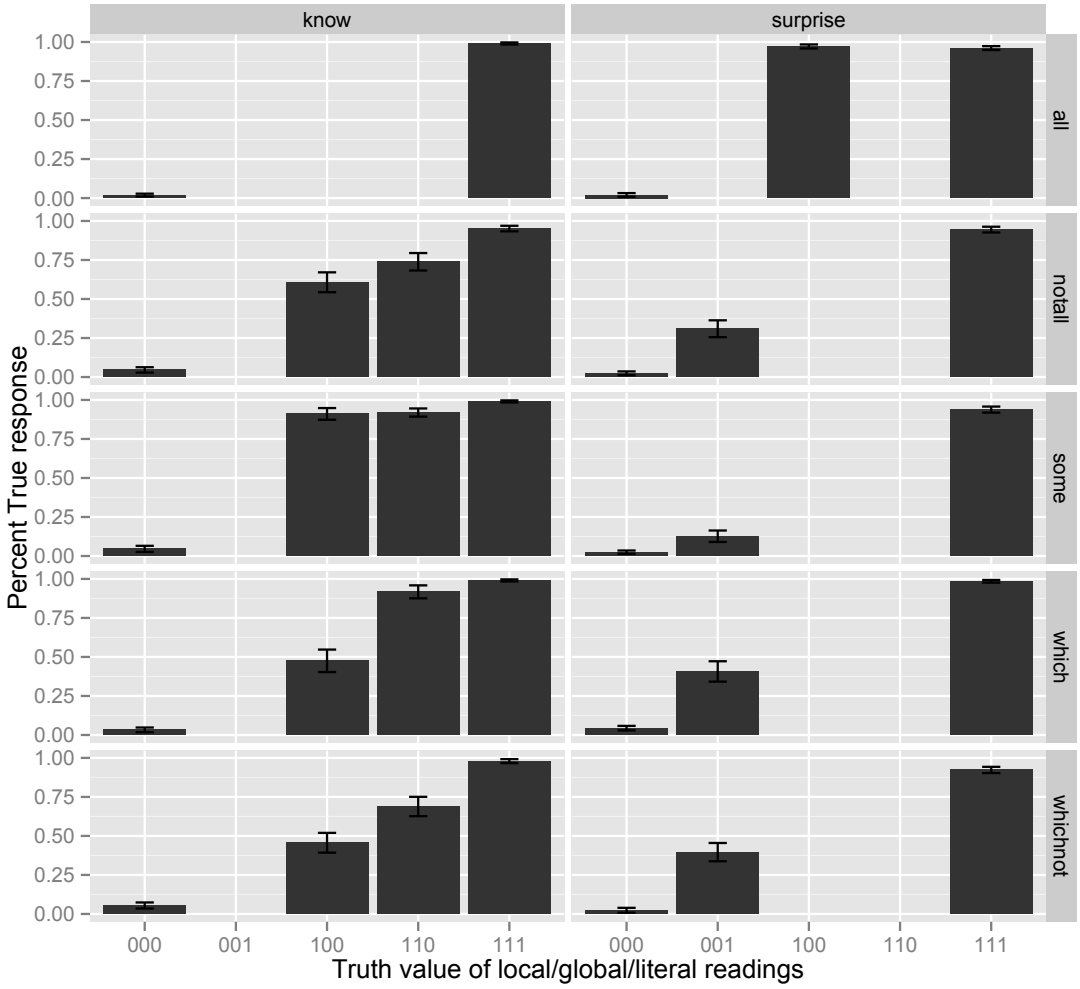

Figure 2: Percent 'True' responses to the different items in Experiment 1. For simplicity, presupposition failure was aggregated with falsity.

1.5 Discussion

Overall, this experiment confirms the results of previous experiments showing that (a) embedded scalar items can sometimes give rise to embedded implicatures and (b) questions embedded under 'know' are ambiguous between a variety of exhaustive readings.

It also extends these results to new embedding environments. Interestingly, a downward entailing environment like the complement of 'surprise' was shown to give rise to both embedded implicatures and 'strongly exhaustive' readings. The existence of both readings had been disputed for a similar reason: they are usually assumed to be derived from a simpler reading but are weaker than this simpler reading. The possibility to weaken a given reading is highly debated.

Finally, we observed correlations between responses revealing a local implicatures of

'some' and a strongly exhaustive reading for an embedded affirmative question and a local implicature for 'not all' and a strongly exhaustive reading for a negative question, respectively. However, we did not observe any correlation between environments. This suggests that all the correlations we observed are due to *True* or *False* biases in some participants, rather than biases toward a consistent locus of exhaustification.

Finally, 'some' may be a positive polarity item (PPI). This would mean that it must move out the scope of 'surprise', giving rise to yet different readings.

2 Experiment 2: Trying to get more variety with 'some'

2.1 Goal

In the next experiments we decided to focus on upward monotonic environments and tried to find a design in which we could observe a greater variety of readings (i.e. less literal readings for 'some'). We simplified the items with 'some' by avoiding embedding under a factive verb (to avoid the complex interactions between presuppositions and implicatures) and we used 'believe' as a replacement. This simplification came at the price of a less minimal comparison between scalar items and questions.

We kept a similar task but decided to create a more realistic situation. We did so by using real playing cards instead of the color squares and we elaborated on a little story.

2.2 Material

Participants read the following story:

> Peter is trying to get better at a kind of poker game in which each player gets dealt a hand of 4 cards. Peter has a very bad memory so he keeps forgetting the cards in his hand.
> You will see the hands that Peter got in each round, and what he remembers after a first look. He often mistakes the suits (for instance he remembers a 3 of clubs instead of a 3 of spades), and sometimes completely forgets some cards.
> You will have to judge whether sentences about Peter's memory are true or false.

To make the situations more plausible, the errors were always a confusion between two red cards (a heart and a diamond) or two black cards (a club and a spade).

The experiment contained two types of target sentences:

(10) Peter knows which of his cards are [spades].

(11) Peter believes that some his cards are [spades].

Each of these sentence could appear with a picture that made it unambiguously true or unambiguously false. They also appeared in a context which made the literal (WE) reading true but other readings false and a context which made the literal (WE) and global (IE) readings true but the local (SE) reading false. The anti-presupposition of 'believe' were always verified (i.e. there was no spades in the actual cards for all situation illustrating sentence (11))

Each of these conditions was repeated 4 times, yielding 32 items of interest. We also included 12 fillers which role was to emphasize the alternatives relevant for the exhaustification ('believe' as an alternative to 'know' and 'all' as an alternative to 'some').

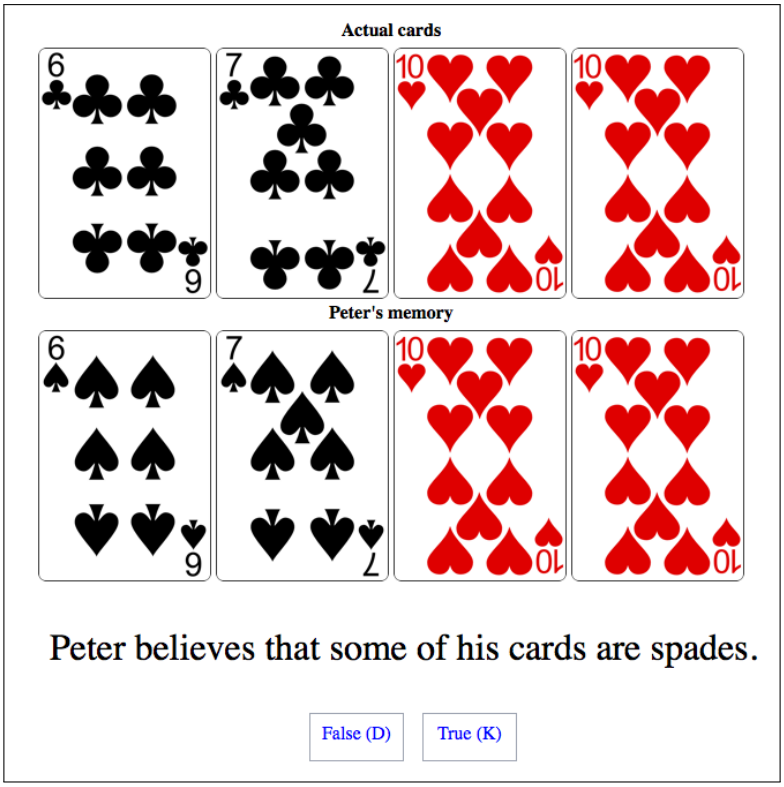

Figure 3: A true control from Experiment 2.

2.3 Participants

20 participants were recruited on Mechanical Turk (all native speakers of English). 2 of them were removed from the analyses because their error rates exceeded the mean error rate by one standard deviation (threshold: 16%).

2.4 Results

Figure 4 displays the percentage of *True* responses on targets. As one can see, the rate of literal responses to scalar items was still very high and we could not detect any embedded implicatures. Although there may have been some embedded implicature readings, they were hidden by the low rates of *True* responses on the true baseline (these are most likely errors since this item was rather difficult to judge).

2.5 Discussion

In this pilot, we did not manage to achieve our goal, which was to obtain a more balanced distribution of the different readings for sentences with an embedded scalar item. Possible

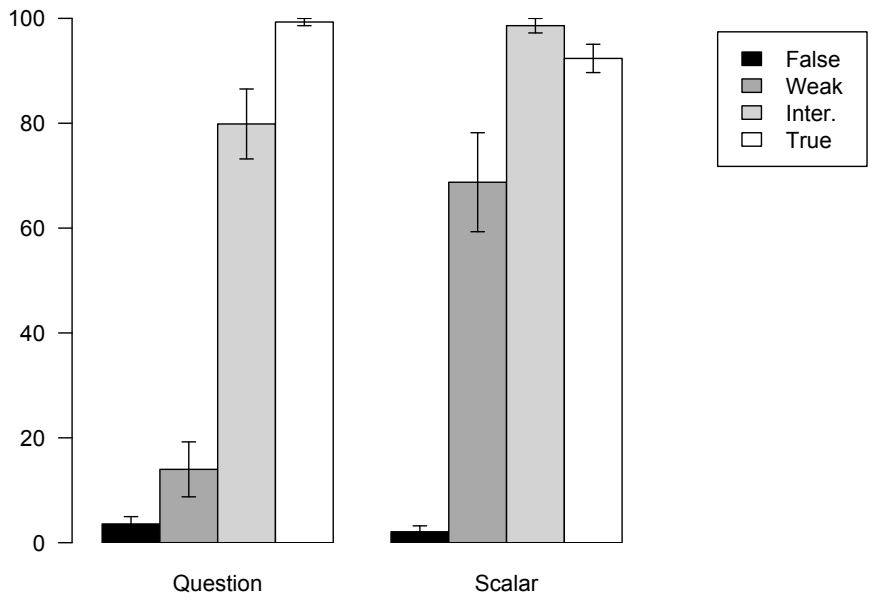

Figure 4: Percent 'True' responses to the different items in Experiment 2.

issues are the low relevance of the alternatives with 'all' and a baseline for *True* responses which turned out to be too difficult.

3 Experiment 3: Increasing the number of cards and investigating the role of focus

3.1 Goal

In order to make the contrast between a subset and the whole set of cards, we increased the number of cards per hand to 7. We also compared a version which was identical to the previous experiment with a version where the scalar item 'some' was written in upper case (thus suggesting a prosody with focus on the scalar item). We also improved our baseline for *True* responses

3.2 Materials

The test items were built as in the previous experiment, but we increased the number of repetition to 8 (yielding 64 items of interest) and increased the number of fillers to 64 as well (including some 'which not' items, as in Experiment 1). The experiment also began with 4 easy training items.

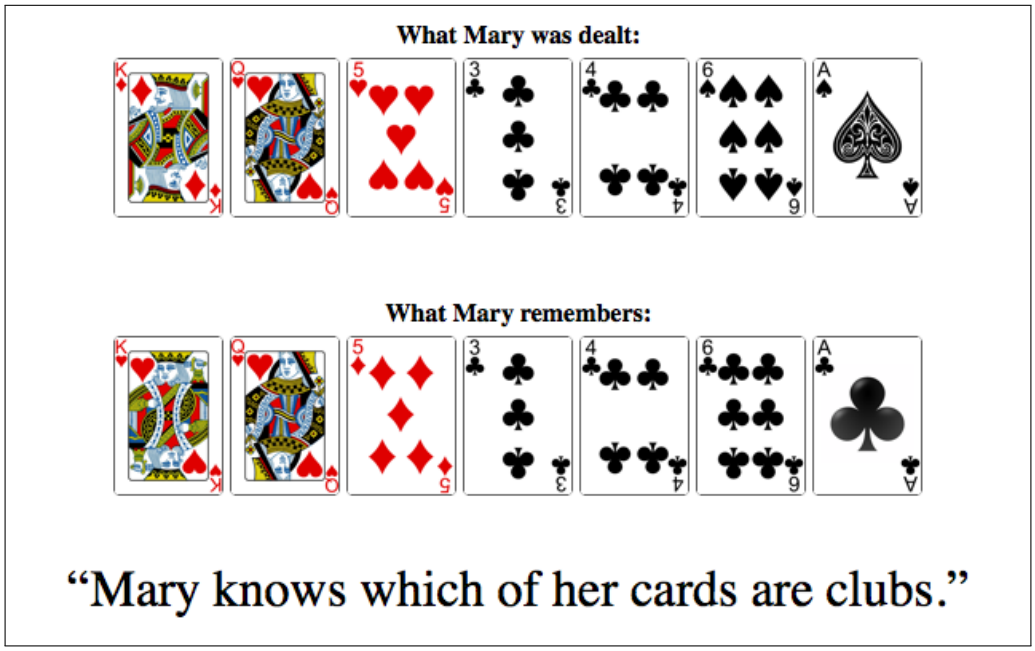

Figure 5: A WE test item from Experiment 3.

3.3 Participants

61 participants were recruited (all native speakers of English). 31 took the simple version (*NoFocus*) and 30 took the version with upper case for the scalar item (*Focus*). 5 participants were removed because their error rates exceeded the mean + one standard deviation threshold (14%, but all 5 were above 20% anyway).

3.4 Results

The version without focus on 'some' did not greatly differ from previous experiments (mostly literal readings), although we observed some local readings. Participants who saw the stressed 'SOME' had less literal readings. They mostly accessed a global reading.

Participants from the two versions accessed the same readings for the embedded questions. In both cases, they had few weakly exhaustive readings and a relatively large proportion of strongly exhaustive readings.

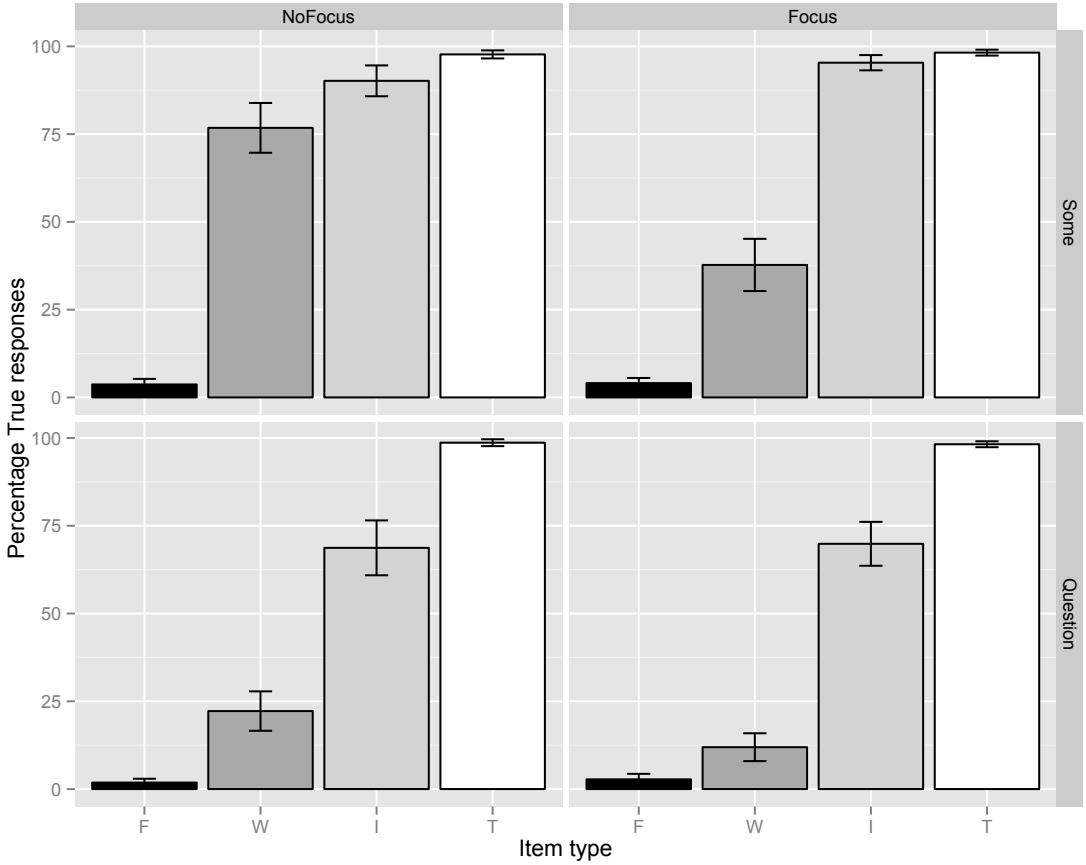

Figure 6: Percent 'True' responses to the different items in Experiment 3. **F** is false under any reading, **T** is true under any reading. **W** is true only under the non-exhaustified reading, while **I** is true under the non-exhaustified and globally exhaustified readings but not under the locally exhaustified reading.

3.5 Discussion

In this last experiment, we were finally able to obtain a varied distribution of readings for embedded 'some'. However, a manipulation which had a dramatic effect on the readings of the embedded scalar item left the responses to embedded questions unchanged.

4 General discussion

Overall, we were not able to observe any clear correlation between the readings participants access for sentences with embedded scalars and sentences with embedded questions.

Some results may be interesting by themselves (e.g., the existence of a SE reading with 'surprise', various results about the interaction between scalar items and factive verbs).

References

Chierchia, G., Fox, D., and Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Portner, P., Maienborn, C., and Heusinger, K. ., editors, *Handbook of Semantics*. Mouton de Gruyter, New York, NY.

Fox, D. (2007). Free choice disjunction and the theory of scalar implicature. In Sauerland, U. and Stateva, P., editors, *Presupposition and Implicature in Compositional Semantics*, pages 71–120. Palgrave Macmillan, New York, NY.

Klinedinst, N. and Rothschild, D. (2011). Exhaustivity in questions with non-factives. *Semantics and Pragmatics*, 4(2):1–23.

Percus, O. (2006). Antipresuppositions. In Uyema, A., editor, *Theoretical and Empirical Studies of Reference and Anaphora: Toward the Establishment of Generative Grammar as an Empiricial Science*, pages 52–73. Report of the Grant-in-Aid for Scientific Research (B), Project No. 15320052, Japan Society for the Promotion of Science.

Sauerland, U. (2008). Implicated presuppositions. In Steube, A., editor, *Sentence and Context: Language, Context, and Cognition.*

Spector, B. and Sudo, Y. (2014). Presupposed ignorance and exhaustification or: How scalar implicature and presupposition interact. Ms., Institut Jean Nicod and UCL.