On the semantics of embedded questions

Alexandre Cremers

École Normale Supérieure

On the semantics of embedded questions

Thèse en Sciences du Langage soutenue le 24 mars 2016

Auteur:
Alexandre CREMERS

Directeur:
Emmanuel CHEMLA

Rapporteurs:
Richard BREHENY
Yael SHARVIT

Jury:
Maria ALONI
Floris ROELOFSEN
Benjamin SPECTOR

Laboratoire de Sciences Cognitives et Psycholinguistique
École Doctorale 158 Cerveau Cognition Comportement

Abstract

Two important questions arise from the recent literature on embedded questions. First, Heim (1994) proposed that embedded questions are ambiguous between a weakly and strongly exhaustive reading. Spector (2005) recently proposed an intermediate exhaustive reading as well. Second, adverbs of quantity such as 'mostly' can quantify over answers to an embedded questions (Berman, 1991). An analysis of this phenomena reveals an analogy between embedded questions and plural determiner phrases, and suggests a fine-grained structures for the denotation of questions (Lahiri, 2002).

The first part of the dissertation consist of three psycholinguistic studies on the exhaustive readings of questions under 'know' in English, the acquisition of these readings under 'savoir' by French 5-to-6-ear-olds, and the properties of emotive-factive predicates such as 'surprise'. The second part presents a theory of embedded questions built on Klinedinst and Rothschild's (2011) proposal to derive exhaustive readings as implicatures, although it differs in the fine-grained structure it adopts for questions denotations in order to account for plurality effects as well. The theory solves problem raised by B. R. George (2013) and makes predictions for a larger range of sentences.

Keywords: Questions, Formal semantics, Psycholinguistics, Language acquisition.

Résumé

Suivant la proposition de Tarski (1936), la sémantique vériconditionnelle associe à une phrase déclarative des conditions de vérité. Ainsi, comprendre le sens de la phrase “Il pleut”, c’est pouvoir dire après avoir regardé par la fenêtre si elle est vraie ou fausse. Toutefois, ceci ne permet de rendre compte que des phrases déclaratives, et pas des questions puisqu’aucune situation ne rendra jamais la question “Qui a appelé ce matin ?” vraie ou fausse. Hamblin (1973) propose la première théorie des questions dans le cadre de la sémantique véri-conditionnelle, et propose de leur associer des conditions de résolutions, c’est-à-dire des ensembles de réponses. Comprendre le sens de la question “Qui a appelé ce matin ?” c’est alors savoir que “Jean a appelé” est une réponse possible, tandis que “il pleuvait” n’en est pas une.

Très rapidement, l’étude de la sémantique des questions s’est tournée vers les questions enchâssées dans des phrases déclaratives (questions indirectes). En effet, il est beaucoup plus aisé de juger des conditions de vérité d’une phrases déclarative que des conditions de résolution d’une question. Or moyennant des hypothèses sur la sémantique des verbes enchâssant des questions (‘savoir’, ‘oublier’. . .), on peut relier les conditions de vérité d’une phrase déclarative au sens de la question qu’elle enchâsse. Cette approche, proposée par Karttunen (1977), a donné lieu à une littérature théorique très riche.

Dans cette thèse, je m’intéresse à deux questions importantes dans la littérature récente. Premièrement, Groenendijk & Stokhof (1982) ont remarqué que les conditions de vérité prédites par Karttunen pour des questions enchâssées sous ‘savoir’ étaient trop faibles. Depuis, la littérature distingue lectures exhaustives faibles et fortes pour les questions enchâssées (Heim, 1994). Ce débat a été ravivé récemment après qu’une lecture exhaustive intermédiaire a été proposée par Spector (2005). Deuxièmement, Berman (1991) a découvert que les adverbes de quantités tels que ‘pour l’essentiel’ pouvait quantifier sur les réponses à une question enchâssée. L’analyse de ce phénomène a révélé un parallèle entre questions enchâssées et groupes nominaux pluriels et amené à raffiner la structure proposée pour les questions (voir en particulier Lahiri, 2002).

La première partie de la thèse est composée de trois études psycholinguistiques sur les lectures exhaustives. La première établit l’existence des lectures intermédiaires et fortes sous le verbe ‘savoir’, sans exclure la possibilité de lectures faibles. La deuxième, portant sur l’acquisition des questions enchâssées, suggère que les enfants commencent par une lecture faible mais acquièrent entre 5 et 6 ans la lecture intermédiaire qui est majoritaire chez les adultes. La troisième étude s’intéresse aux prédicats dits factifs émotifs, à l’origine de nombreux débats dans la littérature, et en particulier au verbe ‘surprendre’. Les résultats confirment certaines intuitions, notamment l’incompatibilité de ce verbe avec les questions polaires, mais établissent l’existence d’une lecture exhaustive forte.

La seconde partie de la thèse présente deux théories des questions enchâssées, compatibles avec les résultats expérimentaux de la première partie. La première théorie s’inscrit dans la lignée de Klinedinst & Rothschild (2011) qui proposent de dériver les lectures exhaustives intermédiaires et fortes comme des implicatures, mais se base sur une structure fine pour la dénotation des questions afin de rendre compte également des effets de pluralité discutés par Berman (1991) et Lahiri (2002). La théorie ainsi obtenue apporte des solutions aux problèmes soulevés par George (2013) sans nécessiter de redéfinir la dénotation des verbes ‘savoir’ et ‘oublier’ et établit des prédictions pour une gamme de phrases plus large que celles traitées habituellement. La seconde théorie traite des effets de pluralité avec les verbes dits rogatifs ('demander', 'dépendre'), dans la lignée de Beck & Sharvit (2002). Cette théorie pourrait également être complétée pour dériver les lectures fortes et intermédiaires, suivant les principes proposés dans la première théorie.

Mots-clés: Questions, Sémantique formelle, Psycholinguistique, Acquisition du langage.

Acknowledgements

I feel very lucky that this section is so long. Unfortunately, I also probably forgot some people who contributed significantly. I tried to cite everyone only once, but of course many people should be cited in several of the following paragraphs! Each chapter acknowledges specific contributions.

I would first like to thank my committee. When I realized that my defense would essentially be a general attack on the theories supported by the members of my committee, I thought that we would have fun. Although everything went much faster than I expected, the discussion was very fruitful and made us realize how far we still are from understanding question embedding. I would particularly like to thank Maria Aloni and Floris Roelofsen who had read my dissertation so carefully and came from Amsterdam with so many clever questions. I am looking forward to working together in the years to come!

I have had extensive discussions with Yael Sharvit on many occasion before the defense and her feedback helped me a lot. I will always remember my very first meeting with Yael two years ago, when she managed to understand an early version of my theory within a few hours from a badly written draft, while she had just landed from California and was 9-hours-jet-lagged. I promise I will look more carefully at NPI data one day!

I wish to thank Richard Breheny not only for his work as a reviewer, but also for his support throughout the past year. Even though I won't move to London in the end, I learned a lot from writing all these grant applications together. We had great discussions on various topics in experimental pragmatics, and I'm convinced we will collaborate sooner or later, even without sharing an office!

Having Benjamin Spector on my committee was very important to me. He introduced me to formal semantics and pragmatics back in 2008, and he was the first one to show me all these fascinating implicature things. Four years later, he was also the one with the original idea for the Experiment 1 of the first chapter of this dissertation. One thing led to another and I changed my PhD project proposal to work on embedded questions. On any possible topic, Benjamin always has clever original ideas, and no matter what you come up with, he will always suggest something better. The coffees we had together were certainly the most productive.

Nathan Klinedinst could not make it for the defense, but I was honored to have him at my 'pre-defense' last year and he gave me very useful feedback on this occasion. His paper with Daniel Rothschild had an obvious influence on my own work, and was my starting point when I began theoretical work. I remember very well the day I presented it in a reading group at ENS.

I don't think I can express what a wonderful advisor Emmanuel has been. His advices on any possible topic have been so valuable that I don't how I will adapt to making any decision without consulting him first. We have worked together on a daily basis for the past 5 years and I can't remember having any conflict or argument about anything more serious than stats[1]. He has the amazing ability to always answer any email within an hour, no matter the time of the day or night and even from a jungle in Uganda. And he doesn't even need a smartphone to do this. He can understand any crazy idea I may be trying to convey within a minute, but anyone who read stuff I wrote before my PhD will also realize that he somehow taught me how to write in an intelligible way.

[1]The core argument being whether stats should be viewed as something important or not. In the end, I think our views have converged a lot, so I can't see this argument as a serious one anymore.

Beyond Emmanuel, Lyn Tieu has been my main collaborator and is the co-first author of one of the chapter dissertation. I would never have dared trying to run an acquisition study without her. I think we have helped each other to various extent at all possible steps of the studies we ran together (from the ethical committee to the stats, from designing and recording stimuli to testing kids or adults). Lyn became a close friend and I am sure we will collaborate on many more projects no matter how far she moves!

Many other people have collaborated with me over the years and must be credited for their influence on this dissertation as well: Micha Breakstone, Martin Hackl, Danny Fox, Jacopo Romoli, Uli Sauerland, Kazuko Yatsushiro, Manuel Križ, Yimei Xiang, Tim Leffel and Nicole Gotzner.

I also learned a lot from a very different sort of collaboration in the form of teaching. I has been a pleasure to teach or TA with Benjamin, Philippe Schlenker, Denis Bonnay, Paul Egré, Valeria Giardino, Michael Murez, and of course the AMS team: Florence Bouhali, Isabelle Dautriche, Muy-Cheng Peich, Marion Rouault and Thomas Schatz!

I enjoyed being at the intersection of two groups at the DEC. First I would like to thank everyone at LSCP for creating such an inspiring place where usual barriers between discipline or hierarchy are basically non-existent. I would particularly like to thank Anne Christophe and Sharon Peperkamp for their support, Jerôme Sackur for great discussions about stats or epistemology, Radhia Accheb for shielding us all from the French administrative burden at all stages of our research, Michel Dutat and Viréack Ul for their availability to solve any possible technical problem, Anne-Caroline Fievet who helped us so much with testing the children and Florian Pellet who relieved me from a lot of my programming work and was so much better than me at this. Florian has since joined my fellow PhD students during these years: Abdellah Fourtassi, Hernan Anllo, Mikael Bastian, Louise Goupil, Alex Martin, Alex de Carvalho, Yue Sun, Leonardo Barbosa, and of course Thomas Andrillon, my long time friend and fellow coffee czar! I was also part of the Linguae group which has been growing so much over the years that I now have many fellow semanticists to thank for their support, friendship and taste for pizzas. I've already mentioned Lyn and Manuel, but I would also like to thank Mora Maldonado, Jeremy Kuhn, Milica Denić, Amir Anvari, Valentina Aristodemo, Mirko Santoro, Wataru Uegaki, Carlo Geraci, Lucie Ravaud, Vincent Gaudefroy. . . and of course Philippe Schlenker, who not only introduced me to question semantics a long time ago but also provided 14% of my lunches for the past two years!

I would also like to thank linguist friends over the world, who among other things make every conference more than just a sequence of talks: Natasha Ivlieva, Sasha Podobryaev, Yasutada Sudo, Hadas Kotek, Guillaume Thomas, Sam Alxatib, Jérémy Zehr, Dunja Veselinović, Erin Zaroukian, Patrick Grosz, Cory Bill, all the Berlin folks, EGG people. . . and most of the people I've already cited above.

The list of people credit for their support continues with Alex Drummond, whose Ibex has helped me so much, although we never met in person, my Lithuanian professors: Hélène de Penanros, who is also a great linguist, Gražina Čiarnaitė and Goda Klimavičiūtė in Paris, Dainius Vaitiekunas in Vilnius. I would also like to thank Piotr Graczyk, who inspired me to turn to research and also helped me very concretely when I was preparing for ENS.

I thank my family for their support, which extends far beyond this PhD, and especially my mother. I am grateful to my friends in Angers, in Paris and everywhere else, who have made life very enjoyable so far!

Last but not least, I owe a lot to Indrė who supports and understands me no matter what. She encouraged me whenever I needed motivation, changed the topic whenever I was talking too much about linguistics. She probably is the only person who can stand me day after day. With Betty of course!

Contents

Introduction

A very short history of questions in formal semantics

Shortly after Montague (1970), Hamblin (1973) proposed a formal account of the semantics of questions in English. If the meaning of declarative sentences was to be contained in their truth-conditions, Hamblin proposed that the meaning of a question be contained in its answerhood-conditions. Therefore, knowing what "Peter called." means is to know which situations would make this sentence true, while knowing what "Who called?" means is to know that "Peter called." is an appropriate answer but "It was raining." is (usually) not. Formally, we can consider that the denotation of a declarative sentence is a proposition (the set of worlds which make it true), while the denotation of a question is a set of propositions (the appropriate answers).

A few years later, Karttunen (1977) focused on declarative sentences containing questions (*embedded*, also called *indirect* questions). Like all other declarative sentences, these sentences should have truth-conditions. Provided some reasonable assumptions on semantic rules and the meaning of other verbs that embed questions, these truth-conditions should inform us about the meaning of the embedded questions. Karttunen proposed that the denotation of a question be the set of its true answers, and that verbs such as *know* combine with question by simply taking the conjunction of all such answers.

Therefore, in a situation where Ann, Bill and Chris called, (1) is true if and only if Mary knows that Ann, Bill and Chris called.

(1) Mary knows who called.

Groenendijk and Stokhof (1982, 1984) argued that the truth-condtions predicted by Karttunen for (1) were not satisfactory. Indeed, if Mary knows that Ann, Bill and Chris called, but also falsely believes that Emily and Frank did, we would not say that she knows who called. They propose a new semantics for questions which ends up predicting for that (1) is true if and only Mary knows that Ann, Bill and Chris called and that no one else did. On the way they dismiss a possible alternative, which would have been to only require that Mary has no false beliefs regarding Emily and Frank.

Berman (1991) argued that the meaning predicted by Karttunen (1977) was actually correct for verbs other than *know*, and in particular *surprise*. He proposed that the stronger readings discussed by Groenendijk and Stokhof (1982) could simply be attributed to pragmatic effects. Heim (1994) then showed that Groenendijk and Stokhof's (1984) semantics could be derived from Karttunen's and proposed a theory where questions are ambiguous between the two denotations and each embedding verbs select a (possibly different) denotation when embedding questions. Such a view has been adopted in most of the subsequent literature, where Karttunen's predicted reading in (2) is called "weakly exhaustive" while Groenendijk and Stokhof's in (3) is called "strongly exhaustive".

(2) Weakly exhaustive reading: Mary knows that Ann, Bill and Chris called.

(3) Strongly exhaustive reading: Mary knows that Ann, Bill and Chris called and that no one else did.

And now?

Recently, Spector (2005) renewed the interest in the alternative truth-conditions that Groenendijk and Stokhof (1982) had quickly dismissed, namely those in (4). I will refer to them as the "intermediate exhaustive" reading to avoid confusion, but it has sometimes been since as an instance of strongly exhaustive reading (Berman, 1991) or of weakly exhaustive reading (Preuss, 2001; Spector, 2005; Spector & Egré, 2015).

(4) Intermediate exhaustive reading: Mary knows that Ann, Bill and Chris called and she does not believe that anyone else did.

A few theories have now been proposed in order to derive this reading along or instead of the weakly and strongly exhaustive readings. I will distinguish two categories, which I will call the "ambiguity" theories and the "exhaustification" theories.

Ambiguity theories

Spector and Egré (2015)[2] propose a theory which main goal is to account for complex presuppositional effects with embedded questions, while deriving both the strong and the intermediate exhaustive readings (which they call weakly exhaustive). They propose two different rules to compose questions with embedding verb, which yield the two different readings. In this they follow Heim (1994), who proposed that questions could be interpreted weakly or strongly and that embedding predicates select one of the readings, but did not propose what determined which reading a certain predicate receives.

Theiler (2014) and Roelofsen, Theiler, and Aloni (2014) propose a theory in the framework of inquisitive semantics which takes inspiration from Spector and Egré (2015) on a certain number of points. In particular, they do not derive the weakly exhaustive reading as defined in (2) and place the ambiguity between the intermediate and strongly exhaustive readings at the level of the composition between the question and the embedding predicate. They also propose a new way to distinguish the factive presupposition of verbs such as *know* and *forget* from other lexical presuppositions, thus allowing a proper definition for the "non-factive entries" of these verbs used in the definition of intermediate readings.

These theories propose a very accurate description of the facts, at the price of explanatory power. As pointed out by Spector and Egré (2015) themselves, had the data to account for been different, nothing would have prevented them from postulating different compositional rules. In Roelofsen et al. (2014), different version of the answer and question-operators could have been postulated as well. In particular, the authors of these theories chose not to derive the weakly exhaustive reading, but their theories could in principle derive it through a third compositional rule/question operator.

[2] Althought Spector and Egré (2015) is the most recently published theory in this section, I would like to point that the manuscript has circulated since 2007 and therefore influenced all other theories I discuss.

Exhaustification theories

The idea that the strongly exhaustive reading could be derived as an implicatures dates back to Berman (1991). After all, Heim (1994) showed that it can be derived from the weakly exhaustive reading, and one could perfectly imagine that this strengthening is an implicature.

Klinedinst and Rothschild (2011) show that for non-factive verbs, such as *predict*, this is feasible. They adopt a grammatical view of implicatures and demonstrate that modulo a well-chosen focus value for the embedded question, the intermediate reading can be derived as a usual implicature and the strongly exhaustive reading as the result of local exhaustification (hence the grammatical view). Following Spector and Egré (2015), they suggest that such a mechanism could be extended to factive verbs by separating the presuppositional component from the assertive one.

Uegaki (2015) elaborates on this theory and implements the separation of presuppositions and assertion as suggested in Theiler (2014). He also shows that the exhaustification approach could be reconciled with a pragmatic, neo-Gricean view of implicatures, by having the strongly exhaustive reading derived from the intermediate reading through an additional pragmatic step rather than via local exhaustification.

Goal for this dissertation

In the first part of the dissertation, I will present experimental results aiming at establishing the existence of the different readings discussed here, their properties and possibly distinguishing between the different approaches. I will conclude in favor of the exhaustification approach, although rejecting the ambiguity theory on empirical grounds is essentially impossible because its predictions are not so explicit and its inherent flexibility would in principle allow it to account for a large range of possible outcomes.

In the second part of the dissertation I will propose my own theory, adopting an exhaustification approach. In addition to exhaustivity, I will investigate plurality effects such as the quantificational variability effect described in Berman (1991).

Part I

Psycholinguistic investigations

Chapter 1

A psycholinguistic study of the exhaustive readings of embedded questions

Cremers, A. & Chemla, E. (2014). A psycholinguistic study of the exhaustive readings of embedded questions. *Journal of Semantics*.[†]

Abstract

What is the semantic content of a question? As pointed out by Karttunen (1977), declarative sentences that embed interrogative complements (such as "John knows which students called") can provide relatively easy access to the semantics of questions. Recent theories attribute different readings to such sentences and their predictions depend in various ways on the embedding verb ('know' in this example). Through a series of four experiments, we provide quantitative offline data to evaluate critical judgments from the literature. We show that the so-called strongly exhaustive reading is not the only available reading for 'know', providing an argument against approaches inspired by Groenendijk and Stokhof (1982, 1984). We also describe processing data which may further constrain theories, provided hypotheses about the derivation processes are made explicit.

1.1 Introduction

1.1.1 The meaning of questions

Our goal is to provide quantitative data to help decide between different semantics which have been proposed for questions.

We cannot study the meaning of questions, such as (1), as we traditionally study the meaning of declarative sentences, such as (2). Declarative sentences can be described by their *truth conditions* (Frege, 1892; Tarski, 1935, 1956). Knowing the meaning of (2) may be reduced to knowing which situations make this sentence true and which situations make it false. For questions, we cannot define truth

[†]We would like to thank Benjamin Spector, Paul Egré, Manuel Križ, Paul Marty, Salvador Mascarenhas and Lyn Tieu for discussion and important comments, Seth Cable as an editor and two anonymous reviewers for their detailed comments on an earlier version, Nicolas Porot and Aparicio Kozuch for their help with the English materials, Radhia Achheb for her administrative support, Alex Drummond for the IBEX interface and his help using it, and the audiences at the XPrag London Masterclass and the Euro-XPrag 2013 Conference in Utrecht for their attention and comments. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610 and was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

conditions. It would not make any sense to say that (1) is true or false in a given situation.

(1) Which students called?

(2) Mary called.

Hamblin (1958, 1973) proposed that the meaning of a question resides in its *answerhood conditions*. Knowing the meaning of (1) is equivalent to knowing what counts as an answer to (1). Because answers are propositions, with truth conditions, questions can be studied within the general framework of truth-conditional semantics. The problem is then reduced to identifying the set of possible answers to a question. This tells us what to look for, but not how to look for it.

1.1.2 Embedding questions as a way to study them

Some verbs such as *know* can embed questions, as in (3), just as well as declarative sentences, as in (4).

(3) John knows which students called.

(4) John knows that Mary called.

As Karttunen (1977) pointed out, the meaning of (3) must be related to the meaning of the question (1) (and similarly (4) must be related to (2)). More specifically, (3) seems to be true in situations where John is able to answer the question (1). Because (3) is a declarative sentence, it can receive a truth value and we know how to study it. Thereby, sentences like (3) provide some access to the semantics of the question (1). As we will see, this provides an entry point into understanding questions both from a theoretical and from a psycholinguistic point of view.

Among all the predicates that can embed questions, we will focus on those like *know*, which are called *veridical responsive* predicates. As we saw, these verbs can embed both interrogative and declarative complements.[1] They are interesting because the meaning of (3) is usually assumed to be *reducible* to something of the form "John knows that p" for some proposition p (namely, an answer to the question). If this is the case, everything we know about the declarative-embedding *know* can be applied to the question-embedding *know*.[2]

1.1.3 Different readings for embedded questions

Let us first review the bare facts. Several readings have been associated with sentences such as (3) in the literature. We can distinguish between readings which are called *exhaustive* and those which are called *non-exhaustive*. In the first part of this section we will present the different exhaustive readings, which are the focus of our study. Their distribution has been strongly debated in the theoretical literature, and our main goal will be to provide quantitative data that speak to their respective distributions. The second part of this section presents an example of a

[1]Predicates which only embed interrogative complements, such as *wonder*, are called *rogative*. Non-veridical responsive predicates, such as *agree*, can receive both types of complements. They differ from *know* in that they can express a relation to any answer to the question, not only to the true answer. For instance, it may be true that John and Peter agree on which students called while both are wrong.

[2]This reducibility property should not be taken for granted. See Schaffer (2007) and B. R. George (2011) for discussion and Chemla and George (2015) for experimental evidence supporting non-reducibility for the non-veridical responsive predicate *agree*.

non-exhaustive reading. Non-exhaustive readings will not be given a prominent role here (although we will briefly discuss their potential role in our first experiment).

Exhaustive readings: The different exhaustive readings of (3) described in the literature are presented in (5). They are all called *exhaustive* readings because they require that John have complete knowledge about the students who actually called. They differ on the required knowledge about students who did not call. The Strongly Exhaustive (SE) reading described in (5a) requires complete knowledge about both callers and non-callers. It entails the Intermediate Exhaustive (IE) reading (5b), which merely requires absence of false beliefs about non-callers.[3] The IE reading in turn entails the Weakly Exhaustive (WE) reading (5c), which does not require anything beyond 'exhaustivity', that is true beliefs about the students who actually called.

(5) Exhaustive readings of (3):
  a. Strongly exhaustive (SE):
     For each student who called, John knows that she called,
     and he knows that no other student called.
  b. Intermediate exhaustive (IE):
     For each student who called, John knows that she called,
     and John does not have false beliefs about students who didn't call.
  c. Weakly exhaustive (WE):
     For each student who called, John knows that she called.

The existence and co-existence of these readings is debated, judgments vary across theoreticians and certainly across verbs. Klinedinst and Rothschild (2011) present the only quantitative survey of the issue, which results suggest that different native speakers may provide responses coherent with any of the three exhaustive readings for questions embedded under the verb *predict*.

Non-exhaustive readings: Although they are not our focus, non-exhaustive readings have also been discussed. At least for *know*, they only seem to be available under specific circumstances. As an example, (6), repeated from B. R. George (2011), is usually considered true as soon as William is able to name at least one place where Rupert could buy an Italian newspaper. Any exhaustive reading seems too strong, since it would require that William know all the places where Italian newspapers can be purchased. This non-exhaustive reading is usually called the *mention-some* reading. It will not be at the foreground of our inquiry.

(6) William knows where Rupert can buy an Italian newspaper.

1.1.4 Architecture of recent theories of questions and embedded questions

We cannot offer a complete introduction to the theories of questions, but in this section we try to give a sense of what they are made of and how they may vary.

[3] A first description of this intermediate reading can be found in Groenendijk and Stokhof (1982), but the authors quickly reject it in favor of the SE reading (thanks to Benjamin Spector for pointing this out). The intermediate reading resurfaces in Spector (2005). Certainly, the IE reading has been implemented in very few formal approaches.

The main material at the basis of our experimental inquiries is to be found in the *empirical* predictions of the relevant theories. We introduce these empirical aspects independently of their formal and conceptual basis in the following section.

In this section, we focus on theories which follow the general architecture originally defined by Hamblin (1973).[4] These theories are based on three essential ingredients: (a) a denotation of questions which corresponds to a set of answers/propositions, (b) some mechanism to combine such sets of propositions with verbs which normally combine with propositions and possibly (c) some form(s) of strengthening mechanism.

(a) Answer sets. A theory of questions first owes a basic denotation for a question. This is typically a set of answers. We can define two groups of theories, depending on which set of answers is taken as the denotation of the question.

- On the first group of theories, following Hamblin (1973) and Karttunen (1977), the denotation of a question consists of so-called 'weak answers', which can be defined through an example:

  (7) Weak answers for a question such as *Who called?*:
  Mary called, Peter called, Mary and Peter called...

  As a result of the combination of this denotation with the compositional mechanism (see (b) below), the WE reading is primitive. The other readings may then be derived by applying strengthening mechanisms (see (c) below).

  Most recent theories fall into this group (Beck & Rullmann, 1999; Lahiri, 2002; Sharvit, 2002; Guerzoni, 2003, 2007; Guerzoni & Sharvit, 2007; Klinedinst & Rothschild, 2011; Spector & Egré, 2015).[5]

- On the second group of theories, the denotation of a question consists of so-called 'strong answers', see (8), as in Groenendijk and Stokhof (1982, 1984, 1993) and more recently B. R. George (2011).

  (8) Strong answers for a question such as *Who called?*:
  Only Mary called, Only Peter called, Only Mary and Peter called...

  As a result, the SE reading is primitive. If we allow only strengthening mechanisms, as opposed to weakening mechanisms, the other exhaustive readings are simply not available.[6]

[4]We do not discuss theories which strongly divert from the work of Hamblin (1973). Some approaches, such as Ginzburg (1996), put less stringent constraints on the set of possible readings and focus their inquiry more on the role of context in the selection of a reading.

The inquisitive semantics framework has also given rise to much work on questions, partly in reaction to the limitations of partition semantics (see Mascarenhas, 2009), but we are not aware of any explicit implementation of question embedding in an inquisitive framework. Nevertheless, the main conclusion from our study applies to these theories as well: they must all be able to derive some form of the intermediate exhaustive reading.

[5]Strictly speaking, Spector and Egré (2015) use a rule which quantifies over strong answers, so they should fall into the second group. However, they also refer to the weak answers in order to derive the IE reading. Because weak answers cannot be retrieved from the strong ones, it makes sense to consider their theory as a member of the first group.

[6]B. R. George (2011) makes use of the weak answers to derive mention-some readings (non-exhaustive) but his theory cannot derive WE readings. He argues that alleged cases of WE readings, such as the examples of Berman (1991) with the verb *surprise*, are in fact mention-some readings.

(b) Combination with responsive verbs. This is usually the most technical part of a theory: it formalizes how compositional semantics combines the denotation of a question (set of answers, see (a) above) with the semantics of the verb with which it merges. A simple example would be the following rule for the verb *know*: "knowing Q is equivalent to knowing the conjunction of all true answers to Q". So, if $Q = \{a_1, a_2, a_3\}$ and a_1 and a_2 are true but a_3 is false, then $[\![\text{know } Q]\!] = [\![\text{know } a_1 \wedge a_2]\!]$.

Actual theories usually provide a uniform treatment of all responsive verbs by means of a generic abstract rule. These rules are somewhat complex and there is room for variation. For instance Lahiri (2002) considers that the verb is lexically responsible for restricting (or not) the set of relevant answers, e.g., to true answers; Spector and Egré (2015) do not use universal quantification but rather existential quantification (roughly: $[\![\text{know } Q]\!] = \exists a \in Q : [\![\text{know } a]\!]$), and they also treat the assertive and presuppositional dimensions independently.

(c) Strengthening mechanisms. Various semantic and pragmatic mechanisms have been described for strengthening the meaning of an utterance.

- One example is exhaustification, through a grammaticalized EXH-operator as in e.g., Fox (2007) or as a proxy for Gricean strengthening. For our purposes, the potential application of exhaustification can help derive stronger readings from weaker ones, even though weaker readings remain primitives.

- Another example more specific to the theory of questions is given by Heim (1994), who showed that the set of strong answers could be recovered from the set of weak answers (but not *vice versa*). (Note that this role can also be supported by a more general EXH operator as described above, provided that it can be applied at the level of individual answers in the denotation of the question). Such a mechanism effectively makes the first group of theories as described in (a) above capable of encompassing the predictions of the second group of theories.

Let us present a semi-concrete theory that puts together all of these ingredients. We could consider a theory which (a) makes use of weak answers, (b) relies on the 'universal' embedding rule and (c) allows for the EXH-operator to produce stronger readings (as a first approximation, this operator strengthens a proposition by conjoining it with the negation of all its non-weaker alternatives). Such a theory would be very close to the approach of Klinedinst and Rothschild (2011) and would be able to derive all of the exhaustive readings that we discussed. First, WE readings would be primitives (using weak answers and no exhaustification yields a weak reading). Second, IE readings would be obtained by exhaustification at the matrix level (see Klinedinst and Rothschild, 2011 for details). Finally, SE readings would be obtained by local exhaustification, the reason being roughly that this would turn weak answers into strong answers, in the sense of (7) and (8), as suggested below:

(9) EXH('Peter and Mary called', alternatives={'Sue called', 'Jack called', ...}) = $[\![\text{Peter and Mary called}]\!] \wedge \neg[\![\text{Sue called}]\!] \wedge \neg[\![\text{Jack called}]\!] \wedge \ldots$

The combinations of various versions of the three ingredients we presented in this section lead to a variety of possible theories. In the next section, we will focus on the empirical predictions of the current implementations found in the literature. In short, starting with weak readings allows for more flexibility, because various tools can be called for to obtain strengthened meanings from weaker ones (e.g., exhaustification, maximization) and less so in the other direction.[7] Although we will not enter into the details of which theories rely on which tools, we will describe the predictions they make.

1.1.5 Predictions of recent theories for *know*

Let us put aside the technical differences between the different theories and the way they derive the different readings. Here we merely want to provide the set of readings each theory predicts to be available, so that we can later evaluate whether these predictions are borne out. Let us also focus on the verb *know*, which has been discussed in greater detail than any other responsive verb. We thus want to summarize which readings are available for *know*, according to some recent theories from the two groups identified in 1.4(a). This is done in Table 1.1, which indicates whether a given theory can derive a given reading or not.

Let us note that theories also vary as to whether they offer different predictions for different verbs. The comparison between different verbs is in fact crucial to put constraints on the available compositional mechanisms between the verb and the embedded question (see 1.4b). However, this cross-verb variation will not be the focus of our inquiry, which focusses on *know*, and we therefore do not report a complete set of predictions across verbs. Let us immediately note however that in Experiment 2 we also tested the verb *predict*, and that the results did not reveal important differences with *know*.

Theory	Can *know* give rise to...		
	WE readings?	IE readings?	SE readings?
Heim (1994)	yes	*not discussed*	yes
Guerzoni and Sharvit (2007)	yes	*not discussed*	yes
Klinedinst and Rothschild (2011)[8]	yes	yes	yes
Spector and Egré (2015)	*primitive but always enriched*	yes	yes
Groenendijk and Stokhof (1982, 1984)	no	no	yes
B. R. George (2011)	no	no	yes

TABLE 1.1 – Theories make different predictions about which (exhaustive) readings can arise for questions embedded under *know*. Following Groenendijk & Stokhof, B. R. George (2011) takes the SE reading as basic. Others, following Heim (1994), take the WE reading as basic, and derive more readings from it.

1.1.6 Goals of our study

The goal of our study will be to provide constraints on what readings an accurate theory should derive and how. We will test the availability of the different exhaustive readings and compare these results with the predictions in Table 1.1. We will

[7]Furthermore, Heim (1994) proved that under some hypotheses (constant domain) it is logically impossible to retrieve the weak answers from the strong ones.

[8]Klinedinst and Rothschild (2011) argue against the availability of WE readings for verbs like *know*, but their restriction is not derived by the theory; rather it is a further pragmatic constraint that the authors stipulate. As it stands, their theory derives IE readings for non-factive verbs only, but they indicate how it could be modified in order to derive IE readings for all veridical responsive verbs.

also gather online data which may yield extra constraints on possible strengthening mechanisms (for theories *à la* Heim, 1994) when hypotheses about the cognitive processes are made explicit.

In Experiment 1, we show that *know* can give rise not only to SE readings but also to weaker readings. In Experiment 2, using a different paradigm, we show that these weaker readings consist mostly of IE readings (and in fact both for *know* and *predict*). The results of Experiment 3 confirm that we cannot reliably establish the availability of the WE reading for *know*. Finally, in Experiment 4 we compare the processing properties of the IE and SE readings.

1.2 Experiment 1: Existence of WE/IE readings under *know*

1.2.1 Goal

Most theories derive various readings for sentences containing questions embedded under *know* (see Table 1.1). But most theorists report judgments that these configurations only give rise to SE readings. The goal of this experiment was to test this empirical claim. We did not aim to distinguish between WE and IE readings for the time being.

We also collected online data in order to test a prediction derived from Heim (1994). Her theory derives the SE reading from the WE reading. Hence, the SE reading requires an extra derivation step and we may thus be able to detect an extra processing cost.

1.2.2 Methods and Materials

The experiment consisted of a truth-value judgment task: Participants read pairs of sentences (which we call *context* and *utterance*, respectively) and had to judge whether the utterance was true or false in the given context. All experiments were conducted using American English.

Procedure

Participants were directed from Amazon Mechanical Turk to an online experiment hosted on Alex Drummond's Ibex Farm. After filling out a consent form they received instructions (see details below), then completed a short training phase (8 items with feedback) followed by the experimental phase. After completing the experiment, participants filled in a questionnaire (including demographic questions about participants' age, gender and native language). The final step was to validate their participation on Mechanical Turk.

Instructions

The instructions provided a story repeated in Fig. 1.1. The goal of this story was to provide a general context for the experiment such that we could easily manipulate what the truth-value of the various hypothesized readings would be. In a nutshell, it was set up so that various characters (e.g., the male baker) would have total (exhaustive) knowledge about some events (e.g., men buying bread) and partial or no knowledge about other events.

A small town in a remote area has two shopping malls: one for women and one for men.

- Men shop at the men's mall, where all shopkeepers are men.
- Women shop at the women's mall, where all shopkeepers are women.

In each mall there is one of each facility: there is one male and one female baker, one male and one female mechanic, and so on.
Sometimes the shopping malls close.

- Men do not always know whether the women mall is open or not.
- Women do not always know whether the men mall is open or not.

Each shopkeeper knows what he/she sells to whom, but knows nothing about what happens in other shops.

FIGURE 1.1 – The story given to participants in the instructions for Experiment 1

The instructions also included 4 example items, which are given in (10). (10a,c) are true and (10b,d) are false given the story in Figure 1.1. This was explained to the participants.[9]

(10) a. *Situation: Today, the men shopping mall was open.*
"John-the-baker knows whether Bill bought bread."
b. *Situation: Today, both shopping malls were open.*
"Sandra-the-pharmacist knows who bought medicine."
c. *Situation: Today the men shopping mall was closed and everybody, including the women, knows about this.*
"Lisa-the-tailor knows who bought a coat."
d. *Situation: Today, both shopping malls were open.*
"Henry-the-hairdresser knows if Peter got an oil change."

Stimuli

The stimuli were built from 3 types of contexts, as exemplified in (11), and from 3 types of utterances, as exemplified in (12).

(11) a. *Today both shopping malls were open.*
b. *Today the women shopping mall was closed. Everybody knows about this.*
c. *Today the women shopping mall was closed. Men do not know about this.*

(12) a. John the baker knows who bought bread.
b. John the baker knows which men bought bread.
c. John the baker knows which men bought meat.

[9] There was a recurring typo in our materials: in the second part of the instructions and in context descriptions (not in target sentences) we dropped the genitive marker in *(wo)men's shopping mall*. Thanks to an anonymous reviewer for catching this mistake.

The crucial utterance was of the form of (12a). The truth value of its different possible readings varies with context as follows (see Table 1.2 for a summary):

- In context (11a), (12a) is false under any exhaustive reading because John does not know which women bought bread (assuming some women did, as soon as their mall is open). The items obtained from the combination of this utterance and this context were false controls.

- In context (11b), (12a) is true under all readings: For each man, John can tell whether he bought bread or not, and he knows that no woman bought bread that day. Therefore for each person John can tell whether he or she bought bread. The items obtained from the combination of this utterance and this context were true controls.

- In context (11c), (12a) is true under the WE reading because only men bought bread, therefore for each person who bought bread it is true that John knows that this person bought bread. It should also be true under the IE reading, because John has no reason to falsely believe that a given woman bought bread. However, it is false under the SE reading, because John does not know that no woman bought bread. The items obtained from the combination of this utterance and this context formed our targets. If only the SE reading is available, we shouldn't expect differences between (12a) in contexts (11a) and (11c): both items would be false.

Context	WE reading	IE reading	SE reading
(11a): False control	false	false	false
(11b): True control	true	true	true
(11c): Target	true	true	false

TABLE 1.2 – The truth value of the crucial utterance (12a) under each reading depends on the context. Critically, answers in the (11c) context distinguish between the SE reading (false) and the other two (true).

We created variations of each item by using the 16 characters presented in Table 1.3 (8 jobs × 2 genders) and varying the words 'men/women' accordingly in the *contexts* and *utterances*. All 16 possible targets, along with true and false controls, were included. 24 true and 24 false fillers for which the truth value did not depend on the *context* were also created using the utterances (12b) and (12c), respectively. They ensured that participants could not guess the truth value of an item using only the *context*.

As for the display, *context* sentences were displayed for 3s and *utterances* were displayed word-by-word (250ms per word). Each word replaced the previous one, and there was no delay between them to avoid flicker effects. The last word remained on the screen until a response was given.

Participants

40 participants were recruited on Mechanical Turk. 39 of them completed the task (17 females, 22 males). Their age ranged from 19 to 66 years (mean: 33). All of them reported English as their native language.

Trade	Male	Female	Match	Mismatch
baker	John	Mary	bought bread	bought meat
butcher	Peter	Linda	bought meat	bought bread
hairdresser	Henry	Sarah	got a haircut	got an oil change
mechanic	Stan	Kelly	got an oil change	got a haircut
pharmacist	Jack	Sandra	bought medicine	bought a painting
painter	Michael	Emma	bought a painting	bought medicine
optician	Fred	Celena	bought glasses	bought a coat
tailor	Chris	Lisa	bought a coat	bought glasses

TABLE 1.3 – We used 16 characters. For each of the 8 types of shop, there was one male and one female name, and one matching and one mismatching activity. All activities were alternatively matching and mismatching, depending on the shop they were associated with.

1.2.3 Results

Data treatment and Statistical methods

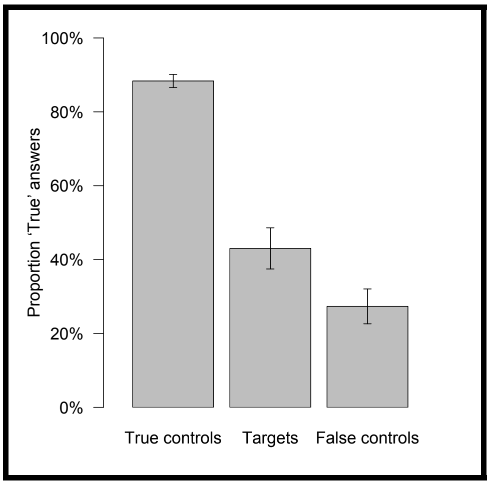
Responses made in less than 100ms or more than 10s were discarded (0.7% of the data). All mixed models were built with a maximal random effect structure based on participants and items as random variables (in the sense of Barr, Levy, Scheepers, and Tily, 2013). Our *item* factor was a value between 1 and 8 corresponding to one line in Table 1.3. This means that two items involving the same shop (e.g., 'baker') were not considered independent measures. For each model we give the estimate of the fixed effects (β) and a p-value with its associated statistics. For linear mixed models we used the t-statistics given in *lme4* and for GLMM the χ^2-statistics with one degree of freedom obtained by comparing the models with and without a given fixed effect.

Analysis of responses

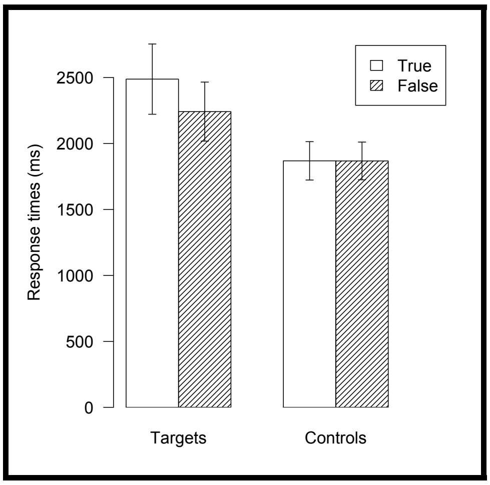
Fig. 1.2a shows the proportions of *True* responses to targets, true controls, and false controls. Overall, the task was well-understood and executed as reflected by high accuracy on true (88%) and false (73%) controls.

On targets, participants gave on average 43% *True* responses and individual rates varied from 0% to 100% (while there was little between-item variation: 36% to 48%). Two logit mixed models were fitted to compare the responses to targets with true or false controls respectively. The results showed that the differences were significant between targets and both true controls ($\beta = 2.7$, $\chi^2(1) = 32$, $p < .001$) and false controls ($\beta = -1.4$, $\chi^2(1) = 8.4$, $p = .003$).

The fact that participants gave fewer *True* responses on targets than true controls suggests that some participants had a SE reading and treated the utterance as false on the target trials. More surprising is the fact that they treated targets differently from false controls. According to Table 1.2, this high rate of true responses reveals the presence of a reading other than the SE reading. The fact that the intermediate rate of true responses corresponded to individual rates from 0% to 100% confirms that participants differed in the way they understood the same sentence. If they were simply at chance because the target sentences were too complicated we might have observed a more homogeneous pattern around 50% across the participants; alternatively, some participants would have been at one end of the spectrum (those that managed to do the task) and others would have been at 50%.[10]



(A) Percentage True to the different conditions. (Targets correspond here to WE+IE targets with the terminology of the following experiments.)



(B) Response times by answer. On targets both responses were acceptable (*True* = WE or IE, *False* = SE). On controls we kept only correct responses so *True* corresponds to true controls and *False* to false controls.

FIGURE 1.2 – Truth value judgments for the target sentence in Experiment 1 and corresponding response times.

Analysis of response times (no clear result)

On targets, *False* responses should reflect SE readings, while *True* responses should reflect WE or IE readings (and possibly some mention-some readings). We can thus compare the response times (henceforth: RT) associated with true and false responses, as indicative of how each reading is processed. We removed 8 participants whose error rate on controls and fillers was at least one standard deviation above the mean (threshold: 28%).[11] Fig. 1.2b displays the response times of the remaining subjects. A model with Answer, Sentence type (Control vs. Target) and their interaction as fixed effects was fitted on the RT to controls and targets (after errors on controls were removed). None of the effects were significant (all t's $< .5$). In subsequent experiments, we will use more powerful methods to investigate this issue.

[10] Alternatively, one may argue that some participants decided to answer true to complicated cases and others decided to answer false. If the reader is not convinced that we are seeing evidence for an ambiguity at this point, a potentially more convincing argument will be presented with the next set of experiments to be described.

[11] These participants were not removed from the analysis of responses to avoid introducing distortions in this analysis. In short, to analyze responses we avoided removing participants based on their responses. More precisely, selecting only participants with a low rate of *True* responses on false controls but putting no constraint on their responses to targets could create an artificial difference between these two conditions that we eventually want to compare. Practically, we argue that the target condition shows high variability, but this can only be assessed against the controls and if the variability of the controls has been removed a priori, this comparison would make no sense. For the RT analysis, since we proved that there is a real difference between these two conditions, we allow ourselves to consider *True* responses to false controls as errors while considering both *True* and *False* responses to targets as reflecting genuine linguistic judgments.

1.2.4 Discussion

The offline results suggest that the SE reading is not the only available reading for questions embedded under *know*. This is surprising, since even theories which derive the WE reading as a primitive usually treat *know* as an exception which only gives rise to SE readings. The findings would thus make a strong point against theories from the second group in our classification (Groenendijk & Stokhof, 1982, 1984, 1993; B. R. George, 2011) since they can only derive SE readings.

Two alternative explanations of these results come to mind. First, one may wonder whether mention-some readings could be called for to explain the appearance of non-SE readings. Note however that mention-some readings should generate true responses in all conditions, including false controls. The rate of true responses to false controls thus provides a baseline of errors *and* mention-some readings,[12] but targets gave rise to *significantly more* true responses, which calls for an independent explanation. Second, the expert reader may wonder whether the relevant false responses are not in fact SE readings in disguise, e.g., as altered by so-called covert domain restriction. We will explain this possibility in further detail in section 1.3.5 and assess it in detail in Experiment 3. For the time being, we will proceed to confirm and refine the current set of results with a more efficient method, which will also help distinguish between two types of non-SE readings: WE and IE readings.

1.3 Experiment 2 : non-SE readings for *know* and *predict*

1.3.1 Goals

Results from Experiment 1 provide reasons to believe that WE and/or IE readings are available for questions embedded under *know*. The main goal of Experiment 2 is to determine which readings participants have when they do not have the SE reading. In this experiment, we also compared different responsive predicates (*know* and *predict*). Klinedinst and Rothschild (2011) in their actual implementation derive IE readings for *predict* but not for *know*, while Spector and Egré (2015) argue that both verbs can support the IE reading.

We also simplified our experimental paradigm into a picture-sentence matching task, which allows finer-grained distinctions between the three exhaustive readings while keeping the differences between the *know* and *predict* versions minimal. Arguably, it also simplifies the task in that the participants can proceed on an item-by-item basis and do not have to memorize a complicated story.

1.3.2 Methods and Materials

The experiment consisted of a picture-sentence matching task: each experimental item consisted of a sentence paired with a picture, as described in the Materials section and illustrated in (13) and (14). Participants had to judge whether the sentence was an accurate description of the picture. The recruitment procedure was similar to the first experiment, except that participants were randomly assigned to one of two versions of the experiment (involving either *know* or *predict*).

[12] A closer look at the data reveals that a few participants consistently answered *True* to false controls. These participants potentially had a mention-some reading.

Instructions

The instructions included a short story, but it did not play a critical role. All the relevant information that was previously given in the story of Experiment 1 was now encoded in the pictures. Here is the story for the *know*-version of the experiment:

> *John is playing a card game to train his memory. Each turn, he picks a card, quickly looks at it and flips it. John tries to recall what he saw on the card. Then he looks at the card again, and checks if he was right. You will see the card, and how John remembered the card before checking it. Using this information you will have to judge whether some sentences about the card and John's knowledge are true or false. All cards are made up of 4 squares, each of which can be red, green or blue.*

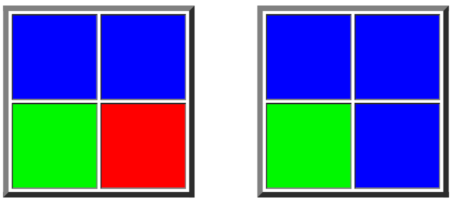
The story for *predict* was similar, except that John was trying to *predict* the content of the card *before* looking at it.

The instructions also included 4 example items (2 true and 2 false) with some explanations on how to understand the pictures, as well as general instructions which were identical to those of the first experiment (e.g., about response buttons).

1.3.3 Stimuli

Examples (13) and (14) show items used in the *know* and *predict* versions of the experiment respectively.

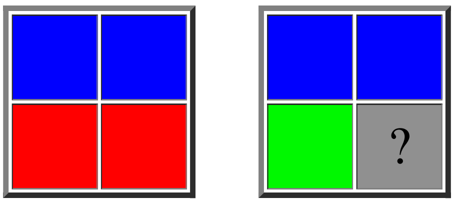
(13) WE target (*know*-version):



The actual card *John's beliefs*

"John knew which squares were blue"

(14) WE+IE target (*predict*-version):



The actual card *John's prediction*

"John predicted which squares would be blue"

Sentences

The sentences were variations on the sentence templates given in (15) and (16). *Position* was one of 'top', 'bottom', 'left' or 'right' and *color* one of 'red', 'green' or 'blue'. The most important sentences were those built on (15a) and (16a).

(15) Sentences for *know*:
   a. John knew which squares were [*color*].
   b. The [*position*] squares were [*color*].
   c. John knew which of the [*position*] squares were [*color*].

(16) Sentences for *predict*:
   a. John predicted which squares would be [*color*].
   b. The [*position*] squares were [*color*].
   c. John predicted which of the [*position*] squares would be [*color*].

Pictures

As seen in examples (13) and (14), the pictures consisted of two parts. In both versions of the experiment, the left side of the pictures represented the actual card that John had picked. The right side of the picture represented John's beliefs about the card, or the prediction he made beforehand, depending on the version. This was explained in the instructions and repeated on each item. In some cases, such as (14), the right part (which represented John's beliefs or predictions) included one or two gray squares with a question mark. This represented the fact that John sometimes had no idea about the color of these squares (*know*-version) or had not made any prediction about these squares (*predict*-version). This was explained in the instructions as well.

Targets

The two examples in (13) and (14) play a central role. As summarized in Table 1.4 and described below, together they allow us to infer which reading a given participant accessed (assuming participants were coherent):

- **WE targets:** Items like (13) are true under the WE reading and false under the stronger IE and SE readings. They are true under the WE reading because for all blue squares on the actual card, John knew that they were blue (respectively predicted that they would be blue). However, such items are false under the IE reading (and SE reading, consequently), because John had a false belief (respectively made a false prediction) about the bottom right square. Therefore, if a participant responds *True* to such items, it is presumably based on WE readings for the sentences.
- **WE+IE targets:** Items like (14) are true under the WE and the IE readings, but false under the SE reading. They are true under the WE and IE readings because for all blue squares on the actual card, John knew that they were blue (respectively predicted that they would be blue), and John did not falsely believe (respectively did not falsely predict) that any other square was blue. However, such items are false under the SE reading because John did not know for sure (respectively make an explicit prediction) that the bottom right square was not blue. Therefore, if a participant responds *True* to such items, it is presumably based on either the WE or the IE reading.

Responses to WE targets	Responses to WE+IE targets	Inferred reading
True	*True*	WE
False	*True*	IE
False	*False*	SE
True	*False*	*inconsistent*

TABLE 1.4 – True responses to WE targets correspond to WE readings, while true responses to WE+IE targets correspond to WE or SE readings so that together responses to the targets distinguish among the three (WE, IE and SE) exhaustive readings.

Controls

In addition to these two types of targets, the experiment included a variety of controls. These controls ensured that (a) each sentence appeared with pictures that made it unambiguously true or false, (b) each picture appeared with sentences that were accordingly unambiguously true or false, and (c) there was an equal balance of true and false controls, and a low proportion of targets (15%).

Of crucial importance were items which involved embedded questions, but were unambiguously true or false. They served as baselines for errors: they can help to prove that a certain category of responses may be rare but still more frequent than actual errors, and that therefore this category of responses corresponds to (rare but genuine) readings (see analyses below for details). Using the terminology above that encodes the set of true readings for a particular condition, these conditions were as follows:

- $\emptyset$ **baseline:** false under each of the three exhaustive readings.
- **WE+IE+SE baseline:** true under each of the three exhaustive readings.

Item generation

The items were varied by rotating the cards in the pictures and adapting the [*position*] words in the sentences accordingly (4 possibilities), or by permutation of the colors in the pictures and in the sentences accordingly (3 possibilities). We also varied the amount of ignorance in WE+IE targets and corresponding controls (one or two gray squares).

Overall there were 240 items in each version of the experiment, including 12 WE targets, 24 WE+IE targets (8 with two gray squares, 16 with one). This gave a 15/43/43% proportion of targets/true controls/false controls respectively.

Participants

30 participants were recruited for each version, but one subject in the *know*-version did not complete the task. One more subject in the *know*-version was removed from the analyses because he reported being a native speaker of a language other than English. The age of the remaining participants ranged from 18 to 59 years (mean: 34), there were 31 females and 27 males. Eight more subjects were removed because their error rate on controls was more than one standard deviation above the mean (threshold for rejection: 12.1%).

1.3.4 Results

Data treatment and Statistical methods

Responses made in less than 100ms or more than 10s were discarded (1.2% of the data).

All mixed models were run with maximal random effect structure when possible, but random slopes for subjects had to be dropped in the logit mixed models of section 1.3.4.[13] Rotations of the pictures were used to encode random effects of the items. This means that two different rotations of the same card were treated as independent measurements, but two items in different conditions under the same rotation were treated as dependent measurements.[14]

For each model we give the estimate of the fixed effects (β) and p-values with the associated statistics. For all mixed models in this experiment, we used the χ^2-statistics with one degree of freedom obtained by comparing the models with and without a given fixed effect.[15]

Analysis of responses

The mean error rate on controls was 5.1%. This confirms that this task was easier than the one used in Experiment 1 (in which the error rate was 15%). No participant had more than a 7% difference in performance between any two different colors and the effects of color on global performance were not significant. If some participants had been red-green colorblind, we would likely have observed bigger differences.

Fig. 1.3a shows the proportion of *True* responses given to all types of items (baselines and targets). First, we compared the WE+IE targets with the WE+IE+SE baseline. These items only differ with regards to the SE reading, which is false on the WE+IE targets but true on the WE+IE+SE baseline. Only participants who accessed SE readings could treat these items differently, while participants with IE or WE readings should answer *True* in both conditions. A logit mixed model (with only random intercepts for subjects, as explained above) was fit on responses to the WE+IE+SE baseline and WE+IE targets with Verb (*know* vs. *predict*), Sentence Type (Baseline vs. Target) and their interaction as fixed effects. There were slightly more *False* responses to targets than to the baseline ($\beta = -0.7$, $\chi^2 = 4.7$, $p = .03$), suggesting that the SE reading was available. This effect interacted with verb: *predict* gave rise to more *False* responses on WE+IE targets ($\beta = -1.4$, $\chi^2 = 11$, $p =< .001$), suggesting that if anything *predict* gives rise to more SE readings than *know*.

Second, Fig. 1.3a reveals that many more true responses were given to WE+IE targets than to WE targets, indicative of the important impact of the IE reading which distinguishes between these two conditions. A logit mixed model confirmed the strong difference between the two types of targets ($\beta = -7.2$, $\chi^2 = 59$, $p < .001$). The interaction with verb was not significant.

Finally, we tested for the presence of the WE reading by comparing WE targets with the $\emptyset$ baselines. These two conditions only differ in that the targets are true under the WE reading, while the baselines are not. A WE reading would thus lead to respond differently in these two conditions, while other readings should

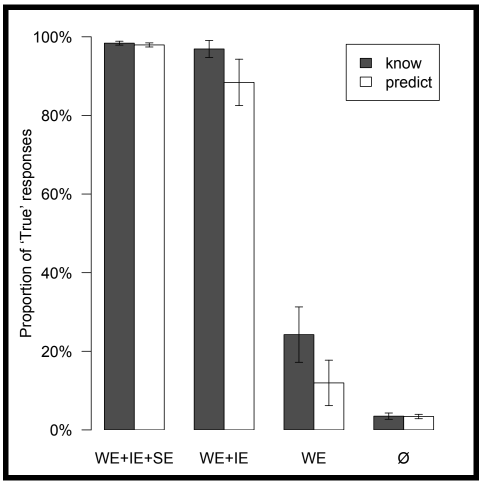
[13] When random slopes associated with the 'baseline vs. target' contrast for Subject were included, the predicted values differed greatly from the actual data, and the distributions of random effects for subjects were clearly not following a centered normal distribution. This seems to be related to a high proportion of subjects who gave exactly 100% *True* responses to WE+IE targets or 0% to WE targets. Dropping the random slope associated with Sentence Type for subject seemed to be the best workaround (Verb is a between-subject factor, hence there is no random effect associated with it).

[14] Note that for controls which referred to a particular row or column in the card (e.g., 'the top squares'), rotation also encoded differences in the sentence (the position words 'top', 'right', 'bottom', 'left'). We may expect differences in decoding pictures where the crucial information is aligned vertically or horizontally, as well as differences due to position words (e.g., 'top' vs. 'bottom' may be easier than 'left' vs. 'right').

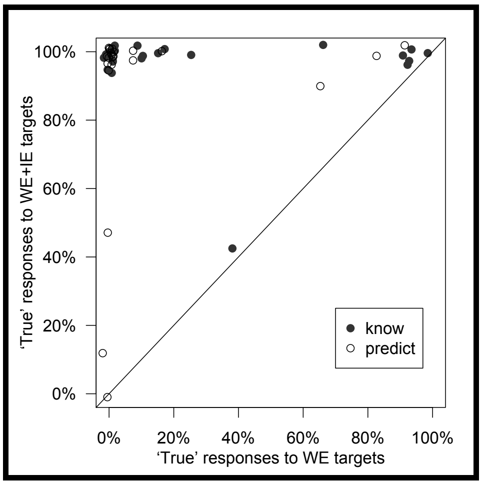
[15] For models with more than one fixed effect, the χ^2 for the simple effects corresponds to the comparison between models restricted to the intercept levels of other effects.

lead to *False* responses in both conditions. A mixed model was fit on responses to ∅ baseline and WE targets with Verb (*know* vs. *predict*), Sentence Type (Baseline vs. Target) and their interaction as fixed effects. We found a significant difference between WE targets and the baseline ($\beta = 2.4$, $\chi^2 = 39$, $p < .001$), suggesting that the WE reading was available. This effect interacted with verb: *predict* gave rise to fewer *True* responses than *know* on WE targets ($\beta = -1.0$, $\chi^2 = 11$, $p < .001$).

Fig. 1.3b presents the responses of each participant combining the two types of targets. It shows that the few *False* responses to WE+IE targets come from a small group of participants who consistently accessed an SE reading (bottom-left corner), and the few *True* responses to WE targets from a group of participants who consistently accessed the WE reading (top-right corner). The rest of the participants mostly answered true to WE+IE targets and false to WE targets consistently (top-left corner), indicative of their accessing an IE reading.



(A) Percentage *True* responses to the different conditions. WE+IE+SE baseline is true under any reading; WE+IE targets are true under the WE or IE readings (and false under the SE reading); WE targets are true under the WE reading (and false under the IE and SE readings); the ∅ baseline is false under all readings.



(B) Percentage *True* responses to WE+IE targets (i.e. estimation of the rate of WE *or* IE readings) as a function of *True* responses to WE targets (i.e. estimation of the rate of WE readings). Each dot represents one subject and a jitter was added to make clusters of similar behaviors more visible.

FIGURE 1.3 – Average and individual responses in Experiment 2. The three exhaustive readings appear to be present.

Analysis of response times

As shown by the analysis of responses, there were very few SE readings in the population. We therefore focused our response time analysis on WE and IE readings, and removed 4 participants who possibly accessed SE readings (criterion: at least 50% *False* responses to WE+IE targets). We categorized 37 of the remaining participants as IE (at least 50% *False* responses to WE targets) and 9 as WE (at least 50% *True* responses to WE targets). We then removed, as errors, responses to WE targets which did not match the category assigned to a participant.

As a result, we sought to analyze false responses for IE participants and true responses for WE participants. To control for the influence of true/false response effects on RT, we included in the model correct responses to conditions in which we expected responses that would match the to-be-analyzed responses for each participant: false responses to the $\emptyset$ baseline for IE participants and true responses to the WE+IE+SE baseline for WE participants. Fig. 1.4 shows RT for each category of participants.

We ran a linear mixed model on responses to WE targets and the appropriate controls ($\emptyset$ baseline for IE participants, WE+IE+SE baseline for WE participants) with Subject category (WE vs. IE), Sentence type (Target vs. Control) and interaction as fixed effects and a maximal random structure for items and subjects.[16] We observed a significant interaction between Subject category and Sentence type ($\beta = -0.3$, $\chi^2 = 14$, $p < .001$), suggesting that WE responses were slower than IE responses, even after controlling for true/false biases.

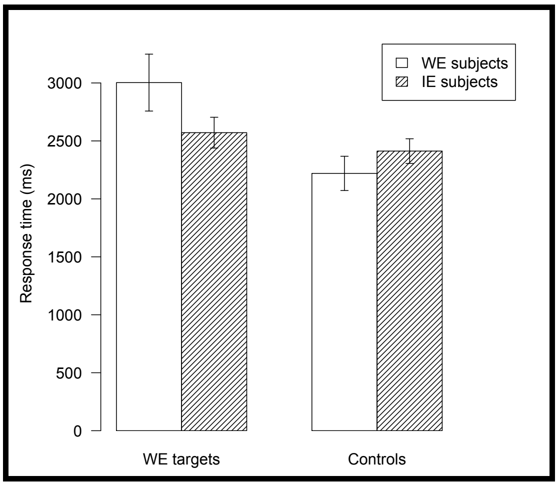


FIGURE 1.4 – Experiment 2, RT on WE targets and baselines by category of subject. IE subjects were faster than WE subjects on targets. This cannot be an effect of response as the two groups did not differ on controls of matched truth value.

1.3.5 Discussion

Discussion of the results. We observed little difference between *know* and *predict*. We confirmed that *know* supports readings other than the SE reading. IE readings were most visibly available, but some participants seem to have accessed the WE reading. We note however that this WE-like behavior could also have emerged from another phenomenon, namely covert domain restriction, as we will now explain.[17]

[16]We first ran a linear mixed model with Subject category (WE vs. IE), Sentence type (Target vs. Control) and Verb (*know* vs. *predict*) and all interactions as fixed effects. We dropped the random effects associated with verbs because the maximal model did not converge. Even then, none of the fixed effects associated with Verb was significant, and a model without these effects was not less powerful ($\chi^2(4) = 4.4$, $p = .35$), so we report the model as above.

[17]Several participants at the Euro-XPrag 2013 conference suggested another source for the appearance of responses seemingly generated by WE readings. There is a parallel between our task and the classical 'Wason selection task' (Wason, 1966): verifying that John is right only about the actual blue squares is equivalent to searching for positive evidence only in Wason's task. We do not investigate this interesting issue further because the results of the next experiment will receive a straightforward interpretation in terms of domain restriction but not in terms of restriction to positive evidence. In fact, domain restriction and restriction to positive evidence have similar practical consequences, but may be generated by different pressures.

Covert domain restriction. Covert domain restriction is a general linguistic phenomenon arising as soon as some form of quantification is involved. As an example, if we look at the dialogue in (17), we understand that B's response in (17a) is not about all the students in the world but only those students who attended the party.

Explicit restrictors usually inhibit covert restriction (see Westerståhl, 1985). For instance, adding the restrictor 'in Europe' to B's response as in (17b) feels awkward as it is now nearly impossible to reinterpret it as "the students who attended the party" (even if the party took place in Europe).

(17) A: "How was the party on Saturday?"
  a. B: "All the students were very happy."
  Possible interpretation: All the students who attended the party were very happy.
  b. B: "All the students in Europe were very happy."
  Unavailable interpretation: All the students (in Europe) who attended the party were very happy.

Now we note that, quite generally, all three exhaustive readings become equivalent if we restrict the domain of quantification to individuals who satisfy the embedded predicate (in Experiment 2, *being blue* or for our introductory example (1), *being a caller*). Abstractly, this is so because the exhaustive readings only differ on the status of individuals which do *not* satisfy the predicate (i.e. non-blue squares, or non-callers), with the WE reading being the one that puts no constraint on these individuals.

Hence, if we restrict attention to individuals who *do* satisfy the predicate, then the three readings collapse: all exhaustive readings become equivalent if the domain is restricted to individuals who do satisfy the predicate. Concretely, all exhaustive readings of (18) can be paraphrased as in (19). Furthermore, these three exhaustive readings under domain restriction end up being equivalent to the WE reading of (20) without domain restriction. Hence, what may look like WE readings of (20) without domain restriction may in fact be any of the exhaustive readings with an underlying domain restriction.

(18) John knew which of the actual blue squares were blue.

(19) For each blue square, John knew that it was blue.

(20) John knew which of the squares were blue.

Therefore, as we explained, what we considered WE readings of (20) in Experiment 2 may in fact be IE or SE readings with covert domain restriction.[18] In the next experiment, we will introduce explicit restrictors to inhibit covert domain restriction (as discussed in (17); see Chemla, 2009 for a similar move). Consequently, most of the alleged WE readings will disappear, suggesting that they were in fact due to covert domain restriction.

[18]One may wonder about the possible import of processing times at this point. Two studies (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Schwarz, 2012) suggest that, at least under some conditions, covert domain restriction is fast, while our suspect responses are not and may thus not be based on domain restriction. But if our suspect responses are WE readings, it would also be natural to expect them to be fast, given that they play the role of primitives in the relevant accounts. Hence, at this point, the RTs are surprising either way and do not provide a reliable source of information about the nature of these responses.

1.4 Experiment 3: WE reading or domain restriction?

1.4.1 Goal

The goal of this experiment was to test whether the WE readings we observed in the previous experiment were due to implicit domain restriction. To reduce the chances that participants would apply a covert restriction of the domain, we used an explicit domain restriction. Apart from this change, the experiment was fully identical to the *know*-version of Experiment 2. We decided to focus on one verb for simplicity and chose *know* because it gave rise to the highest rates of WE responses.

1.4.2 Methods and Materials

Materials

The task and materials were exactly the same as in Experiment 2, except for sentences such as "John knew which squares were blue" in which an explicit domain restriction was added, as shown in (21). If explicit restrictors block covert restriction, participants should only answer *True* to the new WE targets if they access a genuine WE reading. Therefore, if the *True* responses in the previous experiment corresponded to WE readings, adding an explicit restrictor should have no effect. On the contrary, if at least some of these responses corresponded to covert domain restriction, then we predict less *True* responses in this experiment.

(21) John knew which **of the four** squares were blue.

Participants

60 subjects were recruited on Mechanical Turk and completed the task. One was removed from the analysis for reporting to be a native speaker of a language other than English. One more participant was removed from the analyses because his behavior on targets was inconsistent (see Fig. 1.5b, participant at the bottom right). The age of the remaining participants ranged from 19 to 64 years (mean: 40) and there were 24 females and 34 males. 6 participants were removed because their error rate on controls was more than one standard deviation above the mean (the threshold was 10.7%, but all removed participants were above 18%).

1.4.3 Results

Data treatment and Statistical methods

Responses made in less than 100ms or more than 10s were discarded (1.5% of the data). The statistical methods are identical to those of Experiment 2. In particular, we also had to drop random slopes associated with sentence type (target/baseline) for subjects in GLMM.

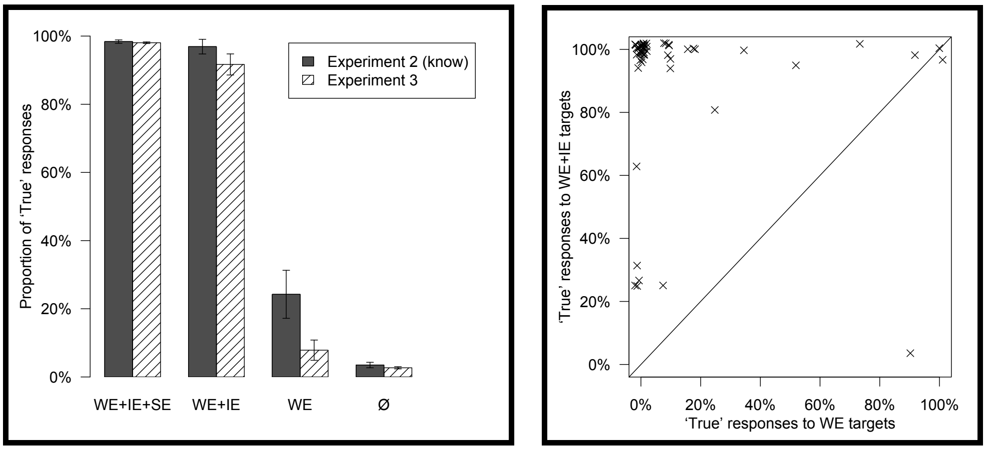
Analysis of responses

Fig. 1.5a shows the rates of *True* responses in the different conditions. It also repeats the results from Experiment 2 for comparison. Fig. 1.5b shows average responses to both types of targets for each participant in Experiment 3.

As in Experiment 2, a logit mixed model was fitted on responses to the WE+IE+SE baseline and WE+IE targets with Sentence Type (Baseline vs. Target) as a fixed effect. The effect of Sentence Type was significant: Participants gave fewer *True* responses to WE+IE targets than to the WE+IE+SE baseline ($\beta = -1.7$, $\chi^2 = 43$, $p < .001$).

A similar model was fitted on responses to the $\emptyset$ baseline and WE targets. The effect of Sentence Type was significant: participants gave more *True* responses to WE targets than to the baseline ($\beta = 1.3$, $\chi^2 = 16$, $p < .001$).

Comparison with Experiment 2



(A) Percentage *True* responses in the different conditions for *know* in Experiments 2 and 3 (see Fig. 1.3a for details).

(B) Proportion *True* responses to targets in Experiment 3 (see Fig. 1.3b for details).

FIGURE 1.5 – Average responses in Experiments 2 and 3, and individual responses in Experiment 3. The availability of the WE reading is much reduced in comparison to Experiment 2, as shown by the difference in responses to WE targets (Fig.2a).

As one can see in Fig. 1.5a, the rate of *True* responses to WE targets seems lower in this experiment than in Experiment 2, whereas the $\emptyset$ baseline does not seem to be affected. This was confirmed by a GLMM on responses to WE targets and $\emptyset$ baselines from Experiments 2 and 3 with Experiment (2 vs. 3), Sentence type (target vs. baseline) and their interaction as fixed effects. The interaction was significant: participants gave similar rates of *True* responses to $\emptyset$ baselines in both experiments, but they gave less *True* responses to WE targets in Experiment 3 ($\beta = 0.8$, $\chi^2 = 13.5$, $p < .001$).[19]

1.4.4 Discussion

Although we still observed a difference between WE targets and the $\emptyset$ baseline, the addition of an explicit restriction greatly reduced it. This confirms that the explicit restrictors we provided inhibit domain restriction (at least to some extent) and that what could have been interpreted as WE reading-based responses in Experiment 2 were probably due to SE/IE readings with covert domain restriction.[20]

[19] The same model restricted to the data for *know* in Experiment 2 gave even stronger results.

[20] One reviewer questioned this hypothesis and mentioned unpublished experimental data revealing an opposite effect with explicit restriction facilitating domain restriction. In our specific control experiment, explicit restriction did not facilitate domain restriction, because as expected the number of *True* responses on WE targets decreased. The circumstances under which explicit restriction facilitates or inhibits domain restriction would be a topic for closer scrutiny.

The residual *True* responses to WE targets are rare. They may be based on genuine WE readings or some remaining possibility for covert restrictions.[21] In our last experiment, we will focus on the comparison between the IE and SE readings.

1.5 Experiment 4: IE vs. SE readings of *know*

1.5.1 Goal

The existence of IE readings for *know* being established, we wanted to compare the processing properties of the IE and SE readings. Therefore we focused on the acquisition of reliable online data about the two readings. This goal was achieved using a design with training. We ignored WE readings because they proved to be rare and hard to identify as such (see previous discussion).

1.5.2 Methods and Materials

Course of the experiment

This experiment was very similar to Experiment 2 (and, therefore, to Experiment 3 as well). The main difference concerned the division of the participants into two groups, dubbed "SE-group" and "IE-group". The training session was longer and the corrective feedback during training was more detailed and adapted to each group. It was designed to bias participants toward the SE reading or the IE reading.

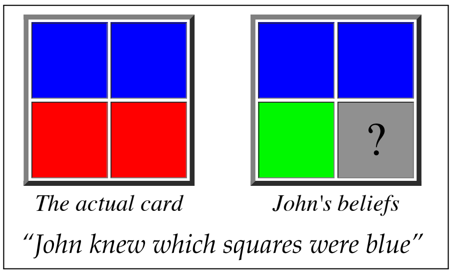
Bott and Noveck (2004) used a similar method in one of their experiments on scalar implicatures. The main advantage is to provide a 50/50 share of responses corresponding to each reading, and thus better online data. There are some differences between our design and that of Bott and Noveck (2004). First, we did not have each participant take the experiment with both types of training, so the training is a between-subject factor in our experiment. Second, we used the same instructions for both training groups (the same as in previous experiments with *know*). The only difference between the two groups was implemented in the feedback they received during the training phase.

Materials

As mentioned above, the instructions were identical to the previous experiment. The training session was longer (28 items), so that subjects would have enough time to get the intended reading, and it now included 6 WE+IE targets. See (22) for an example of a training target and (23) for an example of a WE+IE+SE item (true baseline), both with their associated feedback. The feedback was adapted to the training group. On controls it was triggered by errors, and on WE+IE targets it was triggered by responses which did not match the intended reading: subjects from the SE group received feedback when they answered *True*, whereas subjects from the IE group received feedback when they answered *False*. The training phase also included four WE targets (see 24), on which we trained participants to answer false to fully prevent WE readings. The feedback was always a paraphrase of the reading we wanted the participants to access.

(22) WE+IE target (boldface added to stress differences between the IE and SE feedbacks)

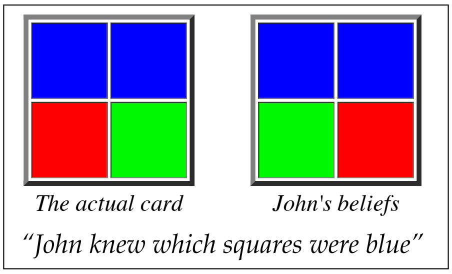
[21] They may also be Wason type errors; see footnote 17.



a. SE feedback: It was **false**. The top squares were blue and he knew it. The bottom squares were not blue **but he didn't remember whether they were blue or not.**

b. IE feedback: It was **true**. The top squares were blue and he knew it. The bottom squares were not blue **and he didn't remember them as blue.**

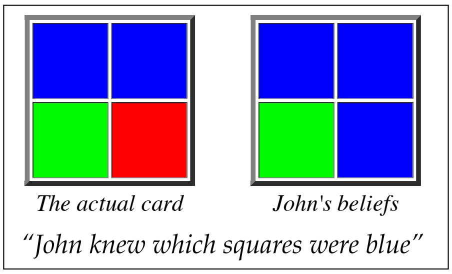
(23) WE+IE+SE baseline item



a. SE feedback: It was true. The top squares were blue and he knew it. The bottom squares were not blue **and he knew they were not.**

b. IE feedback: It was true. The top squares were blue and he knew it. The bottom squares were not blue **nor did he remember them as blue.**

(24) WE target (treated as false controls in this experiment)



a. SE feedback: It was false. John knew that the top squares were blue. **He didn't know that the bottom right square was not blue.**

b. IE feedback: It was false. John knew that the top squares were blue. **He remembered the bottom right square as blue when in fact it was red.**

Participants

Two groups of 40 subjects were recruited on Mechanical Turk. One participant did not complete the experiment and one was removed for not being a native speaker of English. Remaining subjects ranged from 18 to 66 years (mean 32). There were 38 females and 40 males.

For the response time analyses, 8 participants were removed because their error rate on controls was more than one standard deviation above the mean (threshold: 13.4%). We also removed 6 participants whose error rate on targets (with respect to the expected reading for their group) was at least one standard deviation above the mean (threshold: 36.3%). Finally, 2 subjects were removed because they had at least twice more *True* responses to WE targets than errors on controls; therefore we suspected they accessed a WE reading (or covert domain restriction).

1.5.3 Results

Data treatment and statistical methods

As in previous experiments, responses made in less than 100ms or more than 10s were discarded (1.9% of the data).

For response time analyses, we only looked at control sentences which did not involve embedded questions (i.e. generated on the template in (15b)). Indeed, on sentences which embed questions, SE and IE subjects should access different readings. Hence if there is a processing difficulty associated with one reading, it should affect these controls too. Controls with simple sentences on the other hand should not be affected by the training.

Categorical True/False responses were not thoroughly analyzed, relevant differences were simply assessed with t-tests on arcsine-transformed proportions of True responses.

Mixed models on response times were fitted after log transformation, to comply with the homoscedasticity hypothesis. We give the t-statistics and the p-values obtained by treating these t's as z-statistics (this is a reasonable approximation with 40 subjects per condition).

Analysis of responses

The training included 6 WE+IE targets. Fig. 1.6a shows the responses of subjects from both groups on these targets, as a function of their order of presentation during the training phase. Although participants from both groups started with a high rate of *True* responses (they likely accessed an IE reading), one can see that participants in the SE group quickly switched to *False* responses, therefore converging on the intended SE reading.

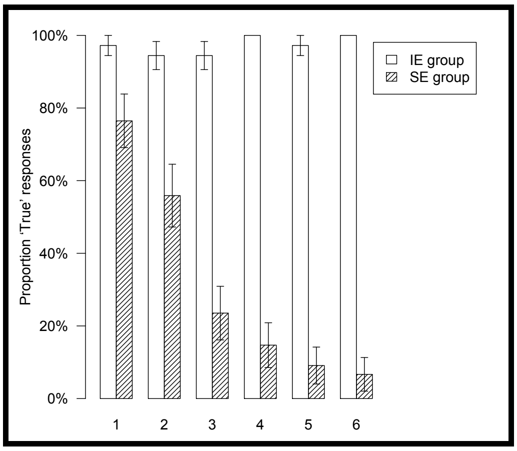
Interestingly, participants in the SE group gave significantly fewer *True* responses than participants in the IE group even on the very first target of the training phase (76% vs. 97%, $t(42) = 2.6$, $p = .012$). This reveals that a quarter of the participants in the SE group switched to the SE reading just by reading the appropriate paraphrase in the feedback of control items, even though the truth value of the SE reading was irrelevant for these items. In short, these participants did not need the feedback on WE+IE targets which explicitly associated these items with a *False* response.[22]

Fig. 1.6b shows the average responses to WE and WE+IE targets in the experimental phase for each participant (including participants with some WE responses or high error rate on targets who were removed from the RT analyses). Most participants fall in the category corresponding to their group and no participant displayed a predominant WE pattern.

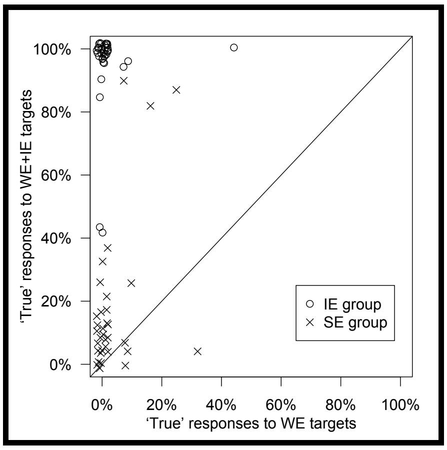
Response times

Response times in a picture-sentence matching task can be affected by features of both the pictures and the sentences, as well as interactions between them. For instance, a sentence such as "John knew which of the top squares were blue" will

[22]Participants in the SE group saw 3 control items (median value) before the first target, and 13 of the 34 SE-participants (38%) made at least one error before encountering their first target. We observed a trend for a correlation between the number of errors made before the first target and the chances of getting the SE reading without explicit feedback: each error on previous controls increased the chance of giving a *False* response to the first target by 18% ($t(32) = 1.8$, $p = .09$).



(A) Training phase: Percent *True* responses on WE+IE targets from the training phase as a function of their order of appearance.



(B) Experimental phase: Proportion *True* responses to targets (see Fig. 1.3b for details).

FIGURE 1.6 – Experiment 4: Most participants started with the IE reading, but participants in the SE-group quickly converged on the SE reading. In the experimental phase, most participants fell in the intended category.

probably draw more attention to the top row in the picture. Pictures in which John was ignorant about one square could have been read differently depending on the sentence associated with them, but also depending on the training. For instance, it may be easy for a participant of the SE group to answer *False* when there is a question mark in the display.

Fig. 1.7 shows response times to targets and different types of control items which did not involve embedded questions. We separated control sentences by truth value and by presence/absence of a question mark in the picture.

We ran a mixed model on control sentences with truth value (true vs. false), presence or absence of a "?" in the display, training (IE vs. SE) and all possible interactions as fixed effects. The random effect structure was maximal for subjects and items. No effect reached significance (all t's< 1.5). If anything, there was a trend for a main effect of truth value (false sentences taking longer than true sentences) and an interaction between the presence of "?" and the training.

We ran a second mixed model in order to compare responses to WE+IE targets with responses to controls. Since targets always had a "?" (a crucial feature to distinguish SE and IE readings), we compared them with controls that also had a "?". The fixed effects were Answer (*True* vs. *False*)[23], Sentence type (control vs. target) and their interaction. Both main effects came out insignificant ($t < 1$), but the interaction was significant ($\beta = .12$, $t = 2.7$, $p = .007$). Targets incurred an extra delay for SE subjects, and this cannot be just an artifact of the presence of "?" in the display (otherwise the effect would not interact with the content of the sentences).

[23]On controls, the answer parallels the truth value of the sentence. On targets it parallels the training of the subject (*True*=IE, *False*=SE). Therefore, Answer is sufficient to encode the training. Since Answer is only used to encode well defined experimental conditions here, it does not fall into the category of "bad controls" for a regression on RT.

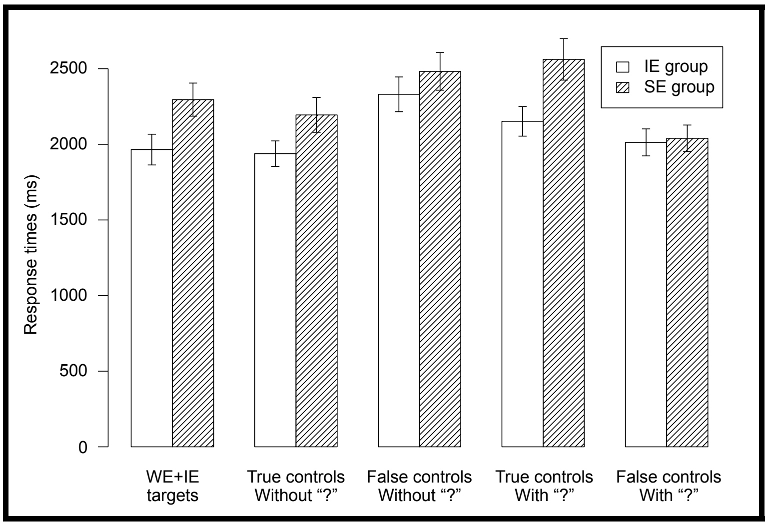


FIGURE 1.7 – Experiment 4: Response times (ms) to targets and control items, as a function of training. Control items are separated by truth value and presence/absence of a question mark in the picture (representing ignorance). WE+IE targets all had a "?" in their display since it is crucial to differentiate IE and SE readings.

1.5.4 Discussion

We saw that the training had a very strong effect on participants in the SE group, who quickly converged on SE behavior. One may worry that the participants' responses in this experiment reflect some artifact induced by the training procedure more than genuine linguistic judgments revealing the existence of SE readings. Let us provide two arguments to alleviate such a worry before discussing the consequences of our results. First, we can show that participants in the SE group did not generalize the explicit training on WE+IE targets ('respond *False*') to other experimental items with similar displays. One specific and possibly salient feature of the display in the WE+IE condition is the presence of a question mark, but the SE participants accurately responded *True* on true controls where the display contained a question mark (94%), showing that participants did not simply respond *False* based on the mere presence of this visual cue. Second, on training items which were not WE+IE targets participants received feedback which consisted of paraphrases of the reading we intended to convey to their group even though the expected response was the same in both groups. We saw that a significant proportion of the participants adopted an SE pattern of behavior before being told explicitly to treat WE+IE targets as false items, thus suggesting that the paraphrases played an important role in the training. This would not be the case if participants' responses were only due to association between a feature of the display and a given response.

Overall, the offline data from this experiment (training and experimental phases) confirm that the IE reading is very salient in this paradigm, but the fact that a quarter of the participants in the SE group switched to the SE reading before receiving explicit feedback suggests that the SE reading is naturally available as well.

The online data indicate that SE responses took longer than IE responses, after controlling for effects of the display and truth value. A first interpretation is that the derivation of the SE reading is more complex, or even that it is derived from the IE reading. However the sustainability of such a conclusion is weakened by the asymmetry between the salience of the two readings in this paradigm. In fact, an alternative interpretation of the online results would be that they simply reflect the salience of the IE reading and an effect of training. Participants in the IE group may not even be aware of the SE reading. Participants in the SE group, on the other hand, must be aware of both readings since they started the training with the IE reading and switched to the SE reading. Therefore the extra processing cost for the latter group may reside in an ambiguity made salient for them between a true and a false reading.[24] Let us note however that if we take the online data to reflect the salience of the IE reading, it strengthens the main conclusion we want to draw from the offline data, namely that the IE reading is naturally available.

1.6 General discussion

1.6.1 Summary of the results

In this series of experiments we showed that the SE reading is not the only reading available for *know* (Experiment 1). It seems that the most prominent reading in this task is the intermediate IE reading proposed by Spector (2005) (Experiments 2, 3, 4). Overall, we found fairly small traces of the WE reading, and we showed how this evidence was polluted by the possibility of covert domain restriction (Experiment 3).

The relative proportions of the different readings differed slightly for the verbs *predict* and *know*, but the general pattern was the same. Crucially, the IE reading was available for both verbs (Experiment 2).

Finally, the online results indicate that the IE reading is faster to access than the SE reading in our task. At the very least, this confirms the existence of the IE reading. This processing asymmetry may merely reflect the offline preference for the IE reading in our paradigm, which would then count as a confirmation of the conclusions obtained from the offline results. But, more tentatively, it may also serve as the basis for developing theoretical models informed by processing results, such that the operations leading to the IE readings could be proved to be simpler than (or a subset of) the operations leading to the SE readings.

1.6.2 Consequences for the theories

The availability of exhaustive readings beyong the SE reading contradicts the predictions of theories we categorized in the second group (see section 1.4(a)). These correspond to theories that rely on *strong* answers as the basic denotation of a question and can only derive strong readings for embedded questions (Groenendijk and Stokhof, 1982, 1984, 1993; B. R. George, 2011). On the other hand, theories from the first group, relying on *weak* answers derive the WE readings as basic and may, depending on the specifics of how the other parameters are set, derive IE and SE readings (Heim, 1994; Beck & Rullmann, 1999; Guerzoni & Sharvit, 2007; Klinedinst & Rothschild, 2011; Spector & Egré, 2015). Currently, the only two theories which explicitly implement IE readings, which were found to be salient, are those by Spector and Egré (2015) and Klinedinst and Rothschild (2011). In both

[24]This hypothesis is consistent with a wide range of processing approaches, including sequential sampling process models of decision making (e.g. Ratcliff, Gomez, & McKoon, 2004; Ratcliff, 2002, 1979; Tuerlinckx, 2004; Busemeyer & Rapoport, 1988; Vickers, 1979; Pike, 1973, 1968, 1966); it may be directly linked to classical results in cognitive psychology which show that when distinct aspects of a given stimulus push subjects in different directions, their decision process takes longer, regardless of the actual outcome.

cases, the way to obtain these IE readings for factive verbs such as *know* is to rely on a separation between the assertive and presuppositional components of the verb (Spector and Egré implement this separation between assertion and presupposition explicitly, while Klinedinst and Rothschild only provide guidelines on how this could be done).

The current results thus provide constraints for future theories of embedded questions. They suggest that (1) weak answers are primitives and (2) IE readings exist. This latter result argues for the possibility of including in the semantic system the tools to decouple presuppositional and assertive content.

To conclude, let us mention two potential questions for future research. First, question-embedding predicates can vary along many dimensions (e.g., presuppositionality but also polarity). Such properties play a different role in current theories of embedded questions, such that one should next test a wider variety of verbs using comparable methods. Such data would inform us on the plausible mechanisms for combination between a responsive verb and a question (see section 1.4(b)). Second, we have reported on the first processing data concerning embedded questions. But at this point it is hard to determine completely what our data reflect and what theories actually predict in terms of processing. Further processing data and explicit implementations of the different theories provide a direction for future research. We hope to have shown that these are not out of reach.

Additional discussion

Since the publication of this experiment, other theories have been proposed to derive IE and SE readings. Uegaki (2015) follows Klinedinst and Rothschild (2011) and derives the IE reading as a strengthening of the WE reading. He departs from their theory when it comes to SE readings however. Rather than local exhaustification, he suggests that SE readings are the result of a second strengthening step, akin to neg-raising. The processing results of Experiment 4, if confirmed, would provide a strong argument in favor of such approaches. Roelofsen et al. (2014) propose a theory which shares more features with Spector and Egré (2015), in that they assume an ambiguity between an IE and a SE question operator.

The experiment presented in the next chapter will look at the acquisition of exhaustive readings. A prediction of Klinedinst and Rothschild (2011) and Uegaki (2015) is that children should acquire exhaustive readings the way they acquire implicatures, that is beginning with the 'literal' WE reading (however, it is unclear what prediction ambiguity theories would make here). The data from adult controls will also provide a replication of Experiment 2 with very different stimuli.

Chapter 2

Children's exhaustive readings of questions

Cremers, A., Tieu, L., & Chemla, E. (2016). *Children's exhaustive readings of questions*. Accepted for publication in Language Acquisition †

Abstract

Questions, just like plain declarative sentences, can give rise to multiple interpretations. As discussed by Spector and Egré (2015), among others, questions embedded under *know* are ambiguous between *weakly exhaustive* (WE), *intermediate exhaustive* (IE), and *strongly exhaustive* (SE) interpretations (for experimental evidence of this ambiguity, see Cremers and Chemla 2014). These three interpretations are related in terms of strength. The SE reading entails both the IE and WE readings, and the IE reading entails the WE reading. Certain proposals derive the stronger readings from weaker ones through the same process of enrichment that underlies scalar implicatures, in particular through comparison with alternatives (Klinedinst and Rothschild 2011). Given previous developmental studies of scalar implicatures that suggest children typically perform this enrichment less often than adults do (Noveck 2001; Chierchia, Crain, Guasti, and Thornton 2001; Papafragou and Musolino 2003, among many others), such proposals lead us to expect that children may initially prefer weak readings of embedded questions. The present study revealed that 5-year-olds were sensitive to the multiple readings of questions embedded under *savoir* 'know'. Compared to adults, however, children were more tolerant of weak readings. These findings relate scalar implicatures and exhaustive readings of embedded questions from a developmental perspective, and are consistent with a close connection between the two: in both cases, children are sensitive to the various possible interpretations but favor the weaker one more than adults do.

† For helpful feedback and discussion, we would like to thank Paul Egré, Martin Hackl, Valentine Hacquard, Jeffrey Lidz, Jacopo Romoli, Philippe Schlenker, Benjamin Spector, Alexander Williams, and audiences at the 6th Generative Approaches to Language Acquisition in North America conference, the Ecole Normale Supérieure, the University of Ulster, and Harvard University. We would also like to thank Anne-Caroline Fievet for coordinating testing at local preschools, and Florian Pellet for implementing the adult version of the task. We are also grateful to the families of the children who participated in our study, as well as the director and teachers at the École Maternelle Publique Lyonnais, who graciously allowed us to conduct our study in their school. The research leading to this work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610, and by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

2.1 Introduction

2.1.1 Strengthened interpretations in child language

The question of how the child learner arrives at adult-like interpretations for sentences that allow for multiple interpretations in different contexts has been the focus of many developmental studies. Such studies, which document the meanings that children assign to different structures and how these may differ from those of adults, are insightful because they allow us to better understand the developmental path children may take towards the adult linguistic system, and ultimately inform us about the components of this mature system itself (one such example corresponds to the vast literature on scope ambiguity, e.g., Musolino 1998; Musolino, Crain, and Thornton 2000; Musolino and Lidz 2003, 2006; Gualmini 2004; Gualmini and Crain 2005; Musolino and Lidz 2006; Miller and Schmitt 2004; Gualmini, Hulsey, Hacquard, and Fox 2008; Conroy, Lidz, and Musolino 2009; Zhou and Crain 2009; Notley, Zhou, Jensen, and Crain 2012; Crain, Goro, Notley, and Zhou 2013).

One highly productive line of research in the developmental literature has highlighted differences between children's and adults rates of derivation of *scalar implicatures*, such as the one that arises in (1) (Noveck 2001; Chierchia et al. 2001; Gualmini, Crain, Meroni, Chierchia, and Guasti 2001; Papafragou and Musolino 2003; Barner, Brooks, and Bale 2011; Katsos and Bishop 2011, among many others).

(1) Some of the horses jumped over the fence.
  a. Some or all of the horses jumped over the fence.
  b. Some but not all of the horses jumped over the fence.

Scalar implicatures arise as the result of taking into account alternative phrases that could have been uttered but were not. Assuming the speaker is being as informative as she can (Grice 1975), the fact that the speaker uttered the assertion containing *some* (1) and not the stronger alternative containing *all* (i.e. *All of the horses jumped over the fence*), can lead us to conclude that the stronger alternative must be false (for modern versions see, e.g., van Rooij and Schulz 2004; Sauerland 2004; Spector 2007; Franke 2011).

Alternatively, one may capture the strengthened meaning in (1b) through the application of a grammaticalized exhaustification operator EXH, which would be roughly equivalent to a silent *only* (Fox 2007; Chierchia, Fox, and Spector 2011), as schematized in (2).

(2) EXH(Some of the horses jumped over the fence) =
Some of the horses jumped over the fence and NOT(all of the horses jumped over the fence)

Although this is a topic that has generated massive interest, we would like to make it clear from the outset that our study will not address the debate between neo-Gricean and grammatical approaches to scalar implicatures. Instead, we will be interested in whether acquisition studies can help to determine whether scalar implicatures are related to another phenomenon, namely the interpretations of embedded questions. We will therefore abstract as much as possible away from the differences between the alternative perspectives on the two phenomena. For convenience, we may use the more flexible EXH notation from the grammatical approach (which can also be taken as a proxy for a neo-Gricean post-semantic operation, as in original approaches à la Groenendijk and Stokhof 1982 or van Rooij and Schulz 2004).[1] We remain non-committal, however, and the presentational choices that our notation implies should be translatable from one framework to another without affecting our main conclusions.

With this in mind, let us return to the situation in the context of the *acquisition* of scalar implicatures. Although there is some hetereogeneity across previous studies, a generally robust finding has been that children compute fewer implicatures than adults. While there is an extensive literature characterizing children's performance on different kinds of scalar implicatures (e.g., involving the scalar quantifiers *some/all*, the modals *may/must*, connectives *or/and*, numerals, etc.), there have only been a handful of detailed proposals for why children might differ from adults, and perhaps more importantly, how they might eventually arrive at an adult-like ability to compute enriched meanings.

Katsos and Bishop (2011), for example, propose that children are more pragmatically tolerant than adults. On their proposal, children are sensitive to the difference between weak and strong forms, and are competent with informativeness. Where they differ from adults lies in the degree to which they are tolerant of pragmatic infelicity. An alternative proposal attributes children's failures to compute scalar implicatures to a difficulty accessing the required alternatives. For example, to compute the strengthened meaning of (1), i.e. (1b), children must be able to access the lexical item *all* as a stronger alternative to *some* (for relevant discussion, see Gualmini et al. 2001; Chierchia et al. 2001; Reinhart 2006; Barner et al. 2011; Singh, Wexler, Astle, Kamawar, and Fox 2015; Tieu, Romoli, Zhou, and Crain 2015). This line of proposals derives in part from the observation that children tend to perform more on a par with adults once the relevant scalar alternatives are made explicit.

The issues that arise in the developmental study of scalar implicatures are far from resolved, and much work remains to be done in terms of fully working out the predictions and consequences of existing proposals. Here we will capitalize on the empirical findings on the topic and use them to conduct a parallel investigation of a phenomenon that has been proposed in the theoretical literature to be related to implicature. The multiple interpretations of embedded questions, to be introduced in the next section, have not been addressed in any previous studies of language acquisition. Given recent theoretical proposals concerning a connection between these questions and scalar implicatures, a developmental study will provide a useful complement to the existing developmental literature on scalar implicatures, and may ultimately be informative about how these phenomena should be analyzed in both child and adult language.

2.1.2 The different interpretations of embedded questions

Like the scalar implicature example in the previous section, questions also give rise to multiple readings. And just as in the case of scalar implicature, studying the comprehension of such structures can shed light on key differences between

[1] We may also talk about *ambiguity*, which should not be understood as representing any particular bias in favor of a grammatical perspective, despite the term being more commonly used in the context of lexical, syntactic, or semantic ambiguity, than in pragmatic or neo-Gricean terms. We also use the term *reading*, which is the dominant terminology in the literature on embedded questions, but the user may replace it with *interpretation* if it creates difficulties.

the child learner and the adult. Let us introduce the various possible interpretations on the basis of example (3), which contains a *wh*-question (i.e. *which toys are in the box*) embedded under the attitude predicate *know*. Questions such as (3) have been claimed to potentially have at least three interpretations: a *weakly exhaustive* reading (WE), an *intermediate exhaustive* reading (IE), and a *strongly exhaustive* reading (SE). These readings are logically related: the SE reading entails the IE reading, which in turn entails the WE reading.

(3) Jack knows which toys are in the box.

a. . *Weakly exhaustive (WE) reading*:
For each toy in the box, Jack knows that it is in the box.

b. . *Intermediate exhaustive (IE) reading*:
For each toy in the box, Jack knows that it is in the box,
and Jack does not have false beliefs about toys that are not in the box.

c. . *Strongly exhaustive (SE) reading*:
For each toy in the box, Jack knows that it is in the box,
and he knows that no other toys are in the box.

Concretely, consider a situation in which there are toy trains and building blocks in the box, and toy boats on the floor. For the WE reading in (3a) to be true, all that is required is that Jack have true beliefs about the toys that are in the box, namely the trains and the blocks. It may be the case that Jack mistakenly believes that the boats are also in the box. The WE reading nevertheless comes out true. For the IE reading in (3b) to be true, we require not only that Jack have true beliefs about the toys that are in the box, i.e. the trains and the blocks, but that he furthermore have no false beliefs about toys that are not in the box, i.e. the boats. It may be the case that Jack has no idea where the boats are. The IE reading would nevertheless come out true. Finally, for the SE reading in (3c) to be true, we require that Jack have complete exhaustive knowledge about the toys that are in the box. In other words, Jack must correctly believe that the trains and blocks are the only toys in the box.

Theoretical background

The availability and distribution of the WE and SE readings have been the subject of extensive debate in the theoretical linguistics literature. On the one hand, Groenendijk and Stokhof (1984) and B. R. George (2011) treat the SE reading as basic and do not derive any weaker readings. On the other hand, Heim (1994), Beck and Rullmann (1999), Sharvit (2002) and Guerzoni (2007) treat the WE reading as basic, and derive the SE reading with a strengthening mechanism. Spector (2005) reintroduces the IE reading (which was quickly discussed in Groenendijk and Stokhof 1982 and Preuss 2001), reviving the debate over exhaustive readings. More recently, Cremers and Chemla (2014) provide experimental evidence for the existence of this IE reading. Only recent theories are able to derive the IE reading. These can be split between theories in the tradition of Heim (1994), which treat the WE reading as basic and derive IE and SE readings from it (Klinedinst and Rothschild 2011; Uegaki 2015), and theories that treat IE and SE readings on a par and do not derive the WE reading at all (Roelofsen et al. 2014; Spector and Egré 2015).

A concrete implementation

Most relevant for our purposes is Klinedinst and Rothschild's (2011) proposal (and those that follow it), which treats WE readings as primitive, following Heim (1994). This proposal derives stronger readings through an enrichment process directly imported from theories of scalar implicatures. Let us briefly sketch one such way of deriving the interpretations in (3) (for details, see Klinedinst and Rothschild 2011 and Uegaki 2015).

First, we can take the denotation of a question such as *Which toys are in the box* to correspond to the set of propositions of the form "x is in the box" where x is a toy. Next, we must combine the denotation of our question with the semantics of the verb *know*. We can assume that 'knowing Q' is equivalent to knowing the conjunction of all true members of Q (Karttunen 1977; Berman 1991; Heim 1994). Take our example above, where the set of possible answers to Q (*Which toys are in the box*) corresponds to {The trains are in the box, The blocks are in the box, The boats are in the box}.[2]

Only the first two propositions are true in our toy scenario. Knowing *which toys are in the box* thus amounts to knowing the conjunction of the two propositions *The trains are in the box* and *The blocks are in the box*. In other words, on the WE reading, Jack knows which toys are in the box by knowing that the trains and the blocks are in the box.

Treating the WE reading as basic, we can show that with an appropriate choice of alternatives, the IE and SE readings can be derived just like scalar implicatures. First, we can assume that the alternatives for *know* are obtained from the scale {*know*, *believe*} (as discussed in the literature on a so-called 'anti-presupposition' of *believe*, e.g., Percus 2006; Sauerland 2008), and second, that the alternatives of *Which toys are in the box* correspond to the set of all propositions 'x is in the box', true or false. Simplifying over Klinedinst and Rothschild (2011), we show how the IE and SE readings can be obtained through the application of an exhaustivity operator at different positions in the structure. The WE reading in (3a) is primitive, and is derived compositionally through the combination of the semantics of *know* and the denotation for the embedded wh-question (4a). The IE reading is obtained by adding to the WE meaning the negation of all the alternatives of the form "Jack believes that y is in the box", where y is a toy that is *not* in the box (4b). The SE reading is obtained by enriching the embedded part "x_1, x_2... are in the box" with the negation of the alternatives "y is in the box", and then composing this with the embedding verb, as illustrated in (4c).[3]

[2] Strictly speaking, our definition requires that the alternatives be of the form "x is in the box", where x is some individual toy, e.g., the blue toy train on the carpet. For simplicity, however, we lump together toys of the same type, e.g., toy trains. This move does not affect the proposal in any relevant way, as long as the predicate has the same value for all toy trains (e.g., they are all in the box or all out of the box). This will also make it easier to carry over the examples described here to the experimental conditions described in Section 2.2.1.

[3] As we mentioned, the presentation above is framed using an EXH operator. For questions, it is also the case that dominant theories are phrased in these terms, partly for reasons of notational convenience but also for non-trivial reasons: as the reader may have noted, in order to derive the SE reading, EXH appears in an embedded position; the possibility of embedding is often seen as the hallmark of the grammatical nature of such a strengthening operation. As in more standard cases of seemingly embedded scalar implicatures, however, alternative ways of deriving this reading have been proposed. For instance, one might posit global exhaustification, complemented by further epistemic reasoning about the resulting interpretation. For details along these lines regarding standard cases of scalar implicatures, we refer the reader to Russell (2006) or Spector (2006). A natural candidate for the implementation of similar ideas in the domain of questions can be found in Uegaki's

(4) a. *WE reading*
⟦Jack knows which toys are in the box⟧ =
Jack knows that $x_1, x_2 \ldots$ are in the box (for all x that are actually in the box).

b. *IE reading*
⟦EXH(Jack knows which toys are in the box)⟧ =
Jack knows that $x_1, x_2 \ldots$ are in the box (for all x that are actually in the box),
and NOT(Jack believes that y_1 is in the box),
and NOT(Jack believes that y_2 is in the box)... (for all y that are *not* in the box)

c. *SE reading*
⟦Jack knows EXH(which toys are in the box)⟧ =
Jack knows that...
$x_1, x_2 \ldots$ are in the box (for all x that are actually in the box),
and NOT(y_1 is in the box),
and NOT(y_2 is in the box)... (for all y that are *not* in the box)

Summary

What is relevant for our purposes is that stronger interpretations of questions can be seen as the result of applying the strengthening mechanisms involved in classical scalar implicatures (independently of the semantic or pragmatic nature of these mechanisms). Given previous findings that children access strengthened meanings less often than adults do in the case of scalar implicatures, proposals like Klinedinst and Rothschild (2011) and Uegaki (2015) lead us to expect that children may initially display a preference for the WE reading in the case of (3), just as they prefer a weak reading in the case of (1). The predictions of competing theories of questions such as Theiler (2014), Roelofsen et al. (2014), and Spector and Egré (2015), on the other hand, are very different. First, these theories do not derive the WE reading at all, so children are certainly not expected to access it. Second, these theories treat the availability of different interpretations as a choice between different lexical rules and operators, and therefore do not offer predictions as to which readings should be preferred or more difficult for children.

In the next section, we describe an experiment designed to investigate children's comprehension of embedded questions such as (3).

2.2 Experiment

2.2.1 Method

Our experiment was conducted in French, but the materials will be described here in English. The original French materials can be found in the Appendix.

(2015) theory, which proposes to derive the SE reading from the IE reading through further pragmatic reasoning (but see also difficulties mentioned therein for a plain neo-Gricean interpretation and implementation of this work). Abstracting away from these details, recall that our interest here lies in the comparison between scalar implicatures and exhaustive readings of embedded questions from the perspective of acquisition, and this issue is to a large extent independent of the particular implementation of the proposed underlying strengthening mechanism, as long as the components of this implementation can be translated from one domain to the other.

Participants

We tested 35 French-speaking children (19 female) between the ages of 4;03 and 6;04 (mean age 5;08) in the ENS Babylab and at a local preschool in Paris. All child participants were acquiring French as a first language. We also tested 23 adult native speakers of French using a web-based version of the experiment, which included all the same materials that children saw. Adult participants were recruited through the online platform FouleFactory and were paid 2.50 € for participating.

Procedure

We designed a question-answer task involving a puppet named Zap. Participants were introduced to Zap and told that they would play a game with him on the computer. Pre-recorded videos of Zap's utterances created the ruse that he was participating live via webcam. Participants saw a series of cartoon images on a laptop computer. Each image contained two sets of objects on the screen, for example, a set of toy trucks and a set of toy boats. Participants were told that Zap too could see the image. Zap was then instructed to put on a blindfold, so that he could no longer see the image (see Figure 2.1). He was then asked to recall where each of the sets of objects was in the picture, e.g., through the prompts, *Where are the trucks, Zap? Where are the boats, Zap?* Children were asked to repeat each of the puppet's answers, to ensure that they had correctly heard and understood the puppet's utterances, and could recall where the puppet thought each of the sets of objects was. At the end of each trial, the participant was asked a question about Zap's knowledge, given what he had reported in the two previous sentences, e.g., *Does Zap know which toys are in the box?* Children received a scorecard booklet, in which they had to place a rubber stamp in colored boxes based on their responses to the final question (Figure 2.2), i.e. yellow for *yes* and blue for *no*.[4]

All children were tested individually in a quiet room away from their peers. Their responses were videorecorded for later analysis. Adult control participants were tested on the same materials but through a web-based version of the task. Importantly, adults saw the same pictures and videos of the puppet as the children did, in the same order. Instead of providing oral responses and stamping on a scorecard, however, adults provided their responses by clicking on appropriate yes/no buttons.

Materials

There were six conditions in total. Each participant saw three trials per condition, for a total of eighteen experimental trials, presented in one of two pseudorandomized orders. These eighteen trials were preceded by two training items: on one practice trial, the puppet uttered an obviously true belief, and on the second, the puppet uttered an obviously false belief. The purpose of the practice trials was to show that the puppet was capable of uttering both true and false statements, and more generally to familiarize the participants with the task.

[4]We chose two neutral colours rather than, for example, a checkmark and an X, or a happy face and a sad face. This was to encourage the child to evaluate the puppet's knowledge objectively, without any interfering desires to punish or reward the puppet. For example, we did not want the child to associate a *yes*-response to the critical question with good performance on the part of the puppet, or conversely, a *no*-response with poor performance on the part of the puppet.

FIGURE 2.1 – Zap the puppet.

FIGURE 2.2 – Children's scorecards.

WH test conditions There were four test conditions involving an embedded *wh*-question such as *Does Zap know which toys are in the box?* As described in (5), these conditions varied in which exhaustive readings (WE, IE, SE) were made true. Figure 2.3 provides an example test image. Table 2.1 contains the corresponding puppet statements in relation to this test image, as well as the expected responses on each reading, for each condition.

(5) WH test conditions

a. **Baseline ∅ condition**
Zap was wrong about the location of both sets of objects. All participants were expected to provide *no*-responses in this condition.

b. **WE condition**
Zap knew the location of the first set of objects but was wrong about the location of the second set, compatible with a WE reading only.

c. **WE+IE condition**
Zap knew the location of the first set of objects, but did not know about the location of the second set. Crucially, he had no false beliefs about the second set of objects. This condition was compatible with the WE and IE readings, but not with the SE reading.

d. **WE+IE+SE condition**
Zap knew where both sets of objects were, thus all readings were made true. All participants were expected to provide *yes*-responses in this condition.

In short, participants' responses would indicate the strength of their exhaustive readings of the embedded *wh*-questions: depending on the readings accessed, a participant would provide *yes*-responses to a different subset of the ordered conditions above.[5]

[5]An anonymous reviewer points out that the elliptical "I don't know" in the WE+IE condition could have given rise to alternative interpretations, such as "I don't know whether the boats are in the box (or not)" or "I don't know whether the boats are in the box or on the carpet", as opposed to "I don't know where the boats are." Note, however, that the "I don't know" was always uttered in response to a specific question, e.g., "Where are the boats?" such that proper ellipsis resolution should force the interpretation, "I don't know where the boats are." Note furthermore that even if a child for some reason accessed one of the alternative interpretations raised by the reviewer, this

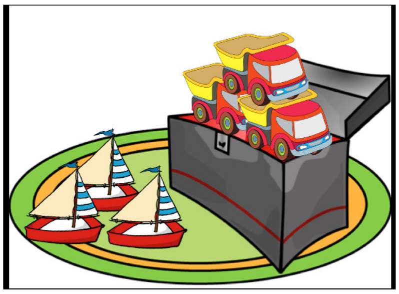


FIGURE 2.3 – Image accompanying the test question *Does Zap know which toys are in the box?*

Condition	Puppet's statements	WE	IE	SE
∅	The trucks are on the carpet. The boats are in the box.	*no*	*no*	*no*
WE	The trucks are in the box. The boats are also in the box.	*yes*	*no*	*no*
WE+IE	The trucks are in the box. The boats, I don't know.	*yes*	*yes*	*no*
WE+IE+SE	The trucks are in the box. The boats are on the carpet.	*yes*	*yes*	*yes*

TABLE 2.1 – WH conditions: Table displays puppet's statements for each condition (in relation to Figure 2.3), along with expected responses on each interpretation.

'*Know-that*' control conditions Given that there were always two sets of objects in each picture, we wanted to ensure that children were indeed responding to the critical wh-questions on the basis of Zap's beliefs about *both* sets of objects. For example, if children rejected WE and WE+IE targets, this might be because they were accessing a stronger reading. However, we wanted to rule out the additional possibility that children might reject as soon as the puppet said something incorrect (as in the WE trials) or explicitly said, "I don't know" (as in the WE+IE) trials. Relatedly, we wanted to rule out the possibility that children might reject based only on Zap's beliefs about the second (most recent) set mentioned, e.g., the boats in the example above.

Conversely, if children accepted WE and WE+IE targets, we wanted to ensure that it was not simply because they were basing their responses only on the puppet's first statement (which always corresponded to a true belief), or were simpy responding 'yes' as soon as the puppet uttered at least one correct statement.

To address these concerns, we always had children repeat each of the puppet's statements. Additionally, we included two '*know-that*' control conditions, which were similar to the test conditions except that the final question specifically targeted one of the sets of objects, e.g., *Does Zap know that the trucks are in the box?* or *Does Zap know that the boats are on the carpet?* These were designed to ensure that children were not basing their responses to the embedded wh-questions on only

would not pose a problem in our experiment because of the way we set up the contexts. In all cases, there were only two salient locations available, e.g., in the box or on the carpet, and so the relevant interpretations above come out equivalent in response to the general question "Where are the boats?"

one of the sets of objects. The controls were therefore selected on the basis of children's responses to the critical test trials: WE '*know-that*'-controls followed WE test trials, and WE+IE '*know-that*'-controls followed WE+IE test trials.

For example, a child who heard the puppet's statements in (6a) or (6b) and who responded *no* to the question in (6c), would see a corresponding '*know-that*' control trial like (7), targeting the first set of objects mentioned (7c) (*yes*-target). A child who responded *yes* in response to (6c) however, would receive the question in (7d), targeting the second set of objects mentioned (*no*-target).

(6) a. "The trucks are in the box. The boats are also in the box." (*WE*)
b. "The trucks are in the box. The boats, I don't know." (*WE+IE*)
c. EXPERIMENTER: Does Zap know which toys are in the box?

(7) a. "The bumblebees are on the leaf. The ladybugs are also on the leaf." (*WE*)
b. "The bumblebees are on the leaf. The ladybugs, I don't know." (*WE+IE*)
c. EXPERIMENTER: Does Zap know that the bumblebees are on the leaf? (target: *yes*)
d. EXPERIMENTER: Does Zap know that the ladybugs are on the branch? (target: *no*)

Participants saw a total of three WE '*know-that*' control trials (each following one of the three WE test trials), and three WE+IE '*know-that*' control trials (each following one of the three WE+IE test trials). Because the control trials were selected dynamically based on responses to critical trials, they also allowed us to maintain an overall balance of 'yes' and 'no' responses for each participant.

2.2.2 Results

Control conditions

Children and adults performed as expected on ∅, WE+IE+SE, and '*know-that*' control trials. Participants had to pass at least five out of the six '*know-that*' control trials in order to be included in the analysis. All participants passed, and consequently no participants were excluded from analysis. The percentage of *yes*-responses on the '*know-that*' control conditions are reported in Figure 2.4a. Performance on controls was near perfect, though participants provided fewer *yes*-responses to WE controls where the target was *yes* (χ^2-test on the distribution of errors: $\chi^2(3) = 20$, $p < .001$). This may suggest that participants found it more difficult to answer 'yes' when one of the puppet's statements was false. Note that this could be expected to have biased these participants towards IE and SE readings on the test trials. Importantly, however, we did not observe any difference between children and adults on the controls.

Children's accuracy on the control conditions indicates not only that they were able to pay attention to, understand, and answer questions targeting Zap's knowledge about specific sets of objects. It also suggests that they were not responding to the critical test questions solely on the basis of one of the sets of objects described by Zap (i.e. the first or the second). In other words, accuracy on these control conditions allows us to rule out a response strategy whereby the participant answered *yes* as soon as Zap uttered a true statement, or *no* as soon as Zap uttered a false statement.[6] Finally, because we dynamically selected the control items to elicit the opposite responses to those provided on the test trials, we can be assured that children did not complete the task with a general *yes* or *no* bias.

Test conditions

The results from the four WH conditions are presented in Figure 2.4b. As expected, all participants accepted WE+IE+SE trials and rejected baseline ∅ trials. The difference in the percentage of *yes*-responses between the ∅ condition and the WE condition indicates that both groups accessed the WE reading ($\chi^2(1) = 53$, $p < .001$), and the difference between the WE and the WE+IE conditions likewise indicates that both groups accessed the IE reading ($\chi^2(1) = 180$, $p < .001$).[7]

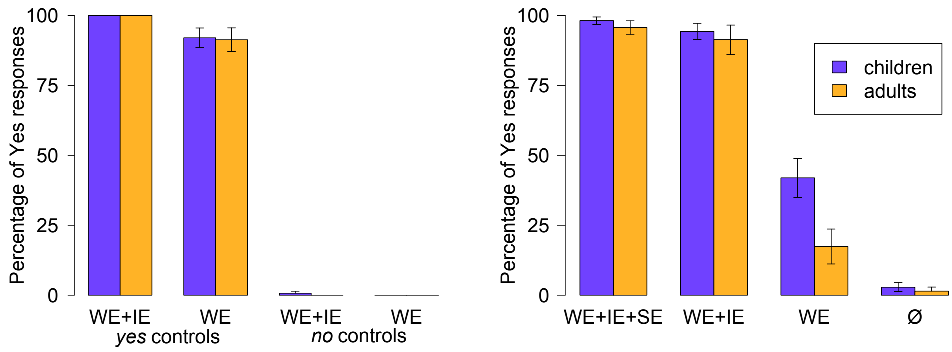
Children differed from adults, however, in how accepting they were of the WE reading. A logit mixed-model revealed that children gave more true responses than adults in the WE condition ($\chi^2(1) = 7.8$, p≈.005), but did not differ from adults in the WE+IE condition ($\chi^2(1) = 0.02$, p≈.9).

That children tended to respond more 'weakly' than adults is also evidenced by the individual results, plotted in Figure 2.5. Of the 35 children: (i) two were categorized as 'SE responders', rejecting both WE and WE+IE items; (ii) 16 'IE responders' rejected when only the WE reading was true, but accepted once the IE reading was also made true; (iii) 14 'WE responders' accepted in both WE and WE+IE conditions, suggesting they considered Zap to "know" which toys were in the box as soon as he knew that the trucks were in the box.[8]

[6]While we do not specifically address the development of Theory of Mind in this paper, we note here that children's target-like performance on the 'know-that' controls provides evidence that 5-year-olds are capable of attribution of both true and false beliefs. While some previous research has suggested children younger than four have difficulty differentiating between different mental state verbs such as *think* and *know* (Macnamara, Baker, and Olson 1976; Johnson and Maratsos 1977; Abbeduto and Rosenberg 1985; Moore and Davidge 1989), recent results by Dudley, Orita, Hacquard, and Lidz (2015) reveal that with less metalinguistically demanding tasks, children as young as three years of age can be shown to distinguish such mental state verbs. In fact, studies using implicit measures have also demonstrated understanding of false beliefs in infants (Onishi and Baillargeon 2005; Song, Onishi, Baillargeon, and Fisher 2008; Southgate, Senju, and Csibra 2007). Crucially, our experiment was conducted with children who were older than the age at which children have previously been shown to have difficulty with mental state verbs. Furthermore, our results from both the test and control conditions suggest that 5-year-olds are also able to distinguish false beliefs from ignorance, and moreover, to make use of this information to provide fine-grained linguistic judgments.

[7]In the WE+IE condition, Zap first correctly stated, for example, that the trucks were in the box, and then stated that he did not know where the boats were. An anonymous reviewer points out that out of the blue, such a discourse would strongly suggest that Zap knew that the boats were not in the box but didn't know their exact location. A participant who would interpret Zap's utterance in this way could then consider Zap to know which toys were in the box, even if they accessed an SE reading; this is because they would assume that Zap knew that the boats were not in the box, even if he didn't know they were on the carpet. In our task, however, there were only ever two salient locations, e.g., the box and the carpet, so knowing that the boats were not in the box was contextually equivalent to knowing that they were on the carpet. Zap should therefore not have been able to utter "I don't know" to mean "I don't know where they are, outside of the box". Furthermore, the presence of the WE condition would make it difficult to infer from Zap's true statement about the trucks' location that he necessarily knew what was in the box, given he would sometimes continue by incorrectly stating that the boats were in the box too.

[8]We do not discuss here so-called *mention-some* readings of embedded questions. Under such a reading, the sentence is true as soon as Zap knows of one animal that is on the couch that it is on the couch. This reading is even weaker than the WE reading. There are reasons to believe, however, that such readings should not appear in the present cases. For instance, Fox (2013) argues that mention-some readings require the presence of an existential modal. Furthermore, B. R. George (2013) and Xiang (2015a) show that mention-some readings can also come with a 'no-false-belief' constraint. We will assume, with Roelofsen et al. (2014), that the distinction between exhaustive and mention-some



(A) '*know-that*' control conditions, split by target response

(B) WH test conditions

FIGURE 2.4 – Percentage of *yes*-responses on (a) '*know-that*' control conditions and (b) WH target conditions.

Given the IE reading entails the WE reading, we observe that no participant displayed an inconsistent pattern: *yes*-responses to WE items but *no*-responses to WE+IE items. In sum, children were more or less equally split between WE and IE responders. By contrast, most adults were IE responders, with the exception of two SE and two WE responders.

Fourteen of the 35 children also spontaneously gave some 'both yes and no' responses, suggesting that Zap both knew and didn't know (which toys were in the box). Ten of these children did so on WE targets, and four of them did so on WE+IE targets. χ^2-tests revealed that the distribution of 'both'-responses across the six conditions is significantly different from chance ($p < .001$) and nearly significant when WE targets are excluded (p=.05). This 'both'-type behaviour may be taken as further evidence that some children are sensitive to the ambiguity of embedded questions. We suggest that these children responded in this way because they were sensitive to the tension between two possible readings (a weaker and a stronger one), in situations in which one was true and the other was false. An example of the kind of justification that children gave for such a response is provided in (8).

(8) CHI: Oui et non, il a dit que les bananes sont dans le panier et les oranges je ne sais pas.
'Yes and no, he said that the bananas are in the basket and the oranges I don't know.' (C05-A, age 5;06,01, WE+IE trial)

Follow-up justifications

Follow-up justifications were elicited after each trial, in order to ascertain that children were providing the responses they did for the expected reasons. Justifications for *no*-responses to the baseline ∅ condition, for example, made reference to Zap's false beliefs about the two sets of objects:

(9) Justifications for *no*-responses to baseline ∅ condition

readings is orthogonal to the distinction between weak, intermediate, and strong readings. In fact, given that mention-some readings, like exhaustive readings, also come in weak, intermediate, and strong flavours, we view our experiment as targeting the distinction between these three flavours, rather than specifically targeting the distinction between mention-some and exhaustive readings. We leave to future research a more targeted investigation of children's knowledge of mention-some readings of questions.

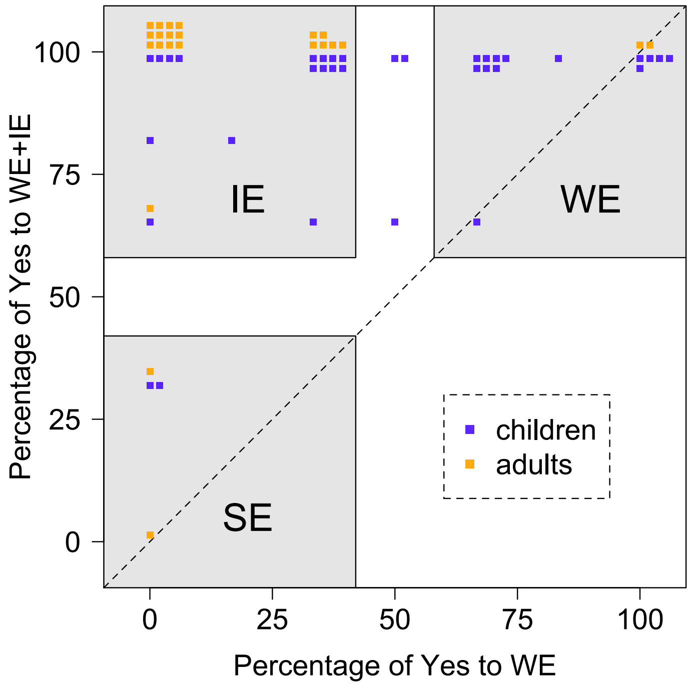


FIGURE 2.5 – This figure represents each individual as a combination of his/her acceptance rate on WE targets (x-axis) and WE+IE targets (y-axis). As indicated in the figure, (i) the top right corner in this space corresponds to participants who accessed the WE reading consistently (i.e. they gave *yes*-responses to both WE and WE+IE targets); (ii) the top left corner corresponds to IE-responders (who rejected the WE targets but accepted the WE+IE targets) and (iii) the bottom left corner corresponds to SE-responders (who rejected both the WE and WE+IE targets). Given that each participant saw three repetitions of each target condition, we can expect to observe individuals falling into one of 4×4 possible positions within the graph (with the additional possibility of 'both'-responses for children, which were encoded as 0.5); for better visibility, overlapping individuals are represented within small groups next to their actual position.

a. CHI: Parce qu'il a dit les bananes sont dans le panier et les pommes sont dans la nappe.
'Because he said the bananas are in the basket and the apples are on the cloth.' (C02-B, age 6;04,18)

b. CHI: Il a dit les voitures sont sur le tapis et les balles sur l'étagère.
'He said the cars are on the carpet and the balls on the shelf.' (C05-A, age 5;06,01)

c. CHI: Les pommes il sait pas c'est dans le panier, et il sait pas que les bananes, ils sont dans le panier.
'He doesn't know the apples are in the basket, and he doesn't know about the bananas, they're in the basket.' (C10-B, age 5;09,16)

Justifications for *yes*-responses on the WE+IE+SE control condition made reference to Zap's true beliefs about the two sets of objects:

(10) Justifications for *yes*-responses to WE+IE+SE condition

a. CHI: Beh qu'il vient de le dire parfait, il a dit les bateaux dans la boîte et les avions sur le tapis.
'Well he just said it perfectly, he said the boats in the box and the airplanes on the carpet.' (C07-A, age 5;08,04)

b. CHI: Parce qu'il a dit que les coccinelles étaient sur la feuille et les chenilles étaient sur la branche.
'Because he said that the ladybugs were on the leaf and the caterpillars were on the branch.' (C18-B, age 5;09,22)

c. CHI: Il a dit que les poires sont dans le panier et les fraises sont sur la nappe.
'He said that the pears are in the basket and the strawberries are on the cloth.' (C04-B, age 5;11,04)

Justifications for *yes*- and *no*-responses in the WE condition are provided in (11) and (12). Children who accepted in the WE condition made reference to Zap's true beliefs about the first set of objects (11). In contrast, children who rejected in the WE condition made reference to Zap's false beliefs about the second set of objects (12).

(11) Justifications for *yes*-responses to WE condition

a. CHI: Il a dit que les avions sont dans la boîte.
'He said that the airplanes are in the box.' (C13-A, age 5;03,12)

b. CHI: Parce qu'il a dit les ananas sont dans le panier.
'Because he said the pineapples are in the basket.' (C06-B, age 5;09,13)

(12) Justifications for *no*-responses to WE condition

a. CHI: Il y a que des moutons dans la maison et les cochons non.
'There are only sheep in the house, not pigs.' (C08-B, age 5;08,00)

b. CHI: Parce qu'il a dit que les cochons ils étaient dans la maison alors qu'ils sont pas dans la maison.
'Because he said that the pigs were in the house when they aren't in the house.' (C14-B, age 5;07,15)

c. CHI: Beh parce que les camions sont pas dans la boîte.
'Because the trucks are not in the box.' (C15-A, age 5;08,21)

d. CHI: Parce qu'il a mélangé tous les deux ensemble dans le panier, alors que c'était pas ça.
'Because he mixed up the two together in the basket, but it wasn't like that.' (C26-B, age 6;04,29)

Examples of justifications for *yes*- and *no*-responses in the WE+IE condition are provided in (13) and (14). Children who accepted in the WE+IE condition made reference to Zap's true beliefs about the first set of objects. In contrast, children who rejected in the WE+IE condition made reference to Zap's lack of knowledge about the second set of objects.

(13) Justifications for *yes*-responses to WE+IE condition

a. CHI: Il a dit que les oranges elles sont dans le panier.
'He said that the oranges are in the basket.' (C13-A, age 5;03,12)

b. CHI: Parce qu'il a dit les bananes sont dans le panier, et elles sont dans le panier.

'Because he said the bananas are in the basket, and they are in the basket.' (C26-B, age 6;04,29)

c. CHI: Parce qu'il a dit que les oranges étaient dans le panier et que c'est vrai.
'Because he said that the oranges were in the basket and this is true.' (C26-B, age 6;04,29)

(14) Justifications for *no*-responses to WE+IE condition

a. CHI: Non, les oranges sont dans le panier et les pommes il sait pas les pommes.
'No, the oranges are in the basket and the apples he doesn't know about them.' (C08-B, age 5;08,00)

b. CHI: Parce qu'il a dit les oranges il sait pas.
'Because he said the oranges he doesn't know.' (C17-A, age 5;09,12)

2.3 Discussion

Our first main finding is that 5-year-old children are sensitive to the rich range of interpretations that are available for embedded *wh*-questions. Children access WE, IE and SE readings, as adults do. In fact, some children explicitly reported their sensitivity to the tension between readings, further suggesting that 5-year-olds are aware of the ambiguity of embedded questions.

Our second main finding is that children are more accepting than adults of weak readings of embedded questions. This is evidenced by two aspects of our results. First, children provided more *yes*-responses in the WE condition than adults. Second, in the context of their individual responses, children were roughly split between WE and IE responders, whereas adults were almost all IE responders.

This main finding that children were more accepting than adults of weak meanings is reminiscent of much of the previous literature on children's development of scalar implicatures. There too, children have been shown to access weaker meanings more than adults. Our results would therefore appear to be consistent with a scalar implicature approach to the strengthened readings of embedded questions.[9]

It may be helpful to consider *why* children appear to access more weak readings of embedded questions than adults do. To do so, we can extend a proposal from the developmental literature on scalar implicatures. According to this line of research (see Gualmini et al. 2001; Chierchia et al. 2001; Reinhart 2006; Barner et al. 2011; Singh et al. 2015; Tieu et al. 2015), children have difficulty computing

[9]It would be interesting to compare the rate of derivation of IE and SE readings with the rate of derivation of global and local scalar implicatures, respectively, since Klinedinst and Rothschild (2011) establish a clear parallel between the two. Preliminary results by Bill, Pagliarini, Romoli, Tieu, and Crain (2015) suggest that embedded exhaustification may be accessible to children, possibly even more so than global exhaustification. Although the cases may not be directly comparable (because IE readings here involve multiple replacements and additional alternatives), such results would lead one to expect that children should predominantly access SE readings rather than IE readings. Alternatively, Uegaki (2015) proposes that SE readings are derived from IE readings, through a second step of strengthening that involves agent opinionatedness, via an Excluded Middle presupposition or implicature (following literature on *neg-raising* phenomena) (see footnote 3). Note that on this approach, children's acquisition of neg-raising could be a predictor of their access to SE readings, though only after they have already mastered IE readings. Before further testing the availability of SE readings in children, however, it would be useful to find a task in which the SE reading surfaces more clearly in adults.

scalar implicatures such as (2b) because they are less able than adults to access the scalar alternative required to compute the implicature. Consider the ingredients that would be required to compute the implicature in an adult-like way. The child must first acquire (i.e. lexicalize) the co-scalar status of the quantifiers *some* and *all*. When presented with the test sentence in 1 containing *some*, the child must then be able to perform lexical retrieval of the stronger alternative *all*. She must then be able to compare the weak and strong forms, and exhaustify with respect to the alternatives. Failure or difficulty with any of these steps along the way may cause children difficulty, and consequently result in a non-adult-like tendency to accept weaker meanings. Unsurprisingly, when children are provided with the required alternatives explicitly, or when they can easily retrieve these alternatives from the context, they perform more on a par with adults.

How can our data on embedded questions fit within such a story? Recall that on the proposal roughly outlined in Section 2.1.2, the derivation of the IE reading requires two sets of alternatives: those for the wh-question, i.e. the set of propositions {x is in the box}, and the alternative to *know*, namely *believe*. While it can be argued that children could retrieve the first set of alternatives from the context, e.g., {The trains are in the box, The boats are in the box}, performing lexical retrieval of *believe* as an alternative to *know* is entirely parallel with the case of retrieving *all* as an alternative to *some*. It may therefore not be surprising that children display a similar non-adult-like preference for weak meanings in the two cases. In short, assuming a theory along the lines of Klinedinst and Rothschild (2011) for questions, we can extend the alternatives-based explanation for children's difficulty with scalar implicatures to the present case of questions, and explain why children gave the WE response-pattern more than adults. Note that the explanation provided here for children's performance makes the prediction that if we were to make the *believe* alternative more salient to children, more IE responses might be observed. We leave the testing of this prediction to future research.

Finally, we would like to comment on a further consequence of the child data reported here, which we believe can serve to further constrain linguistic theories of questions. Among the theories that are able to derive the IE reading, only the strengthening-based theories additionally derive the WE reading. The fact that we observed many WE-responders among children, and even some among adults, provides evidence in support of theories that can derive the WE reading (although see Cremers and Chemla 2014 for discussion of some intricacies concerning the availability of WE readings). Moreover, these theories make further predictions for the acquisition of exhaustive readings. They treat the WE reading as basic, and derive the IE and SE readings through a process of strengthening that is known to cause difficulties for children. This is very much in line with what we observed in our experiment. In sum, our results can be taken to provide empirical support for strengthening-based theories of questions, which are currently the only ones that can derive both WE and IE readings.

2.4 Conclusion

Our findings provide novel evidence that 5-year-old children are sensitive to the ambiguity of questions embedded under *savoir* 'know'. They also show, however, that children answer according to the stronger readings of questions less often than adults. Our results are consistent with the findings from previous studies of children's acquisition of scalar implicatures, which also reveal a preference for weaker meanings in children compared to adults. According to recent proposals in the developmental literature, children's difficulty accessing strengthened meanings may lie in a difficulty with the lexical retrieval of the necessary alternatives. This interpretation of the data allows for a parallel between children's ability to access strengthened readings of embedded questions and their ability to compute classical scalar implicatures, and further supports a link that has recently been made explicit in certain theoretical accounts of embedded questions (Klinedinst & Rothschild, 2011; Andreea Cristina Nicolae, 2013; Uegaki, 2015).

Appendix

(15) Puppet's statements, followed by experimenter's question, for WE and WE+IE test conditions

a. WE condition (target: *yes* on WE reading, *no* on IE and SE readings)

i. Les avions sont dans la boîte. Les camions aussi sont dans la boîte.
Est-ce que Zap sait quels jouets sont dans la boîte ?

ii. Les moutons sont dans la maison. Les cochons aussi sont dans la maison.
Est-ce que Zap sait quels animaux sont dans la maison ?

iii. Les ananas sont dans le panier. Les poires aussi sont dans le panier.
Est-ce que Zap sait quels fruits sont dans le panier ?

b. WE+IE condition (target: *yes* on WE and IE readings, *no* on SE reading)

i. Les oranges sont dans le panier. Les pommes, je ne sais pas.
Est-ce que Zap sait quels fruits sont dans le panier ?

ii. Les chevaux sont dans la mare. Les vaches, je ne sais pas.
Est-ce que Zap sait quels animaux sont dans la mare ?

iii. Les bananes sont dans le panier. Les oranges, je ne sais pas.
Est-ce que Zap sait quels fruits sont dans le panier ?

(16) Puppet's statements, followed by experimenter's question, for ∅ and WE+IE+SE control conditions

a. ∅ control condition (target: *no*)

i. Les moutons sont dans la maison. Les poules sont sur l'herbe.
Est-ce que Zap sait quels animaux sont dans la maison ?

ii. Les bananes sont dans le panier. Les pommes sont sur la nappe.
Est-ce que Zap sait quels fruits sont dans le panier ?

iii. Les balles sont sur l'étagère. Les voitures sont sur le tapis.
Est-ce que Zap sait quels jouets sont sur l'étagère ?

b. WE+IE+SE control condition (target: *yes*)

i. Les coccinelles sont sur la feuille. Les chenilles sont sur la branche.
Est-ce que Zap sait quels animaux sont sur la feuille ?

ii. Les bateaux sont dans la boîte. Les avions sont sur le tapis.
Est-ce que Zap sait quels jouets sont dans la boîte ?

iii. Les poires sont dans le panier. Les fraises sont sur la nappe.
Est-ce que Zap sait quels fruits sont dans le panier ?

(17) Puppet's statements, followed by experimenter's question, for NP control conditions

a. WE NP-control condition

i. Les lapins sont sur le fauteuil. Les singes aussi sont sur le fauteuil.
Est-ce que Zap sait que:
α . les lapins sont sur le fauteuil ? (target: *yes*)
β . les singes sont par terre ? (target: *no*)

ii. Les cubes sont sur l'étagère. Les voitures aussi sont sur l'étagère.
Est-ce que Zap sait que:
α . les cubes sont sur l'étagère ? (target: *yes*)
β . les voitures sont sur le tapis ? (target: *no*)

iii. Les abeilles sont sur la feuille. Les coccinelles aussi sont sur la feuille.
Est-ce que Zap sait que:
α . les abeilles sont sur la feuille ? (target: *yes*)
β . les coccinelles sont sur la branche ? (target: *no*)

b. IE NP-control condition

i. Les singes sont sur le fauteuil. Les souris, je ne sais pas.
Est-ce que Zap sait que:
α . les singes sont sur le fauteuil ? (target: *yes*)
β . les souris sont par terre ? (target: *no*)

ii. Les camions sont dans la boîte. Les bateaux, je ne sais pas.
Est-ce que Zap sait que:
α . les camions sont dans la boîte ? (target: *yes*)
β . les bateaux sont sur le tapis ? (target: *no*)

iii. Les vaches sont dans la mare. Les canards, je ne sais pas.
Est-ce que Zap sait que:
α . les vaches sont dans la mare ? (target: *yes*)
β . les canards sont sur l'herbe ? (target: *no*)

Additional discussion

This experiment showed that children's acquisition of stronger exhaustivity is parallel to their acquisition of scalar implicatures, and suggests that WE readings can be taken as basic, even though it is usually strengthened for adults. Although this result does not invalidate the theories of ambiguity, I will take it as an argument in favor of a theory of exhaustification for the derivation of SE and IE readings. The next chapter addresses an other empirical questions, namely the behavior of questions embedded under emotive-factive predicates, which have been argued to give rise to WE readings only.

Chapter 3

Experiments on the acceptability and possible readings of questions embedded under emotive-factives

Cremers, A. & Chemla, E. (2016). *Experiments on the acceptability and possible readings of questions embedded under emotive-factives*. (submitted) †

Abstract

Emotive-factive predicates, such as *surprise* or *be happy* are a source of empirical and theoretical puzzles in the literature on embedded questions. Although they embed *wh*-questions, they seem not to embed *whether*-questions. They have complex interactions with negative polarity items such as *any* or *even*, and they have been argued to preferentially give rise to weakly exhaustive readings with embedded questions (in contrasts with most other verbs, which have been argued to give rise to strongly exhaustive readings). We offer an empirical overview of the situation in three experiments collecting acceptability judgments, monotonicity judgments and truth-value judgments. The results first straightforwardly confirm the special selectional properties of emotive-factive predicates. More interestingly, the results reveal the existence of strongly exhaustive readings for *surprise*. The results also suggest that the special properties of emotive-factives cannot be solely explained by their monotonicity profiles, which was not found to differ from the profiles of other responsive predicates.

3.1 Emotive-factive predicates and Questions

Emotive-factive predicates are, as their name indicates, a class of factive predicates of propositional attitudes involving emotions such as surprise, happiness or regret. In English, although some verbs of this class enter in traditional SVO constructions such as (1a), they also enter in constructions like (1b) and (1c), or adjectival constructions (1d). Some constructions may also leave the emotion holder implicit, as in (1e). The label 'emotive-factive' thus primarily refers to a semantic property and does not apply to a homogeneous syntactic class of verbs.

(1) a. John regrets breaking the vase.

†We wish to thank the Attitude Ascriptions & Speech Reports group at SIASSI Berlin, Angelika Kratzer, Florian Pellet, Yael Sharvit, Benjamin Spector, Lyn Tieu and Wataru Uegaki. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.313610 and was supported by ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC.

b. It surprised Peter that Mary came.
c. Peter was surprised that Mary came.
d. Mary is happy that Peter came.
e. It is amazing that John fixed the vase.

Emotive-factive predicates have drawn a lot of attention in the literature on embedded questions, because they are associated with several puzzles that we quickly review below. Given the recent dispute about the possible readings of embedded questions, there now exists some quantitative surveys on this topic, but they only concern predicates which are not emotive-factives (Cremers and Chemla, 2014 on *know* and *predict*, Chemla and George, 2015 on *agree*, Phillips and George, 2015 on *know*, Xiang, 2015c on *know* and *tell*). The goal of this paper is to offer systematic empirical coverage of the behavior of questions embedded under emotive-factive predicates. In this section, we first introduce the phenomena and puzzles of interest. We will observe that judgments and facts are sometimes difficult to assess based on the current claims in the literature, hence our proposed wrapping up survey.

3.1.1 Two puzzles regarding questions and emotive-factives

Puzzle 1: *Whether*-questions

The first puzzle we want to introduce dates back to Karttunen (1977): although emotive-factives generally embed *wh*-questions, they do not embed *whether*-questions, as exemplified by the contrast between (2) and (3a,b). Grimshaw (1979) proposed the first analysis of this fact, but many possible factors have been put forward since then to explain the ungrammaticality of (3a,b). Some attribute it to the competition between the embedded questions in (3a,b) and the embedded *that*-clauses in (4a,b) made equivalent by a presupposition stronger than factivity ('speaker-factivity' or 'super-factivity', see Guerzoni and Sharvit, 2007 and Sæbø, 2007) or by anaphoric properties of emotive-factives (Herbstritt, Roelofsen, and Aloni, 2015). Other analyses propose that questions come with a determinate strength of exhaustivity and that *whether*-questions force a strongly exhaustive reading which is incompatible with emotive-factives (see Andreea C Nicolae, 2015 and Guerzoni and Sharvit, 2014). Abels (2007) proposes that the polar question in (3b) systematically fails to satisfy the presupposition of *surprise* (but does not account for 3c). Finally, Romero (2015) proposes that the focus-sensitivity properties of emotive-factives are at the source of the deviance of (3b,c).

(2) It is amazing what they serve for breakfast.

(3) a. * It is amazing whether they serve breakfast.
b. * It is amazing whether they serve TEA or COFFEE for breakfast.

(4) a. It is amazing that they serve breakfast / that they do not serve breakfast.
b. It is amazing that they serve tea / that they serve coffee for breakfast.

The different proposals also differ on the characterization of the puzzle itself. While most authors take the unacceptability of (3a,b) as a case of plain ungrammaticality, Sæbø (2007) argues that under some circumstances *whether*-questions are acceptable with emotive-factives, based on examples like (5). Unlike others, Herbstritt et al. (2015) attribute the ungrammaticality of the polar question (3a) and the alternative question (3b) to different mechanisms, which may lead us to expect differences in judgments when assessing the acceptability of the corresponding sentences.

(5) Don't read this installment before seeing the episode if you want to be surprised at whether or not Hercules makes it.

Puzzle 2: Exhaustive readings

Focussing on well-formed constructions, another dispute concerns the possible meanings of constructions involving emotive-factives. Several readings have been proposed for sentences with questions embedded under verbs like *know*, as in (6). Karttunen (1977) first proposed the reading in (6a) which was later named "weakly exhaustive" (WE), while Groenendijk and Stokhof (1982, 1984) argued for a stronger reading, namely the "strongly exhaustive" (SE) in (6b).

(6) Mary knows who was at the party.
  a. For each person who was at the party, Mary knows that he/she was.
  b. For each person who was at the party, Mary knows that he/she was and she knows that no one else was.

While *know* was traditionally considered to only (or predominantly) receive SE readings (Groenendijk and Stokhof, 1982, 1984), Berman (1991) argued that *surprise* must only convey a WE reading. This introspective judgment has been endorsed by most authors following Berman, and notably Heim (1994) who took it as the main argument for a theory in which embedded questions are ambiguous between a SE and a WE reading.[1] A crucial empirical difference between *know* and *surprise* would be that, in a situation where Mary knows who the students are, the inference in (7) seems valid while the inference in (8) does not. Since the SE but not the WE reading makes the two questions "which students called" and "which students didn't call" equivalent, we may conclude that *know* but not *surprise* receives a SE reading.

(7) Mary knows which students called.
$\Rightarrow$ Mary knows which students didn't call.

(8) It surprised Mary which students called.
$\not\Rightarrow$ It surprised Mary which students didn't call.

3.1.2 Monotonicity as a key to Puzzles 1 and 2?

It has been suggested that some of the puzzling facts related to questions under emotive-factive predicates may be explained by their specific entailment patterns (Guerzoni, 2003, 2007; Guerzoni & Sharvit, 2007; Uegaki, 2015), which were first studied by Wilkinson (1996) to explain the interaction between emotive-factives and negative polarity items (NPI). As an example, some theories derive SE readings of embedded questions as implicatures (following Klinedinst and Rothschild, 2011), and therefore predict that they should not arise in the scope of downward entailing predicates. *Whether*-questions also have specific interactions with NPIs, and some accounts relate the distribution of NPIs and *whether*-questions under responsive predicates (Guerzoni, 2003; Guerzoni and Sharvit, 2007, 2014).

[1] In Heim's (1994) theory, full sentences with embedded questions are *not* ambiguous because the predicate selects the appropriate reading of the question it embeds.

However, the different studies diverge on the monotonicity they attribute to *surprise* (downward-entailing or non-monotonic) and little has been said about other emotive-factive predicates (although see Uegaki, 2015, on *be happy*). We therefore tested the Strawson-entailment patterns associated with some of these predicates in the same way that we collected data for their selectional properties and possible readings because (a) monotonicity plays a role in some popular accounts, (b) experimental data would help arbitrate between conflicting introspective judgments[2] and (c) the experimental literature suggests that 'perceived monotonicity' may be the relevant factor (Chemla, Homer, and Rothschild, 2011) and that it differs from actual monotonicity (e.g., because downward monotonicity is poorly evaluated in experimental tasks, Geurts and van der Slik, 2005).

3.1.3 Summary

Given the interest for emotive-factive predicates and the large variety of theories proposed to account for their properties, the main goal of this project was to gather quantitative data in a theory-neutral perspective (as much as possible). As the short review above justifies, we focussed on testing which types of questions can be embedded under emotive-factives, what (exhaustive) readings would these embedded questions carry and what monotonicity properties are associated with emotive-factives. We focussed on testing various widespread claims from the literature rather than predictions of specific theories, but we will still discuss precise theoretical consequences of our data along the way.

3.2 Experiment 1: Selectional properties of different attitude predicates

3.2.1 Goals

The main goal of this experiment was to test the selectional properties of emotive-factives. As we explain below, this experiment was also an opportunity to address a few other empirical questions regarding the selectional properties of various attitude predicates.

Selectional properties

The main goal of this experiment was to test the unacceptability of *whether*-questions under emotive-factive verbs (Karttunen, 1977; Grimshaw, 1979). We also tested the selectional properties of other attitude predicates, and compared different complements. In Lahiri's (2002) typology, emotive-factive predicates fall in the category of *responsive* predicates, which take both declarative and interrogative complements (e.g., *know*). Lahiri distinguishes responsive verbs from *rogative* verbs, which only embed interrogatives (e.g., *wonder*), and verbs which only embed declaratives (e.g., *believe*). For simplicity, we will refer to the latter category as "non-responsive" predicates. We decided to compare the (un)acceptability of *whether*-questions under emotive-factives with constructions where they should be clearly acceptable

[2] Note that a theoretical characterization of the monotonicity of a predicate depends crucially on its denotation. Uncertainty about the monotonicity of emotive-predicates amounts essentially to uncertainty about their denotations. Experimental explorations of the semantic properties of a predicate may thus inform theoretical semantics by constraining conceivable denotations.

or unacceptable. We also included the verb *regret* which has been claimed to be an exception among emotive-factive for not embedding questions at all.

Degrees of unacceptability

In addition to answering a general question about which predicates can embed which complements, gathering such quantitative results may also allow us to evaluate finer-grained differences in the degree of acceptability of different constructions, which may in turn distinguish various theories. For instance, Guerzoni (2007), Guerzoni and Sharvit (2007) and Sæbø (2007) propose that *whether*-questions are ruled out under emotive-factives because of the competition with declaratives at the pragmatic level. In contrast, Andreea Cristina Nicolae (2013), Andreea C Nicolae (2015) and Guerzoni and Sharvit (2014) propose that *whether*-questions are encoded as strongly exhaustive while emotive-factives select weakly exhaustive questions only. The former predict mere pragmatic oddness for *whether*-questions under *surprise*, while the latter predict a stronger grammatical incompatibility. In fact, within Herbstritt et al. (2015)'s approach, one may even expect to find a difference in the status of the incompatibility when it involves polar questions and when it involves alternative questions. Indeed, their approach derives the infelicity of alternative questions under emotive-factives from a contradictory presupposition, whereas polar questions are eliminated through competition with a declarative complement (using the maxim of manner).[3] We may expect all these differences to translate into different degrees of unacceptability, even though the link between the strength of incompatibilities and their nature (syntactic, semantic or pragmatic) is not entirely clear.

Quantificational variability

Quantificational Variability Effects (Berman, 1991) are another phenomenon related to embedded questions which has given rise to conflicting judgments in the literature. It describes the reading of (9) where Mary knows of most students who called that they called.

(9) For the most part, Mary knows which students called.

The availability of QVE has been debated for rogative verbs. On the one hand, Lahiri (2002) argues that they can never receive QVE readings and his theory predicts a semantic type mismatch in any possible structure. On the other hand, Beck and Sharvit (2002) argue that in some cases QVE readings may be available. They then propose a semantics which can derive QVE for any rogative verb, and propose that the unavailability of this reading in Lahiri's (2002) examples be due to an independent, softer constraint (subject to contextual variation). We took this experiment as an opportunity to gather data on this issue. The idea was to use the acceptability of sentences where a predicate modified by a quantity adverb embeds a *wh*-question as a proxy for the availability of QVE. We did not test which reading participants had for the target sentences, but our intuition was that if QVE

[3] However, Herbstritt et al. (2015, fn16) discuss the possibility that this infelicity has been grammaticalized over time. One difference between their account and Guerzoni and Sharvit (2007) is that they predict perfect synonymy between the polar question *whether-p* and the *positive* declarative *that-p*, whereas the competitor of the question in Guerzoni and Sharvit (2007) depends on which of p or $\neg p$ is true in the world of evaluation.

readings are available, these sentences should be more natural and receive a higher rating.[4]

3.2.2 Methods

Task and Instructions

Participants were asked to provide acceptability judgments for different sentences. They provided a continuous response with a horizontal slider which ends were labelled 'weird' and 'natural', as illustrated in Figure 3.1.

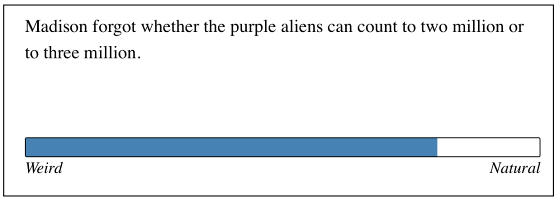


FIGURE 3.1 – Example of an item from Experiment 1, with an alternative question under *forget*.

The instructions introduced a context for the sentences. To make sure that world knowledge or other pragmatic factors would not interfere with the naturalness of the stimuli, all sentences were about aliens visiting the Earth. To encourage participants to judge the grammaticality of the sentences, and in particular to focus on selectional properties of the verbs, the instructions provided the two examples in (10) which were described as 'natural' and 'odd' respectively.

(10) a. Peter saw a fluffy purple alien playing the piano.
b. Peter went a fluffy purple alien playing the piano.

At the beginning of the experimental phase, participants were given 4 items comparable with the sentences presented in the instructions and of no interest to us, in order to let them familiarize with the task.

Design and Stimuli

Two factors were crossed: Embedding predicate (17 predicates from 4 broad categories, see Table 3.1) and Complement type (5 levels, see Table 3.2). Each combination was repeated 3 times, for a total of 255 experimental items.

Each item was generated by drawing randomly from lists of proper **names** for the agents, **adjectives** characterizing the aliens and **predicates** describing what the aliens do (complete lists are presented in Appendix A). For each complement type, the embedded clause (interrogative or declarative) was inserted in the COMP argument position of the schematic structures in Table 3.1. In the case of the Adv+*Which*-Question, the *which*-question was in the COMP position but the adverbial phrase was appended in sentence-initial position. The Polar and Alternative questions only differed in their PRED complement: a specific version of each PRED with a disjunction was used to generate alternative questions. Examples for each complement type are given in (11–15), embedded under various predicates.

[4] Lahiri (2002) shows that another reading may be available for rogative verbs with quantity adverbs (the *focus affected* reading), but this reading is intuitively harder to get and may require a specific prosody. In any case, the focus-affected reading should be equally available for all question-embedding predicates, so any difference we would see between two predicates cannot be attributed to this reading.

Predicate Type	Predicate	Schematic structure
Non-responsive	*believe*	X believes COMP.
	think	X thinks COMP.
	assume	X assumes COMP.
Rogative	*wonder*	X wonders COMP.
	depend	COMP depends on genetics.
	ask	X asked COMP.
Responsive	*know*	X knows COMP.
	remember	X remembers COMP.
	forget	X forgot COMP.
	misunderstand	X misunderstands COMP.
	disregard	X disregarded COMP.
	agree	X_1 and X_2 agree (on) COMP.
	guess	X guessed COMP.
Emotive-factive	*surprise*$_1$ (double object)	It surprised X COMP.
	surprise$_2$ (passive)	X is surprised (by) COMP.
	regret	X regrets COMP.
	be happy	X is happy (about) COMP.

TABLE 3.1 – List of embedding predicates tested in Experiment 1, and the sentence structures associated. Prepositions in parentheses were present with interrogative complements but absent with declarative complements. *Depend* did not require any proper name, *agree* required two names and we ensured that they would always be different.

(11) *surprise*$_1$ + Declarative:
"It surprised Mary that the fluffy aliens play the piano with their wings."

(12) *wonder* + Polar question:
"Peter wonders whether the hollow aliens can eat 5 pounds of licorice."

(13) *know* + Alt. question:
"Grace wonders whether the red aliens drink soda with a straw or with a spoon."

(14) *agree* + *Which*-question:
"Alex and Madison agree on which aliens write poems about the moon."

(15) *believe* + Adv+*Which*-question:
"For the most part, Jacob believes which aliens ride tall purple horses."

Participants

50 participants were recruited on Amazon's Mechanical Turk and were paid $2 for their participation (age range: 21–64, 29 males).

Statistical methods

The results were plotted using `ggplot2` (Wickham, 2009). For statistical analyses, responses were centered and normalized by participants. These normalized responses were analyzed with mixed-effects linear models in `R`, using package `lme4` (R Core Team, 2014; Bates, Mächler, Bolker, & Walker, 2015). We followed the procedure proposed in Bates, Kliegl, Vasishth, and Baayen (2015), who suggest to first

Type	Structure
Declarative	...that the ADJ aliens PRED.
Polar Question	...whether the ADJ aliens PRED.
Alternative Question	...whether the ADJ aliens $\text{PRED}_\vee$.
Which-Question	...which aliens PRED.
Adv+*Which*-Question	For the most part, ...which aliens PRED.

TABLE 3.2 – List of complement types tested in Experiment 1. For each type, the clause in the second column was inserted in the COMP argument position of the schematic structures in Table 3.1, except for the Adv+*Which*-Question in which the *which*-question was in the COMP position but the adverbial phrase was appended in sentence-initial position. The Polar and Alternative questions only differed in their PRED complement. A specific version of each PRED with a disjunction was used to generate alternative questions (see Appendix A).

try to fit maximum random effect structures by participants and by item (as suggested in Barr et al., 2013), and then eliminate useless components in the random effects structure to avoid over-parametrization.[5]

For each relevant parameter, we report the 95% confidence interval (CI) rescaled to the original 0-100 scale used in the graphs. The theoretically relevant comparisons all yield very clear results, as the graphs will show.

3.2.3 Results

Figure 3.2 presents a summary of the results by predicate type (see Table 3.1). For a break down by predicate and results from the sentences with an adverb of quantity, see Figure 3.3. In section 3.2.3, we first discuss the different categories of predicates one by one, ignoring sentences with adverbs of quantity and differences between the two types of *whether*-questions, we turn to the more specific questions these conditions raise in a second stage of analysis in section 3.2.3.

Misunderstand, *disregard* and *regret* were removed from the analyses, since these predicates did not reach 75% acceptability in any condition.

Predicate categories

Under non-responsive predicates, declarative complements were well-accepted (CI: [79, 86]), while interrogative complements were degraded (CI: [34, 43] for *whether*-questions and [38, 48] for *which*-questions).[6]

[5] More precisely, we always fitted models with maximal by-subject and by-item random structures, although they often did not converge. Then we used the function `rePCA` provided in the package `RePsychLing` to remove all components which explained less than a 0.1% of the variance explained by the main component. This was done independently for the subject and item random effects structures, because random effects associated with items tend to be smaller. We also removed correlations between random effects in the updated "parsimonious" model. A second pass ensured that no component under the threshold remained in the new model (this could happen when the PCA had been done on a maximal model which had not converged). All parsimonious models converged without requiring any further simplification, and whenever the maximal models had converged, we found that they did not explain more variance than the parsimonious models.

[6] On declarative sentences, there was no difference between *believe* ([82, 88]), *think* ([79, 88]) and *assume* ([75, 85]). With *wh*-questions, *assume* was rated higher ([55, 69]). A quick search on the internet

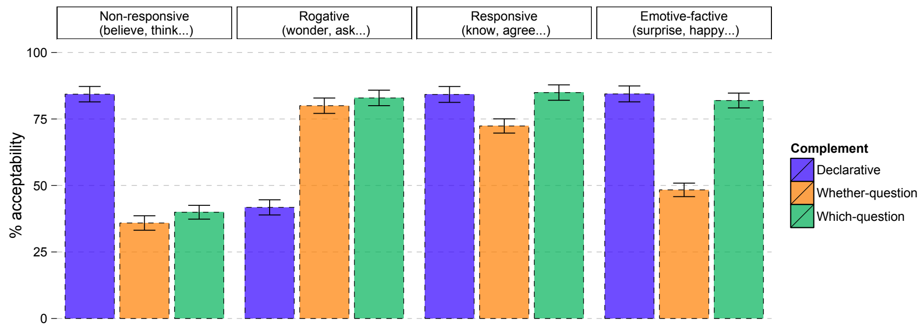


FIGURE 3.2 – Aggregated acceptability by predicate category and complement category in Experiment 1. Error bars correspond to standard error of the mean. The verbs *regret*, *misunderstand* and *disregard* were excluded from the analyses (see Figure 3.3 for details). *Whether*-questions include both polar and alternative questions. Sentences with an adverb of quantity are not shown in this figure.

Under rogative predicates, we observed the opposite pattern. Declarative complements were clearly degraded compared to all interrogative complements (CIs: $[39, 49]$ vs. $[75, 81]$). This difference showed up for the three verbs we tested, although the contrast was slightly reduced for *ask*.

With responsive predicates, all complements were clearly above 50%, although *whether*-questions were slightly degraded. Leaving aside *misunderstand* and *disregard*, all these verbs were qualitatively similar: equally compatible with declarative complements ($[79, 85]$) and *wh*-questions ($[80, 86]$), and almost as acceptable with *whether*-questions ($[69, 74]$). Veridical (*know, forget, remember*) and non-veridical (*agree, guess*) responsive predicates did not differ significantly: a model with veridicality, complement type and their interactions as fixed effect showed no effect of veridicality and no interaction (all $t < 1.5$, estimated effects below 1% acceptability).

Under emotive-factive predicates (excluding *regret*), declarative and *which*-questions were as acceptable as with responsive verbs ($[79, 86]$ and $[78, 83]$ respectively).[7] This is in contrast with *whether*-questions which were significantly degraded compared to their acceptability under other responsive predicates ($[46, 54]$ vs. $[69, 74]$) but still more acceptable than under non-responsive predicates ($[34, 43]$). There was no significant difference between the overall acceptability of *surprise*$_1$, *surprise*$_2$ and *be happy*. *Regret* was very degraded, even with declarative complements.

Specific questions

Rogative predicates and QVE: Under responsive, non-responsive and emotive-factive predicates, we found that the presence of an adverb slightly reduced the acceptability of a sentence with an embedded *wh*-question ($[64, 67]$ vs. $[68, 71]$). The effect of the adverb did not interact with verb category ($\chi^2(4) = 5.0, p = .29$) but it had an important effect within the class of rogative predicates: sentences with adverbs of quantity were judged as unacceptable as sentences with declaratives

indeed revealed some examples of questions under *assume*, such as: "People like to assume which of my parents is black or which of my parents is white."

[7] There is actually a statistically significant difference, but it is qualitatively very small: $[0.8, 3]$.

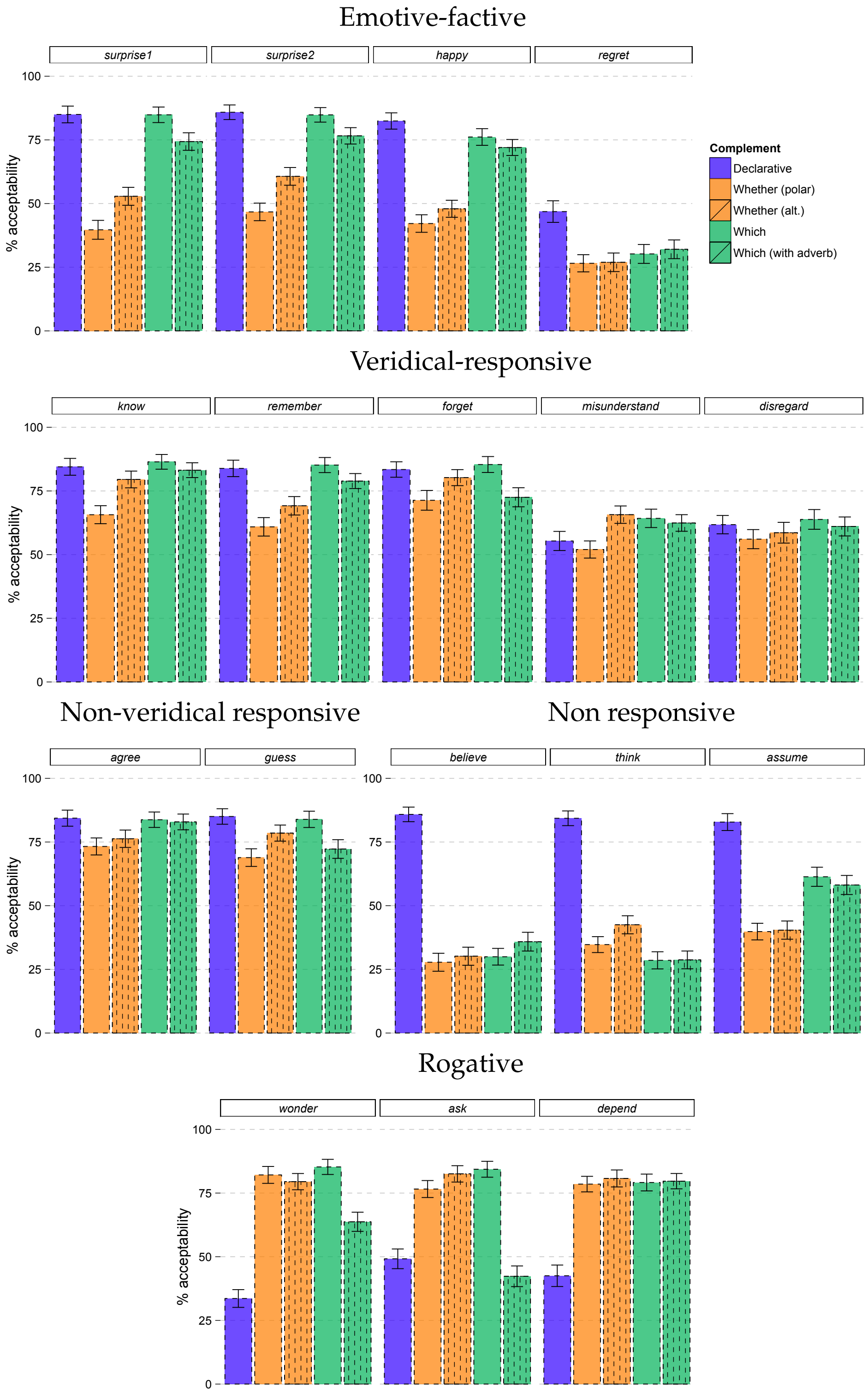


FIGURE 3.3 – Acceptability for each predicate and complement in Experiment 1, by predicate category.

for *ask* ([36, 52]), as acceptable as sentences with interrogatives for *depend* ([75, 82]) and somewhere in between for *wonder* ([58, 69]).

Polar and Alternative questions: As is visible on Figure 3.3, polar questions were less acceptable than alternative questions across all predicates ([62, 67] vs.

$[69, 73]$). This may seem surprising, because alternative questions were strictly more complex than polar questions (they only differed in that they had an extra disjunct appended). From discussion with a dozen informants, it seems that some native speakers, mostly in western regions of the USA, dislike embedded *whether*-questions without a disjunction (they would need to add "or not" in polar questions). With 50 participants, from only 26 states in the USA, and no precise information regarding their linguistic background, we did not have the adequate resources for a proper study of American dialects. Nevertheless, it turned out that longitude was a significant predictor of individual differences in the acceptability of polar and alternative questions ($t = -2.5$, $p = .02$). Latitude did not have a significant effect, but there was a trend for an interaction ($t = 1.7$, $p = .09$).[8] We would like to point out that Guerzoni and Sharvit (2014) assume that polar questions are underlyingly alternative questions with a silent "or not". The fact that some dialects of American English require an overt "or not" suggests that such an analysis of polar questions is on the right track.

3.2.4 Discussion

Our results confirm most judgments reported in the literature (see also Sprouse and Almeida, 2013; Sprouse, Schütze, and Almeida, 2013 on the reliability of introspective judgments): the four classes of predicates we discussed correspond to a natural typology based on acceptability of declarative and interrogative complements. The predicates we labelled as "non-responsive" are bad with any interrogative complement, while rogative predicates display the opposite pattern (good with interrogatives, bad with declaratives). Responsive predicates are good with any complement. Emotive-factive predicates are equally good with declarative and *wh*-questions, but are clearly degraded with *whether*-questions.

Turning to new conclusions we can draw from this experiment, we observed that *whether*-questions under emotive-factives, although degraded, are not as unacceptable as plain ungrammatical sentences. This suggests that their unacceptability is not due to a rigid constraint, at least not as rigid as the ban on questions under verbs like *believe*. This is in line with pragmatic accounts such as Guerzoni (2007) and Sæbø (2007), who argue that the unacceptability of *whether*-questions under emotive-factives is due to a contextual competition with *that*-clauses (note however that Sæbø proposes a similar argument for questions under *believe*, which is not supported by the current results).

We did observe a contrast between polar and alternative questions, but it was not specific to emotive-factive predicates and if anything it went in the opposite direction from what Herbstritt et al. (2015) may have predicted. Most likely, this contrast can be explained as a syntactic constraint on the construction of polar questions in some western dialects of American English.

Regret was degraded with interrogative, but also with declarative complements. The first fact is a known puzzle in the literature (Lahiri, 2002; Egré, 2008), the second one is more surprising. It may be that *regret* is better with propositions related to the agent. In our stimuli, the propositions denoted by the embedded declarative were facts about aliens, completely independent from the agents. Alternatively, the

[8] We would like to thank Aparicio Kozuch, who insisted on adding "or not" to polar questions when we were eliciting judgments for another project, thus suggesting this hypothesis. We would also like to thank the many American linguists we met during the summer 2015 who kindly provided judgments and Florian Pellet who included the question about state when programming the experiment and thus allowed us to test the hypothesis.

class of “emotive-factive” may not be as homogeneous as previously assumed and should maybe be split into two classes, with predicates like *surprise* and *be happy* on one side and verbs like *regret* and *resent* which do not embed questions on the other side.

Finally, the results on sentences headed by ‘for the most part’ suggest variation in the availability of QVE for rogative verbs. Of course, since we did not probe the readings participants accessed but only their acceptability of the various sentences, we cannot be sure that the availability of QVE is what drives the differences. However, we would like to point out that (a) QVE is probably the only salient reading for these sentences (the alternative would be what Lahiri, 2002, calls the focus-affected reading, which would not explain the differences between predicates) and (b) the differences we observed go exactly in the direction of Beck and Sharvit's (2002) judgments (good with *depend on*, bad with *ask*, unclear with *wonder*, probably context-dependent).

Before addressing the question of strong exhaustivity, we will first investigate the monotonicity of emotive-factive predicates. This will inform us on the semantics of the predicates themselves, without which we cannot even be sure of what their exhaustive readings are.

3.2.5 Conclusions for Experiment 1

We collected naive speakers judgments pertaining to the selectional properties of embedding predicates and the results mostly confirm existing judgments in the literature. In the remainder of this paper, we move to more semantic investigations, focusing on constructions that were judged grammatical in this experiment. We will first investigate the monotonicity of emotive-factive predicates, simply when they embed straight declarative complements. This will inform us on the semantics of the predicates themselves and is in fact a prerequisite to studying exhaustive readings they could give rise to when they embed questions, an issue we turn to in the final experiment.

3.3 Experiment 2: On the monotonicity of responsive predicates

3.3.1 Goal

The goal of this experiment was to determine the monotonicity of some emotive-factive predicates. This is important for at least two reasons. First, various special properties of these verbs (e.g., NPI licensing abilities, selectional properties, available readings) have been linked to their monotonicity profiles. Second, judgments in the theoretical literature diverge. Guerzoni and Sharvit (2007) consider that *surprise* is Strawson-downward entailing, whereas Uegaki (2015) takes *surprise* to be Strawson-non-monotonic. Furthermore, most of the literature focuses on *surprise*, and we thought it would be interesting to test *be happy* which, we thought, could be judged 'less' downward entailing.

The relevant concept for theories is not classical monotonicity but Strawson-monotonicity, which describes the entailment patterns between complements of a predicate, provided all presuppositions are satisfied. In order to test Strawson-monotonicity, we thus tested monotonicity like inference patterns, augmented with premises which would guarantee that the presuppositions of all relevant propositions are true.

3.3.2 Methods

Task and Instructions

This experiment was an inferential task with continuous responses, in which we tested the inference patterns of various environments that embed propositions. As illustrated in Figure 3.4, experimental items consisted of two premisses and one conclusion.

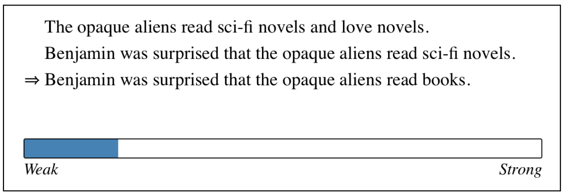


FIGURE 3.4 – Example of a direct item in Experiment 2. One can see that the predicate in the second premise (*read sci-fi novels*) is stronger than the predicate in the conclusion (*read books*). The fact (first premise) ensures that the opaque aliens read sci-fi novels, thus validating the factive presuppositions of both the second premise and the conclusion.

The instructions introduced a minimal context for the sentences, as in Experiment 1 (all sentences were about aliens who spent the last week on Earth). The instructions provided the two examples in (16/17)a followed by the explanations in (16/17)b.

(16) a. There are pink aliens and blue aliens.
Less than 50 pink aliens play the piano.
⇒ Less than 50 pink aliens play the piano with their left hand.

b. In this case, you should put the bar close to the right *[strong]*. Indeed, if in total less than 50 pink aliens play the piano, there cannot be more than 50 pink aliens who play the piano with their left hand.

(17) a. The blue aliens live in tree houses 10 feet above the ground with rope ladders.
Peter knows that the blue aliens live in tree houses.
⇒ Peter knows that the aliens live in tree houses 10 feet above the ground.

b. Here you should put the bar close to the left *[weak]*, because even if Peter knows that the blue aliens live in tree houses, nothing indicates that he knows that their tree houses are 10 feet above the ground.

The two examples in the instructions were given as the first actual items so that participants would familiarize with the presentation and task before getting to actual items of interest.

Design and Stimuli

We tested direct inferences, where the proposition embedded in the premise entailed the proposition embedded in the conclusion, and indirect inferences, where the proposition embedded in the second premise was asymmetrically entailed by the proposition embedded in the conclusion. We crossed two factors: Environment (8 levels) and Direction of the inference (2 levels). Participants saw 8 repetitions of each combination plus 16 filler items, for a total of 144 items.

Most sentences tested in this experiment were minor modifications of sentences which received very high ratings in Experiment 1.

Participants were asked to indicate how strongly the conclusion followed from the premisses using a horizontal slider which ends were labelled 'weak' and 'strong'. The first premise was always a "fact" which ensured that the presuppositions of the second premise and the conclusion were satisfied, thus allowing us to test Strawson-entailment rather than classical entailment. To make the sequence less repetitive and more natural, the fact was always a conjunctive statement, of which one conjunct was the presupposition (of the second premise or of the conclusion, depending on which one was the strongest). The second premise and the conclusion only differed in the predicate of their embedded clause. These predicates could be "strong" (PRED_+) or "weak" (PRED_-), and this allowed us to test both directions of inference. Direct items tested the inference from the strong to the weak, which is only valid in Strawson-upward entailing (SUE) contexts. Indirect items tested the inference from the weak to the strong, which is only valid in Strawson-downward entailing (SDE) contexts. They were obtained by switching the conclusion with the second premise in a direct item. Strawson-non-monotonic (SNM) contexts make both the direct and the indirect inferences invalid.

Table 3.3 presents the attitude predicates tested in Experiment 2, and Table 3.4 presents the structure for each item type. We tested the emotive-factive *surprise* and *be happy*. To keep the sentences with emotive-factives superficially similar to those using other attitude predicates, we used the SVO-like construction: "Mary is surprised/happy that…". We compared these to three other responsive verbs: *know* and *agree*, which we expected to be SUE, and *forget* which we expected to be

SDE. For *agree*, we used the structure "Peter agrees with Mary that...".[9] We used three control constructions without attitude verbs to serve as a baseline for UE, DE and NM environments.[10] We created valid fillers to counterbalance the fact that NM controls made all inferences invalid. All items were in the past tense, except for *agree*, where the main verb was in present tense.

Each participant saw 8 repetitions of each condition, obtained by picking a triplet of instantiations for PRED, ADJ and proper names, with the constraint that direct items and corresponding Indirect items were matched for each participant (i.e. constructed on the same 8 triplets). Each participant thus saw a total of 144 items made of 8 (environment) × 2 (direct vs. indirect) × 8 (repetitions) and 16 direct *only*-fillers.

Predicate Type	Predicate	Schematic structure
Emotive-factive	*surprise*	X was surprised that...
	be happy	X was happy that...
Other responsive	*know*	X knew that...
	forget	X forgot that...
	agree	X_1 agrees with X_2 that...

TABLE 3.3 – List of attitudes tested in Experiment 2, and the sentence structures associated. *Agree* required two names and we ensured that they would always be different. We used the *agree with* form this time, because it introduces a presupposition without being a factive verb.

Participants

50 participants were recruited on Amazon's Mechanical Turk and were paid $2 for their participation (age range: 19–58, 29 males).

Statistical methods

For statistical analyses, responses were centered and normalized by participants. The responses to targets were further transformed, as discussed in section 3.3.3 (we projected the two dimensional response space <direct, indirect> onto two new dimensions determined by the responses to controls). As in the previous experiment, we used parsimonious mixed-effects models, following Bates, Kliegl, et al. (2015). We fitted random effects for PRED which was the most crucial random variation between items (we did not fit random effects for ADJ since it would have increased

[9] There were a few reasons for this move. The collective form with a plural subject, which we used in the previous experiment, should not contribute to a decrease in acceptability compared to other predicates with a singular subject, but its semantic effects are not fully understood. The construction 'John agrees with Mary that p' allows for a singular subject (and therefore simplifies issues of cumulative readings). It also has a better understood presupposition (that Mary believes p), which allowed the use of a uniform underlying structure to construct the material (in which the first premise validates a presupposition of the second premise). Note that in a semantic task, Chemla and George (2015) found no difference between these two constructions when they embed questions. They did not test embedded declaratives however.

[10] The reader may notice that our NM controls are actually not SNM (they were constructed with *only* so they are SDE). However, since the 'fact' for these items did not validate the UE presupposition of *only*, the end-result was an environment which made both directions of inference invalid (was non-monotonic). We preferred *only* over genuine SNM expressions such as 'exactly n' which were too hard to provide a proper baseline. The results confirmed our intuition that *only* is easily understood as non-monotonic.

Predicate	Sentence	Schematic structure
Factive attitude V	Fact Strong/Weak	The ADJ aliens PRED$_\wedge$ X V that the ADJ aliens PRED$_{+/-}$
agree	Fact Strong/Weak	X_2 believes that the ADJ aliens PRED$_\wedge$ X_1 agrees with X_2 that the ADJ aliens PRED$_{+/-}$
UE	Fact Strong/Weak	There are ADJ$_1$ aliens and ADJ$_2$ aliens. The ADJ$_1$ aliens PRED$_{+/-}$
DE	Fact Strong/Weak	There are ADJ$_1$ aliens and ADJ$_2$ aliens. The ADJ$_1$ aliens didn't PRED$_{+/-}$
NM	Fact Strong/Weak	There are ADJ$_1$ aliens and ADJ$_2$ aliens. Only 12 ADJ$_1$ aliens PRED$_{+/-}$
Only-fillers	Fact Strong Weak	There are ADJ$_1$ aliens and ADJ$_2$ aliens. Only 12 ADJ$_1$ aliens PRED$_+$ At least 5 ADJ$_1$ aliens PRED$_-$

TABLE 3.4 – Structure for each item type appearing in Experiment 2. The 'fact' was always the first premise. The 'strong' and 'weak' sentences played the roles of second premise and conclusion, alternatively. All attitudes except *agree* were factive and shared the same structure, illustrated in the first line of the table. The *only*-fillers only appeared in direct version (inference from the strong to the weak). See Appendix B for details about the lists from which proper names, ADJ and the different types of PRED were drawn.

the risk of models not converging and was unlikely to have an important effect to begin with).

3.3.3 Results

Figure 3.5 presents responses to Indirect items as a function of responses to Direct items, for each environment. In such a representation, SUE items should be attracted towards the lower-right corner and SDE items towards the opposite upper left-corner. SNM items should fall away from this anti-diagonal and be attracted towards the lower-left corner.

Control items

We fitted a mixed-effects model for each inference type (Direct and Indirect) on the 3 control conditions with full random by-subject and by-item effects (environment is both a within-subject and within-item effect). The first one was fitted on responses to the Direct items, and showed that Direct inferences were judged more valid in UE controls than in NM controls ($t = 6.8, p < .001$), and more valid in NM controls than in DE controls ($t = 2.9, p = .004$). The indirect inferences were judged equally low in UE and NM conditions ($t = -.2, p = .84$) but were significantly more valid in the DE condition ($t = 7.7, p < .001$).

Attitude predicates

Using the control items, we defined two new dimensions along which we analyzed the attitude predicates.

First dimension ("Deviation from monotonicity"): We determined the representative $\overline{UE}$, $\overline{DE}$ and $\overline{NM}$ positions in the bidimensional space of Direct and

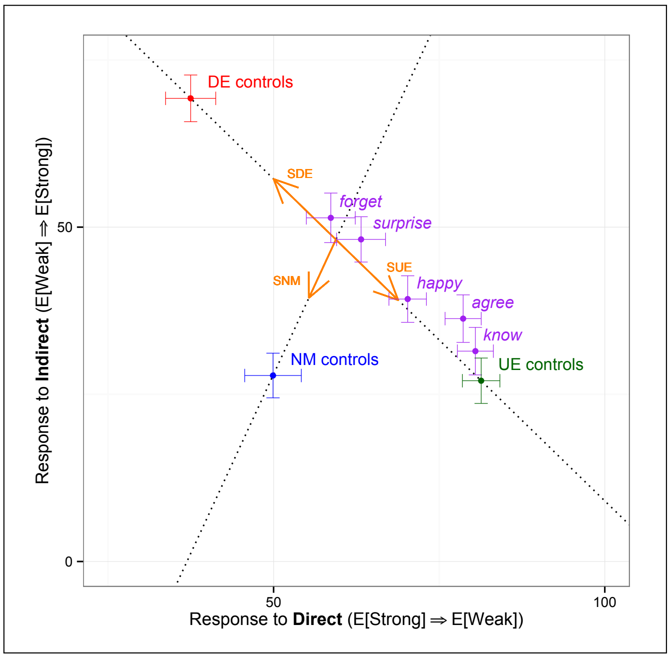


FIGURE 3.5 – Responses to Indirect as a function of responses to Direct items, for each environment. In such a representation, SUE predicates should fall to the bottom right corner (making Direct inferences valid, but Indirect inferences invalid). SDE predicates are expected to reverse this pattern and appear at the top left corner. SNM predicates should make any inference invalid, and therefore fall at the bottom left corner. The two dashed lines represent the dimension on which we projected for the statistical analyses (although this projection was actually done on normalized responses rather than on the raw data). The orange arrows indicate in which directions SUE, SDE and SNM predicates should deviate from the central point.

Indirect responses as the average of the responses to the UE, DE and NM control items. We then computed the projection of each response on the axis starting at the middle of $[\overline{UE}, \overline{DE}]$ segment and going towards $\overline{NM}$. A non-null value on this axis represents a tendency towards non-monotonicity.[11] We then fitted a mixed-effect model on these dependent variables. We included Environment as a fixed effect (5 levels, corresponding to the 5 attitudes tested) and full random by-subject and by-item structure (including slopes for Environment which is within-subject and within-item). None of the fixed effects was significant (all $|t| < .7$) and a model with only an intercept explained as much variance ($\chi^2(4) = .6, p = .96$, note that the intercept itself was very close to zero $t = -.2$). This suggests that no predicate yields non-monotonicity.

Second dimension ("Upwardness"): We fitted a mixed-effect model on the projections on the $[\overline{UE}, \overline{DE}]$ line of responses for each predicate, to determine whether each verb was more SUE or SDE. This time, we found clear differences between the verbs ($\chi^2(4) = 20, p < .001$). As a post-hoc analysis, we compared attitude predicates two by two, following their order on the $[\overline{UE}, \overline{DE}]$ line. Based on Bonferroni-corrected p-values for 4 comparisons, *Know* was similar to *agree*

[11]Negative values are not expected, since they would translate a tendency towards tautology (making too many inferences valid).

($\chi^2(1) = 3.4, p' = .26$), which was slightly more upward-monotonic than *be happy* ($\chi^2(1) = 5.6, p' = .073$). *Be happy* was clearly more upward monotonic than *surprise* ($\chi^2(1) = 11, p' = .003$), which did not differ significantly from *forget* ($\chi^2(1) = 1.3, p' = 1$). Even though *surprise* and *forget* were judged more downward entailing than other verbs, they were still very far from the DE controls.

3.3.4 Discussion

Most verbs ended up being perceived as upward entailing. We may imagine that this is an artifact of the task, which could induce a general "upwardness" bias. This may be due to low-level strategies adopted by some of the participants. A few of them reported basing their answers on the subcategory/super-category relations between complements of the embedded clauses. These kinds of strategies would ultimately lead a participant to ignore the embedding environment and base their responses on the complement clauses alone (which is equivalent to an UE environment). Nevertheless, participants as a group made clear distinctions between the control items, in the expected directions: UE controls made the direct inference valid and the indirect inference invalid, DE controls reversed this pattern, and NM controls made all inferences less valid.

The results on attitude predicates, validated by the clear results on controls, thus suggest that we can distinguish two classes of predicates based on monotonicity: *know*, *agree* and *be happy* on the one hand, which are clearly upward entailing, and *forget* and *surprise* on the other hand, which may be perceived as more downward entailing. We can draw two conclusions from this result. Crucially, the two emotive-factive predicates we tested were not perceived as less monotonic than other verbs. As a result, monotonicity appears to be orthogonal to the selectional properties of the verbs. Indeed, the groups that emerged here both contain verbs that embed *whether*-questions and verbs that do not.

No theory draws a direct connection between monotonicity properties and the acceptability of *whether*-questions. However, it has been suggested that (i) *whether*-questions are only acceptable under predicates for which a strongly exhaustive reading is available (Andreea Cristina Nicolae, 2013; Andreea C Nicolae, 2015; Guerzoni and Sharvit, 2014), and (ii) that emotive-factive predicates do not allow strongly exhaustive readings because of their monotonicity properties (Uegaki, 2015). According to the results of Experiment 2, these hypotheses cannot both be right. In the next experiment, we investigate the availability of strongly exhaustive readings under *know*, *forget* and *surprise*.

3.4 Experiment 3: Strongly exhaustive readings

3.4.1 Motivations and additional background

The main goal of this experiment was to test the availability of strongly exhaustive (SE) readings for questions embedded under *know, forget* and *surprise*. Berman (1991) first argued that *surprise* only gives rise to weakly exhaustive (WE) readings, and most introspective judgments in the literature agree.[12]

[12] Among exceptions, some do not challenge Berman's judgment on the assertive component but argue for different presuppositions. In particular, Abels (2007) and B. R. George (2011) propose a mention-some reading instead of the WE readings (hence a weaker presupposition) while Spector and Egré (2015) propose a WE reading for the assertion, but a SE presupposition. Although everyone acknowledges that the WE reading is available, Klinedinst and Rothschild (2011, fn. 18) argue that a

The (un)availability of SE readings with *surprise* has independently been linked to the monotonicity properties of this verb and to its selectional properties. Nevertheless, *forget* shares the monotonicity properties of *surprise* (Experiment 2), but patterns with *know* when it comes to selectional properties (Experiment 1), so if there is a link between these two properties, it has to be mediated by yet another factor. By testing the availability of SE readings with these three verbs, we should in principle be able to disentangle all possible links between monotonicity, exhaustivity and selectional properties.

Before getting to the experimental part, we will review basic assumptions about *know*, *forget* and *surprise*. Our goal is to obtain a list of conceivable readings for sentences where they embed questions. We will then be in a position to evaluate the availability of each of these readings. More specifically, we will be able to assess whether *surprise* gives rise to a SE reading, factoring out the influence of other possible readings.

Know

Denotation: We will assume a very simple denotation for *know*: "John knows p" presupposes that p is true and asserts that John believes p. There are constraints on what types of beliefs count as (justified) knowledge, but we will set an experiment in which it is clear that the beliefs attributed to agents satisfy these constraints.

Readings: *Know* has received more attention that any other question embedding predicate, and this is reflected in the wide variety of readings proposed in the literature. We will obtain a list of these readings in (19).

Different *exhaustive* readings have been proposed, which have in common that for John to know who was at the party, they at least require that for each person who actually was at the party, John know that they were. The WE reading does not require anything else (Karttunen, 1977), the Intermediate Exhaustive (IE) reading further require that John would not falsely believe that anyone else was at the party (Groenendijk & Stokhof, 1982; Berman, 1991; Preuss, 2001; Spector, 2005), and the most stringent Strong Exhaustive (SE) reading requires John to know that no one else was at the party (Groenendijk & Stokhof, 1982).

It has been argued that questions embedded under *know* can sometimes give rise to weaker, non-exhaustive readings, which do not require the agent to know a complete answer to the question. As an example, (18) may be true as soon as Rupert knows of one place where Italian newspapers can be purchased, without knowing the exhaustive list of all such places. This reading is often called the *mention-some* reading.

(18) Rupert knows where he can buy an Italian newspaper.

As for exhaustive readings, different mention-some readings have been proposed in the literature. In particular, B. R. George (2011, 2013) argues for a stronger notion of mention-some reading, which requires Rupert to know one true answer to the

SE reading is possible in some circumstances, and Theiler (2014) elaborates on this idea by suggesting that emotive-factives are ambiguous between a *literal* reading which only gives rise to WE reading and a *deductive* reading which leads to SE readings.

question, but also not to have any false beliefs regarding the selling of Italian newspaper at other places. We will call it the 'false-belief sensitive mention-some' reading, which we will abbreviate as FBS-MS. By contrast, we will call 'weak mention-some' (WMS) the mention-some reading that requires knowledge of at least one true answer without any further constraint.

Cremers and Chemla (2014) and Xiang (2015c) provide experimental evidence for the availability of all exhaustive readings (WE, IE and SE); Phillips and George (2015) and Xiang (2015c) for the mention-some readings (WMS and FBS-MS). This leads us to the list of readings in (19).

(19) Mary knows which of her cards are spades.
- WE: For each actual spade, Mary knows that it is a spade.
- IE: For each actual spade, Mary knows that it is a spade, and she does not falsely believe that any other card is a spade.
- SE: For each actual spade, Mary knows that it is a spade, and she knows that no other card is a spade.
- WMS: There is an actual spade that Mary knows to be a spade.
- FBS-MS: There is an actual spade that Mary knows to be a spade, and she has no false beliefs regarding other cards.

Let us make a note about presuppositions. Although *know* normally triggers a presupposition, no presupposition arises with embedded questions. Instead, the different readings all relate agents to propositions which are actually true (true answers), and a factive presupposition does not show up beyond this (because, in short, it would simply say that the true answers are true).

Forget

Denotation: Little has been said about *forget*, but the semi-formal denotation in (20) should not be controversial. According to (20), *forget* presupposes that there used to be correct belief (which is different from a factive presupposition, because it is in the past), and asserts that there is no such belief anymore.[13]

(20) $[\![\text{forget}]\!] = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believed}]\!](p)(x)}; \neg[\![\text{believe}]\!](p)(x)\right]$

Readings: Fewer readings have been proposed for *forget*. Most theories predict that it only receives a WE reading. Heim (1994) proposed a systematic way to go between WE and SE readings, which applies here to complete the list in (21).[14]

(21) Mary forgot which of her cards are spades.
- WE: Mary does not remember that all actual spades are spades. = for at least one actual spade, she forgot that it is a spade.
- SE: Mary does not remember that all actual spades are spades and all other cards are not spades. = for at least one card, she forgot whether it is a spade.

[13] We simplify details about time, and simply assume that the presupposition is about a time strictly anterior to the assertion.

[14] There is no systematic definition of an IE reading which would apply blindly to the embedding predicate. Instead, there could be several ways to generalize the IE reading to verbs other than *know* (e.g., adding absence of false belief to the WE reading, or adding negation of a presupposition-less version of the verb), but in practice, for *forget* and downward-entailing predicates in general, all theories block what could correspond to such an IE reading.

Two remarks are in order. First, *forget* presupposes past knowledge and as a result the WE reading is predicted to presuppose that Mary used to know which of her cards are spades in the WE sense, and the SE reading that she used to know it in the SE sense. In the experiment, we ensured that the strongest possible presupposition would always be supported so that this should not play any role. Specifically, the design ensured that the agent had known at some point the suit of each card, which is stronger than any possible presupposition for the relevant sentence. Second, we leave aside issues regarding homogeneity effects among the entities being quantified over in the embedded questions (Cremers, 2015a, 2015b; Križ, 2015c), which may give rise to stronger readings (e.g., "Mary forgot of *all* actual spades that they are spades"). To anticipate on the results, it turns out they did not play any role.

Surprise

Denotation: Several denotations have been proposed for *surprise*. Most prominently, Guerzoni and Sharvit (2007) proposed the Strawson-downward-entailing entry in (22) (from which we dropped the presupposition of speaker-factivity, which is not relevant for our purposes). According to this denotation, 'x was surprised that p' presupposes that p is true and that x knows it, and it asserts that x expected not p to be true. We may also consider the weaker denotation in (23), with which the assertion would merely be that x did not expect p to be true. We refer to this denotation as the 'ne' denotation, for '<u>n</u>ot <u>e</u>xpected'. We note that the wide scope of the negation here is closer to what we see in the denotation we considered for *forget* in (20).

(22) $[\![\text{surprise}]\!]_{\text{G\&S}} = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believe}]\!](p)(x)}; [\![\text{expected}]\!](\neg p)(x)\right]$

(23) $[\![\text{surprise}]\!]_{\text{ne}} = \lambda p_{st}.\lambda x_e.\left[\underline{p \wedge [\![\text{believe}]\!](p)(x)}; \neg[\![\text{expected}]\!](p)(x)\right]$

Uegaki (2015) proposed yet another entry, given in (24). This entry is non-monotonic and given the results from Experiment 2, which showed no sign of non-monotonicity for *surprise*, we disregard this denotation.

(24) $[\![\text{surprise}]\!]_{\text{U}} = \lambda p_{st}.\lambda x_e.[\underline{p \wedge [\![\text{believe}]\!](p)(x)}; \wedge$
$\forall w' \in DOX_x^w, Sim_{w'}(\neg p) <_{x,w}^{\text{exp}} Sim_{w'}(p)]$
$\approx$ "x expected $\neg p$ more than she expected p"

Readings: The WE and SE readings corresponding to each of the denotations in (22) and (23) are given in (25). There is a debate however as to how the presupposition of *surprise* surfaces when it embeds questions (Abels, 2007; B. R. George, 2011; Spector & Egré, 2015). In the experiment, we will abstract away from these difficulties by making sure that the strongest possible presupposition is supported by the evaluation context, as we did for *forget*. Specifically, the setup will ensure that the agent eventually knows the suit of each card.

(25) Peter is surprised by which of his cards are spades.
WE: Peter expected some actual spades not to be spades.
SE: Peter expected some actual spades not be spades,
or some other cards to be spades.
neWE: It is not the case that Peter expected all actual spades to be spades.
neSE: It is not the case that Peter expected all actual spades to be spades
and all other cards not to be spades.

Summary

We have now listed possible readings for sentences with embedded questions under *know* (19), *forget* (21) and *surprise* (25). We will now present the experiment we set up to assess the availability of each of these readings, focussing eventually on the controversial SE reading for *surprise*.

3.4.2 Methods

Task

We adapted the truth-value judgment task that Cremers and Chemla (2014) used to detect exhaustive readings for questions embedded under *know*. The task requires participants to evaluate the truth of sentences with embedded questions against a general background (given in the instructions) and a picture representing both a situation in the actual world and the mental state of the subject of the sentence in this situation. The relevant mental states were beliefs for *know* and *forget* and expectations for *surprise*. Different situations then introduced various discrepancies between the agent's mental state and the reality, which would make different readings of the relevant sentences true and false. This will allow us to run a reading detection analysis as we explain below.

Instructions and training phase

At the beginning of the experiment, participants received the following instructions:

> A group of friends is playing a kind of poker game in which each player gets dealt a hand of 7 cards. When they receive their cards, they only have a quick look at them and try not to show any emotion, while still remembering the cards.
>
> You will see the hands that one of them got in each round, and either which suits they expected or what they remember about the cards. The players often mistake the suits (for instance they remember clubs instead of spades), and sometimes they completely forget what some of their cards were.
>
> You will have to judge whether sentences about their memories or expectations are true or false.

As we mentioned in the introduction, all verbs tested have a factive presupposition which is trivially satisfied when they embed questions, but *forget* and *surprise* have additional presuppositions: *forget* presupposes that the agent used to know the answer to the question, while *surprise* presupposes that she now knows it. While there may be some debates regarding how much the agent had to know/came to know, the story in the instructions ensured that the agent saw all the cards at some point, and therefore ensured that the context would support even the strongest presupposition.

The instructions also included 4 example items with explanations about why the correct answer was **True** or **False** in each case. After the instructions, participants were presented with 5 training items (in random order) to help them familiarize with the task. After answering a training item, participants received feedback (a green "Correct!" message 350ms if the answer was correct, or otherwise a message in red 8s with explanations on why their answer was incorrect, the difference in time created an incentive to be accurate).

Design and Stimuli

Each item consisted of two rows of seven playing cards and a sentence that participants were asked to judge true or false, as can be seen in the examples presented in Figure 3.6. Stimuli were based on two crossed factors: Predicate (3 levels), which was pertinent (mostly) for the sentence, and Situation type (5 levels), which was pertinent for the picture.

Sentences: The sentences were all of the form "NAME VERB which of [his/her] cards are SUIT", where NAME was drawn randomly from a list of 20 male and 20 female names, VERB was one of 'knows', 'forgot' or 'was surprised by', the possessive pronoun agreed with the name and SUIT was one of 'hearts', 'diamonds', 'clubs' and 'spades'. For instance, a sentence for *forget* could be "Mary forgot which of her cards are clubs". Our goal was to explore the truth-conditions of the sentences by displaying them in various conditions, introduced through pictures.

Pictures: Pictures were made of two rows of playing cards. The first row was meant to represent the 'actual' situation. The second row represented the relevant mental state of an agent about each of the cards in the first row. For *know* and *forget*, the relevant type of mental state is belief and each card in the second row could thus be a match with the first row (correct belief), a mismatch (false belief) or a special type of card representing absence of belief by displaying two shapes simultaneously. For *surprise*, the relevant type of mental state concerns expectations and the card could thus be a match (satisfied expectation), a mismatch (failed expectation) or a special card representing absence of firm expectation by representing two shapes simultaneously. For simplicity, we present the rest of this paragraph in terms of beliefs only, thus focussing on *know* and *forget*, hoping that the translation to *surprise* is unproblematic.

This card system thus allows one to represent different types of relations between mental states of an agent and the actual world. In particular, one can introduce two types of discrepancies, I(gnorance) or C(onflict) about a particular card. And the discrepancy could concern either a card which is actually of the suit mentioned in the sentence ($\oplus$), or a card which is actually part of a suit different than the one mentioned in the sentence ($\ominus$). Very much in the spirit of Xiang (2015c) then, we constructed five conditions: one with no discrepancy at all (N), and four with conflicts of each of the different 2×2 possible types: $I^\oplus$, $I^\ominus$, $C^\oplus$, $C^\ominus$. When the condition involved a discrepancy of some kind, there were always two cards instantiating this discrepancy.[15]

All cards were taken randomly between Ace and 10 (thus excluding the king, queen and jack which may have been harder to categorize). The actual cards were

[15] More situations could in principle be constructed and tested: some which would mix different types of discrepancies (e.g., an actual false belief on a positive card, and ignorance on a negative card, a $C^\oplus$+$I^\ominus$ situation, say) or some which would cover the cards differently (e.g., contrasting violated expectations on some positive cards vs. violated expectations on *all* positive cards). These possibilities were left aside for the time being. The 'simple' situations were sufficient to test all the readings listed in (19), (21), (25).

always sorted in the following order: diamonds, hearts, clubs, spades. Depending on whether the sentence was about a red suit (heart/diamond) or a black suit (club/spade), there were 5 red cards and 2 black cards, or the opposite. The expectations/memories were only about the suits, not about the numbers, therefore cards in the second row did not have a number but only a suit symbol (this also discouraged purely visual matching of the top and bottom cards). Whenever there was a discrepancy between the two rows, it was always within color (i.e. a player could mistake a heart for a diamond, but not for a spade).[16]

Interaction between Sentences and Pictures: Table 3.5 provides the predicted truth value of each reading of each sentence in each condition. The table makes visible a few facts we discussed earlier. First, the WE, IE and SE readings only diverge in "negative" situations $I^{\ominus}$ and $C^{\ominus}$, i.e. cases where the discrepancies between a player's mental state and the actual world concern cards which are not of the target suit (the "negative" part of the predicate). Second, we can see what responses are predicted. If we assume the 'not-expect' denotation for *surprise* in (23), we obtain the exact same predictions as we had for *forget*. If we assume the more standard denotation for *surprise* in (22), we then obtain different predictions in the I conditions ($\ominus$ or $\oplus$): If Chris has no specific expectation regarding which suit a card will be, we predict that he cannot be surprised when it turns out to be a diamond (with *forget*, if Chris has no idea whether his third card is a heart or a diamond, we *can* say he forgot the suit of this card).

Participants

47 participants were recruited on Amazon's Mechanical Turk and were paid $1.8 for their participation (age range: 19–65). Two participants were excluded from the analysis because they did not report English as their native language. Three more participants were excluded because their error rate exceeded the average by more than one standard deviation (threshold: 34%; error rates were calculated on uncontroversial items for which all readings lead to the same truth value in Table 3.5).

Analytical and statistical methods

The results were submitted to a reading detection analysis (which has a precursor in Chemla and Spector, 2014). We optimized models of the response data, using as predictors the possible readings (as listed in (19) and (25)). Hence, a positive coefficient in such a model is evidence for the availability of a particular reading (when other plausible readings are taken into account). For each parameter (i.e. each reading), we give the statistics obtained from model comparison as well as a

[16]This, and the fact that Ignorance was represented as an hesitation between 2 suits of the same color, helped control for possible non-monotonic results of *surprise*, corresponding to the denotation proposed in Uegaki (2015). Indeed, he suggests that $[\![\text{surprise}]\!](p)(x) = 1$ if and only iff x expected $\neg p$ more than p. If we were to represent Ignorance with a question mark as in Cremers and Chemla (2014), there would be 4 alternatives (corresponding to the 4 suits). Assuming they are all equally likely, this would yield a probability of $3/4$ for $\neg p$ ("the card is not a spade") and only $1/4$ for p ("the card is a spade"), thus making Ignorance practically equivalent to Conflict. Here we made sure there were only 2 alternatives, ensuring equal probability to p and $\neg p$.

confidence interval in percentage[17] which we interpret as an approximation of the proportion of responses that can be attributed to the corresponding reading.

The fastest and slowest 2% responses were removed. Remaining responses were analyzed with mixed-effect logit models with random effects for participants and items (encoded by suits). Because logit mixed models are computationally more demanding than the linear mixed models we used in previous experiments, we often had to drop several random effects before obtaining models such that at least the parsimonious models derived from them would converge (Bates, Kliegl, et al., 2015). We tried to fit maximal random effects structures for participants, but only random intercept for items (the effects were usually very small for items).

3.4.3 Results

Figure 3.7 presents the raw responses for each condition and for each verb.[18] Table 3.5 presents the results of the statistical analyses using readings as predictors as described above, and the corresponding estimation of the role of each reading in terms of explained proportion of responses for *know* and *surprise*.

We provide more detailed analyses for each of the verbs below, but we can immediately summarize the results. For *know*, we obtain clear evidence for SE readings, WE readings and FBS-MS readings. The results for *forget* are best understood as the mirror image of the results for *know*, as if the readings of *forget* were the exact negations of the readings of *know* (more on this below). For *surprise*, there was clear evidence for the WE reading, as expected, as well as for the more controversial SE and neSE readings.

Know

We fitted a logit mixed model on responses to *know* items with predictors corresponding to the 5 readings in (19). Note that the WMS reading is true in all 5 conditions, therefore it corresponded to the intercept of the model, i.e. the rate of **True** answers when all predictors/other readings are false, which also happens to correspond to the $C^{\oplus}$ condition. As such, and in the absence of a baseline for errors, we could compute a confidence interval but no informative p-value.

The results, reported in Table 3.5a, indicate that all readings but the IE reading were significantly detected. As suggested by the low rate of **True** responses in the $C^{\oplus}$ condition, the WMS reading was estimated below 6% and may still be conflated with simple errors.

Forget

We ran a similar model for *forget*. As indicated in Table 3.5b, both potential readings (WE and SE) played a significant role. Yet, looking at Figure 3.7, it is clear that this model does not provide a satisfying explanation of the results on *forget* items. Due to the very limited number of readings proposed for questions embedded under *forget*, it cannot explain all the differences we observed, e.g., between the $I^{\ominus}$ and $C^{\ominus}$ conditions, nor between the $I^{\oplus}$ and $C^{\oplus}$ conditions.

[17] These confidence intervals were obtained by estimating confidence intervals on the logit scale, transposing them by the estimated intercept and converting back to the 0 – 100% scale.

[18] As the figure suggests, we note that all simple 2-by-2 comparisons between conditions for each verb turned out significant ($\chi^2(1) > 5$), except for the $C^{\ominus}$ and $I^{\oplus}$ conditions on *know* and *forget* items and the $I^{\oplus}$ and $I^{\ominus}$ conditions on *surprise* items (all three $\chi^2(1) < 1.3$).

Reading	N	$I^\ominus$	$C^\ominus$	$I^\oplus$	$C^\oplus$	Confidence interval	Statistics
SE	1	0	0	0	0	38% – 88%	$\chi^2(1) = 37, \quad p < .001$
IE	1	1	0	0	0	<10%	$\chi^2(1) = .1, \quad p = .721$
WE	1	1	1	0	0	8% – 20%	$\chi^2(1) = 29, \quad p < .001$
FBS-MS	1	1	0	1	0	10% – 24%	$\chi^2(1) = 43, \quad p < .001$
WMS	1	1	1	1	1	2% – 6%	Could not be tested

(A) Readings, predicted responses and estimated share of participants' responses for each reading of "X knows which of his/her cards were clubs". The proportion of WMS readings could not be statistically tested because we lacked a false baseline to compare the $C^\oplus$ condition to.

Reading	N	$I^\ominus$	$C^\ominus$	$I^\oplus$	$C^\oplus$	Confidence interval	Statistics
SE	0	1	1	1	1	56% – 85%	$\chi^2(1) = 115, \quad p < .001$
WE	0	0	0	1	1	9% – 29%	$\chi^2(1) = 24, \quad p < .001$

(B) Readings, predicted responses and estimated share of participants' responses for each reading of "X forgot which of his/her cards were clubs". Due to the limited number of readings proposed for *forget*, this model could not explain any difference between the $I^\ominus$ and $C^\ominus$ conditions, nor between the $I^\oplus$ and $C^\oplus$ conditions.

Reading	N	$I^\ominus$	$C^\ominus$	$I^\oplus$	$C^\oplus$	Confidence interval	Statistics
SE	0	0	1	0	1	8% – 55%	$\chi^2(1) = 17, \quad p < .001$
neSE	0	1	1	1	1	4% – 48%	$\chi^2(1) = 7.6, \quad p = .006$
WE	0	0	0	0	1	5% – 29%	$\chi^2(1) = 15, \quad p < .001$
neWE	0	0	0	1	1	?	Could not be tested

(C) Readings, predicted responses and estimated share of participants' responses for each reading of "X was surprised by which of his/her cards were clubs".

TABLE 3.5 – In these tables, we list the various readings tested for *know*, *forget* and *surprise*, the expected responses for a participant who would access each of these readings, and the estimated proportions of responses that can be attributed to each of them in our data.

A simple look at the raw data in Figure 3.7 shows that *know* and *forget* look like the mirror image of each other. More specifically, responses to *forget* were strongly anti-correlated to responses to *know*, across conditions and participants (Pearson's product-moment correlation: $\rho = -.94$, $t(208) = -39$, $p < .001$).

We ran a logit mixed model on responses to *know* and *forget*, after flipping all responses to *forget* items (encoding **True** responses as **False** and *vice versa*, in order to directly compare the rates of **True** responses to *know* with the rates of **False** responses to *forget*). The model included Condition (5 levels), Verb (2 levels) and their interactions as fixed effects, and random slopes for Verb per participant, but not for Condition (would not converge otherwise), and random intercepts for Items (suits). Overall, the factor Verb explained very little variance ($\chi^2(5) = 7.2, p = .20$) and judging from the estimated z-values, only one interaction corresponding to the $C^{\oplus}$ condition approached significance, and only so before correction for multiple comparison ($z = 2.2, p = .03$, all other z's $< .5$).

Surprise

On responses to *surprise* items, we fitted a model with the 4 readings in (25) as predictors, as well as an intercept. Given that none of the tested readings was true in the P condition, the intercept roughly corresponded to the P condition and could be interpreted as the baseline rate of **True** responses due to errors. Dropping either the WE or the neWE reading had little effect on the model's fit (both $\chi^2(1) < 1.6, p > .2$), yet dropping both had a very significant effect ($\chi^2(2) = 19, p < .001$). It turned out that the two predictors were strongly correlated (estimated correlation of the fixed effects: $-.85$).

Going back to the raw data, as previously mentioned, there was no significant difference between the $I^{\oplus}$ and $I^{\ominus}$ conditions, which only differ on the neWE reading ($\chi^2(1) = .2, p = .6$), but there was a major difference between the $C^{\oplus}$ and $C^{\ominus}$ conditions which differ on the WE and neWE readings ($\chi^2(1) = 9, p = .003$). Therefore we decided to drop the neWE reading from the model. In the second model, reported in Table 3.5c, all three remaining readings were significantly detected (all $\chi^2(1) > 7, p < .01$).

3.4.4 Discussion

SE readings for all predicates

First, we partially replicate Cremers and Chemla's (2014) results about *know*, in that we observed WE and SE readings. We also found a significant proportion of FBS-MS readings (which have the no-false-belief constraint of the IE without the exhaustivity). However, we did not obtain evidence of the IE reading, which was the predominant reading found in (Cremers & Chemla, 2014).[19]

The results on *forget* are somehow puzzling. At first, one would think that provided the presuppositions of both verbs are satisfied, we expect the proposition-embedding *forget* to be equivalent to the negation of *know*. Assuming the question-embedding entries are reducible to the proposition embedding entries would predict the WE readings of *know* and *forget* to be negations of each other (and possibly so for their SE readings). However, this would *not* explain why we observe a reading of *forget* which seems to be the negation of the FBS-MS reading of *know* (because

[19] The current task elicits more SE readings for *know* than Cremers and Chemla's (2014) task. Note that if this was due to a bias towards **False** answers, the chances to detect SE readings with *surprise* would be reduced.

this reading is crucially a case of non-reducibility, see B. R. George, 2011, 2013).[20] The most plausible interpretation of the result for *forget* then is a task-specific strategy, according to which participants would deal with the *forget* sentences by first evaluating corresponding, simpler sentences with *know*, and then flip their answers. Such an effect may for instance be reduced in a between-subject design.

For *surprise*, we detected WE and SE readings. We also detected the neSE reading, which suggests that some participants considered that an agent could be said to be surprised even when she did not exclude that the actual outcome was possible. Since this neSE reading does not seem to correspond to the intuitive meaning of *surprise* it is tempting to attribute this result to a low-level strategy again and some possibilities come to mind. For instance, one could try to defend the idea that participants come to accept sentences with *surprise* in all conditions but N, or to treat *surprise* as a negation of *know* with a strongly exhaustive reading. Yet another, more sophisticated possibility is that (i) participants had the non-monotonic reading in (24) and (ii) in a case of absence of expectations, they considered that the agent had stronger expectations for the wrong suit. This second assumption basically reduces the neSE to a genuine SE reading with Uegaki's denotation (24). What this leads us to anyway, is that we found clear SE reading (with the standard denotation) and then maybe on top of this other SE readings with an ne denotation or with Uegaki's denotation (plus auxiliary, task related assumptions). In the next section, we will focus on the SE reading associated with the most standard denotation, as evidenced by the very high proportion of **True** responses in the $C^{\ominus}$ condition, and discuss in more detail possible alternative explanations to this important fact.

SE readings with *surprise*, alternative interpretations?

We interpret the high rate of **True** responses in the $C^{\ominus}$ condition as evidence that *surprise* can give rise to SE readings. This conclusion would go against a long history of introspective judgments. An alternative view on this result has been suggested to us.[21] Imagine that participants interpreted or read "Jacob was surprised by which of his cards are clubs" as (26). It seems intuitive that the latter is true in the $C^{\ominus}$ condition, which would thus explain the high rate of **True** responses in that condition. But if this is correct, this would still be evidence in favor of a SE reading here, according to Beck and Rullmann (1999) who predict the readings in (27).

(26) Jacob was surprised by how many of his cards are clubs.

(27) a. WE: Jacob expected a lower number of clubs among his cards.
b. SE: Jacob expected a different number of clubs among his cards.

[20] The only existing theories we are aware of which derive the FBS-MS reading are Roelofsen et al. (2014) and Xiang (2015a). However, there is a general agreement that its derivation should parallel that of the IE reading on the exhaustive side. Yet, no theory perdicts IE readings for *forget*. For the theories in the exhaustification tradition (Klinedinst & Rothschild, 2011; Uegaki, 2015), the IE reading is a pragmatic strengthening of the WE reading and it is vacuous when applied to *forget*. For alternative theories (Spector & Egré, 2015; Roelofsen et al., 2014), the IE reading is obtained through a different composition rule or operator, which is blocked for downward-entailing predicates.

[21] Thanks to the audience at SIASSI Berlin 2015, and in particular Angelika Kratzer for this suggestion. We would like to point that the objection would apply equally to our experimental results and to the example presented by Klinedinst and Rothschild (2011, fn. 18).

Hence, even under this re-interpretation view, we obtain evidence for a SE reading of questions embedded under *surprise*, albeit maybe for *how-many* rather than *which* questions.[22]

Finally, one could imagine other reinterpretations of the embedded question. The logic behind the *how-many*-reinterpretation was that in the context of a card game it may not matter which cards are clubs, but simply how many of each suit you got. We showed that even under this reinterpretation, the **True** responses were underlyingly SE readings, but this may not be the case in general. We will not discuss this broader objection because it is too vague to be rejected. Indeed, with no limit on which reinterpretations must be considered, one can always build a sentence which WE reading corresponds to the SE reading of the original sentence (for instance: "It surprised Mary which cards are clubs [and which are not]").

3.4.5 Conclusion/summary for Experiment 3

In Experiment 3, we found that all verbs we tested gave rise to SE readings. While the interpretation of the results for *forget* is obscured by an apparent low-level strategy of the participants, we are confident that the results with *surprise* do indicate genuine SE readings. The availability of SE readings is a challenge for many theories which are either designed to derive the unavailability of this reading (Guerzoni & Sharvit, 2007; Romero, 2015) or take it as the cause for the unacceptability of *whether*-questions (Andreea Cristina Nicolae, 2013; Andreea C Nicolae, 2015; Guerzoni & Sharvit, 2014).

3.5 Conclusion

3.5.1 Summary of the results

The acceptability judgments collected in Experiment 1 confirmed the soundness of the typology proposed in the literature (in line with previous studies on the convergence of quantitative methods and introspective judgements, Sprouse and Almeida, 2013; Sprouse et al., 2013). Each class of verbs allowed embedding of the type of complements they were expected to allow, and there was little variation within each class. In particular, we confirmed the widespread introspective judgment regarding the unacceptability of *whether*-questions under emotive-factive predicates. However, the unacceptability seems to be 'soft', in that *whether*-questions are not as unacceptable under emotive-factives as they are under non-responsive verbs (in line with Sæbø, 2007). We also confirmed the judgements

[22] One may wonder whether the reading described as a SE reading in (27b) can be obtained differently, and in particular by assuming directly an exact reading for numerals (as opposed to an at least reading, as Beck and Rullmann, 1999, do). It is in fact possible, but we then cannot obtain the other reading (27a). A strong argument for a theory that can derive this reading of *how-many* questions comes from the asymmetry between (i) and (ii), already noted by Beck and Rullmann (1999). With exact readings only, this asymmetry cannot be captured.

(i) It surprised Mary how many guests showed up.
Interpretation: it surprised her that *so many* guests showed up.

(ii) It surprised Mary how many eggs are sufficient to bake this cake.
Interpretation: It surprised her that *so few* eggs are needed.

Summing up, there is strong evidence for at least readings of *how many* questions. The most straightforward explanation for the possibility of the reading in (27b) then is that it is a SE reading (and not an independently existing exact reading for *how many* questions).

of Beck and Sharvit (2002) regarding rogative predicates and QVE, and found an interesting contrast between polar and alternative questions which seems to reflect dialectal variation within American English and may support Guerzoni and Sharvit's (2014) derivation of polar questions as alternative questions.

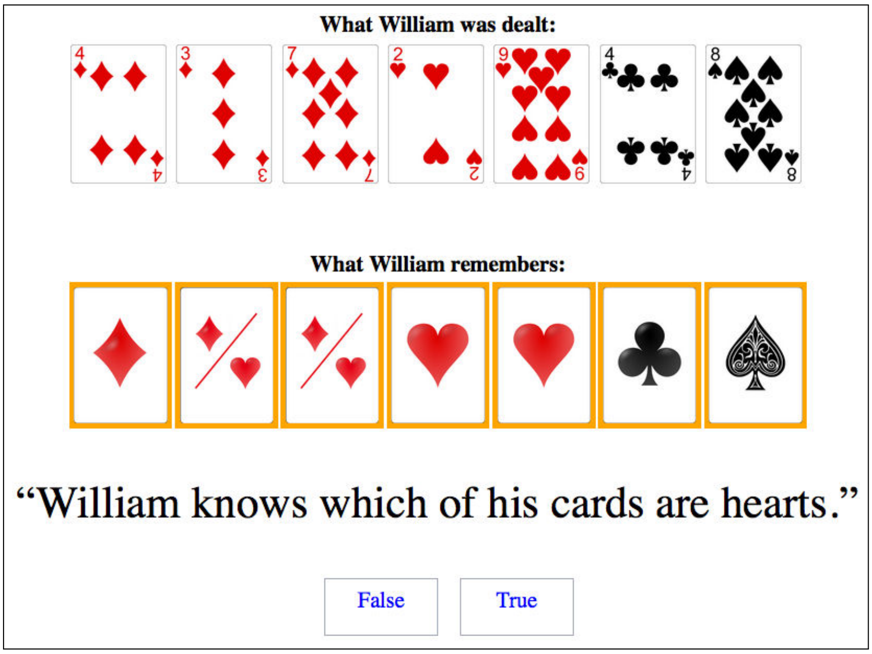
Experiment 2 collected inferential judgments between sentences with various responsive predicates embedding declarative complements. In this experiment, *forget* and *surprise* patterned together, and were judged more downward entailing than *know*, *agree* and *be happy*. These results suggest that monotonicity is independent from selectional properties, since the distinction between upward/downward monotonicity is orthogonal to the typology from Experiment 1. All responsive predicates were judged monotonic.

Experiment 3 was a truth-value judgment task. A reading detection analysis revealed that both *know* and *surprise* give rise to SE readings when embedding questions. The fact that *surprise* does give rise to SE readings goes against most introspective judgments in the literature. It also suggests that the unacceptability of *whether*-questions under emotive-factive cannot be explained by the unavailability of a SE reading.

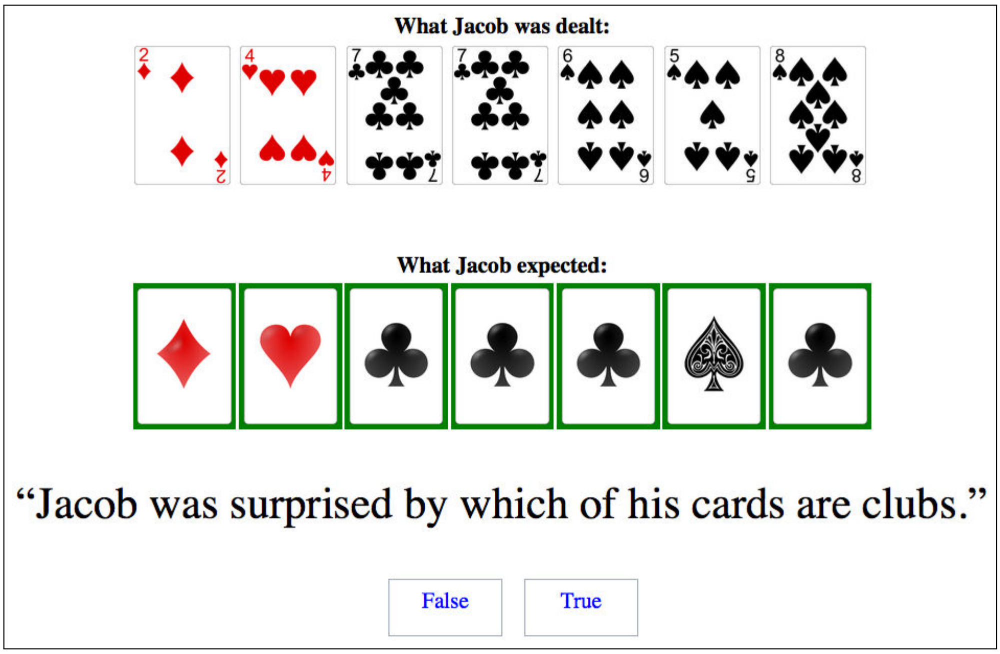
3.5.2 Conclusions

Although no theory has directly linked selectional properties of emotive-factive predicates with their monotonicity, it has been argued that the selectional properties may be explained by the unavailability of SE readings, which in turn may be explained by monotonicity properties . The results of Experiments 1 and 2 showed that monotonicity and selectional properties vary independently, so the two connections cannot both hold. The results of Experiments 1 and 3 suggest that the broken link is (at least) the connection between strong exhaustivity and the acceptability of *whether*-questions (contra Andreea Cristina Nicolae, 2013; Andreea C Nicolae, 2015; Guerzoni and Sharvit, 2014). So far, strong exhaustivity may still be linked to monotonicity (or perceived monotonicity, Chemla et al., 2011), however. Indeed, *surprise* was not judged as downward entailing as the DE controls in Experiment 2, and it may be that the participants who had a SE reading for *surprise* in Experiment 3 did not access a clearly Strawson-downward-entailing meaning for this verb.

In sum, the present studies provide a range of empirical facts about emotive-factives and other question embedding predicates, describing acceptable constructions, monotonicity properties and potential readings. These studies immediately contribute to some empirical debates in the literature, in particular by providing the strongest evidence to date in favor of the existence of SE readings for questions embedded under *surprise*.



(A) An I$^{\ominus}$ target for *know*. Under a SE reading, we expect a 'False' answer. Under a WE or an IE reading we expect a 'True' answer.



(B) A C$^{\ominus}$ target for *surprise*. Under a SE reading, we expect a 'True' answer. Under a WE reading we expect a 'False' answer.

FIGURE 3.6 – Two items from Experiment 3.

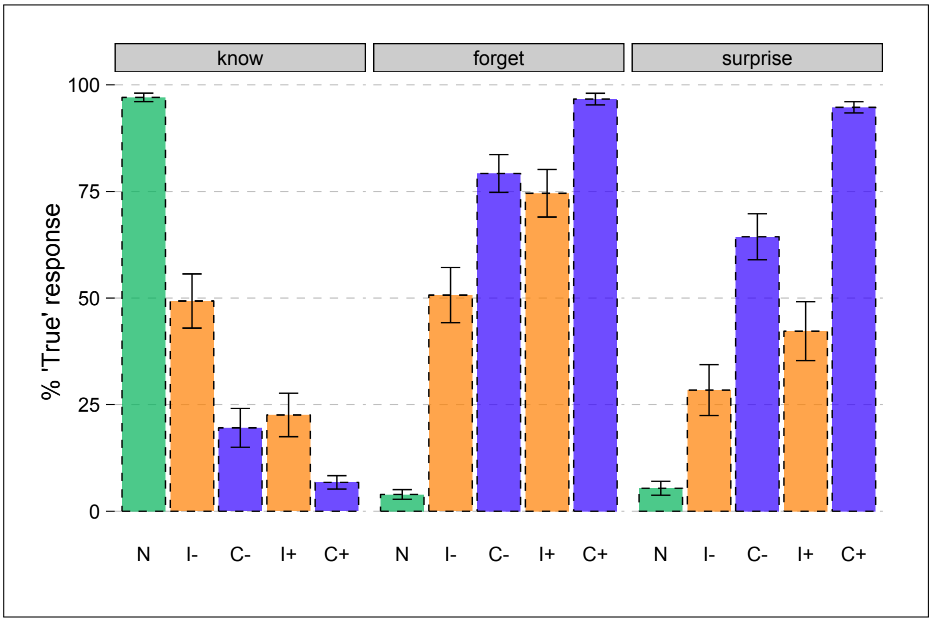


FIGURE 3.7 – Percentage of '**True**' responses for each verb and condition in Experiment 3. In a 2-by-2 comparison, the only differences which turned out non-significant were between the $C^{\ominus}$ and $I^{\oplus}$ conditions under *know* and *forget*, and between the $I^{\ominus}$ and $I^{\oplus}$ conditions under *surprise*.

Appendices

A Lists for Experiment 1

20 proper names were taken from the most frequent male and female names in the US. We selected 28 adjectives which described appearance or physical properties for aliens, to avoid interaction with the activity predicates as much as possible.

Names: *Alex, Anthony, Ashley, Ava, Benjamin, Chloe, Chris, Elizabeth, Emily, Emma, Ethan, Grace, Jacob, James, John, Madison, Michael, Olivia, Sarah, William*

ADJ: *blue, cubical, flexible, fluffy, fuzzy, green, hollow, large, opaque, orange, pink, purple, red, round, slender, slim, small, smooth, speckled, spherical, spiky, spiny, spotty, stringy, striped, transparent, watery, yellow*

For activity predicates, we created the 20 predicates in the **PRED** column of Table A.1. We also used versions with a disjunction to generate alternative questions, in the $\textbf{PRED}_\vee$ column.

PRED	$\textbf{PRED}_\vee$
play the piano with their wings	play the piano with their wings or with their feet
can eat 5 pounds of licorice	can eat 5 pounds or 10 pounds of licorice
drink soda with a spoon	drink soda with a straw or with a spoon
write poems about the moon	write poems about the moon or about the sun
ride tall purple horses	ride purple horses or blue horses
read 18th century books	read 18th century or 19th century books
visit archeology museums	visit archeology museums or geology museums
sleep with their head down	sleep with their head up or with their head down
lay translucent eggs	lay translucent eggs or opaque eggs
speak several African languages	speak Asian languages or African languages
can count to two million	can count to two million or to three million
freeze certain tropical plants	freeze or fry certain tropical plants
talk with their noses	talk with their noses or with their ears
hibernate every three years	hibernate every three years or twice a year
believe in green unicorns	believe in green unicorns or blue dragons
listen to classical music	listen to classical music or to techno music
watch movies from the 30's	watch movies from the 30's or from the 40's
use smartphones to cook pasta	use smartphones to cook pasta or to serve wine
drive old cars	drive old cars or rusty motorbikes
use the Korean alphabet	use the Korean alphabet or the Greek alphabet

TABLE A.1 – List of **PRED** used in Experiment 1

B Lists for Experiment 2

We used the same **names** as in Experiment 1. We used the same 28 adjectives, but each was paired with another related adjective in case it would appear in one of the control conditions (which required two difference adjectives). The pairing is presented in Table B.1. The predicates used in this experiment are presented in Table B.2, in their weak ($\textbf{PRED}_-$), strong ($\textbf{PRED}_+$) and conjunctive ($\textbf{PRED}_\wedge$) versions.

ADJ$_1$	**ADJ**$_2$	(continued)	
blue	red	slender	slim
cubical	spherical	slim	stringy
flexible	rigid	small	large
fluffy	spiky	smooth	fuzzy
fuzzy	fluffy	speckled	slender
green	purple	spherical	cubical
hollow	watery	spiky	spiny
large	small	spiny	speckled
opaque	transparent	spotty	striped
orange	blue	stringy	round
pink	orange	striped	spotty
purple	yellow	transparent	opaque
red	green	watery	hollow
round	flexible	yellow	pink

TABLE B.1 – List of **ADJ** used in Experiment 2

PRED$_-$	**PRED**$_+$	**PRED**$_\wedge$
burn flowers	burn roses	burned roses and tulips
buy clothes	buy shirts	bought shirts and trousers
color trees	color pines	colored pines and oaks
compliment humans	compliment children	complimented children and teenagers
destroy musical instruments	destroy violins	destroyed violins and guitars
drink sodas	drink coke	drank coke and lemonade
drive cars	drive Toyotas	drove Toyotas and Fords
eat at restaurants	eat at Mexican restaurants	ate at Mexican and Chinese restaurants
eat meat	eat pork	ate pork and beef
kiss animals	kiss dogs	kissed dogs and cats
play with toys	play with toy cars	played with toy cars and toy soldiers
read books	read sci-fi novels	read sci-fi novels and love novels
read magazines	read news magazines	read news magazines and sports magazines
see birds	see doves	saw doves and crows
taste cookies	taste chocolate cookies	tasted chocolate cookies and caramel cookies
throw balls	throw tennis balls	threw tennis balls and soccer balls
use coins	use quarters	used quarters and dimes
use the internet	send emails	sent emails and visited websites
visit museums	visit French museums	visited French museums and Italian museums
watch sports matches	watch baseball matches	watch baseball matches and football matches

TABLE B.2 – List of weak, strong and conjunctive versions of **PRED** used in Experiment 2

Additional discussion

Part II

Theoretical contribution

Chapter 4

Plurality effects and exhaustive readings of embedded questions

This chapter presents work in progress, which is susceptible to change a lot in the months to come. Please be indulgent if you spot any incoherence and feel free to contact me or to check my website for up-to-date versions.[†]

Abstract

The goal of this chapter will be to describe similarities between (a) definite descriptions and (b) questions embedded under responsive predicates, and then suggest an analysis which would allow us to extend theories of (a) to (b), building largely on Dayal (1996) and Lahiri (2002). In a second part, I will show how this is compatible with a theory of strong and intermediate exhaustivity using a strengthening mechanism similar to implicatures, as suggested in Klinedinst and Rothschild (2011). We will then see how the theory explains a few puzzles from previous literature and discuss new predictions and shortcomings.

4.1 Plurality effects with definite descriptions and questions

Questions have been shown to behave like (definite) plurals in many aspects. Each of the phenomena affecting definite plurals in (1-3a) has an equivalent with embedded questions (1-3b).

(1) Quantificational Variability Effect (QVE)[1] / Modification by an adverb of quantity (Berman, 1991):

a. "The students mostly arrived"
⇔ Most students arrived.

b. "John mostly knows which students arrived"
⇔ For most students who arrived, John knows it.

(2) Semi-distributive / cumulative readings (Lahiri, 2002):

[†]I would like to thank the audiences at NYU Department of Linguistics, MIT Department of Linguistics & Philosophy, the Workshop "Sentences and Embedded clauses" at IHPST (Paris) and at Sinn und Bedeutung 20 (Tübingen). For additional discussion and feedback, I would like to thank Maria Aloni, Lucas Champollion, Emmanuel Chemla, Danny Fox, Benjamin George, Nathan Klinedinst, Hadas Kotek, Manuel Križ, Friederike Moltmann, Yael Sharvit, Benjamin Spector, Anna Szabolcsi, Wataru Uegaki and Yimei Xiang.

[1]Berman (1991) makes an analogy with modification of an indefinite by adverbs of frequency, as in "A professor usually likes his students". This led to a theory of questions as free variables lacking a quantificational force. Although I won't pursue such an analysis, I will stick to the widely accepted acronym 'QVE'.

a. "The boys talked to the girls"
⇔ Each boy talked to some girl and each girl "was talked to" by some boy.
b. "The witnesses knew which klansmen were present at the lynching"
⇔ Each witness knows a partial answer and each klansman who was present is known to be so by at least one witness.

(3) Homogeneity (Xiang, 2015a; Križ, 2015c):
a. "The students didn't come to the class"
⇔ No student came to the class.
b. "Mary doesn't know which students arrived"
⇔ Mary has no idea which students arrived.

While (1) has been studied at length since its discovery by Berman (1991) up to Lahiri's (2002) major contribution, only Lahiri proposed an analysis of (2) (the data was mentioned in Dayal, 1996), and it is only very recently that (3) received any attention.

A few words on Homogeneity:

- Intuitively: (4a) is true if all students arrived, false if none did, but neither true nor false otherwise. (4b) is true if Mary knows a complete answer to the question "Which students arrived?", false if she doesn't know any answer, and neither true nor false if she has partial knowledge regarding which students arrived.
- The examples in (5) show that the inferences in (3) cannot be explained by having the definite description/question take scope over the negation. Indeed, the binding of the pronoun forces it to stay in the scope of the negative DP, yet the homogeneity effect remains. As illustrated with example (19) discussed in the next section, questions embedded under *know* cannot take scope over the subject of the sentence anyway.
- As noted by Löbner (1985, 2000), homogeneity effects disappear with adverbs: if some but not all students smoke, (6a) is plain false and (7a) is true. Similarly, if Mary know for some, but not all students who arrived that they did, (6b) is plain false and (7b) is true.

(4) a. The students arrived.
b. Mary knows which students arrived.

(5) a. No professor$_i$ talked with the students she$_i$ likes.
b. No professor$_i$ knows which of her$_i$ students arrived.

(6) a. The students all smoke.
b. Mary completely knows which students arrived.

(7) a. The students don't all smoke.
b. Mary doesn't completely know which students arrived.

4.2 Embedded questions as definite plurals in the literature

Several proposals have already linked questions with plural nouns. In this section, we will introduce the standard semantics for plural nouns and the theories of questions which draw on it.

4.2.1 Plurals and definite descriptions

Link (1983) proposed a very influential account of plurality in natural languages. Let us present the main assumptions of this proposals.[2]

- The domain of individuals, D_e, contains not only *atomic* individuals ($a, b, \ldots$), but also *plural* individuals, which we will write $a \oplus b$. D_e is assumed to be closed under the sum operation ($\oplus$).
- An order relation can be defined on elements of D_e (i-part): $x \preceq y$ iff $x \oplus y = y$ (for instance, 'Ann' is a part of 'Ann and Bill')
- A $\star$-operator transforms predicates of atomic individuals (e.g., *girl*) into plural predicates (*girls*):

(8) $^{\star}P(X) = 1$ iff $\forall y \preceq X, Atomic(y) \rightarrow P(y) = 1$
$[\![\text{girls}]\!] = {}^{\star}[\![\text{girl}]\!]$: 'girls' is true of any plurality which atoms are all girls.

The plural definite article *the* takes a plural predicate and returns its maximal element (the biggest plural individuals which satisfies the predicate):

(9) $$\begin{aligned} [\![\text{the } {}^{\star}P]\!] &= \sigma x.{}^{\star}P(x) \\ &= \iota x.{}^{\star}P(x) \wedge \forall y[{}^{\star}P(y) \rightarrow y \preceq x] \end{aligned}$$

While this is not frequent in the literature on plural definites nor very useful at this point, we can see *the* as an overt type-shifter from predicates (<e, t>) to individuals (e).

4.2.2 Some previous theories of plurality in questions

Conjunction and material implication for propositions allow an analogy with sum-formation and i-part relation for individuals. The first proposal to apply Link's (1983) model of plurality to questions is Dayal (1996). In order to account for multiple *wh*-questions and the unicity presupposition of singular *which*-questions, she argued that we must distinguish singular *which*-questions from plural-marked *which*-questions and proposed to do so by translating the plural marking on 'which student**s**'. However, she did not consider the implications for the plurality effects presented in the previous section, except for a quick discussion of sentences like (2b).

Lahiri (2002) introduced a sophisticated theory of questions, designed to explain plurality effects (although he does not discuss homogeneity). In his theory, questions map onto *Proposition Conjunction Algebras*, a structure similar to that of a starred predicate in Link's (1983) theory, and this structure allows QVE and semi-distributive readings.

In (10), we see more concretely how the plural morphology on a *which*-phrase can be translated with Link's (1983) $\star$-operator, in a Hamblin (1973) style denotation. Note the correspondence in (11a) between '$\oplus$' in the domain of the *which*-phrase and '$\wedge$' in the domain of answers, and the correspondence in (11b) between the i-part relation '$\preceq$' and the material implication '$\leftarrow$'.[3]

[2]Here and in the rest of this chapter, we will only consider the individual sum $\oplus$ and the i-part relation which we will write '$\preceq$'. We will not discuss the material fusion and the m-part relation. Indeed, we will focus on *which*-questions. *Which* only combines with count-nouns, and the i-part relation will translate into to the implication relation between answers to such questions. For simplicity, we will not distinguish between the operators in the object language and their translations.

[3]Crucially, these correspondences are only possible because *arrive* is distributive. If the *which*-phrase was associated with a collective predicate, the correspondences would disappear.

(10) $[\![$Which students arrived$]\!] = \lambda p.\exists X\,[{}^{\star}[\![\text{student}]\!](X) \wedge p = \lambda w.{}^{\star}[\![\text{arrived}]\!](w)(X)]$
$= \{$Mary arrived, Peter arrived, Peter and Mary arrived…$\}$

(11) a. $[\![$Peter$\oplus$Mary arrived$]\!] \equiv [\![$Peter arrived$]\!] \wedge [\![$Mary arrived$]\!]$
b. Peter$\preceq$ Peter$\oplus$Mary
$[\![$Peter arrived$]\!] \leftarrow [\![$Peter $\oplus$ Mary arrived$]\!]$

4.2.3 Answer operators

Most theories of embedded questions share a common feature: they make use of an answer operator to combine questions with declarative-embedding predicates such as *know*. Several examples from influential theories can be found in (12).

(12) a. Heim (1994): $ans_1(Q, w) = \cap(Q(w))$
b. Dayal (1996): $Ans(Q, w) = \iota p.[p \in Q \wedge p(w) \wedge \forall p' \in Q, [p'(w) \rightarrow p \subset p']]$
c. Beck and Rullmann (1999): $answer_1(w)(Q) = \cap\{p : Q(w)(p) \wedge p(w)\}$

Beyond their differences, these operators have a few common properties.

- They restrict the set of answers to true answers (if this is not done already in the denotation of the question, as proposed by Karttunen, 1977),
- They type-shift the question from a set of propositions (or a set of sets of propositions) to a proposition,
- They usually do so by selecting the most informative proposition among the true answers or by building it (taking the conjunction of all true answers).

Some versions of the answer operator are clearly transpositions of the denotation for the definite article *the* proposed in (9). This is particularly true of Dayal (1996), who was the first to use 'ι' in her definition, and it is explicit in Fox (2012) who defines Ans as "*The True*".

4.3 A possible implementation

4.3.1 Hypotheses

Building on Hamblin (1973), Dayal (1996), Lahiri (2002) and Fox (2012), we will use the question denotation in (13). I will assume that this denotation is derived as proposed in Kotek (2015). I will assume that this question denotation combines with the answer operator in (14), which takes a second argument C.

(13) $[\![$Which students arrived$]\!] = \lambda p.\exists X[{}^{\star}\text{student}(X) \wedge p = \lambda w.{}^{\star}\text{arrived}(w)(X)]$

(14) $[\![Ans]\!](Q)(C) = \iota p.[Q(p) \wedge C(p) \wedge [\forall p', Q(p') \wedge C(p') \rightarrow p \subseteq p']]$

Concretely, C will be a restrictor provided by the embedding verb and $Ans(Q)(C)$ can be understood as "The C p in Q". In this, I am following Lahiri (2002).[4] For veridical responsive verbs such as *know*, C is simply the set of all true propositions, and we recover Fox's answer operator: $Ans(Q)(C)$ = "The true Q". For pair-list

[4]For further arguments in favor of the lexically specified restrictor and discussion of a wider variety of embedding predicates, see 5.4 in the next chapter.

readings of multiple-*wh* questions, we need to recursively define a generalized answer operator for higher types of questions (Dayal, 1996; Fox, 2012; Kotek, 2014, 2015).

(15) $[\![Ans]\!](Q_{<\sigma,t>})(C) = \forall P \in Q, [\![Ans]\!](P)(C)$

The last thing we need is a definition of 'atomic answers', just like there are 'atomic individuals'. Although we showed that a strong analogy can be drawn between definite descriptions and questions, it is not perfect. Indeed, if we follow Link (1983) in assuming that D_e is an atomic Boolean algebra, we can always retrieve the atoms from a plural individual. However, we cannot do so from an arbitrary proposition, because there is no such thing as an 'atomic proposition' in general (the domain D_{st} is not an atomic Boolean algebra). Nevertheless, denotations of usual questions do form atomic algebras on D_{st}. Therefore, atomic answers can be defined, but only in relation to a question.

At this point, we will simply assume that some information about the algebra of answers is preserved after application of Ans, but the second line of (17) below is ill-defined. In the next section, we will define a focus value for $Ans(Q)(C)$ which will contain all the information lost in the ordinary value after application of Ans. In footnote 8 we will update the definition of atoms into a fully compositional one. For convenience, we will define At operators for both sets and pluralities.

(16) $At_e(P) = \lambda x.[P(x) \wedge \forall y.[(P(y) \wedge y \preceq x) \rightarrow y = x]]$
$At_e(X) = At_e(\lambda x_e.x \preceq X)$

(17) $At_{st}(Q) = \lambda p_{st}.\left[p \in Q \wedge \forall p' \in Q, [(p \subseteq p') \rightarrow (p = p')]\right]$
$At_{st}(Ans(Q_{<st,t>})(C)) = At_{st}(Q) \cap C$

Our definition of At for questions must be refined if we want to consider multiple *wh*-questions (see Lahiri, 2002, p.202). In order to get rid of propositions of the form "$a \oplus b$ love $c \oplus d$", the right definition would be:

(18) $$At_{st}(Q) = \lambda p_{st}.\Big[p \in Q \wedge \forall r \in Q, [(p \subseteq r) \rightarrow (p = r)] \wedge \bigwedge \left\{q \in Q | q \neq p \wedge \forall r \in Q, [(q \subseteq r) \rightarrow (q = r)]\right\} \not\subseteq p\Big]$$

We now have all the tools needed to extend existing theories of plurality effects to questions:

- QVE can be derived as in Lahiri (2002), although I will suggest a few modifications.
- On the simple cases we consider in this paper, the different theories of cumulative readings do not make different predictions. I will propose one implementation for concreteness, but the reader is free to choose another one if they prefer (see Champollion, 2015 for a review).
- Homogeneity can be implemented with supervaluationist theories (Spector, 2013b; Križ & Spector, 2015; Križ, 2015a). Other theories of homogeneity exist, but they may not all be compatible with the theory.[5]

[5] Löbner (1985, 2000) propose to treat homogeneity as a presupposition and this approach may be compatible with the present theory. Magri (2013) propose to treat it as an implicature. This will not be compatible with the theory proposed in section 4.4.

4.3.2 Application

Quantificational Variability

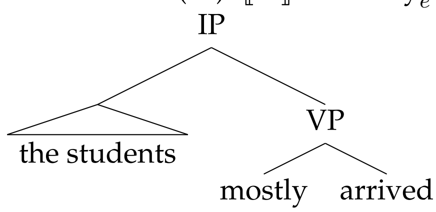
My theory of QVE will essentially be an adaptation of Lahiri's (2002). A main feature of this theory is that questions embedded under responsive predicates must undergo movement to the restrictor position of an adverb (overt or covert). This movement, modeled after Quantifier Raising (QR), is called Interrogative Raising (IR). I will stick to the idea that QVE involves IR, but to a position lower than assumed in Lahiri (2002). Indeed, the example in (19) suggests that the question cannot take scope over the subject of the sentence (the unavailable reading is what Lahiri would predict).

Following Preuss (2001), I will assume the structure in (20b), where IR targets Spec,VP. Her reason for this move was to let the embedding verb determine the strength of exhaustivity (as first suggested by Heim, 1994). Considering the data gathered in Part I, I will treat exhaustivity as independent from the embedding verb, yet I will assume that the embedding verb does determine the lexical restrictor.

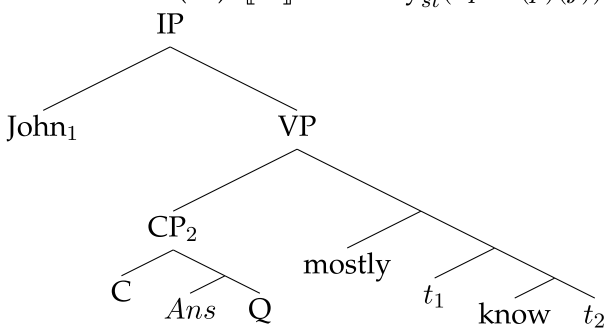
(19) A professor knows which students cheated.
$\not\leadsto$ For each student who cheated, a (different) professor knows it.

(1) a. The students mostly arrived.
b. John mostly knows which students arrived.

(20) Adapting the structure proposed in Lahiri (2002)/Preuss (2001):
a. Structure for (1a): $[\![\text{IP}]\!] = \text{mostly}_e(\lambda x.\text{arrived}(x))(s_1 \oplus \dots)$



b. Structure for (1b): $[\![\text{IP}]\!] = \text{mostly}_{st}(\lambda p.K(p)(j))(Ans(Q)(C))$



My semantics for *mostly* relies on the cardinality function # which simply takes a set (of propositions or individuals) and returns the number of elements in it. The denotation also contains a free parameter α. At this point, we can simply assume that its value is fixed to $1/2$, but in the next section we will treat it as a vague threshold.

(21) Flexible semantics for *mostly*, where τ is either e or st:
$[\![\text{mostly}_\tau]\!] = \lambda P_{<\tau,t>}.\lambda X_\tau.\left[\#(At_\tau(X) \cap P) > \alpha.\#At_\tau(X)\right]$

Putting everything together, we arrive at the following denotations for (1a,b), which correspond to the usual meaning for these sentences (of course, in the case of embedded questions, this is the WE reading). For simplicity, I use Heim's (1994) notation in (23): Awx stands for $[\![\text{arrived}]\!]^w(x)$, and similarly with S for students.

(22) $[\![1a]\!] = \#(\{x_e | [\![\text{student}]\!](x) \wedge [\![\text{arrived}]\!](x)\}) > \alpha.\#([\![\text{student}]\!])$
"The proportion of students who arrived is more than α"

(23) $$\begin{aligned}[\![1b]\!] &= \#(\{p_{st}|K(p)(j) \wedge At_{st}(Ans(Q)(C))(p)\}) > \alpha.\#(At_{st}(Ans(Q)(C))) \\ &= \#(\{x_e|K(\lambda w.Awx)(j) \wedge Aw_0x \wedge Sw_0x\}) > \alpha.\#(\{x|Aw_0x \wedge Sw_0x\})\end{aligned}$$
"Among students who arrived, the proportion of those for whom John knows that they arrived is more than α."

Cumulative readings

Cumulative readings arise when a transitive verb takes two pluralities as its arguments (Scha, 1981). In the usual case, the two pluralities are simply plural definite descriptions (see 2a). The same phenomenon happens when some responsive verbs take a plural individual and a plural-marked question (see 2b). Many theories have been proposed for (2a) (Kratzer, 2008; Champollion, 2010; Scha and Winter, 2014; Champollion, 2014, among many others), but their predictions do not differ on the basic cases we are interested in. For a concrete implementation, I will generalize the cumulation operator '**' as defined in Sternefeld (1998) by extending his definition from 2-place predicates defined on $E \times E$ products to arbitrary $E \times F$ products.

(24) Given a two place predicate R defined on $E \times F$ where both E and F are Boolean algebras, $^{**}R$ is the smallest two place predicate such that:
  a. $\forall x \in E, \forall y \in F, Rxy \rightarrow {}^{**}Rxy$
  b. $\forall x_1, x_2 \in E, \forall y_1, y_2 \in F, ({}^{**}Rx_1y_1 \wedge {}^{**}Rx_2y_2) \rightarrow {}^{**}R(x_1 \oplus_E x_2)(y_1 \oplus_F y_2)$

Let us see how this applies to Lahiri's (2002) example repeated below (assuming 2 witnesses and 3 klansmen present at the lynching in the actual world):

(2) The witnesses knew which klansmen were present at the lynching

(25) $$\begin{aligned}[\![(2)]\!] &= {}^{**}[\![\text{knew}]\!](\iota X^{*}[\![\text{witness}]\!](X))(Ans(Q)(C)) \\ &= {}^{**}[\![\text{knew}]\!](w_1 \oplus w_2)(PL(k_1) \oplus PL(k_2) \oplus PL(k_3))\end{aligned}$$

Given our definition for '**', we correctly derive that each witness must know at least one of the three propositions, and each proposition must be known by at least one witness.

The only puzzle left at this point is the fact that responsive verbs may not all allow cumulative readings. In fact, Lahiri (2002) argues that non-veridical predicates never do, possibly because their lexical restrictor is dependent on the subject which creates problem when we have a plural subject.

(26) Lahiri (2002): John is certain about Q = John is certain about every answer to Q that he considers possible.
More formally: $C_j = \lambda p_{st}.(Dox(j) \cap p \neq \emptyset)$

(27) The witnesses were certain about which klansmen were present at the lynching
Available distributive reading: Each witness was certain about which klansmen were present at the lynching (different witnesses may not agree)
Possibly available reading: All witnesses were certain about the same set of

klansmen, and they were all certain that any other klansman was absent.[6]
Unavailable cumulative reading: For each klansman that some/all witnesses considered to be possibly present, at least one witness was certain that he was present, and each witness was certain about the presence of at least one klansman. (compatible with each witness being uncertain about some klansmen).

Nothing in principle prevents us from deriving the cumulative reading for cases like (26), but I see two ways to rationalize the constraint. First, we could assume that there simply is no convention regarding how the common restrictor should be built (union, intersection?). Second, we could assume that there is a way to build the restrictor in principle, but the complex dependency makes it cognitively too demanding.

Nevertheless, the reading may become available when there is a natural pairing between attitude holders and propositions, as illustrated by the example below (Ben George, p.c.). Under a distributive reading, (29) is clearly false in the provided scenario, yet the sentence seems to be a correct description of the situation.

(28) Scenario:
Each expert is an expert on a different genus, and she is certain about which of the species in her genus are capable of asexual reproduction. Suppose further that every genus has an expert associated with it, and that no two experts are opinionated about the same species (each one explicitly reserves judgment and defers to the appropriate other expert).

(29) The experts are certain about which of the species are capable of asexually reproduction.

I will leave this as an open empirical issue. If anything, Lahiri's generalization should be tested with more predicates before we try to derive it.

Homogeneity

We can implement a simple supervaluationist theory of homogeneity. The basic idea is rather simple but I refer to Križ and Chemla (2014) and Križ (2015a) for a derivation of the projection behavior from arbitrary embeddings.

- Given a distributive predicate $P_{<\tau,t>}$ applied to a plurality X_τ ($\tau = e/st$):
  - P is (super)true of X if P is true of all atoms of X
  - P is (super)false of X if P is false of all atoms of X
  - P is a truth-value gap otherwise.
- $\neg P$ is true iff P is false, and false iff P is true.
- Universal adverbs of the appropriate type make the sentence false in all gap situations (see Križ, 2015b)

Note that the complex predicate of propositions which combines with the question in (20b) is not distributive: $\lambda p.[mostly_{st}(\lambda p'.K(p')(j))(p)]$. This is sufficient to explain why we do not see homogeneity in the presence of an adverb of quantity.

[6] Whether this is a different reading or just a specific case of distributive reading is unclear, because it entails the previous one. If we were to derive it, we would probably need the question to take scope over the subject, contra the argument we built on (19).

Alternative theories

Many theories of QVE exist, but none implements intermediate exhaustive readings, which we will introduce in the next section (Preuss, 2001 is an exception, but she does so by postulating a specific combination rule rather than deriving it from more general principles).

For an alternative theory of homogeneity with questions, see Križ (2015c). He shows that, provided some reasonable hypotheses on the projection of truth-value gaps, the homogeneity of questions derives immediately from the fact that answers which involve plural individuals are trivalent themselves. His theory explains how an adverb *within* the question can also remove homogeneity in languages like German (see also examples in Beck and Rullmann, 1999), which is a puzzle for my theory.

The same idea could be applied to cumulativity in principle. We may be able to derive cumulative readings with questions by simply extracting the plural noun which combines with the *which*-phrase (or doing so at a more abstract level if we were to account for cumulative readings with neutral *wh*-words such as *who*). This would essentially require a cumulation operator that can apply to arbitrarily complex abstract representations, which is a debated question in the recent literature.

It is therefore not impossible that one could show that the right approach to all plurality effects with questions is to derive them directly from the fact that some answers involve plural individuals, without resorting to plural-like structures of propositions. At this point I will stick to my implementation which has the advantage of providing a uniform account of plurality effects with questions. My impression is that the next sections of this paper could easily be adapted to a theory of plurality based on individuals rather than propositions.

4.4 Incorporating stronger exhaustivity

In this section, we introduce some background on strongly and intermediate exhaustive readings, before implementing their derivation in the theory introduced in the previous section.

4.4.1 Different exhaustive readings of embedded questions

Different readings have been proposed sentences with embedded questions such as (30). In the previous section, we focused on the Weakly exhaustive (WE) reading, which corresponds to the theory of Karttunen (1977). However, at least two other readings have been discussed in the literature: the Intermediate exhaustive (IE) and Strongly exhaustive (SE) readings. All three readings are illustrated in (31).

(30) John knows which students arrived

(31) WE: For all students who arrived, John knows that they did.
IE: For all students who arrived, John knows that they did, and he does not believe that any other student arrived.
SE: For all students who arrived, John knows that they did, and he knows that no other student arrived.

The SE and IE readings were first discussed by Groenendijk and Stokhof (1982), although they quickly rejected the IE reading as a real possibility. The IE reading has then been discussed several times and theories focusing on the SE reading have sometimes assimilated it to the WE reading (Spector, 2005; Spector & Egré, 2015; Roelofsen et al., 2014) while theories focusing on the WE reading associated it with the SE reading (Berman, 1991; Preuss, 2001). Cremers and Chemla (2014) established its existence as a reading independent from both the WE and the SE readings.

Klinedinst and Rothschild (2011, henceforth: K&R) show it is possible to derive IE and SE readings from WE for non-factive predicates, using an EXH operator and some well-chosen alternatives (for an early attempt at deriving these readings as implicatures, see Berman, 1991, p. 180). The view that stronger exhaustivity is the result of a process independent from question embedding *per se* will make it easier to implement on top of the structured denotations we introduced for plurality effects, and this view is also supported by acquisition data from Cremers et al. (2016). I will therefore follow K&R for the implementation of the IE and SE readings in the theory introduced in the previous section. I first present their theory before introducing my own implementation.

4.4.2 Klinedinst and Rothschild (2011)

Presentation

K&R focus on veridical non-factive predicates, such as *predict* (although see Spector and Egré, 2015, for arguments against this classification). They propose the readings in (33) for sentence (32), which are supported by a survey of their own and experimental results from Cremers and Chemla (2014).

(32) John predicted which students arrived

(33) WE: For all students who arrived, John predicted that they did.
IE: For all students who arrived, John predicted that they did,
. . . and he did not predict that any other student arrived.
SE: For all students who arrived, John predicted that they did,
. . . and he predicted that no other student arrived.

The core idea of K&R is to use a set of alternatives to the embedded question, from which they derive as implicatures the 'second lines' in the paraphrases of the IE and SE readings in (33). Their proposal assumes that the denotation of the question in a world w is simply its Karttunen-answer in this world (34). They further assume that the focus value of the question (the set of alternatives) is obtained by varying the world variable, as illustrated in (35). Crucially, the focus value $Alt(Q(w))$ contains false answers. To compute the implicatures, they use Fox's (2007) operator EXH as defined in (36). For the basic cases it is sufficient to assume that $[\![\varphi]\!]_{F+}$ is the set of alternatives which are stronger than the prejacent (this will be revised this in the next section).

(34) $Q(w)$ is the true (weak) complete answer in w.
Example: $Q(w) = [\![\text{Ann arrived and Bill arrived}]\!]$

(35) Any potential complete answer to the question is an alternative:
$Alt(Q(w)) = \{Q(w')|w' \in W\}$
$Alt(Q(w) = \{\text{Ann arrived}, \text{Bill arrived}, \text{Chris arrived}, \ldots \text{A, B and C arrived}, \}$

(36) $[\![\text{EXH}\varphi]\!] = [\![\varphi]\!] \wedge \left(\bigwedge_{p\in[\![\varphi]\!]_{F+}} \neg p\right)$

This simple proposition is sufficient to derive the three readings of (32). As an example, let us see how this works in a situation w_0 where out of three students Ann and Bill arrived, but Chris did not.

- the WE reading is simply the result of combining $Q(w_0)$ with the verb:

$$[\![(32)]\!]^{w_0} = \text{predict}'(j)(Q(w_0)) = \text{predict}'(j)(A(a) \wedge A(b))$$

- the IE reading is derived through global (matrix) exhaustification:

$$[\![\text{EXH}(32)]\!]^{w_0} = \text{predict}'(j)(A(a) \wedge A(b)) \wedge \neg\text{predict}'(j)(A(a) \wedge A(b) \wedge A(c))$$

- SE reading derived through local exhaustification:

$$[\![\text{John predicted EXH[which students arrived]}]\!] = \text{predict}'(j)\big((A(a) \wedge A(b)) \wedge \neg(A(a) \wedge A(b) \wedge A(c))\big)$$

Comments

K&R's proposal can capture the WE, IE and SE readings of non-factive veridical verbs. It can be extended to factive verbs by adopting a two-dimensional view of meaning with independent assertive and presuppositional contents (see Spector and Egré, 2015; Uegaki, 2015). However, in order to capture plurality effect, I need more structure than what K&R propose. The goal of the next section will be to retrieve their focus values for questions from the structured denotation introduced in section 4.3.

4.5 An exhaustification theory for plural questions

The goal of this section is to update the theory of plurality effects proposed in section 4.3 so that it would account for strong and intermediate exhaustivity and not just weak exhaustivity. To achieve this goal, we will need to retrieve the alternatives of K&R from the richer structure that was required for plurality effects. At this point, we will not discuss multiple-*wh* questions.

4.5.1 Hypotheses

Generating alternatives answers

In my theory, responsive verbs do not combine directly with questions but with the result of applying the answer operator to the question and to a lexical restrictor C, so I need to define alternatives of $Ans(Q)(C)$. For simplex questions (only one *wh*-phrase), $Ans(Q)(C)$ is just the maximal element of $Q \cap C$, so Q and C play symmetrical roles: $Ans(Q)(C) = Ans(C)(Q)$. This would be different for pair-list readings of multiple-*wh* questions, where Q would be of higher type and Ans would have to be recursively defined to cover arbitrary complex question types (see Dayal, 1996; Fox, 2012; Kotek, 2014, 2015), while C would remain of type $<st, t>$. For veridical verbs, $C = \lambda p.p(w_0)$ with w_0 the actual world, and $Ans(Q)(C)$ is equivalent to K&R's $Q(w_0)$. Their proposal is to generate alternatives by replacing w_0 with any possible world in this denotation. This is equivalent to replacing C with any

$C' = \lambda p.p(w')$ in my theory, and so I will adopt this approach. One important property of such restrictors is that they are closed under conjunction. I will assume that even non-veridical verbs have restrictors which are closed under conjunction.[7]

In sum, we will update our denotation of Ans to add a non-trivial focus value. The ordinary value is repeated below for convenience (basic denotation and recursive definition for higher-types questions). The focus value is defined in (37). Note that there is no need to define it recursively since it is based on the generalized $[\![\text{Ans}]\!]$.[8]

(14) $[\![\text{Ans}]\!](Q_{<st,t>})(C) = \iota p.[Q(p) \wedge C(p) \wedge [\forall p', Q(p') \wedge C(p') \rightarrow p \subseteq p']]$

(15) $[\![Ans]\!](Q_{<\sigma,t>})(C) = \forall P \in Q, [\![Ans]\!](P)(C)$

(37) $[\![\text{Ans}]\!]_F(Q)(C) = \{[\![Ans]\!](Q)(C') | \exists w : C' = \lambda p.p(w')\}$

One could imagine alternative ways to derive the alternatives of the complete answer. In particular, we could directly postulate a focus value for the question itself. However, this would not be compatible with the compositional semantics of Kotek (2015) for pair-list readings of multiple questions (the reader is free to check that all parses would lead to a flat, single-pair denotation).

I assume that alternatives project by pointwise function-application and combination with alternatives from any other source. This means that I put no restriction on multiple replacement but let independent rules rule out problematic alternatives (see Chemla and Romoli, 2016).

Exhaustivity operator

I will use a version of Fox's (2007) EXH operator to derive (secondary) implicatures, because its notion of "innocently excludable" alternatives allows an adaptation to the current problem without changing the basic definition. Unlike K&R, I will assume that EXH can only combine with nodes of type $<s,t>$, so I will not need to generalize the notions of negation, entailment and conjunction to complex types. However, since most question-embedding verbs come with presuppositions, I will need to specify how this operator interacts with presuppositions, which I model with a trivalent logic.[9,10] As we will see, it will not be possible to provide a proper treatment of QVE without acknowledging that *most* is vague, so I will also need to specify how EXH interacts with vagueness.

[7]The restrictor of collective *agree* may be an exception to this rule, as pointed out to me by Wataru Uegaki.

[8]The definition of the atoms of an answer as proposed in section 4.3 was not compositional. We can now propose an updated compositional version, compatible with multiple-*wh* questions.

(i) $At_{st}([\![X]\!]) = \lambda p_{st}.\big[(p \in [\![X]\!]_F) \wedge ([\![X]\!] \subset p) \wedge \forall r \in [\![X]\!]_F, [(p \subseteq r) \rightarrow (p = r)]$
$\wedge \bigwedge \{q \in [\![X]\!]_F | q \neq p \wedge \forall r \in [\![X]\!]_F, [(q \subseteq r) \rightarrow (q = r)]\} \not\subseteq p\big]$

The atoms of $[\![\text{Ans}]\!](Q)(C)$ are now properly defined.

[9]Gajewski and Sharvit (2012) and Spector and Sudo (2014) discuss the projection of presuppositions out of EXH. With my set of alternatives, I cannot adopt Spector and Sudo's (2014) view of EXH as a presupposition hole (otherwise, any instance of "which students arrived" embedded under a factive verb would presuppose that all students arrived). My definition of EXH is therefore somewhere between theirs and what they call EXH^3 (it differs from EXH^3 in that I use a strong notion of consistency). I leave for future research the possibility to give a unified account.

[10]Note that I already treated homogeneity using a trivalent logic. I will ignore most of the complications due to interactions between homogeneity and presuppositions, but it would require at least 4 values if we consider that homogeneity is not a presupposition. The only place where I will discuss potential interactions between homogeneity and presupposition is the case of *forget*, which I handle with the simple extra assumption in (53). For examples of multivalued logics applied to interactions between linguistic phenomena, see Zehr (2014), Spector (2015).

(38) $[\![\text{EXH } \Phi]\!]^w = 1$ iff $\big([\![\Phi]\!]^w = 1$ and $\forall P \in \mathbb{IE}(\Phi, Alt(\Phi)), P(w) \neq 1\big)$

(39) $\mathbb{IE}(\Phi, A) = \bigcap \{B \subset A | B$ is a maximal set in A s.t. $B^\neg \cup \{\Phi\}$ is s-consistent$\}$
$B^\neg = \{\neg p | p \in B\}$

(40) A set of propositions B is s(trongly)-consistent if there is a world which makes all propositions in B *clearly true* (i.e. neither presupposition failures nor borderline vague cases).

This definition of EXH has a few important properties:

- It negates non-weaker alternatives, and not just stronger ones.
- It does not project presuppositions from stronger alternatives (but see fn. 9).
- It blocks implicatures which would make a vague sentence borderline.

In the rest of this paper, I will often write EXH($[\![\Phi]\!]$) instead of the proper $[\![\text{EXH } \Phi]\!]$, thus using the same symbol for the exhaustive operator in the object language and its translation.

Extra assumptions on specific lexical items

***Know*:** I will assume that *believe* is a lexical alternative to *know*, in line with Percus (2006) and Sauerland (2008) who assume that *know* is an alternative to *believe* for the computation of anti-presuppositions. This will allow me to derive the IE reading of *know* without having to assume that the assertive content of a verb can be accessed independently from its presuppositional content, as Klinedinst and Rothschild (2011), Spector and Egré (2015), Uegaki (2015) do.[11]

***Most*:** Intuitively, *most* does not exactly mean "more than 50%" but is a vague quantifier with a context-dependent threshold. I will model this fact as uncertainty regarding the value of the α parameter in the denotation (21), repeated below.[12]

(21) $[\![\text{mostly}_{st}]\!] = \lambda P_{<st,t>}.\lambda X_{st}.\big[\#(At_{st}(X) \cap P) > \alpha.\#At_{st}(X)\big]$

Furthermore, *mostly* most likely has scalar alternatives, such as *completely*.

4.5.2 First application

It is now time to check that the proposed theory can derive the correct readings for the simple sentence (30). Following K&R, we will consider the parses in (41). We will replicate their prediction that the WE is the literal reading and that the IE and SE readings are the results of global and local accommodation, respectively.[13]

[11] I am very grateful to Emmanuel Chemla, who made this suggestion and later forgot it was his.

[12] Without going into too much detail, we could imagine an underlying distribution for the parameter α. Then, we could either specify the meaning of "clearly true" in (40) by specifying a threshold of certainty, or assume the probability of deriving an implicature to be a continuous function of the certainty of its consistency with the prejacent, so that when the probability of the negation of an alternative given the prejacent is very low, the probability to derive this implicature is also very low (Chemla & Romoli, 2016). For probabilistic models of implicatures and vagueness, see Franke (2011), Lassiter and Goodman (2014, 2015), Qing and Franke (2014)

[13] Note that one more position is available for EXH insertion, above VP: $(\lambda x.\text{EXH}[K(j)(Ans(Q)(C))])(j)$. However, it is equivalent to (41c) when the subject is not quantificational and does not introduce alternatives. Since we assumed that EXH only combines with nodes of type st, no other position is available.

(30) John knows which students arrived.

(41) Semi-formal LF for each reading:
  a. WE reading: $K(j)(Ans(Q)(C))$
  b. SE reading: $K(j)(\text{EXH}[Ans(Q)(C)])$
  c. IE reading: $\text{EXH}[K(j)(Ans(Q)(C))]$

Again, let us imagine a situation with 3 students where Ann and Bill are the students who arrived, while Chris is a student who didn't arrive. We'll write A for the proposition 'Ann arrived', $A \oplus B$ for 'Ann and Bill arrived', and so on.

WE reading

In the given situation, $Ans(Q)(C) = A \oplus B$. Then (41a) simply means that John knows that Ann and Bill arrived, which is the WE reading of (30).

SE reading

In (41b), the only source of alternatives for EXH is the question (and the restrictor). The derivation is similar to that of K&R:

$$Ans(Q)(C) = A \oplus B$$
$$Alt(Ans(Q)(C)) = \{A, B, C, A \oplus B, B \oplus C, C \oplus A, A \oplus B \oplus C\}$$

The alternatives A and B are entailed by the prejacent. The alternatives C, $A \oplus C$, $B \oplus C$ and $A \oplus B \oplus C$ are all non-weaker and innocently excludable. Therefore:

$$\text{EXH}[Ans(Q)(C)] = (A \oplus B) \wedge \neg C$$
$$K(\text{EXH}[Ans(Q)(C)])(j) \equiv K(A \wedge B \wedge \neg C)(j)$$

This is the SE reading: it is true if and only if John knows that Ann and Bill arrived and that Chris did not.

IE reading

In (41c), the sources of alternatives for EXH are the question and the verb (because we assumed that <*believe*,*know*> form a scale). In (42) we present a subset of the alternatives for 'know($A \oplus B$)', in a semi-formal format.

(42)
$$\begin{array}{ccccccc} K(A)(j) & \leftarrow & K(A \oplus B)(j) & \leftarrow & \cancel{K(C)(j)} & \leftarrow & \cancel{K(A \oplus B \oplus C)(j)} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ B(A)(j) & \leftarrow & B(A \oplus B)(j) & \leftarrow & B(C)(j) & \leftarrow & B(A \oplus B \oplus C)(j) \end{array}$$

Importantly:

- 'know(A)', 'believe(A)', 'believe($A \oplus B$)' are entailed.
- 'know(C)' and 'know($A \oplus B \oplus C$)' are innocently excludable alternatives, but they are trivially not-true in any world where $A \oplus B$ is the K-answer.
- 'believe(C)' and 'believe($A \oplus B \oplus C$)' are excludable and their negation yields a non-trivial implicature.

As a result, (41c) does correspond to the IE reading because it is true if and only if John knows that Ann and Bill arrived and does not believe that Chris arrived:

$$\textsc{exh}[K(Ans(Q)(C))(j)] = K(A \oplus B)(j) \wedge \neg B(C)(j)$$

4.5.3 False answer-sensitive readings beyond *know*

We saw how the theory derives IE reading for *know*, but what about other factive verbs?

First, one should note that there is no consensus on the definition of "intermediate reading". Everyone agrees on the definition of the IE reading of *know*, but for other verbs most authors simply call 'IE reading' the reading predicted by their theories when the mechanism they propose for *know* applies to another verb. Therefore, there are as many definitions of 'intermediate readings' as there are ways to derive the IE reading of *know*. I will favor the term 'false-answer sensitive' reading (henceforth FAS) introduced by Xiang (2015c), which covers all readings making reference to false answers of the embedded question. This definition includes the IE and SE readings of *know*. To distinguish the two, I will call 'global FAS reading' any FAS reading which differs from the SE reading.

Second, to derive the IE reading of *know* I assumed that it has a non-factive lexical alternative (*believe*). This assumption does not come out of the blue, a similar assumption has been made in the literature on anti-presupposition (Percus, 2006; Sauerland, 2008; Chemla, 2008). For other factive verbs however, such natural non-factive alternatives are not available, and alternatives built on false-answers will always be presupposition failures in the world of evaluation. Therefore, I derive the generalization in (43). Whether this generalization is supported is an open empirical issue which will not be easy to address, because (i) SE readings can obscure the data and (ii) determining which verbs are alternatives to each other can be challenging.

(43) **FAS generalization:** A factive verb receives global FAS readings if and only if it has a lexical non-factive alternative.

Note that even if no global FAS reading is predicted for a verb, global exhaustification is not necessarily vacuous. In particular, downward-entailing or non-monotonic verbs *forget* and *surprise* will be discussed in the next section.

4.6 Application to new cases

In this section, we explore predictions of the theory regarding cases that have not been discussed by K&R.

4.6.1 QVE sentences

Let us first turn to the sentence with an adverb of quantity which we used to introduce QVE. In section 4.3.2 we showed how the theory derives the WE reading of sentence (1b) repeated below.

(1b) John mostly knows which students arrived.

(44) a. John$_j$ [which students arrived]$_i$ mostly $\textsc{exh}[t_j$ knows $t_i]$.

b. John$_j$ $\textsc{exh}$[which students arrived]$_i$ mostly [t_j knows t_i].

	mostly		*completely*	
	know	*believe*	*know*	*believe*
$A' = A_0$	entailed	entailed	$\in \mathbb{IE}$	$\in \mathbb{IE}$
$A' \prec A_0$	*see footnote 14*		$\notin \mathbb{IE}$	$\notin \mathbb{IE}$
other cases	*negation entailed by (45b)*			

TABLE 4.1 – Alternatives for the sentence "John mostly knows which students arrived" can be characterized by (a) whether *mostly* has been replaced by its alternative *completely* or not, (b) whether *know* has been replaced by its alternative *believe* or not and (c) by the relation between the complement of the verb $A' \in Alt(Ans(Q)(C))$ and the actual answer A_0. Each cell in this table corresponds to a type of alternative, with parameters (a) and (b) set by the column and parameter (c) set by the line. We indicate whether the alternative is innocently excludable ($\in \mathbb{IE}$) or not ($\notin \mathbb{IE}$).

c. John$_j$ EXH[[which students arrived]$_i$ mostly [t_j knows t_i]].

d. EXH[John$_j$ [which students arrived]$_i$ mostly [t_j knows t_i]].

The structure we proposed in (20b) provides several positions where EXH could be inserted, listed in (44). However, (44a) is vacuous and the atoms of EXH$[Ans(Q)(C)]$ are undefined so it cannot be fed to *mostly* in (44b). This leaves (44c,d) as the only options, and they are equivalent when the subject is a proper name.

There are now 3 sources of alternatives for the sentence: the scales <*know*,*believe*> and <*mostly*,*completely*>, and the focus value of the answer. Since I imposed no restriction on multiple replacements, this generates a lot of alternatives. For readability, they are presented by category in Table 4.1.

I leave giving a formal proof for later, but the global-FAS reading obtained can be approximated by the conjunction of (45a,b,c).[14]

(45) a. For most students who arrived, John knows that they did.

b. He has no false belief regarding students who did not arrive.

c. It is not the case that he completely knows which students arrived.

Preuss (2001, p145–155) proposed a strongly exhaustive reading for QVE sentences which roughly corresponds to the one I proposed here: John knows a significant proportion of true answers and does not have any false belief regarding false answers. She argues that this reading is the most salient for the following examples:

(269) a. John uncovered to a large extent who took bribes.

b. John confessed to a large extent who he had cheated in poker.

c. John found out to a large extend who had cheated on the exam.

[14] To be precise, (45c) should be refined. The theory as it stands predicts a stronger inference: Let m be the minimum number of students who arrived that John has to know in order to clearly know for the most part which students arrived. Then the number of students who arrived such that John knows that they did does not exceed m. With a Bayesian semantics/pragmatics and a probabilistic account of vagueness (Franke, 2011; Frank and Goodman, 2012; Lassiter and Goodman, 2014; Qing and Franke, 2014), I would not need such clearcut distinctions between alternatives which are innocently excludable and alternatives which are not. Rather, the probability of negating an alternative (i.e., of deriving an implicature) would be proportional to the compatibility of this implicature with the assertion, which would then be a probability rather than a binary value (see also Chemla and Romoli, 2016).

4.6.2 Cumulative readings

(2b), repeated below, was argued to be ambiguous between a distributive and a cumulative reading. In section 4.3 however, we only considered weakly exhaustive readings, which correspond to a parse without any EXH. I will now discuss parses with EXH in different positions.

As we saw in the case of QVE, local exhaustification is predicted to block plurality effects because the strongly exhaustive answers is not a member of the algebra of answers and its atoms are not defined. Therefore, local exhaustification in (2b), repeated below, would only yield the distributive strongly exhaustive reading in (46a). Similarly, we can derive an intermediate distributive reading with the parse in (46b).

With global exhaustification, the ambiguity between a distributive and a cumulative interpretation resurfaces. We will first look at the cumulative reading, which I assume to correspond to the parse in (46c) with a cumulative operator. In this case, the implicature is that no witness had any false beliefs (otherwise, the group of witnesses would cumulatively believe a false answer). For the distributive parse (46d), we obtain a weaker reading, because for an alternative of the form "The witnesses *each* believed p" to be false, it is sufficient that one witness did not believe p.

(2b) "The witnesses knew which klansmen were present at the lynching"

(46) a. The witnesses knew [EXH which klansmen were present . . .]
"Each witness knew which klansmen were present and which were not"

b. [The witnesses]$_1$ EXH [t_1 knew which klansmen were present . . .]
"Each witness knew which klansmen were present and no witness had false beliefs"

c. EXH [The witnesses **knew which klansmen were present . . .]
"The witnesses together knew which klansmen were present, and no witness had false beliefs regarding klansmen who were not present."

d. EXH [The witnesses *each* **knew which klansmen were present . . .]
"Each witness (weakly) knew which klansmen were present, and there is no klansman that all witnesses falsely believed to be present."

To summarize, we predict a WE and an IE cumulative reading, and four different distributive readings (WE, SE and two types of IE). Of course, pragmatics should play an important role in selecting appropriate readings. We also did not take into account homogeneity, which may complexify the situation a bit.

4.6.3 Mention-some questions

Some questions, such are "Where can I get gas?", are known to receive readings which are not even exhaustive. Indeed, naming one place where the speaker can get gas seems to be a satisfactory answer to the question. These questions are usually called 'mention-some'. Although mention-some questions are not directly related to plurality effects, the structure we introduced allows for a treatment of these questions with minimal modification.

Mention-some questions have no clear equivalent in definite plurals[15], but they received a lot of attention in the recent literature on questions.

(47) Mary knows where John can get gas.
$\rightsquigarrow$ Mary knows at least one place that sells gas, but not necessarily all of them.

Fox (2013) proposes that mention-some questions involve a distributivity operator (silent *each*) in the scope of an existential modal, as in (49). This LF breaks the correspondence between sum-formation in the domain of individuals and conjunction in the algebra of answers, because although it is possible that x_1 or x_2 chair the committee, it is usually impossible to have $x_1 \oplus x_2$ chair the committee (together).

(48) Who can chair this committee?

(49) $Q = [\text{who } \lambda X[\Diamond\, \textit{each}(X)([\text{chair this committee}])]]$
$= \{\Diamond(x_1 \text{ chair this committee}), \Diamond(x_2 \text{ chair this committee})\}$

The denotation of the question is not closed under conjunction, and the answers are now independent. Therefore, all answers are maximally informative. In such a situation, the answer operator I introduced in (14) presupposes that there is a unique true answer. Fox (2013) proposes instead existential quantification over maximally informative (true) answers, but we would immediately lose the uniqueness presupposition of simple singular *wh*-questions. Fox (2013) proposes two solutions for this problem, see further discussion in Xiang (2015b).

B. R. George (2013), Xiang (2015a) show that mention-some questions also display false-answers sensitivity. If we adopt a theory à la Fox (2013), we can account for this immediately.

(50) Janna knows where she can buy an Italian newspaper
$\Leftrightarrow$ Janna knows at least one place where she can buy an Italian newspaper, and she does not falsely believe that she can buy one in a place that actually does not sell Italian newspaper.

(51) EXH[Janna knows Ans(where $\lambda X.{}^{\star}place(X) \wedge [\Diamond\, \textit{each}(X)([\beta])$)(C)]

Let $a_1, a_2, \ldots$ be places where Janna can actually buy an Italian newspaper and $b_1, b_2, dots$ places where she cannot. Then the weakly exhaustive reading of (50) is that Janna knows that she can buy an Italian newspaper in a_1 or she knows that she can buy one in a_2, and so on. The exhaustification automatically rules out false beliefs about the b_j-places. Indeed, let w_j be a world where Janna can only buy a newspaper in b_j. The alternative with *believe* and $C_j = \lambda p.p(w_j)$ is innocently excludable and its negation expresses the fact that Janna does not believe that she can buy an Italian newspaper in b_j. Note that alternatives regarding the a_i-places are not innocently excludables, nor are alternatives mixing a_i and b_j places, so exhaustification does nothing else than banning false beliefs regarding b_j-places.

$\rightarrow$ B. R. George (2013) argued from example (50) that knowledge-*wh* is not reducible to knowledge-*that*. I gave an account based on the regular declarative-embedding lexical entry for *know*, using lexical competition between *know* and *believe* as the only source of "non-reducibility".

[15] A few differences may explain this asymmetry: definite articles compete with indefinites. Mention-some usually require number-neutral wh-words (not *which*-phrases), which are somehow equivalent to pronouns, but pronouns do not take complements which would embed a modal.

4.6.4 *Forget*

Forget is an interesting verb because it provides a good illustration of homogeneity effects with questions. Let us assume the informal semantics in (52) for declarative-embedding *forget*. We will assume that the lexical restrictor for *forget* is simply the set of true propositions.[16]

(52) John forgot p iff p and John used to know p and it's not the case that John believes p.

Not all propositions in the restrictor satisfy the presupposition, hence we need some hypotheses on the interaction between homogeneity and presupposition. We will assume universal projection with the hypothesis (53).

(53) If a predicate P with a presupposition ρ applies to a plurality X, $P(X)$ presupposes that each atom of X satisfies ρ.

We now have everything we need to apply the theory to questions embedded under *forget*, as in (54).

(54) Sue forgot which students arrived.

(55) a. Sue forgot [which students arrived]
"Sue used to know which students arrived (weakly), and for each student who arrived, Sue forgot that (s)he arrived."

b. Sue forgot EXH[which students arrived]
"Sue used to know which students arrived (strongly), and for at least one student, Sue forgot whether (s)he arrived."

c. EXH[Sue forgot which students arrived]
EXH *is vacuous, result equivalent to (a)*

With no EXH operator, we get a WE reading. However, homogeneity ensures that Sue had exhaustive knowledge in the past *and* exhaustive oblivion in the present. Note that these truth conditions cannot be reduced to 'Sue forgot that p', with p an atomic proposition, as was noticed by B. R. George (2013). Yet, we see that this reading can be derived from the usual denotation for declarative-embedding *forget*, so it does not challenge reducibility.

Local exhaustification would lead to a SE presupposition (stronger than previously) but the SE assertion is actually very weak because *forget* is downward-entailing. I also argued that local exhaustification blocks plurality effects, including homogeneity. However, theories of exhaustification usually block it in downward-entailing contexts (Chierchia et al., 2011. The availability of this (very weak) reading is thus an empirical issue.

We explained how matrix exhaustification could not yield a FAS reading for factive verbs other than *know*, because all alternatives built on answers which are actually false are presupposition failures in the actual world, hence do not contribute any implicature (see generalization 43). In the case of *forget*, global exhaustification is simply vacuous. Indeed, alternatives built on answers entailed by the actual answer are not innocently excludable: they would lead to contradictions such as "Sue forgot that a, b and c arrived, but she didn't forget that a arrived, that b arrived and that c arrived".

[16] *Forget* provides a strong argument against using the presuppositions of the verb as a restrictor (intermediate accommodation). Indeed, this would make wrong predictions for the meaning of (54) because we do want to derive the presupposition that John used to know which students arrived. With intermediate accommodation no presupposition would project to the global context and the sentence would have very weak truth conditions (roughly: John forgot what he used to know).

4.6.5 Multiple *wh*-questions

Adopting the syntax/semantics of Kotek (2015) allows us to treat multiple-*wh* questions. For sentences like (56) where a multiple-*wh* question is embedded under *know*, the result of global exhaustification is rather straightforward: we simply derive a no false belief constraint. Local exhaustification is also very simple and derives the expected SE reading (Jane knows which student read which book, and that no other student read any other book).

(56) Jane knows which student read which book.

I will not address plurality effects with multiple-*wh* questions (nor with questions with quantifiers). For a detailed analysis of QVE with such questions, I refer to Preuss (2001).

4.6.6 Primary implicatures

Grammatical theories of implicatures such as Fox (2007) usually assume that primary implicatures (ignorance inferences) are derived independently from secondary implicatures (which are the result of EXH). It would be natural to assume that the alternatives I introduced for the derivation of stronger readings also participate in primary implicatures. For most cases, this will simply make the literal reading odd, because if the speaker knows that John falsely believe that all students smoke, asserting that John knows who smoke would be a blatant violation of the ignorance inferences. Rather, we predict that a good situation to detect WE readings would be one in which the speaker (a) knows that John knows that all student smokers smoke and (b) does not believe that John falsely believes that any other student smoke. Note that (b) is weaker than the secondary implicature at the base of the IE reading, because the speaker is not sure that John has no false beliefs.

In the case of mention-some questions, the predicted primary implicatures are more than a weaker version of the secondary implicatures. Indeed, they add a requirement that the speaker would not know which mention-some answer(s) the agent knows (unless there is only one true answer). Whether this prediction is correct is a complex empirical question.

4.6.7 A few unwelcome predictions

The theory makes correct predictions in several cases and provides reducible solutions to the challenges proposed by B. R. George (2013). However, it also makes unwelcome predictions in a few simple cases.

Quantificational subjects

In all the examples we treated, the subject of the sentence was always a proper noun or a definite plural. K&R note that their theory makes too strong predictions for sentences like (57) if EXH is allowed to apply globally. In particular, it predicts an implicature that no student made any false prediction.

(57) EXH[At least one student predicted who came]
$\Rightarrow$ No student made any false prediction regarding who came.

K&R suggest that the scope of EXH be limited to the VP level, thus preventing it from interacting with quantificational subjects. Of course, this would mean that the exhaustification operator responsible for the strong readings of questions is different from the operator for scalar implicatures. I will come back to this discussion in the conclusion of the paper.

Negative sentences

In the introduction, we discussed (3b), repeated below, as an example of homogeneity effect with embedded questions. However, we did not look at the possible parses with EXH for this sentence. Local exhaustification would simply lead to a SE reading, which would be the exact negation of the SE reading for the affirmative sentence. Global exhaustification is in principle available and as the theory stands, it predicts a very odd reading.

(3b) EXH [Mary doesn't know which students arrived]

If we allow alternatives of the form "Mary doesn't believe A'" with A' consisting of false answers, we get the truth-conditions in (58) for (3b). This reading is clearly unattested.

(58) EXH [Mary doesn't know which students arrived]
"For all students who arrived, Mary doesn't know that they did *(homogeneity)* and for all students who didn't arrive, Mary falsely believes that they did *(implicature)*."

The solution suggested by K&R for the previous problem would also partially solve this one: If the scope of EXH is limited to the VP level, below negation, we would not get the false-belief implicatures. Rather, we would get a reading equivalent to the negation of the IE reading.

A second solution to the problem would be to accept alternatives obtained by deletion of the negation (e.g., "Mary does believe A'"), because this would create symmetry and none of these alternatives would be innocently excludable. While this is in principle possible in the theory of Fox and Katzir (2011), it is usually considered a shortcoming of this theory because it disallows so-called 'indirect implicatures' from "Not all students arrived" to "Some students arrived". I will leave it as an open issue for the moment. Hopefully, these alternatives can be neutralized by an independent principle (see Trinh and Haida, 2015 for recent discussions on the generation of alternatives, and Chemla and Romoli, 2016 for an alternative possible solution).

Surprise

The behavior of questions embedded under *surprise*, as in (59), is not fully captured by my theory. First, this sentence does not seem to give rise to homogeneity effects. Indeed, it seems sufficient that one student surprised Peter for (59) to be true, whereas homogeneity would predict that the sentence is true only when Peter expected none of the students to show up (as we correctly predict for *forget*). One solution is to use a denotation for *surprise* which is not distributive. This seems to be correct because we can easily build a scenario in which (60) does not entail that Peter was surprised by the fact that John or Mary showed up but only by the fact that they both came to the party (e.g., John and Mary have a baby and Peter expected one of them to stay home).[17] One such denotation has been proposed by Uegaki (2015).

(59) It surprised Peter which students arrived.

(60) It surprised Peter that John and Mary came to the party.
$\not\Rightarrow$ It surprised Peter that John came to the party and it surprised Peter that Mary came to the party.

The second problem is that matrix exhaustification of (59) yields the very strong reading in (61), which is clearly not attested. Unlike *forget*, *surprise* does not validate the inference in (62). By contrapositive, the negation of "It surprised Peter that John came" and "It surprised Peter that Mary came" is consistent with "It surprised Peter that John and Mary came to the party". Therefore, all alternatives built on answers which are asymmetrically entailed by the actual answer can be negated.

(61) EXH [It surprised Peter which students arrived]
Peter is surprised that this particular combination of students arrived, but there is no subgroup of students whose arrival surprises him.

(62) It surprised Peter that John and Mary came to the party.
$\not\Rightarrow$ It surprised Peter that John came to the party or it surprised Peter that Mary came to the party.

I will not address this issue in detail, but point at a possible solution. A degree semantics for *surprise* has been proposed in Villalta (2008) and more recently Romero (2015). This denotation, plus our hypotheses on the interaction of exhaustivity and vagueness, would presumably block these alternatives from playing a role (indeed, the reading is perceived as deviant because it is very unlikely to be surprised by the specific combination of students who arrived without being surprised by any subgroup).

Note that my theory can derive SE readings for *surprise* if local exhaustification is allowed. Data in Cremers and Chemla (2016) suggest that this is a good prediction.

4.7 Conclusion

I showed that embedded questions can be treated as definite descriptions of their answers in order to account for plurality effects. The same idea has already been proposed to account for the uniqueness presupposition of singular questions (Dayal, 1996) and QVE (Fox, 2012). The analogy between definite descriptions and embedded questions extends easily to cumulativity and homogeneity.

Such a theory can be complemented with a module that derives false-answer sensitive readings as implicatures by postulating the right set of alternatives. I showed that this extension is very powerful in that it makes predictions for a wide range of sentences and explains phenomena for which an explanation had to be postulated so far (intermediate QVE readings, non-reducible readings under *know*

[17] Note that *surprise* does give rise to quantificational variability effects, as illustrated in (i) below. However, because I followed Lahiri (2002) in assuming that QVE requires movement of the question out of the scope of the verb, we can derive QVE even for predicates which are not distributive with respect to their propositional complement.

(i) For the most part, it surprised Peter which students arrived.

and *forget*). This theory also has the advantage to be compatible with psycholinguistics data on questions embedded under *know*. In particular, it explains why children begin with a WE reading and why the most prominent reading for adults is the IE reading, which corresponds to regular implicatures, rather than the SE reading, which is derived through the more controversial local exhaustification.

A few issues stand unresolved (exhaustification of negative sentences and sentences with *surprise*), however there is a good hope that they could be solved with an improved notion of innocent excludability for exhaustivity (in particular, a better treatment of complex alternatives derived from multiple replacements and vagueness).

Finally, I would like to discuss the nature of the EXH-operator used in this theory and the relation between strengthened readings of questions and other implicatures. There is a strong intuition, even among proponents of the exhaustification approach to exhaustivity with questions, that this operator is not the same as the one responsible for usual implicatures (an in particular scalar implicatures). Possible limitations on the scope of this operator are an important difference (Klinedinst & Rothschild, 2011). Another one has to do with the overt counterpart of EXH. Indeed, if *only* can be an overt counterpart to the operator responsible for scalar implicatures, it seems that the counterpart of the EXH-operator for questions is *exactly* (and the position of *exactly* is also more constrained than that of *only*). Finally, Xiang (2015b) argues again the exhaustification approach by contrasting the no-false-beliefs inference with scalar implicatures. In particular, the no-false-beliefs inference is harder to cancel than scalar implicatures, persist in downward entailing environments and does not seem to be computed from above matrix negation. Therefore, if we want to save the exhaustification approach it seems necessary and natural to assume that there may be more than one EXH-operator, and that the one responsible for strengthened readings of questions is different from the operator responsible for scalar implicatures.

I would like to suggest that although strengthened readings of questions are different from scalar implicatures, we may be able to maintain a unified account if we acknowledge that there is variety between different sources of implicatures. First, note that exact readings of numerals share most of the properties that distinguish strengthened readings of questions from scalar implicatures. They are hard to cancel, easy to embed and the analogy even extends to the example of K&R: From "At least one student talked to 5 professors", the predicted implicature that no student talked to 6 or more professors seems too strong. Yet, they can still be analyzed in an exhaustification based theory (Spector, 2013a, see). We could thus imagine that the different properties we discussed are correlates of one underlying property (e.g., easily accessible alternatives) but do not necessarily reflect the presence of different operators. Variation along this dimension would explain the difference with scalar implicatures, just like some underlying properties of scale may explain differences between different scalar implicatures (van Tiel, van Miltenburg, Zevakhina, & Geurts, 2013). Note that accessibility of the alternatives has been proposed as an explanation for children's specific difficulty with scalar implicatures (in contrast with other implicatures which do not require accessing lexical alternatives).[18]

It will be hard to reach a clear conclusion in this debate without quantitative data (for instance correlation between the derivation of implicatures, strengthened

[18]Note that my theory predicts that lexical alternatives are still involved in the computation of the IE reading of *know*. The fact that children need to access the lexical alternative *believe* in order to compute this reading would explain their difficulty in Cremers et al. (2016) experiment.

readings of questions and exact readings of numerals). This may be an interesting topic for future research in the psycholinguistics of embedded questions.

Additional discussion

The theory presented in this chapter is rather flexible and can make prediction for a large range of sentences. However, it only applies to responsive predicates. In the next chapter, I propose a theory which allows plurality effects with rogative predicates as well, building on earlier work by Beck and Sharvit (2002). It will also be the occasion to discuss lexical restrictors in greater detail (justify their use over intermediate accommodation and explore a wider variety of responsive predicates).

Chapter 5

Homogeneity and Quantificational variability with Embedded Questions

Cremers, A. (2015a). *Homogeneity and quantificational variability with embedded questions*. Under Revision for *Natural Language Semantics* †

Abstract

Quantificational variability effects (QVE) with questions have been studied extensively since Berman (1991). Lahiri (2002) provided a particularly detailed account of QVE. Beck and Sharvit (2002) showed that this effect could arise with a wider variety of verbs than previously thought. Embedded questions also give rise to homogeneity effects reminiscent of those arising with definite plurals (Fodor, 1970; Löbner, 1985), but these effects have almost never been studied. In this manuscript, I propose an extension of Beck and Sharvit's (2002) theory which will treat QVE and homogeneity as two consequences of questions being inherently pluralities, and predict that these effects could in principle arise with all interrogative-embedding predicates. Finally, I discuss predictions of the theory for a wide variety of predicates.

Embedded questions are known to interact with adverbs of quantity to yield a reading where the adverb seems to quantify over answers to the questions (Berman, 1991). This effect was first related to the Quantificational Variability Effect (henceforth QVE) noticed by Lewis (1975), which descrives the interaction between an indefinite and an adverb of frequency to a reading that has little to do with time or frequency, as illustrated in (1) below.

(1) A quadratic equation usually has two different solutions.
$\rightsquigarrow$ Most quadratic equations have two different solutions.

A parallel analysis of QVE with questions and indefinites was appealing since many theories treat questions as some sort of indefinites (this view is supported by morphological similarities between indefinites and *wh*-phrases in many languages, among other things). However, questions which give rise to QVE were independently related to *plural* nouns, and embedded questions to definite descriptions (Dayal, 1996). What we call QVE for questions may have little to do with the effect Lewis discovered with singular indefinites, but the name *QVE* remains. The goal of this paper will be to extend the view of questions as pluralities presented in Beck and Sharvit (2002) to explain another effect arising with plural nouns: homogeneity.

† ...

In a first section, we will introduce homogeneity effects as defined in the literature on plural definite descriptions and show that very similar effects arise with embedded questions. As we will see homogeneity, just like QVE, has a deep link with plurality, so in the second section, we will review a few proposals from the rich literature on QVE with questions. The third section will present a new account of QVE which treats homogeneity effects in a very similar way. The last section will explore a wider variety of question-embedding predicates and discuss predictions of the new theory.

5.1 Homogeneity effects

5.1.1 Plural definite descriptions and homogeneity effects

The data: As first noticed by Fodor (1970), plural definite descriptions interact with negations in a particular way, which is illustrated in (2). (2a) is usually understood as "John talked with all of the students", while (2b) can be understood as "John talked with none of the students", a reading which is stronger than the mere negation of (2a). This effect has often been dubbed *homogeneity* because it seems that a predicate has to apply uniformly on the definite description: either all individuals must satisfy the predicate or none of them, and any intermediate case is somehow deviant.

(2) a. John talked with the students.
b. John didn't talk with the students.

Using a bound pronoun and a negative QNP, we can show that the effect cannot simply be explained by having the object DP take scope over the negation. Indeed, (3) is usually understood as meaning that no professor talked with **any** of the students she likes, but the co-reference prevents the object DP from taking scope over the quantifier *No* in subject position.

(3) No professor$_i$ talked with the students she$_i$ likes.

Let me add immediately for later comparisons that, as first noted by Löbner (1985), homogeneity effects disappear in the presence of an overt universal adverb. This is illustrated in (4a,b), where we get a weaker reading which corresponds to the logical negation of a universal quantifier.

(4) a. John didn't talk with all the students.
b. The students didn't all talk with John.

Theoretical accounts: Several theories have been proposed to explain this homogeneity effect. Schwarzschild (1993), Löbner (2000) and Gajewski (2005) treat it as a form of presupposition. They suggest that $S =$"the students P" has a universal assertive meaning, but that it also introduces a presupposition of the form "all students P or no student P". Depending on the sentence, the presupposition would be satisfied by making the first or the second disjunct true. Magri (2013) proposed that 'the' is basically an existential quantifier, but that 'some' is a scalar alternative to 'the'. Since 'all' is in turn an alternative to 'some', recursive exhaustification of 'the' would lead to the meaning "some students P and it is not the case that some but not all students P", which is equivalent to "all students P". The exhaustification would be blocked in downward entailing environments, just like usual scalar implicatures, explaining why the negative sentence (2b) is understood as the negation of an existential sentence. Finally, supervaluationnist theories (Spector, 2013b; Križ and Spector, 2015) treat plurals more like vague items and develop the *Strongest meaning hypothesis* of Krifka (1996). In this view, 'the students' can be seen as a quantifier with a vague quantificational force and the sentence S is *super-true* if both $S_\forall$ = "all students P" and $S_\exists$ = "some students P" are true. It is *super-false* if both $S_\forall$ and $S_\exists$ are false, and corresponds to a truth-value gap in other cases.

5.1.2 Homogeneity with embedded questions?

Similar facts: The sentences presented in (5) suggest that embedded *which*-questions behave exactly as plural definite descriptions. To begin with, (5a) suggests that John knows about all students who called, while (5b) suggests that he knows for none of them.

As with definite descriptions, (5c) shows that an account in terms of inverse scope is not possible, and (5d) suggests that the effects disappears in the presence of an adverb.[1]

(5) a. John knows which students called.
b. John doesn't know which students called.
c. No professor$_i$ knows which of her$_i$ students called.
d. John doesn't completely know which students called.

Theoretical accounts: Even though some of the facts described here have already been noticed (e.g., B. R. George, 2011, p.109), no theoretical account has been proposed (a few recent exceptions being work by Yimei Xiang or Manuel Križ).

5.1.3 Summary

Homogeneity effects seem to arise for both definite descriptions and embedded questions. The introspective data presented above are supported by quantitative data from Križ and Chemla (2014) and Xiang (2015a)[2] for definite descriptions and embedded questions, respectively. The results from both studies do not reveal any difference between the two cases (although the methodologies are very different). In particular, the observed rates for projection of homogeneity out of negation are similar (about 50%).

The goal of this paper will be to show how current theories of embedded questions on the one hand and of homogeneity on the other hand can be combined to provide an account of the data. We will then evaluate new predictions derived from this proposal.

[1] Obviously, *completely* affects the meaning of the sentence in a much more complicated way than what I assume here. However, the crucial observation is that (5d) does not give rise to homogeneity effect. A reasonable assumption would be that sentences with *completely* are ambiguous between the QVE reading and a 'degree of knowledge' reading. In affirmative sentences, the QVE reading is equivalent to the meaning of the sentence without any adverb, hence the 'degree of knowledge' reading is more salient. In negative sentences however, the QVE reading becomes more salient. As we will see in further sections, the effect of the adverb in such sentences will be to remove homogeneity effects by forcing a universal quantification on the question, which under negation yields a weaker reading than what we would get without the adverb. Furthermore, it is likely that *completely* will compete with other adverbs of quantity in this case, giving rise to implicatures. The pragmatic QVE reading for (5d) could be paraphrased as "John doesn't know for all student callers that they called, but he knows for some/most of them".

[2] What she calls *projection of completeness* corresponds to what I call *homogeneity*.

5.2 Quantificational variability effects and questions as pluralities

In the discussion of homogeneity effects, we showed that these effects disappear when the sentence contains an overt adverb. Even though the homogeneity effects due to questions have not been studied, the interaction between embedded questions and adverbs of quantity has given rise to an extensive literature. The theory presented in section 5.3 will provide a unified account of these two things, building mostly on previous theories of *Quantificational variability effects*.

5.2.1 What is Quantificational variability?

Quantificational variability effects (henceforth QVE), refer to the interaction between an adverb of quantity or frequency and an embedded questions or an indefinite DP respectively. The effect with embedded questions, illustrated in (6), was first described in Berman (1991).[3] The adverb *mostly* can yield the reading (6a), which Lahiri (2002) calls the focus-affected reading, but I will only be interested in the reading (6b), where the adverb seems to quantify over answers to the questions.

(6) John mostly knows which students called.
  a. Most of what John knows is the answer to the question 'which students called'.
  b. For most students who called, John knows that they called.

QVE with questions involve adverbs of quantity (e.g., *mostly*, *in part*) and these adverbs usually take as a restrictor something which has the structure of a plurality (similar to Link's 1983 starred predicates). Because of this, recent accounts of QVE treat questions with a plural *which*-phrase as pluralities of some kind (an idea that goes back to Dayal, 1996).[4] Since QVE has been studied extensively and gave rise to many theories which treat questions as pluralities, it makes a lot of sense to try to account for homogeneity and QVE together.

The rest of this section will consist in a quick presentation of the theories of questions I will build on to add an account of homogeneity: Lahiri (2002) and Beck and Sharvit (2002).

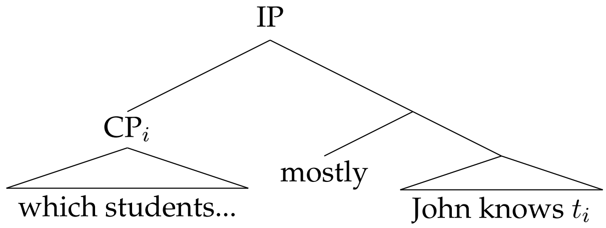
5.2.2 Lahiri (2002)

Lahiri (2002) derives QVE for *responsive* verbs (verbs such as *know* or *agree* which can receive interrogative as well as declarative complements). The main idea is that these verbs take arguments of type $<s,t>$, propositions, while questions are of a higher type (usually $<st,t>$, sets of propositions). Because of this type mismatch, the question in (6) has to move to the restrictor position of the adverb *mostly*, leaving a trace of type st. This movement is dubbed *interrogative raising* (henceforth IR). This leads to the structure sketched in (7), because Lahiri assumes that adverbs combine with their nuclear scope before combining with their restrictor.

[3] Some speakers find the sentence with *mostly* odd. They usually find the sentence better if *mostly* is replaced with *for the most part*. I will stick to *mostly* for simplicity, but the reader should feel free to replace every occurrence with the adverb s/he prefers, since I will not make any distinction between the two.

[4] Another phenomenon noticed by Lahiri (2002), cumulative (or semi-distributive) readings, provides yet another argument for analyzing questions as pluralities. I will not discuss this phenomenon here.

(7) Lahiri's (2002) structure for (6b):



Before combining with the adverb, the set of propositions denoted by the question must be restricted, and the restriction is dependent on the embedding verb. For instance, veridical predicates like *know* only relate the agent to the true answers, while a non-veridical verb like *be certain* relates the agent to a different set of answers (the answers that the agent considers possible, according to Lahiri). Lahiri (2002) argues that this restriction cannot simply be accommodation of the verb's presupposition, contra Lahiri (1991). He suggests that each verb comes with a lexical restrictor which combines with the question. Section 5.4 presents arguments for this approach and further characterization of the restrictors.

Lahiri (2002) also makes use of a type-shifting operator, which takes a question and returns a single proposition. Among other things, this allows questions to be interpreted *in-situ* under responsive verbs, when no overt adverb is present. We will come back to this in section 5.3.1, where we introduce a similar operator.

5.2.3 Beck and Sharvit (2002)

Lahiri's theory only derives QVE for responsive verbs. However, Beck and Sharvit (2002) observed that rogative verbs (which can receive interrogative but not declarative complements), also give rise to QVE readings in some cases. As an example, (8) below has the reading (8a), which could be dubbed 'degree of dependence', but also has the reading (8b), which seems very similar to (6b). Nevertheless, the predicate *depend on* does not embed propositions. When one adds *exclusively*, (8b) even becomes the only available reading for the sentence.[5]

(8) Who will be admitted depends for the most part (exclusively) on this committee.

a. This committee is an important factor in deciding who will be admitted.

b. For most candidates, it depends (exclusively) on this committee whether they will be admitted.

They propose a theory which shares many features with Lahiri's (2002), but has quantity adverbs quantify on subquestions rather than answers to the question. I will not present the details of this analysis here to avoid redundancy, since the theory I present in section 5.3 below draws heavily on the proposal of Beck and Sharvit (2002). Rather, I will point the differences when my account diverges significantly from theirs.

[5] Lahiri (2002, pp. 224–230) argues that although *depend on* can combine with adverbs of quantity, this is not an instance of QVE. Instead, he proposes that the adverbs combine with *depend on* to provide a new meaning postulate, which inherits its 'degree of dependency' from the force of the adverb. However, the version of (8) with the adverb *exclusively* provides an argument against this analysis. Indeed, assuming that *exclusively* affects the degree of dependency is probably the right analysis, but it does not explain how the resulting meaning can still combine with *for the most part*.

5.3 A unified theory of plural embedded questions

In this section, I will present a theory of questions which accounts uniformly for QVE and the homogeneity effects discussed in section 5.1, and discuss a few new predictions. Following Beck and Sharvit (2002), the theory derives these effects with both responsive and rogative verbs. However, the theory presented here will not derive the strong exhaustive reading (nor the *intermediate* exhaustive reading, Spector, 2005). Implementing a derivation of these readings would be possible by extending the proposal of Klinedinst and Rothschild (2011), but this would make the current proposal very technical and blur the main message: QVE and homogeneity can be unified under a view of embedded questions as pluralities.

5.3.1 Ingredients of the theory

Denotation of questions and operators: Following Hamblin (1973), I will assume that questions denote sets of propositions. More precisely, a question such as (9) will denote all propositions of the form "x called" with x in the denotation of *students*. Following Dayal (1996), I will assume that the question inherits the plural morphology of *students*, translated with Link's (1983) star (⋆) operator. Formally, this will yield the denotation in (10).[6]

(9) $\mathcal{Q}$ = Which students called?

(10) $[\![\mathcal{Q}]\!] = Q = \lambda p_{st}.\, \exists x \in [\![^{\star}\text{ student}]\!] : [p = \lambda w.[\![\text{called}]\!]^{w}(x)]$
$Q = \{C(p), C(m), C(p \oplus m), \ldots\}$ (with $C = [\![\text{called}]\!]$ and p, m denoting some students)

I will define a few operators on these question denotations. Following most previous authors, I will define a supremum operator σ, as in (11). Because it is based on the definite description operator ι, σ will have a *uniqueness* presupposition, implying that σQ is only defined if Q has a unique maximal element. Questions such as (9), which are schematically of the form: "which A's are B" (with A a count noun and B a distributive predicate), will always have a maximal element because the set of propositions they denote is closed under conjunction.[7]

(11) $\sigma = \lambda Q_{<st,t>}.\, \iota p.\, \left[p \in Q \wedge \forall q \in Q, (q \subseteq p) \rightarrow (q = p)\right]$

Secondly, I will define an operator At, which will return all the minimal elements in the denotation of the question (the *atoms*).

(12) $At = \lambda Q_{<st,t>}.\, \lambda p_{st}.\, \left[p \in Q \wedge \forall p' \in Q, [(p \subseteq p') \rightarrow (p = p')]\right]$

I will define an operator Sub which applies to a question denotation and returns a set of subquestions which corresponds to a *division* of the question, in the sense of Beck and Sharvit (2002).[8] Concretely, Sub will return all *whether*-questions formed on the atoms of Q, hence the denotation in (13). Since answering all the questions in $Sub(Q)$ provides a complete answer to Q, $Sub(Q)$ is indeed a division of Q in the sense of Beck and Sharvit (2002).

[6] I will not discuss the question of *de re/de dicto* readings. With minor modifications, this could be implemented as an ambiguity regarding which world fills the w argument in $[\![\text{student}]\!]^{w}$. For further discussion, see Sharvit (2002).

[7] This is the case because ⋆A is closed by sum and B satisfies: $B(a \oplus b) \equiv B(a) \wedge B(b)$.

[8] Beck and Sharvit (2002) proposed that the division of the question be determined from the context. I will favor a more rigid approach here: each question is associated to a unique division, which is fully determined by the operator Sub.

(13) $Sub = \lambda Q_{<st,t>}.\lambda q_{<st,t>}.\exists p \in At(Q)\colon q = \textit{whether-p} \quad (\textit{whether-p} = \{p, \neg p\})$
$Sub(\llbracket(9)\rrbracket) = \{\llbracket\text{whether Peter called}\rrbracket, \llbracket\text{whether Mary called}\rrbracket, \ldots\}$
Note that: $\llbracket\text{whether Peter}\oplus\text{Mary called}\rrbracket \notin Sub(\llbracket(9)\rrbracket)$

As defined here, $Sub(Q)$ corresponds to the strong version of Beck and Sharvit's (2002) $Part(Q)$. This is so because our denotation for questions is not restricted to true answers, therefore $Sub(Q)$ contains *all* questions of the form "whether x called". In particular, if Bill is a student who did not call, ⟦whether Bill called⟧ will be part of $Sub(Q)$, so answering all the questions in $Sub(Q)$ will require affirming that Bill did not call. Nevertheless, we will see in the next paragraph that Sub will not usually apply directly to Q, and we will often see weak exhaustive answers.

Another point about Sub that we will discuss in 5.3.3 is that this operator does not distinguish singular and plural questions (since they only differ on their non-atomic propositions).

Lexical restrictors: Each responsive verb V will be assumed to come with a restrictor C. For veridical responsive verbs, such as *know*, C is simply the set of propositions which are true in the context. More generally, I will make the following assumptions:

(14) a. C is closed under conjunction.
b. If q is a set of propositions, a structure "$V q$" can be interpreted as $V(\sigma(q\cap C))$.
c. If a question embedded under V undergoes IR, a copy of C projects with the question.

As mentioned above, $Sub(Q)$ corresponds to the strong version of Beck and Sharvit's (2002) $Part(Q)$, but $Sub(Q\cap C)$ corresponds to the weak version of $Part(Q)$. To see this, let us come back to our previous example with Bill, a student who did not actually call, and imagine that C is the set of true propositions. The atomic answer "that Bill called" is in Q but not in C, since it is false. Therefore, the question denoted by "whether Bill called" will not be part of $Sub(Q \cap C)$. As a matter of fact, only questions about students who actually called will be part of $Sub(Q \cap C)$. Therefore, answering all these questions will only provide Karttunen's weak exhaustive answer.

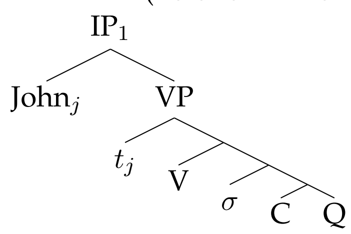
As a consequence, the theory only derives weak exhaustive readings for responsive verbs, although it still derives strong exhaustive readings for rogative verbs (which do not introduce a restrictor).

One could imagine making the rule (14c) optional. This would lead to a theory which can derive Heim's (1994) weak and strong exhaustive readings (including in QV sentences). Nevertheless, deriving the intermediate exhaustive readings (Klinedinst & Rothschild, 2011; Spector & Egré, 2015; Cremers & Chemla, 2014) would require more than simple parameters tweaking, and results from Cremers and Chemla (2014) suggest that this reading is the most salient exhaustive reading. Another argument for not making the rule (14c) optional is that the strong exhaustive reading it would predict for QV sentences is intuitively wrong: all previous literature agrees on the fact that QVE is quantification on the *true* answers, and a strong exhaustive QVE reading should add an independent constraint on the false answers, but not affect the domain of quantification. In the following, I leave the issue of exhaustivity aside since it is mostly orthogonal to the questions discussed in this manuscript and would require sophisticated adjustments to be fully addressed.

Possible structures: When the sentence contains an overt adverb, the question must undergo IR to fill the restrictor position of the adverb. When the sentence does not contain any overt adverb, the question can be interpreted *in situ* or undergo IR to the restrictor position of a silent adverb, which I will write $\mathcal{H}$. The two resulting structures for responsive verbs are presented in (15). Rogative verbs involve similar structures, except they do not introduce a lexical restrictor nor σ-operator.

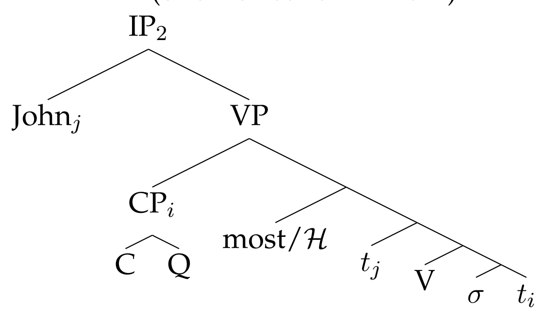
(15) Possible structures for "John $(Adv)VQ$" with V a responsive verb and Q a question.

a. *in situ* LF (no overt adverb)



$[\![\text{IP}_1]\!] = V'(j)(\sigma(C \cap Q))$

b. LF with IR (overt or covert adverb)



$[\![\text{IP}_2]\!] = \text{Adv}'(Sub(C \cap Q))(\lambda q.V'(j)(\sigma(C \cap q)))$

Note that we need a rule like Rule AB (Lahiri, 2002, p85) to get the right LF in (15b).

Homogeneity: Instead of $\mathcal{H}$, Lahiri (2002) proposes a null adverb in (15b), which is resolved as either universal closure (yielding an exhaustive reading) or existential closure (yielding a *mention-some* reading), depending on context. The present theory will depart from this view and treat the silent adverb as the source of homogeneity effects.

We will adapt the theory of homogeneity proposed in Spector (2013b), building on Krifka's (1996) *Strongest Meaning Hypothesis* (SMH). The idea is rather simple: we assume, following Lahiri (2002) that $\mathcal{H}$ is somehow ambiguous between a universal and an existential closure. However, instead of assuming an ambiguity resolved by context, we will apply a supervaluationist version of the SMH: the sentence will be true only if it would be true under both an existential and a universal closure, false if it would be false under both. In any other case, the sentence will be a truth value gap. This idea can be reflected in the entry for $\mathcal{H}$ presented in (16).

(16) Interpretation rule for $\mathcal{H}$:

$$[\![\mathcal{H}]\!](A)(B) = \left|\begin{array}{lllll} 1 \text{ iff} & \forall q.[A(q) \to B(q)] & \wedge & \exists q.[A(q) \wedge B(q)] \\ 0 \text{ iff} & \neg\forall q.[A(q) \to B(q)] & \wedge & \neg\exists q.[A(q) \wedge B(q)] \\ \# & \text{otherwise} \end{array}\right.$$

Note that, assuming A is not empty, one of the conjuncts is always entailed by the other. In fact, the only contribution of the existential conjunct to the truth-conditions is to force the existence of at least one A-element. The same point can be made about the contribution of the universal conjunct to the falsity-conditions. A sentence with $\mathcal{H}$ will therefore always be a truth-value gap if the restrictor is empty.

We refer to Križ and Spector, 2015 for details regarding the projection of this truth value gap in quantified sentences, but for the simple cases we are interested in (mostly embedding under negation), we can simply use Kleene's strong logic (K_3).

5.3.2 Application to simple cases

We will now see how the theory introduced above fares on simple sentences where a *which*-question is embedded under *know*, possibly with sentential negation or an adverb of quantity.

Simple affirmative sentence with *know*

(5a) John knows which students called.

Since the sentence contains no overt adverb, it is ambiguous between the LF (15a), where the question is interpreted *in situ*, and (15b) where it undergoes IR to the restrictor of $\mathcal{H}$.

***In situ* interpretation:** The LF is presented in (17). In the case of *know* (and other veridical responsive verbs) C restricts Q to only the true answers, and σ returns the strongest of these. Since the set $C \cap Q$ is closed under conjunction, its strongest element is simply the conjunction of all, hence $\sigma(C \cap Q)$ will return the conjunction of all true answers, i.e. Karttunen's true complete answer.[9] The meaning can be paraphrased as "John is in the *know*-relation with the conjunction of all true answers to the question", which corresponds to the usual weakly exhaustive reading of (5a).

(17) $[\![(5a)]\!] = \text{know}'(j)(\sigma(C \cap Q)))$
$[\![(5a)]\!] = \text{know}'(j)\left(\lambda w. \bigwedge_{x \in \text{student}' \cap \text{called}'} [\![\text{called}]\!]^w(x)\right)$

Interrogative raising: If the question undergoes IR to the restrictor position of $\mathcal{H}$, homogeneity comes into play (i.e. application of rule 16). As shown in (18), the sentence receives truth-conditions equivalent to the previous LF (the usual weakly exhaustive reading). Nevertheless, the falsity-conditions are more demanding than the mere negation of this: for the sentence to be plain false, John has to be ignorant about all students who called. Any intermediate case (in which John knows about some of the student callers but not all) leads to a truth-value gap.

(18) $[\![(5a)]\!] = [\![\mathcal{H}]\!](Sub(C \cap Q))(\lambda q.\text{know}'(j)(\sigma(C \cap q)))$
$[\![(5a)]\!] = 1 \leftrightarrow \left(\forall x \in \text{student}' \cap \text{called}', \text{know}'(j)(\lambda w.[\![\text{called}]\!]^w(x))\right)$
$[\![(5a)]\!] = 0 \leftrightarrow \left(\neg\exists x \in \text{student}' \cap \text{called}', \text{know}'(j)(\lambda w.[\![\text{called}]\!]^w(x))\right)$

Simple negative sentence with *know*

(5b) John doesn't know which students called.

Again, the sentence is ambiguous between two LFs, depending on whether the question remains *in situ* or undergoes movement.

[9] In cases where there is no true answer (no student called), this would just return a presupposition failure. I see two possible solutions to this problem, which I will not implement. The first one would be to modify the definition of σ so that it returns a stronger proposition in such cases (following Karttunen, 1977 or Heim, 1994). Another one would be to assume that in such cases, pragmatics favors a strongly exhaustive reading of the sentence because the other readings are too weak.

***In situ* interpretation:** In this case, we simply obtain the logical negation of (5a), which means that for at least one student caller, John does not know that she called.

Interrogative raising: We will assume that negation interacts with trivalence following the rules of Kleene's K_3 logic, in (19). The situation is symmetric compared to (5a): as shown in (20) the falsity conditions are as usually predicted, but the truth conditions are stronger. The sentence is true only if John knows nothing about the students who called. Note that the scope of the negation has no effect here. If $\mathcal{H}$ was to take scope over negation, we would retrieve the same truth- and falsity-conditions by exchanging the roles of the "$\forall\ldots$" and "$\exists\ldots$" conjuncts.

(19) Truth table of the negation in K_3:

p	$\neg p$
1	0
#	#
0	1

(20) $[\![(5\text{b})]\!] = \neg[\![\mathcal{H}]\!](Sub(C \cap Q))(\lambda q.\text{know}'(j)(\sigma(C \cap q)))$
$[\![(5\text{b})]\!] = 1 \leftrightarrow \neg\big(\exists x \in \text{student}' \cap \text{called}', \text{know}'(j)(\lambda w.[\![\text{called}]\!]^w(x))\big)$
$[\![(5\text{b})]\!] = 0 \leftrightarrow \neg\big(\neg\forall x \in \text{student}' \cap \text{called}', \text{know}'(j)(\lambda w.[\![\text{called}]\!]^w(x))\big)$

Sentence with an adverb of quantity with *know*

(6) John mostly knows which students called.

If the adverb is overt, for instance *mostly*, the question must undergo IR to fill the restrictor position of the adverb. In this case, the theory derives the usual QV reading. The sentence (6) is predicted to be true if and only if for most students who actually called, John knows that they did. The equivalence is shown in (21).[10]

(21) $[\![(6)]\!] = \text{mostly}'(Sub(C \cap Q))(\lambda q.\text{know}'(j)(\sigma(C \cap q)))$
$[\![(6)]\!] = \text{mostly}'(\lambda q.\exists p \in At(C \cap Q) : q = ?p)(\lambda q.\text{know}'(j)(\sigma(C \cap q)))$
$[\![(6)]\!] \equiv most_{st}\, p \in At(C \cap Q), \text{know}'(j)(p)$
$[\![(6)]\!] \equiv most_e{}'\, x \in \text{student}' \cap \text{called}', \text{know}'(j)(\lambda w.[\![\text{called}]\!]^w(x))$

Summary: The theory predicts the correct reading for simple QVE sentences such as (6), and an ambiguity in the case of sentences which embed questions without an overt adverb of quantity, such as (5a,b). If the question is interpreted *in situ*, the sentence receives the usual weakly exhaustive reading and no homogeneity. If the question undergoes IR to the restrictor position of a silent adverb $\mathcal{H}$, homogeneity effects come into play. In the case of (5a) the difference is subtle because it only shows up in the falsity conditions. For (5b), the sentence which motivated the implementation of homogeneity in the first place, the truth-conditions are affected (the homogeneous reading has stronger truth-conditions). This is in line with experimental results by Xiang (2015a).

The theory also inherits from Spector (2013b) the prediction that homogeneity projects universally in sentence (5c) and accounts for the fact that (5d) does not show homogeneity (because the adverb *completely* replaces the silent adverb $\mathcal{H}$).

[10] I use informal quantifiers '$most_\tau$' on types τ: "$most_\tau\, x_\tau \in P, P'(x)$" iff $|P \cap P'| > \alpha|P|$. For simplicity, we can assume $\alpha = 1/2$ but the threshold is likely to be vague.

5.3.3 New predictions and puzzles

We have seen that the theory derives the correct readings for simple sentences. We will now inspect its predictions on a wider variety of cases (in particular, with more question-embedding verbs).

Rogative verbs

I adopted Beck and Sharvit's (2002) idea of quantification over subquestions rather than answers in order to derive QVE for rogative verbs as well. The theory treats homogeneity as the counterpart of QVE in sentences which lack an overt adverb. Therefore, it predicts that any verb which allows QVE should also give rise to homogeneity effects. In the case of rogative verbs, there seems to be some restrictions on QVE (in particular, in the case of *wonder*). Without discussing these restrictions, the theory at least predicts a correlation between the availability of QVE and homogeneity.

***Depend on*:** Based on the sentence (8) repeated below, Beck and Sharvit (2002) argue that *depend on* allows QVE (at least in its first argument). This predicate also seems to give rise to homogeneity effects. At least under a certain reading, (22a) suggests that the committee has no influence on any admission. As in previous examples, the adverb *completely* in (22b) blocks homogeneity (and probably introduces a scalar implicature as well), as predicted by the current theory.

(8) Who will be admitted depends for the most part (exclusively) on this committee.

(22) a. Who will be admitted doesn't depend on this committee.
b. Who will be admitted doesn't completely depend on this committee.

***Wonder*:** This predicate usually does not to give rise to QVE: it cannot be modified by adverbs like *completely* and QV readings are not available in sentences such as (23a). However, in a sentence like (23b), *wonder* seems to allow QVE. Without a clear characterization of the sentences where *wonder* allows QVE, we cannot extract clear predictions about the distribution of homogeneity effect. Even though a definite argument would be hard to build, we can still exhibit some sentences which may be problematic for a theory that does not allow homogeneity with *wonder*.

(23c) intuitively feels a bit contradictory. This suggest that one cannot wonder about a question which is mostly resolved. Hence we can reject an existential semantic for *wonder*. However, (23d) intuitively implies that no professor wonders about any of her students.[11] The simplest way to account for these sentences may be to allow *wonder* to sometimes give rise to homogeneity, as predicted by the theory.

(23) a. John mostly wonders which students called.
b. John mostly knows, but still partly wonders which students called.
c. ??John mostly knows, but still wonders which students called.

[11] For some reason, *wonder* under negation sounds odd. One exception would be contrastive focus, as in:

(i) John doesn't wonder which candidates were admitted, he KNOWS which candidates were admitted.

d. No professor wonders which of her students called.

To conclude, *wonder* seems to give rise to QVE or homogeneity under certain conditions. The current theory predicts that the two should be parallel, but this prediction is hard to evaluate.

Surprise

(24) *It surprised Mary whether Peter called.

(25) It surprised Mary which students called.

(26) For the most part, it surprised Mary which students called.

Surprise does not embed *whether*-questions, as shown in (24). It does not seem to exhibit any homogeneity effect either. Otherwise, (25) would mean that every student who called surprised Mary, while it is clear that only a few unexpected callers are sufficient to make the sentence true.[12] In the theory I presented, it should be easy to link these two facts: if *surprise* cannot embed *whether*-questions, it should not allow structures with IR, hence it should not give rise to homogeneity nor QVE. Nevertheless, (26) naturally receives a QVE reading.

Other types of questions

So far we varied the embedding contexts, but we only discussed questions such as (9), which have a few important features. First of all, they only contain one *wh*-phrase. Second, this *wh*-phrase is plural-marked. Third, the verb in the question (e.g., *call*) is a distributive predicate. Let me first note that replacing a plural *which*-phrase with a neutral *wh*-word, such as *who* or *what*, does not affect any of the judgments. As a consequence, the domain for these words should include pluralities. We will now discuss what happens with other types of questions, and how the theory must be updated to account for all this.

Multiple *wh*-questions and questions with quantifiers: Under a pair-list reading, multiple *wh*-questions exhibit QVE effects (see Fox, 2012 for a recent discussion). As pointed out by Lahiri (2002), if we want to account for these questions, the definition of At given in (12) must be refined. In order to get rid of propositions of the form "$a \oplus b$ read $c \oplus d$", the updated definition in (27) is necessary. Unsurprisingly, the prediction that homogeneity correlates with QVE extends to multiple *wh*-questions. This seems correct, as illustrated in (28).

QVE for questions with quantifiers may require a slightly complex denotation, but it is possible (see Preuss, 2001). However, the intuitions about (29b) suggest that it does not display homogeneity. This may be because the quantified DP *every boy* is able to take scope on its own (see example 90 in Fox, 2000, p64), or because it can trigger a scalar implicature that John knows for some boys which book they read. In any case, this may be an argument in favor of Križ (2015c)'s account of homogeneity in questions.

[12] Note that a denotation *à la* Guerzoni and Sharvit (2007) would make *surprise* very similar to *forget*, since it has a positive presupposition and a negative assertion. Nevertheless, the two verbs behave very differently. We will see in section 5.4.2 that *forget* even provides a further argument for homogeneity.

(27) Updated definition of the At operator (see Lahiri, 2002, p.202):

$$At = \lambda Q_{<st,t>}.\lambda p_{st}.\Big[\ p \in Q \wedge \forall r \in Q, [(p \subseteq r) \rightarrow (p = r)] \wedge \bigwedge \{q \in Q | q \neq p \wedge \forall r \in Q, [(q \subseteq r) \rightarrow (q = r)]\} \not\subseteq p\Big]$$

(28) a. For the most part, John knows which boy read which book.
b. John doesn't know which boy read which book.
$\rightsquigarrow$ For no boy x, John knows which book x read.

(29) a. For the most part, John knows which book every boy read.
b. John doesn't know which book every boy read.
$\not\rightsquigarrow$ For no boy x, John knows which book x read.

Singular *which*-phrase: Dayal (1996) noted that questions with a singular *which*-phrase, as $\mathcal{Q}$ in (30a), trigger a uniqueness presupposition (only one student called). The number-marking can be translated in the denotation.

(30) a. $\mathcal{Q}$ = Which student called?
b. $[\![\mathcal{Q}]\!] = \lambda p_{st}.\,\exists x \in [\![\text{student}]\!] \colon [p = \lambda w.[\![\text{called}]\!]^w(x)]$

Crucially, the denotation in (30b) does not contain any proposition of the form $\lambda w.[\![\text{called}]\!]^w(p \oplus m)$, which involves plural individuals. This means that $Q \cap C$ will usually not be closed under conjunction or contain a maximal element and $\sigma(Q \cap C)$ will not be defined, unless $Q \cap C$ is a singleton set. In the case of veridical predicates, this correctly derives the uniqueness presupposition. For other predicates, the presupposition depends on the restrictor.

However, the theory as it stands does not prevent the question from scoping out of σ. Sub, unlike σ, does not presuppose anything, and it does not distinguish between singular and plural marked questions, so the theory incorrectly predicts that the presupposition can be obviated. Several solutions are conceivable: we could either encode in Sub some presupposition about its argument, block IR for singular questions, or require that the presupposition of σ be satisfied before IR. I will pursue the first solution, and simply assume that Sub presupposes that its complement has a maximal element.

Collective predicates: Collective predicates, such as *lift the piano*, are predicates which take as arguments genuine pluralities of individuals. In particular, they do not distribute over atomic parts when applied to a plural individual, so the inference in (31) is not valid.

(31) Peter and his friends lifted the piano.
———
Peter lifted the piano.

(32) $\mathcal{Q}$ = Which students lifted the piano?
$[\![\mathcal{Q}]\!] = \lambda p_{st}.\,\exists x \colon [[\![{}^\star\text{student}]\!](x) \wedge p = \lambda w.[\![\text{l.t.p.}]\!]^w(x)]$

Since $[\![\text{l.t.p.}]\!](a \oplus b)$ is not equivalent to the conjunction of the propositions $[\![\text{l.t.p.}]\!](a)$ and $[\![\text{l.t.p.}]\!](b)$, the denotation of the question will not be closed under conjunction and will usually not have a maximal element. However, the presupposition of (32) will be easier to satisfy or accommodate. Indeed, (32) will presuppose that there is a unique plurality of students who lifted the piano, with no constraints on how many students composed this plurality.

Since there is in general no maximal element, we predict no homogeneity or QVE effect here. However, some questions with collective predicates seem to be ambiguous between a question about a single plurality or about a set of pluralities, thus allowing more than one answer. As an example, it seems possible to ask “Which students are siblings?” in a situation where one expects several sibling groups, and sibling groups are the atoms for quantification in one reading of the sentence (33). Furthermore, even when a question has a unique plural answer, it seems to allow quantification on atomic individuals involved in the unique answer, as in (34).

(33) John knows, for the most part, which students are siblings.

(34) John knew, for the most part, which students lifted the piano, and he gave them $5 for their help.

5.4 More on lexical restrictors

5.4.1 On the necessity of lexical restrictors

Most veridical responsive verbs happen to be factive, and conversely, all factive verbs that embed questions are veridical responsive. For this reason, Berman (1991), Lahiri (1991) assumed that the restrictor for QVE is determined by accommodating the presuppositions of the verb. Following Lahiri (2002), I departed from this view and assumed that each responsive verb comes with a restrictor that is a lexical property of the verb.

One argument of Lahiri (2002) against the presupposition accommodation solution has to do with the controversial status of so-called *intermediate accommodation* (see Beaver, 1995), but recent literature seemed to reach a consensus for the availability of such a mechanism (Geurts & van der Sandt, 1999; Singh, 2008, 2009). However, two simple facts challenge the intermediate accommodation solution for question embedding. First, communication verbs such as *tell* or *predict* are not factive nor even veridical when embedding propositions, but are usually veridical when embedding a question (this argument is already present in Lahiri, 2002). Second, complex factive verbs such as *forget* or *discover* have stronger presuppositions than *know*, but they are simply veridical when embedding questions.

Spector and Egré (2015) provide arguments against the first point. As we will see in section 5.4.3, they show that communication verbs (a) are factive in some contexts and (b) are not always veridical when embedding questions. The second point however has not been challenged. We will now look at complex factive verbs in detail.

5.4.2 Complex factive verbs

We will focus on the case of *forget*, as in the sentence (35), for a concrete example.

The most natural entry for *forget that p*, illustrated in (35a), introduces two presuppositions (factivity and knowledge in the past) and asserts that the agent no longer knows *p*. According to the presupposition accommodation approach, this implies that (35) should receive the truth-conditions in (35b), which seem clearly wrong. (35c) looks like a much better paraphrase of (35). One way to maintain the presupposition accommodation approach would be to attribute different statuses to the two presuppositions of *forget*, so that only factivity would play a role in accommodating the restrictor.[13]

In the lexical restrictor approach, we need to postulate that the restrictor is a simple veridical restrictor (the same as *know*). This makes the factivity presupposition trivially satisfied. However, the 'knowledge in the past' presupposition remains, and the underlined part of (35c) is presumably a presupposition.

(35) John forgot which students called.

a. John forgot p: p and John used to know p, John does not know p (any more).

b. # For each student x who called and such that John knew that x called, John forgot that x called.

c. For each student x who called, John knew that x called and John forgot that x called.

Note in passing that this sentence provides one more argument for a decomposition of the embedded question into smaller units. Indeed, the most salient reading attributes exhaustive knowledge to John in the past, but also exhaustive oblivion. This would not be possible in an account which relates John to a single proposition. For instance, if a, b and c are the students who called, the true complete answer will be 'a, b and c called'. If John forgets about c, but still remembers about a and b, he would still count as having forgotten the proposition that a, b and c called. Conversely, a disjunctive proposition 'a, b or c called' would give the right assertion but a weak presupposition.

If we look at other complex factive verbs (e.g., *remember*, *discover*, *find out*), the same observation holds: all these verbs are simply veridical.

5.4.3 Communication verbs

Communication verbs were long thought to be veridical when embedding questions, but Spector and Egré (2015) recently proposed that they are ambiguous between a veridical and a non-veridical reading when embedding questions, and between a factive and a non-factive reading when embedding declaratives. The latter is supported by data from Schlenker (2007), repeated in (36). From (36a) we infer that Sue is indeed pregnant, and this inference projects out of negation or questions, a projection pattern typical of presuppositions. In support of the former point, Spector and Egré (2015) provide examples such as (37). Furthermore, they argue that the veridical readings correspond exactly to the factive entries of communication verbs and provide data from Hungarian in support of this view. Hungarian makes a distinction between factive and non-factive *tell*, and they show that the veridical reading of *tell* corresponds to the factive entry.

[13] Such an analysis is sketched in Roelofsen et al. (2014) and Uegaki (2015), following a suggestion of Theiler (2014), although none of them apply it to factive verbs beyond *know* and emotive-factives. The idea is to derive the factivity of declarative entries from the presence of an operator which is responsible for the veridicality of the corresponding question-embedding predicates. Therefore, the factivity presupposition is not hard-coded in the semantics of the verb, but is rather a by-product of the verb being *extensional*, in Groenendijk and Stokhof's (1984) sense. Other presuppositions, such as the 'knowledge in the past' presupposition of *forget* would remain purely lexical and would not play any role in the restriction of the question. This approach seems promising, but it may not work well with non-veridical responsive predicates. As we will see, the restrictors for verbs like *be certain* or *agree* seem arbitrary and would, if anything, correspond to lexical presuppositions of these verbs.

(36) a. Sue told someone that she is pregnant.
b. Sue didn't tell anyone that she is pregnant.
c. Did Sue tell anyone that she is pregnant?

(37) Every day, the meteorologists tell the population where it will rain the following day, but they are often wrong.

(38) John told me which students called.
a. For each student x who called, John told me that x called.
$\rightarrow$ John told me the complete answer to Q
b. For each student x such that John believes that x called, John told me x called.
$\rightarrow$ John told me what he believes to be the complete answer to Q

The restrictor for the veridical entry must be $C = \lambda p.p(w_0)$, which yields the veridical reading in (38a). Spector and Egré's (2015) lexical rule involves existential quantification, hence the non-veridical reading they derive is simply "John told me some answer to Q" (most of recent work in the field also assumes some form of existential quantification for non-veridical predicates). In the theory we proposed, in the absence of any restrictor, (38) would mean "John told me that every student called", which is clearly wrong (and IR does not help here since Sub also relies on the restrictor). Non-veridical entries require a restrictor as well, and a reasonable option for non-veridical *tell* and other communication verbs is to use the set of propositions that John believes: $C = \lambda p.B_j(p)$. This yields the reading (38b).

5.4.4 Other non-veridical predicates

As we have just seen, every responsive verb must come with a lexical restrictor. In the case of non-veridical predicates, there is no obvious solution and the restrictor must be postulated. The solution adopted by Lahiri (2002) associates restrictors which are dependent on the agent, as exemplified by the restrictor for *be certain* in (39) and *agree on* in (40).

(39) $[\![x \text{ is certain about } Q]\!] = [\![\text{be certain}]\!](\sigma(C_x \cap Q))(x)$
$C_x = \lambda p_{st}.\exists w \in Dox_x(w_0) : p(w)$ ("x considers it possible that p")

(40) $[\![x \text{ and } y \text{ agree on } Q]\!] = B_x(\sigma(Q \cap C_{xy})) \wedge B_y(\sigma(Q \cap C_{xy}))$
$C_{xy} = \lambda p.\left[Dox_x(w_0) \subseteq p \vee Dox_y(w_0) \subseteq p\right]$ ("x or y believes p")

(41) Peter and Mary disagree on who called.
$[\![x \text{ and } y \text{ disagree on } Q]\!] = \neg B_x(\sigma(Q \cap C_{xy})) \vee \neg B_y(\sigma(Q \cap C_{xy}))$

Interestingly, this restrictor for *agree on* is not necessarily closed under conjunction, thus contradicts hypothesis (14a)[14]. For instance, Mary may believe that only Ann called, while Peter may believe that only Bill called. In such a case, 'Ann called' and 'Bill called' would be in C, but 'Ann and Bill called' would not. This seems wrong because in such a situation, (41) sounds true and not a presupposition failure. Therefore, we can assume that the hypothesis (14a) holds, and that the right restrictor for *agree* and *disagree* is the closure under conjunction of the restrictor in (40). In this case the maximal answer is 'Ann and Bill called', which both Peter and Mary believe to be false (note that they also disagree on every atom, so the sentence also satisfies homogeneity).

[14]Thanks to Wataru Uegaki for pointing this to me.

Now consider a situation where Mary believes that only Ann and Bill called, while Peter believes that only Bill and Celine called. The closure of the union of Peter and Mary's beliefs contains 3 atomic propositions, call them a, b and c, and the maximal answer is $a \wedge b \wedge c$. In this case, the assertion 'Peter and Mary agree on who called' is predicted to be false if the question is interpreted *in situ*, but not super-false, since Peter and Mary agree on b. Conversely, 'Peter and Mary disagree on who called' would be true under the *in situ* reading, but would violate homogeneity under the silent adverb reading.

5.4.5 *Believe* and embedded questions

A surprising fact about responsive verbs is that *believe* is not one of them, as shown by the ungrammaticality of (42). This has been a recurrent topic in the literature and several explanations have been proposed, none of which is fully satisfying.[15] The current framework allows for a new explanation for the fact, although it will remain speculative.

(42) *John believes which students called.

If *believe* was a responsive verb, we would need to determine its restrictor. We could imagine using a veridical restrictor, but then *believe* would be almost equivalent to *know*. In fact, *believe* does sometimes embed questions, and is then veridical with respects to its interrogative complement, as shown by the example (43a) from Egré (2008). However, this is only possible in negative, exclamative sentences, and *believe* then receives a semantic closer to an emotive-factive predicate than to a cognitive predicate. Egré (2008) shows that the parallel with emotive-factives extends to the fact that *believe* cannot embed *whether*-questions, even in negative exclamative sentences, as indicated by the ungrammaticality of (43b). Finally, (43c) seems to indicate that the corresponding declarative-embedding *believe* is indeed factive.

(43) a. Peter will never believe who came to the party!
b. *Peter will never believe whether Mary came to the party!
c. Peter will never believe that Mary came to the party!
$\rightsquigarrow$ Mary came to the party.

Apart from the veridical restrictor, we could imagine using the default restrictor we used for non-veridical communication verbs: $C_x = \lambda p. [\![\text{believe}]\!](p)(x)$. Unfortunately, this would make (42) a tautology: "for each student x such that John believes that x called, John believes that x called".

Most theories treat non-veridical responsive predicates as existential quantifiers on possible answers, so they predict that – provided it was grammatical – (42) would mean something like "John believes some answer to the question". In a theory where responsive verbs *must* have a restrictor, such a weak reading is impossible. Of the two plausible restrictors, one gives a tautology and the other returns a veridical entry which is only available in very specific contexts and with an emotive-factive semantics, thus escaping the competition with *know*. Hence we may assume that the only special property of *believe* is that it is the default propositional attitude, and as such cannot embed questions without leading to a tautology, unless we add another dimension to its meaning.

[15] Vendler (1972) and Ginzburg (1995a, 1995b) argue for a distinction between *facts* and *propositions*, but this misses the fact that *regret* cannot embed questions although it embeds facts just like *know*. Egré (2004) proposes a generalization for French based on a new class of verbs (*indicative factive* predicates), but this predicts that all responsive predicates are veridical, missing the fact that *agree* or *be certain* can embed questions.

5.5 Conclusion

Embedded questions give rise to quantificational variability effects which have been studied extensively since Berman (1991) and are usually treated as a consequence of the plurality feature of some questions. There are other effects related to plurality with embedded questions (cumulative readings, homogeneity), but they have not been studied in so much detail. Here I proposed a general theory of QVE and homogeneity, building on previous accounts (Lahiri, 2002 and Beck and Sharvit, 2002, for the most part) and on the super-valuationist theory of homogeneity (Krifka, 1996; Spector, 2013b). This theory extends existing theories of questions and makes a few welcome predictions about homogeneity effects.

Following Lahiri (2002), I used a context variable C for restricting questions and assumed that its value is lexically determined by the embedding verb. This solution is less minimal than the presupposition accommodation approach of Berman (1991), Lahiri (1991), and it raises non-trivial questions about the acquisition of these lexical restrictors. However, it seems necessary if we want to derive the correct meaning for complex factive verbs. It also predicts stronger meaning for non-veridical responsive predicates than the existential quantification on answers which is usually proposed. This is a desirable feature when we look at predicates like *be certain* or *agree*. Finally, the lexical restrictor hypothesis points toward a new explanation for why *believe* does not embed questions.

Three points I did not discuss in this paper may have puzzled some readers. The first one has to do with what we call *plurality*. In this theory, questions denote sets of propositions and the Sub operator returns a set of questions. Strictly speaking, Link's (1983) pluralities are not just sets. Sets of propositions which are closed under conjunction (Lahiri's Proposition Conjunction Algebras) have the same structure as plural predicates of individuals, which are closed under sum-formation. Conjunction plays the role of a sum operator, and we can define everything we would define for regular pluralities (supremum operator, atoms). Sets of questions by contrast do not come with a natural ordering or binary operator. This makes the theory stipulative regarding why questions give rise to plurality effects.

The second point has to do with the $\mathcal{H}$ operator. The super-valuationist theories of homogeneity do not make use of such a thing. In the nominal field (at least in English), homogeneity arises when a plural noun is combined with a definite article. The phrase "the $\text{NP}_{\text{plural}}$" denotes a single plural individual, and combines with a (plural) predicate of individuals, of type $<e, \tau>$. In the case of questions, the situation is a bit more complicated because we have rogative and responsive predicates. We could account for homogeneity with responsive verbs in a much simpler way by giving up the null adverb and making our σ-operator responsible for the homogeneity effects (after all, it is a form of definite description). However, this would not work with rogative verbs, because they must combine with the whole question and not just its supremum.[16]

Thirdly, the current theory does not derive strong or intermediate exhaustive readings. I assume that such exhaustivity effects can be dealt with independently. This in particular can be attempted following Klinedinst and Rothschild's (2011) proposal. The key for such an approach is to define proper focus values for questions, so that they would percolate into alternative sets for sentences which embed questions. An exhaustivity operator can then strengthen these sentences and this

[16] A symmetric problem arises if we consider that questions denote sets of (proto)-questions instead of sets of propositions. This would work well with rogative verbs, but would make the restriction problem very complicated for responsive verbs.

is how we would derive the intermediate exhaustive reading. A few technicalities which make this extension non-trivial here explain why I considered that it was out of the reach of the present contribution, but I refer to Cremers (2015b) for a similar implementation.

Additional discussion

We saw how this second theory can derive QVE and homogeneity effects for rogative predicates. Yet, we did not talk about stronger exhaustive readings. One solution would be to add a mechanism to generate alternatives on top of what we have, as I did in the previous chapter. However, this requires additional modifications. First, the denotation of *whether*-questions in Sub must be updated so that $[\![\text{whether } p]\!] = \{p, \top\}$ (p and the tautology, rather than p and $\neg p$). Then "Mary knows whether p" would receive a WE reading equivalent to "if p, Mary knows p". The usual reading of *whether*-questions would be derived as a SE reading (through local exhaustification), and we would predict an additional IE reading (if p, Mary knows that p, otherwise, she does not falsely believe p).[17] This theory matches most of the predictions of the first theory. Since rogative predicates do not embed Ans operators and do not have lexical restrictors, they are not affected by this move (they always relate the agent to all the sub-questions).

[17]Note that this would only affect the subquestions introduced by Sub. We could assume that the actual word *whether* triggers obligatory exhaustification (just like complex disjunction "either..or", Spector, 2014). For a similar view on *whether*-questions, see Andreea Cristina Nicolae (2013), Andreea C Nicolae (2015).

Bibliography

Abbeduto, L. & Rosenberg, S. (1985). Children's knowledge of the presuppositions of know and other cognitive verbs. *Journal of Child Language*, *12*(3), 621–641.

Abels, K. (2007). Deriving selectional properties of'exclamative'predicates. *Interface and interface conditions*, 115–140.

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

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.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using `lme4`. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Beaver, D. I. (1995). *Presupposition and assertion in dynamic semantics* (Doctoral dissertation, University of Edinburgh).

Beck, S. & Rullmann, H. (1999). A flexible approach to exhaustivity in questions. *Natural Language Semantics*, *7*(3), 249–298.

Beck, S. & Sharvit, Y. (2002). Pluralities of questions. *Journal of Semantics*, *19*(2), 105–157.

Berman, S. R. (1991). *On the semantics and logical form of Wh-clauses* (Doctoral dissertation, University of Massachusetts, Amherst).

Bill, C., Pagliarini, E., Romoli, J., Tieu, L., & Crain, S. (2015). *Children's interpretation of sentences with multiple scalar terms*. Ms., Macquarie University.

Bott, L. & Noveck, I. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language*, *53*, 437–457.

Busemeyer, J. R. & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*(2), 91–134.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.

Champollion, L. (2010). *Parts of a whole: distributivity as a bridge between aspect and measurement* (Doctoral dissertation, University of Pennsylvania).

Champollion, L. (2014). *Each vs. all: distributivity, cumulativity, and two types of collectivity*. Talk at the 4th Cornell Workshop in Linguistics and Philosophy.

Champollion, L. (2015). *Distributivity, collectivity and cumulativity*. To appear in Wiley's companion to semantics.

Chemla, E. (2008). An epistemic step for anti-presuppositions. *Journal of Semantics*, *25*, 141–173.

Chemla, E. (2009). Presuppositions of quantified sentences: experimental data. *Natural Language Semantics*, *17*(4), 299–340.

Chemla, E. & George, B. (2015). Can we agree about 'agree'? *Review of Philosophy and Psychology*, 1–22. doi:10.1007/s13164-015-0278-8

Chemla, E., Homer, V., & Rothschild, D. (2011). Modularity and intuitions in formal semantics: the case of polarity items. *Linguistics and Philosophy*, *34*(6), 537–570. doi:10.1007/s10988-012-9106-0

Chemla, E. & Romoli, J. (2016). *The role of probability in the accessibility of scalar inferences*. Ms.

Chemla, E. & Spector, B. (2014). *Distinguishing typicality and ambiguities, the case of local scalar implicatures*. Ms. LSCP & IJN.

Chierchia, G., Crain, S., Guasti, M. T., & Thornton, R. (2001). 'Some' and 'or': A study on the emergence of logical form. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Boston University Conference on Language Development (BUCLD) 24* (pp. 22–44). Somerville, Massachusetts: Cascadilla Press.

Chierchia, G., Fox, D., & Spector, B. (2011). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *An international handbook of natural language meaning volume 3*. Berlin: Mouton de Gruyter.

Conroy, S., Lidz, J., & Musolino, J. (2009). The fleeting isomorphism effect. *Language Acquisition*, *16*(2), 106–117.

Crain, S., Goro, T., Notley, A., & Zhou, P. (2013). A parametric account of scope in child language. In S. Stavrakaki, M. Lalioti, & P. Konstantinopoulou (Eds.), *Advances in language acquisition* (pp. 63–71). Cambridge: Cambridge Scholar Publishing.

Cremers, A. (2015a). *Homogeneity and quantificational variability with embedded questions*. Under Revision for *Natural Language Semantics*.

Cremers, A. (2015b). *Plurality effects and exhaustivity with embedded questions*. In preparation.

Cremers, A. & Chemla, E. (2014). A psycholinguistic study of the exhaustive readings of embedded questions. *Journal of Semantics*.

Cremers, A. & Chemla, E. (2016). *Experiments on the acceptability and possible readings of questions embedded under emotive-factives*. (submitted).

Cremers, A., Tieu, L., & Chemla, E. (2016). *Children's exhaustive readings of questions*. Accepted for publication in Language Acquisition.

Dayal, V. (1996). *Locality in wh quantification: questions and relative clauses in hindi*. Dordrecht: Kluwer Academic Publishers.

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-olds' understanding of know and think. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions* (Chap. 11, Vol. 45, *8*, pp. 241–262). Studies in Theoretical Psycholinguistics. Springer.

Egré, P. (2004). *Attitudes propositionnelles et paradoxes épistémiques* (Doctoral dissertation, Université Paris 1 & IHPST).

Egré, P. (2008). Question-embedding and factivity. *Grazer philosophische studien*, *77*(1), 85–125.

Fodor, J. D. (1970). *The linguistic description of opaque contexts* (Doctoral dissertation, MIT).

Fox, D. (2000). *Economy and semantic interpretation*. Cambridge, MA: MITWPL and MIT Press.

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

Fox, D. (2012, September). *Lecture notes on the semantics of questions*. Class notes (MIT Seminars).

Fox, D. (2013, February). *Mention-some readings of questions*. Class notes (MIT Seminars).

Fox, D. & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*, 87–107.

Frank, M. C. & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M. (2011, June). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*(1), 1–82. doi:10.3765/sp.4.1

Frege, G. (1892). On sense and reference. *Ludlow (1997)*, 563–584.

Gajewski, J. R. (2005). *Neg-raising: polarity and presupposition* (Doctoral dissertation, Massachusetts Institute of Technology).

Gajewski, J. R. & Sharvit, Y. (2012). In defense of the grammatical approach to local implicatures. *Natural language semantics*, *20*(1), 31–57.

George, B. R. (2011). *Question embedding and the semantics of answers* (Doctoral dissertation, University of California Los Angeles).

George, B. R. (2013). Knowing-'wh', mention-some readings, and non-reducibility. *Thought: A Journal of Philosophy*, *2*(2), 166–177.

Geurts, B. & van der Sandt, R. (1999). Domain restriction. *Focus: Linguistic, cognitive, and computational perspectives*, 268–292.

Geurts, B. & van der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics*, *22*(1), 97–117. doi:10.1093/jos/ffh018

Ginzburg, J. (1995a). Resolving questions, i. *Linguistics and Philosophy*, *18*(5), 459–527.

Ginzburg, J. (1995b). Resolving questions, ii. *Linguistics and Philosophy*, *18*(6), 567–609.

Ginzburg, J. (1996). Interrogatives: questions, facts and dialogue. *The handbook of contemporary semantic theory. Blackwell, Oxford*.

Grice, P. (1975). Logic and conversation. In D. Davidson & G. H. Harman (Eds.), *The logic of grammar* (pp. 64–75). Encino, CA: Dickenson Publishing Company.

Grimshaw, J. (1979). Complement selection and the lexicon. *Linguistic inquiry*, 279–326.

Groenendijk, J. & Stokhof, M. (1982). Semantic analysis of wh-complements. *Linguistics and Philosophy*, *5*(2), 175–233.

Groenendijk, J. & Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*. Doctoral Dissertation, University of Amsterdam.

Groenendijk, J. & Stokhof, M. (1993). Interrogatives and adverbs of quantification. In K. Bimbo & A. Mate (Eds.), *Proceedings of the 4th symposium on logic and language* (pp. 1–29). Budapest: Aron Publishers.

Gualmini, A. (2004). *Some* knowledge children don't lack. *Linguistics*, *42*, 957–982.

Gualmini, A. & Crain, S. (2005). The structure of children's linguistic knowledge. *Linguistic Inquiry*, *36*(3), 463–474.

Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. T. (2001). At the semantics/pragmatics interface in child language. In R. Hastings, B. Jackson, & Z. Zvolenszky (Eds.), *Semantics and Linguistic Theory (SALT) 11* (pp. 231–247). Cornell University, Ithaca, NY: CLC Publications.

Gualmini, A., Hulsey, S., Hacquard, V., & Fox, D. (2008). The Question-Answer Requirement for scope assignment. *Natural Language Semantics*, *16*, 205–237.

Guerzoni, E. (2003). *Why even ask?: on the pragmatics of questions and the semantics of answers* (Doctoral dissertation, Massachusetts Institute of Technology).

Guerzoni, E. (2007). Weak exhaustivity and 'whether': a pragmatic approach. In *Proceedings of salt* (Vol. 17, pp. 112–129).

Guerzoni, E. & Sharvit, Y. (2007). A question of strength: on NPIs in interrogative clauses. *Linguistics and Philosophy*, *30*(3), 361–391.

Guerzoni, E. & Sharvit, Y. (2014). *Whether or not anything* but not *whether anything or not*. In L. Crnič & U. Sauerland (Eds.), *The art and craft of semantics: a festschrift for irene heim* (Vol. 1, pp. 199–224). Cambride, MA: MIT Working Papers in Linguistics.

Hamblin, C. L. (1958). Questions. *Australasian Journal of Philosophy*, *36*, 159–168.

Hamblin, C. L. (1973). Questions in Montague English. *Foundations of Language*, *10*(1), 41–53.

Heim, I. (1994). Interrogative semantics and Karttunen's semantics for "know". In *IATL 1* (Vol. 1, pp. 128–144). Hebrew University of Jerusalem.

Herbstritt, M., Roelofsen, F., & Aloni, M. (2015). *Embedded questions: the* wh/*whether *puzzle*. Manuscript, submitted for publication in the Questions in Discourse volume.

Johnson, C. & Maratsos, M. (1977). Early comprehension of mental verbs: think and know. *Child Development*, 1743–1747.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and philosophy*, *1*(1), 3–44.

Katsos, N. & Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

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

Kotek, H. (2014). *Composing questions* (Doctoral dissertation, Massachusetts Institute of Technology).

Kotek, H. (2015). On the semantics of wh-questions. In *Proceedings of* Sinn und Bedeutung 20*, tübingen, germany*.

Kratzer, A. (2008). On the plurality of verbs. *Event structures in linguistic form and interpretation*, 269–300.

Krifka, M. (1996). Pragmatic strengthening in plural predications and donkey sentences. In *Proceedings of salt* (Vol. 6, pp. 136–153).

Križ, M. (2015a). *Aspects of homogeneity in the semantics of natural language* (Doctoral dissertation, University of Vienna).

Križ, M. (2015b). Homogeneity, non-maximality, and 'all'. *Journal of Semantics*, ffv006.

Križ, M. (2015c, May). Homogeneity, trivalence and embedded questions. In *Proceedings of the 20th amsterdam colloquium*. Presentation at ZAS (Berlin), May 6th.

Križ, M. & Chemla, E. (2014). Two methods to find truth-value gaps and their application to the projection problem of homogeneity. *Natural Language Semantics*, 1–44.

Križ, M. & Spector, B. (2015). *A supervaluationist theory of homogeneous plural predication*. In preparation.

Lahiri, U. (1991). *Embedded interrogatives and predicates that embed them* (Doctoral dissertation, Massachusetts Institute of Technology).

Lahiri, U. (2002). *Questions and answers in embedded contexts*. Oxford Studies in Theoretical Linguistics 2. New York: Oxford University Press.

Lassiter, D. & Goodman, N. D. (2014). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory* (pp. 587–610).

Lassiter, D. & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*.

Lewis, D. (1975). Adverbs of quantification. *Formal Semantics of Natural Language*, 219–240.

Link, G. (1983). The logical analysis of plurals and mass terms: a lattice-theoretical approach. *Rainer Bauerle, Christoph Schwarze, and Arnim Von Stechow (eds)*.

Löbner, S. (1985). Definites. *Journal of semantics*, *4*(4), 279–326.

Löbner, S. (2000). Polarity in natural language: predication, quantification and negation in particular and characterizing sentences. *Linguistics and Philosophy*, *23*(3), 213–308.

Macnamara, J., Baker, E., & Olson, C. (1976). Four-year-olds' understanding of "pretending", "forget", and "know": evidence for propositional operations. *Child Development*, 62–70.

Magri, G. (2013). An account for the homogeneity effects triggered by plural definites and conjunction based on double strengthening. In S. P. Reda (Ed.), *Semantics, pragmatics and the case of scalar implicatures*. Palgrave Macmillan.

Mascarenhas, S. (2009). *Inquisitive semantics and logic* (Master's thesis, ILLC – Universiteit van Amsterdam).

Miller, K. & Schmitt, C. (2004). An experimental study on child comprehension of Spanish indefinites and bare singulars. *Language Acquisition*, *12*(3-4), 247–256.

Montague, R. (1970). English as a formal language. In B. Visentini (Ed.), *Linguaggi nella societa e nellà tecnica* (pp. 189–223). Mailand.

Moore, C. & Davidge, J. (1989). The development of mental terms: pragmatics or semantics. *Journal of Child Language*, *16*(3), 633–641.

Musolino, J. (1998). *Universal grammar and the acquisition of semantic knowledge* (Doctoral dissertation, University of Maryland).

Musolino, J., Crain, S., & Thornton, R. (2000). Navigating negative quantificational space. *Linguistics*, *38*, 1–32.

Musolino, J. & Lidz, J. (2003). The scope of isomorphism: turning adults into children. *Language Acquisition*, *11*(4), 277–291.

Musolino, J. & Lidz, J. (2006). Why children aren't universally succesful with quantification. *Linguistics*, *44*, 817–852.

Nicolae, A. C. [Andreea C]. (2015). Questions with NPIs. *Natural Language Semantics*, *23*(1), 21–76.

Nicolae, A. C. [Andreea Cristina]. (2013). *Any questions? polarity as a window into the structure of questions* (Doctoral dissertation, Harvard).

Notley, A., Zhou, P., Jensen, B., & Crain, S. (2012). Children's interpretation of disjunction in the scope of 'before': A comparison of English and Mandarin. *Journal of Child Language*, *39*(3), 482–522.

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

Onishi, K. H. & Baillargeon, R. (2005, April). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258.

Papafragou, A. & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, *86*, 253–282.

Percus, O. (2006). Antipresuppositions. In A. Uyema (Ed.), *Theoretical and empirical studies of reference and anaphora: toward the establishment of generative grammar as an empiricial science* (pp. 52–73). Report of the Grant-in-Aid for Scientific Research (B), Project No. 15320052, Japan Society for the Promotion of Science.

Phillips, J. & George, B. R. (2015). *Non-reducibility with Knowledge wh: Experimental Investigations*. (submitted).

Pike, R. (1966). Stochastic models of choice behaviour: response probabilities and latencies of finite Markov chain systems. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 15–32.

Pike, R. (1968). Latency and relative frequency of response in psychophysical discrimination. *British Journal of Mathematical and Statistical Psychology*, *21*(2), 161–182.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*(1), 53.

Preuss, S. M.-L. (2001). *Issues in the semantics of questions with quantifiers* (Doctoral dissertation, Rutgers).

Qing, C. & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: a speaker-oriented model. In *Semantics and linguistic theory* (Vol. 24, pp. 23–41).

R Core Team. (2014). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological bulletin*, *86*(3), 446.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological review*, *111*(1), 159.

Reinhart, T. (2006). *Interface strategies*. Cambridge, MA: MIT Press.

Roelofsen, F., Theiler, N., & Aloni, M. (2014). Embedded interrogatives: the role of false answers. In *7th questions in discourse workshop, göttingen*.

Romero, M. (2015). Surprise-predicates, strong exhaustivity and alternative questions. In *Semantics and linguistic theory* (Vol. 25, pp. 225–245).

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics*, *23*, 361–382.

Sæbø, K. J. (2007). A whether forecast. In *Logic, language, and computation* (pp. 189–199). Springer.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367–391.

Sauerland, U. (2008). Implicated presuppositions. In A. Steube (Ed.), *Sentence and context: language, context, and cognition*.

Scha, R. (1981). Distributive, collective and cumulative quantification. In J. Groenendijk, M. Stokhof, & T. M. V. Janssen (Eds.), *Formal methods in the study of language* (Vol. 2, pp. 131–158). Mathematisch Centrum, Amsterdam, reprinted in Groenendijk et al. (eds), *Truth, Interpretation and Information*, Foris Publications, Dordrecht (1984).

Scha, R. & Winter, Y. (2014). Plurals. *Handbook of Contemporary Semantic Theory, 2nd edn. Oxford: Wiley-Blackwell*.

Schaffer, J. (2007). Knowing the answer. *Philosophy and Phenomenological Research*, *75*(2), 383–403.

Schlenker, P. (2007). Transparency: An incremental theory of presupposition projection. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics*. Palgrave Macmillan.

Schwarz, F. (2012). *Domain restriction and discourse structure - evidence from processing*. Poster presented at the 25th CUNY conference.

Schwarzschild, R. (1993). Plurals, presuppositions and the sources of distributivity. *Natural Language Semantics*, *2*(3), 201–248.

Sharvit, Y. (2002). Embedded questions and 'de dicto' readings. *Natural Language Semantics*, *10*, 97–123.
Singh, R. (2008). *Modularity and locality in interpretation* (Doctoral dissertation, Massachusetts Institute of Technology).
Singh, R. (2009). *Maximize Presupposition!* and informationally encapsulated implicatures. In A. Riester & T. Solstad (Eds.), *Proceedings of* Sinn und Bedeutung 13 (pp. 513–526). Universität Stuttgart.
Singh, R., Wexler, K., Astle, A., Kamawar, D., & Fox, D. (2015). *Children interpret disjunction as conjunction: consequences for the theory of scalar implicatures*. Accepted pending revisions.
Song, H.-j., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? psychological reasoning in 18-month-old infants. *Cognition*, *109*, 295–315.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief in two-year-olds. *Psychological Science*, *18*(7), 587–592.
Spector, B. (2005). *Exhaustive interpretations: what to say and what not to say*. Presentation at LSA Workshop: Context and Content.
Spector, B. (2006). *Aspects de la pragmatique des opérateurs logiques*. Doctoral Dissertation, Université Paris 7.
Spector, B. (2007). Scalar implicatures: Exhaustivity and Gricean reasoning. In M. Aloni, A. Butler, & P. Dekker (Eds.), *Questions in dynamic semantics, current research in the semantics/pragmatics interface* (pp. 229–254). Elsevier.
Spector, B. (2013a). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, *7*(5), 273–294.
Spector, B. (2013b). *Homogeneity and plurals: from the Strongest Meaning Hypothesis to Supervaluations*. Presented at Sinn und Bedeutung 18.
Spector, B. (2014). Global positive polarity items and obligatory exhaustivity. *Semantics and Pragmatics*, *7*(11), 1–61.
Spector, B. (2015). Multivalent semantics for vagueness and presupposition. *Topoi*, 1–11.
Spector, B. & Egré, P. (2015). Embedded questions revisited: an answer, not necessarily the answer. *Synthese*.
Spector, B. & Sudo, Y. (2014). *Presupposed ignorance and exhaustification or: how scalar implicature and presupposition interact*. Ms., Institut Jean Nicod and UCL.
Sprouse, J. & Almeida, D. (2013). The empirical status of data in syntax: a reply to Gibson and Fedorenko. *Language and Cognitive Processes*, *28*(3), 222–228.
Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, *134*, 219–248.
Sternefeld, W. (1998). Reciprocity and cumulative predication. *Natural language semantics*, *6*(3), 303–337.
Tarski, A. (1935). Zur grundlegung der boole'schen algebra I. *Fundamenta Mathematicae*, *24*(1), 177–198.
Tarski, A. (1956). The concept of truth in formalized languages. *Logic, semantics, metamathematics*, *2*, 152–278.
Theiler, N. (2014). *A multitude of answers: Embedded questions in typed inquisitive semantics* (Master's thesis, ILLC, University of Amsterdam).
Tieu, L., Romoli, J., Zhou, P., & Crain, S. (2015). Children's knowledge of free choice inferences and scalar implicatures. *Journal of Semantics*, *doi:10.1093/jos/ffv001*.
Trinh, T. & Haida, A. (2015). Constraining the derivation of alternatives. *Natural Language Semantics*, *23*(4), 249–270.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 702–716.

Uegaki, W. (2015). *Interpreting questions under attitudes* (Doctoral dissertation, Massachusetts Institute of Technology).

van Rooij, R. & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language, and Information*, *13*, 491–519.

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2013). *Scalar diversity*. Unpublished MS.

Vendler, Z. (1972). *Res cogitans*. Cornell University Press.

Vickers, D. (1979). *Decision processes in visual perception*. Academic Press New York.

Villalta, E. (2008). Mood and gradability: an investigation of the subjunctive mood in spanish. *Linguistics and Philosophy*, *31*(4), 467–522.

Wason, P. C. (1966). Reasoning. *New horizons in psychology*, *1*, 135–151.

Westerståhl, D. (1985). Determiners and context sets. *Generalized quantifiers in natural language*, *1*, 45–71.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Springer Science & Business Media.

Wilkinson, K. (1996). The scope of 'even'. *Natural language semantics*, *4*(3), 193–215.

Xiang, Y. (2015a, September). *Complete and true: a uniform analysis for mention-some and mention-all questions*. Talk at Sinn und Bedeutung (SuB) 20, Tübingen.

Xiang, Y. (2015b). Mention-some readings of questions: complexities from number-marking. In *Proceedings of amsterdam colloquium 2015*.

Xiang, Y. (2015c). *Sensitivity to false answers*. Poster at XPRAG 2015, Chicago.

Zehr, J. (2014). *Vagueness, presupposition and truth-value judgments* (Doctoral dissertation, École Normale Supérieure de Paris).

Zhou, P. & Crain, S. (2009). Scope assignment in child language: evidence from the acquisition of chinese. *Lingua*, *119*, 973–988.