# THE QUESTION OF QUESTIONS: RESOLVING (NON-)EXHAUSTIVITY IN *WH*-QUESTIONS

**BY MORGAN C. MOYER**

**A dissertation submitted to the**

**Graduate School—New Brunswick**

**Rutgers, The State University of New Jersey**

**in partial fulfillment of the requirements**

**for the degreee of**

**Doctor of Philosophy**

**Graduate Program in Linguistics**

**Written under the direction of**

**Kristen Syrett, PhD**

**and approved by**

_____

_____

_____

_____

**New Brunswick, New Jersey**

**October, 2020**

**ABSTRACT OF THE DISSERTATION**

# The question of questions: resolving (non-)exhaustivity in *wh*-questions

**by Morgan C. Moyer**

**Dissertation Director: Kristen Syrett, PhD**

Different questions appear to call for different kinds of answers. We can refer to these readings as Mention-Some (MS) and Mention-All (MA), based on their level of exhaustivity (Hintikka 1976, 1978; Karttunen 1977; Asher & Lascarides 1998). For example, (1) is said to require exhaustivity, where Dana knows all of the relevant party-goers. (2) permits non-exhaustivity, where Dana knows some relevant place to find coffee. (3) appear to require non-exhaustivity, where Dana knows at least one way to get to Central Park.

(1) Dana knows who came to the party. #MS/MA

(2) Dana knows where we can find coffee. MS/MA

(3) Dana knows how we can get to Central Park. MS/?MA

MS readings seem to be more tightly constrained than MA readings. However, it has been an open question precisely why this is case. Across the literature, two main hypotheses have emerged. Hypothesis 1: linguistic form constrains MS availability. Three main linguistic form factors have been pinpointed. Ginzburg (1995) and Asher & Lascarides (1998) noted that who-questions favor MA, while others (why, how, and where-questions) favor MS. George (2011), following Heim (1994) argued that the matrix verb know selects for MA. Finally, a number of researchers have pointed out that questions with existential modals/non-finite clauses permit MS (Bhatt 1999; George 2011, Ch 6; Fox 2014; Nicolae 2014; Dayal 2016; Xiang 2016). Hypothesis 2: contextual goals license MS (Groenendijk & Stokhoff 1982, 1984; Ginzburg 1995; Asher & Lascarides 1998; Beck & Rullmann 1999; van Rooij 2003, 2004; George 2011, Ch.2).

Theoretical proposals have taken two different approaches to these observations about MS/MA availability. One strategy posits underlying question ambiguity, housing the variability in the semantics (Beck & Rullmann 1999; George 2011; Nicolae 2014; Fox 2014; Xiang 2016). The second strategy posits a unique semantic representation

that is either MA by default (Groenendijk & Stokhof 1982, 1984; Karttunen 1977), MS (Asher & Lascarides 1998; Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010), or semantically underspecified for either (Ginzburg 1995; van Rooij 2003, 2004). No matter which underlying semantics, context then allows for the hearer to resolve (non-)exhaustivity.

This dissertation tests these two hypotheses concerning the sets of factors licensing MS and MA readings, and thereby weighs in on the theoretical debate concerning the baseline representation of question semantics and the role of pragmatics. I provide quantitative empirical evidence that addresses the role of the linguistic factors, but demonstrate that contextual goals can indeed override those interpretational defaults. Furthermore, I demonstrate that not only MS, but MA readings, too, are subject to contextual constraints (see Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010). I argue that baseline interpretations do not reveal underlying semantics, but rather reflect hearer expectations about why a speaker would utter a given question, given that it surface-underspecifies meaning. Under this view, linguistic factors are defeasible cues to speaker goals, which direct the resolution of (non-)exhaustivity: when the context is informative with respect to discourse goals, linguistic factors are neutralized, both in interpretation and production. This finding resonates with a line of psycholinguistic research on communication and audience design (e.g., Brennan & Clark 1996; Clark & Wilkes-Gibbs 1986; Keyser et al 2000; Ferreira & Dell 2000; Ferreira 2019).

Finally, I show that hearer-specific properties drive (non-)exhaustivity resolution in questions, depending on the extent to which a hearer is more 'literal' or more 'pragmatic'. This finding helps us work toward a novel computational model of question-answer dynamics that incorporates aspects of the question, the questioner, and the hearer.

# Acknowledgements

The months leading up to the completion of this project have been rife with...events. The coronavirus pandemic made it an easy excuse to do nothing but write and establish a daily health and exercise routine. In contrast, the rising tide of social justice made it hard to take writing a dissertation seriously in the grand scheme of things.

I was a college drop-out. This small fact formed a core aspect of my identity for a long time, because it marked me as a failure in the eyes of society. It took me seven years to work my way through community college, and then to the University of Maryland where I finally completed my bachelor's. But this small stuggle has taught me the value of taking my time, and how the uncertainty of the future can bring unexpected and unforseen adventures.

My graduate experience has forged me into a researcher as a sword is forged into a weapon. The ordeal is fire, and a near-constant hammering of the mallet, under close eye of the swordsmith. Many times I felt like I was about to crack. In the end, somewhat shockingly, I have come out in once piece, but different.

First thanks go to my chair, Kristen Syrett. If we're going with the sword metaphor (which will fail at a certain point due to my ignorace of metalugy), I came to the program as a soft, raw ore, and under your direction I've become reinforced steel. You have been readily available for guidance in my moments of need and distress. I cannot accurately quantify the number of emails you've looked over, the many times you pushed me to really think something through, to structure and organize my internal world and reflect on how I conveyed that world to others. I'm not quite sure how you're able to handle two children, and a messy advisee like me, but I'm incredibly grateful.

My (internal) committee, Veneeta Dayal, Yimei Xiang, and Paul Pietroski kept me honest to the standards of both linguistic and scientific methodology. I thank Veneeta for her thorough comments on my drafts, and putting up with my bumbling expositions of semantic theory. I thank Yimei for her patience as I struggled to understand the ins/outs of her theory. I expect it wasn't easy for either of them to have a hard-headed experimentalist as a student. For Paul, I thank for both allowing me to tiptoe the frontier between philosophy-informed linguist and philosophy-too-far-gone linguist; in other words, for reminding me of my goals as a linguist but letting me ask the deeper questions about the mind.

I met Alex Lascarides in Summer 2019 when XPrag was hosted at Edinburgh. I already knew I wanted her to be my external, but after our meeting I was only more convinced. Her comments have both suscinctly summarized what I stuggle to say, and added depth and insight to this work by connecting it to other research in computational linguistics and psycholinguistis.

I would also like to express deep gratitude to Bruce Tesar for his patience, dilligence, and compassion for an experimental semanticicst bumbling through OT phonology, and Ruby. My fondest memories of the rewarding rigor of graduate experience involve the many hours we spent working through the Output-Driven Learner code and talking about programming languages. You have also given me profound guidance for navigating graduate school. Your advice has not only grounded me, but helped me to keep going. I thank you for this.

Paul de Lacy made me feel like I wasn't asking stupid questions. Mark Baker and Ken Safir unintentionally pushed me to work harder.

Ernie Lepore deserves thanks for welcoming me in the philosophy department. You kindness and encouragement helped me greatly. And Liz Camp for entertaining an independent study with a random linguist, and teaching me so much about being a woman in academia, handing difficult male personalities, but all the while with a sense of humor. Also shout out to Frankie Egan for also entertaining a linguist in her philosophy of mind classes.

I'd like to thank the organizers of the Norwegian Summer Institute on Language and Mind, especially Terje Londahl and Nick Allott. What a fantastic two-week experience! I looked forward to it every summer for the past three years. It's really programs like this, at the intersection of Mind and Language, which help seed revolutionary interdisciplinary research.

I'd like to thank the folks over at RuCCS. Sara Pixley has taught me to be a professional, to take deep breaths, that I deserve to be respected. You have been such an unwavering champion, you make me believe I am worth it (self-depricating humor). But seriously, I don't think I could have achieved so much without your support. Brian McLaughlan has also been incredibly supportive and generous.

Shout out to the ladies who really run things. Marilyn Reyes puts up with so much, but doesn't let it weigh her down. Sue Cosentino, Jo'Ann Meli, and Lynn Flannery get it done, are warm and welcoming over at RuCCS. The four of you keep these structures afloat, and deserve so much more recognition and respect than you actually receive.

I must thank my OG linguistics fam at the University of Maryland, who are the real reason I have found myself in graduate school. Valentine Hacquard, Alexander Williams, Jeff Lidz, Tonia Bleam and Peggy Antonisse. Alexander, you encouraged me to take philosophy courses and taught me how to put empirical meat on philosophical issues; Jeff, you trained me to be an experimentalist, but also not to take things so seriously. Tonia and Peggy: you are the backbone of the UMD department and the future of the field. Tonia especially, without your warm encouragement I would have never applied to the Baggett program. Your flexibility as undergraduate major allowed me to to finish my degree without compromising on the classes I felt drawn to—and that I didn't think was possible at the time. You gave me the best piece of advice about grad school: give yourself permission to leave, and if you choose to stay, that will give you the power to finish despite the difficulty. Last but absloutely not least, Valenine. Your mentorship over past 8(?!) years has kept me going.

The former UMD grads who spent hours mentoring me in the profession, and have become friends. I struggle to encapsulate our relationships in a mimimal word count. Kate Harrigan's friendship has been a solid foundation of support from the beginning; Aaron White's was sassy with twist of ghost pepper; Mike Fetters' like a WWE surprise entrance. I didn't expect to find you but there you were. Naho Orita

and Rachel Dudley, my first mentors who taught me to work my ass off. Shevaun Lewis who demands excellence and "hates" hugs. Sol Lago...I can't even express how much I appreciate you. Anamaria Bentea is a classy broad who can balance research and two toddlers. Alexis Wellwood has been a mentor-from-afar. Running into you at a conference makes it ten times more worthwhile. The Baby Lab: Tara Mease and all the other research assistants.

My grad-cohort: Livia Souza, Augustina Owusu, Deepak Alok, and Nick Winter. Though we drifed apart after first year, I treasured all the evenings we spent together working through homework and eating Delhi Garden till our stomachs exploded (or maybe that last part was just me). I love you guys so much. The many other grads at Rutgers, dear Yagmur Sag, Hazel Mitchley, Eileen Blum, Lydia Newkirk, Jess Law. Malihe Alikhani. You. Are. An. Inspiration. One day we will build robots together. Cal Howland never compromise. Vera Gor we will always have Paris. Austin Baker you saved me in Berlin. Cheers to the ones who bring us together in solidarity. I thank Satarupa Das for the long walks.

Thank you to the dear, dear, friends who have brought support, laughter, and happiness to the last six years of my life, and many more to come! Josh Anthony those last months at 103 South 4th would have been unbearable without. Andrew Rubner, you have been a selfless supporter and you deserve the best. Kat Zilka you kept me sane and taught me to love myself. Peter van Elswyk, Baby Pim. You my other family. Tim Dawson: Hail Satan. My Laddiez: Livia Souza, Selen Altiok, Sonia Szczęsna, Arcadia Lee, Gaby Figueredo (and Adam too). Selen, I'll be seeing you VERY soon for shenanigans; we can be angry east coasters in SF together without shame. Livia I've been so lucky to have you as a best friend upon arrival as well as a colleague; we've struggled through this waterfall together and now we're rainbows. Lisette Varon Carvajal and the history ladies taught me what it means to be an intersectional feminist. Lis you supported me when I struggled from the pervasive opression of sexism. You have profoundly shaped my social activism.

Now that we've run through a long list. I'd like to thank my family. My siblings Kim, Bess, Chris, Rachel; my parents. And of course, the light of my life, Cezi. I do this all for you.

To the Haters.

# Dedication

I dedicate this to my parents, for never giving up on me.

# Table of Contents

# Chapter 1

## What is the question of questions?

In nature, information is transmitted between two organisms through at least two means. Primarily, genetic evolution transmits genetic code from parents to offspring through reproduction. On the other hand, social learning allows organisms to rapidly (compared to the evolutionary processes that mold genes) acquire and transmit new information from conspecifics.

Humans are uniquely social creatures from birth. One endowment that appears to set us apart from other species of social learners is our productive and systematic linguistic capacity. Language allows us to interact with the world, coordinate with others, exchange truths (and falsehoods), discuss the weather. More than that, language allows us to make the unobservable observable; to voice our internal beliefs and desires.

Questions in particular are essential to social learning because they serve many communicative functions. In virtue of the productive and systematic descriptions that our linguistic capacity affords us, a speaker can seek out information about almost anything, and a hearer can understand whatever the speaker requests and cooperatively respond (or not).

Often more information is transmitted in the speaker's utterance than can be traced back to the compositional linguistic structure–what is often referred to as the *literal meaning*. For a classic case of this, an English speaker at the dinner table might ask, Can you pass me the salt? (Austin 1967). Under normal circumstances, any English speaker would understand this question to be, not an inquiry about the addressee's *ability*, but a *request* for the salt itself. This additional request goes beyond the literal meaning of the question, and can be referred to as the *speaker's meaning* (Grice's *what is*

*implicated*). The speaker intended for the addressee to pass them the salt, not to answer with a yes or a no.

Because linguistic understanding and production appears quite effortless and is often successful, it is perhaps surprising to discover that the notion of a *literal meaning* is not a straightforward one to establish. Literal meaning, it is often thought, is truth-conditional and context-independent. In most cases however, a speaker's utterance is rife with ambiguity, vagueness, and indeterminacy, and requires additional information.

Such is the case with questions. Questions do not *prima facie* have truth conditions: under what conditions in the world is Can you pass me the salt? true or false? Linguists, philosophers, psychologists, and computer scientists have spent decades searching for the correct analysis of the literal meaning of a question.

Linguists, as mathematicians, have answered that questions denote the sets of their answers. Thus, the meaning of Can you pass me the salt? is essentially the set of two truth-evaluable declarative (answers), { I can pass you the salt, I cannot pass you the salt }. This strategy suggested that the literal meaning of a question is thus context-independent, and seemed to offer a straightforward explanation of the truth conditions of sentences with embedded questions.[1] Thus, the literal meaning of Dana knows whether you can pass me the salt can be paraphrased as, Dana knows the answer to the question, 'Can you pass me the salt?'. Unlike these Polar, or Yes-No, questions that have only two possible answers, Wh-questions like Where can I find coffee? or How do I find the train station? have potentially infinite answers.

There are many ways in which questions have missing information, but this dissertation is particularly concerned with the issue of **(NON-)EXHAUSTIVITY**. (Non-) exhaustivity refers to an interpretational variability in questions: the number of answers to a question a hearer must provide to *answer* the question, or that an agent

[1] Of course, the answers here do contain a context-sensitive expression, the first person indexical pronoun I. The referent of the first person indexical shifts with the speaker. At the same time, it does have a context-independent aspect of meaning, which Kaplan (1989) referred to as a 'character': in any context, I will refer to the speaker. Only the 'content', which we can think of as the pronoun's extension, will change with context.

must know in order to be ascribed knowledge-wh.[2] For our purposes, we will use the term **QUESTION** to refer to both a root question as in 4, and its embedded counterpart as in 5.

(4) Where can I find coffee?
(5) Dana knows where I can find coffee.

When a speaker asks a question as in (4), there are many ways that a hearer may answer her. (6) are all responses a speaker might provide when asked a question like (4), and thus constitute answers to the question, in a broad sense. However, in a more specific sense, they do not all ANSWER the question. Semanticists since the mid-20th century have considered ANSWER to be a technical term referring only to those *semantic answers* which essentially fill-in for the missing information (replacing the where in the sentence I can find coffee where to create a complete (and grammatical) sentence). The reason for this is to define a notion of answer that is compatible with a compositional semantic theory (see Chapter 2 for more discussion).

(6) a. I don't know, I'm not from around here.
b. Ask over there.
c. Somewhere.
d. Around the corner.
e. Hidden Grounds.
f. Hidden Grounds, Peet's, Starbucks, Stumptown, Penstock, ...

Given this technical notion ANSWER, (6a-c) do not ANSWER the question, because they cannot be grammatically substituted for where. (6d-f) resolve the question in different ways. (6d) essentially tells the questioner *how* to find coffee, but does not name a coffee shop. In contrast, both (6e) and (6f) do name coffee shows, but differ in how many they name. The first is called a **NON-EXHAUSTIVE**, or MENTION-SOME (MS) answer because the answerer provides at least one answer to the question; the second is called an **EXHAUSTIVE**, or MENTION-ALL (MA) answer, because the answerer provides the exhaustive list of answers to the question.[3] The phenomenon of (NON-)EXHAUSTIVITY

[2]Or even more generally (cf. Lahiri 2002), for whatever relation to the question denoted by a question-embedding verb, the amount of answers that the subject must be in that relation to.

[3]In Chapter 2 we will discuss the different strengths of exhaustive answers. For the moment, we focus

refers to the fact that root questions may be ANSWERED by either kind of answer, and that sentences with embedded questions as in (5) may be true on an exhaustive reading, and/or on a non-exhaustive reading. For simplicity, when I say that a question is exhaustive or has an exhaustive reading, I intend to mean that a root question is (felicitously) answered by an exhaustive answer, or that an embedded question is true on an exhaustive reading. Likewise, when I say that a question is non-exhaustive or has a non-exhaustive reading, I intend that a root question is felicitously answered by a non-exhaustive answer, or that an embedded question is true on a non-exhaustive reading.

(7) Dana knows where I can find coffee.
   a. Dana knows that I can find coffee at Hidden Grounds.
   b. Dana knows that I can find coffee at Hidden Grounds, Peet's, Starbucks, Stumptown, Penstock....

Questions—at least, the acostic signal—underspecify (non-)exhaustivity. As a result, the resolution of (non-)exhaustivity is highly context dependent. The job of a semanticist is to determine whether this underspecification is located at the level of literal meaning (and how it is underlyingly realized), or whether it is located at the level of speaker meaning (and how a hearer derives the speaker meaning from the underlying literal meaning). There are many logical possibilities one can explore to answer these questions, and the inherent context-dependence of the phenomenon obscures a straightforward answer to any of them.

In studying the semantics-pragmatics interface, or even semantics generally, the line between competence and performance becomes blurred. Determining the truth conditions of a sentence (and thus, as linguists using introspection to determine the output of the semantic grammar, the semantic competence) is intertwined with the performance process of interpretation. As such, when we make judgements about the truth-conditions of sentences, we deploy our interpretive mechanisms for the purpose of introspection. This is especially true of judgements involving "pragmatics".

The goal of this dissertation is to understand how (non-)exhaustivity in both root

on the distinction between exhaustive and non-exhaustive.

and embedded questions is resolved with the above points in mind. To do this, we will start with the theories suggested to us by semanticists as guides for the the predicted answerhood-conditions of particular questions. I will argue that received intuitions about baseline context-independent answerhood conditions fall out from establishing, in a given context, appropriate level of information required to resolve a salient discourse goal. Hearers recruit prior expectations about the likely contexts in which a question (or a declarative with an embedded question) is uttered. When a question is evaluated without explicit context, the performance mechanism of understanding kicks in to fill in the missing, necessary, contextual information. As a result, semantic theories of (non-)exhaustivity in questions are adept at capturing the typical goals, the typical amount of information required/intended by a speaker.

Over the course of the next six chapters, I hope to convince the reader of this. I will attempt, over five experiments, to provide evidence of the relationship that I see between linguistic form (signal) and context (or goal), and along the way, reflect on various methodological considerations about the relationship between semantics and pragmatics.

In **Chapter 2**, I survey the major evidence for (non-)exhaustivity in questions and answers, and discuss how theoretical semantics handles these interpretational variations. There are several different issues I discuss here, but the basic observation is that, most semanticists implicitly assume that non-exhaustivity is more limited in its distribution—that all questions can have an exhaustive reading (or allow exhaustive answers), but not all questions allow a non-exhaustive reading (or non-exhaustive answers).

In the spirit of researchers like Asher & Lascarides (1998), Ginzburg (1995), and van Rooij (2003), I argue that that perceived onmnipresence of exhaustivity should not be mistaken for evidence of underlying semantic exhaustivity for the following reasons. As the first two authors pointed out, when you look at the data that different semanticists appeal to in their theories, you see they only use a small set of example questions, and often these are not provided with explicit linguistic contexts. Those who argue for

exhaustive semantics use almost exclusively who-questions, while those who argue for non-exhaustivity include how- and where- (and why-) questions. These different kinds of questions non-trivially affect our baseline expectations about likely answers. I argue that this determines how (non-)exhaustivity resolved in that question. If questions are semantically exhaustive, why would this only be reflected in who-questions?

My main approach to this issue involves two tactical manoeuvres. First, I hold the linguistic form fixed, and consistently manipulating the surrounding context. Second, I deploy the inverse algorithm: hold the context fixed and systematically manipulate linguistic form. I find that these baseline interpretations disappear when context is systematically manipulated, and others emerge which contradict the received generalizations when more data are considered. Some who-questions are baseline non-exhaustive (e.g., Who has a light?), and how-questions can permit exhaustivity with the appropriate context.

In examining these cases, the following generalization emerges: whether a question is exhaustive or non-exhaustive is a matter of whether discourse goals are exhaustive or non-exhaustive. When no explicit goals are provided, hearers (and linguists too), must reach into their expectations and prior beliefs about the most likely goal that a speaker who uttered that question would have. Thus, hearers must impute contexts where none are provided to determine (non-)exhaustivity. This is a necessary step, because questions are semantically underspecified for (non-)exhaustivity. Thus, "default exhaustivity" is epiphenomenal of both how hearers access their prior expectations, and reflects a general "safe-bet" heuristic to choose the maximally informative message when they are uncertain.

In order to account for non-exhaustive answers/readings, every semantic theory needs some notion of context-sensitivity, disambiguation, or precisification, regardless of the underlying semantic representation the theory assigns to a question. Thus, by systematically examining the relationship between context and linguistic form, we can articulate the facts about this relationship.

In **Chapter 3**, I present initial evidence for my view that questions require context

to manifest discourse goals. We find that, while certain linguistic factors may indeed boost the baseline acceptability of a non-exhaustive reading of an embedded question, when discourse goals are explicitly non-exhaustive, putative linguistic factors are neutralized.

Further, participants rate (non-)exhaustivity on the basis of informational sufficiency: in non-exhaustive contexts, participants rate both singleton answer mention-some ("mention-one") and non-singleton answer mention-some equally good (and rated mention-all significantly lower), while in exhaustive contexts, they rate non-singleton answer mention-some and mention-all equally good (and singleton mention-one significantly lower).

Discourse goals specify how much information is required, thereby resolving (non-) exhaustivity. Linguistic form factors should be thought of as defeasible cues to the discourse goals. As such, a hearer would be expected to rely on them more when the linguistic context is underinformative with respect to a discourse goal, but not necessarily when the context is informative (Wu & Keysar 2006).

In **Chapter 4**, I present two answer rating experiments to quantify the likelihood of (non-)exhaustivity given linguistic form (Experiment 3a) and given explicit discourse goals (Experiment 3b). We find that in the first case, hearers do not rate exhaustive and non-exhaustive answers significantly different based on question form factors on the magnitude that we might expect to see, if there were grammatical restrictions on non-exhaustivity. We see small deviances from ceiling high ratings for expected exhaustive form factor combinations (e.g., non-modal know-who questions), but these do not yield near-floor ratings of un-acceptability or un-likelihood.

We do find significant differences between dependent measures in our experimental tasks. This further suggests to me that (non-)exhaustivity resolution is not a context-independent phenomenon as assumed by many formal semanticists. Task-sensitivity can be viewed as yet another form of goal-sensitivity (cf. Roberts 2018;

Degen & Goodman 2014). Whatever the underlying semantics, accounts of (non-)exhaustivity must acknowledge context, as manifested explicitly in a linguistic context or implicitly in the hearer's prior expectations about the likely context associated with a given question token.

As most semantic theories do not provide explicit contexts for determining (non-)exhaustivity in questions, these theories provide insight into the hearer's baseline expectations about the likely goals associated with the question, rather than a characterization of a context-independent question meaning. These baseline expectations will then be directly dependent on the linguistic form of the question, which hints to the hearer likely goals/contexts.

Can we vindicate this probabilistic account by looking at the relationship between speaker and hearer and quantify the informational content of different question types and contexts that articulate (non-)exhaustivity?

In **Chapter 5**, I present two studies which aim to understand the probabilistic relationship between (non-)exhaustivity and linguistic form from the perspective of the speaker with communicative goals, as well as what kind of information relevant to (non-)exhaustivity resolution would be available as input to the language learner. We might expect that a speaker who is concerned with maximal clarity for the sake of her hearer would produce questions that maximally indicate (non-)exhaustivity by the linguistic form of the question. In other words, that speakers with exhaustive goals should produce questions loaded with exhaustive linguistic cues, while speakers with non-exhaustive goals should produce questions with non-exhaustive linguistic cues.

I present a corpus study that quantifies speakers' naturalistic production of surface-level form cues. We find that the frequency and co-occurrence of cues provides conflicting evidence for (non-) exhaustivity. While how-questions, a non-exhaustive cue, are the most frequent question, FINITE (non-modal) clauses are also the most frequent, and the most frequently co-occurring with how-questions. Thus, the linguistic signal alone does appear to be informative enough to determine (non-) exhaustivity. I next present a production study that aims at quantifying the extent to which form cues are

produced given contextual goals. If a contextual goal is exhaustive, then we might expect speakers to produce questions with more exhaustive surface-level cues; and if a contextual goal is non-exhaustive, then speakers would produce questions with more non-exhaustive cues. Instead, we find that participants do not produce a significant amount of cues generally—their questions are ambiguous/underspecified for (non-) exhaustivity. However, when they do produce cues, those cues align nicely with our contextual manipulation: exhaustive cues are produced significantly more in the exhaustive goal contexts, while non-exhaustive cues are produced more in the non-exhaustive goal contexts.

Pragmatic reasoning as articulated by Grice is often assumed to involve recursive mindreading on the part of the speaker and the hearer. In particular, part of the equation in speech production is the speaker reasoning about possible utterances and the inferences that the hearer will draw from the speaker's choice of utterance. At the same time, research on communication and mindreading reveals a much more subtle relationship between speaker and hearer, utterance meaning and context: speakers actually do not avoid ambiguous, vague, or difficult-to-parse utterances (Clark & Wilkes-Gibbs 1989; Clark & Brennan 1996; Kehler & Rohde 2018; Kehler et al. 2008; Arnold et al. 2004; Ferreira & Dell 2000; Ferreira & Hudson 2011; Ferreira & Schotter 2013; Jaeger 2010, 2011); nor do they always deploy their theory of mind to reason about their interlocutors in the process of communicating (cf. Wu & Keysar 2006; Lin, Keysar, & Epley 2010).

In Chapter 6, I return to the hearer with two goals. First, to test the hypothesis that informative contexts neutralize effects of linguistic form, while underinformative context induce the opposite effect. In a card-game experimental scenario (Cremers & Chemla 2017, Phillips & George 2018), we tested three contexts that manifested an exhaustive goal, a non-exhaustive goal, and the third unspecified for a goal. However, as there are no truly "null contexts", we also expected the card game setting to inherently encode exhaustive goals. The reason is that *typically* the goal of a card game is to gain or win as many points, tricks, chips, as possible. Our hypothesis was not exactly borne

out: participants responded purely based on context and never based on the linguistic form of the question. However as expected, they treated the "null context" on par with our exhaustive context. I argue this provides support for the view that exhaustivity derives primarily from hearer expectations about likely goals, and not primarily from the semantics of the question.

The second goal of this study is to locate context-sensitivity in (non-)exhaustivity resolution using independent measures of hearer "literalness" or "pragmatic-ness". In this correlational analysis, participants' responses to sentences that give rise to a scalar implicature with the existential quantifier some (cf. Grice 1967, 1989; Horn 1972, Gazdar 1979) determined whether they were literal or pragmatic hearers (in a replication of Bott & Noveck 2004). Literal hearers do not calculate the scalar implicature, while pragmatic hearers do. We discovered two things. First, that literal hearers rated mention-some conditions near-ceiling, while pragmatic hearers accepted these conditions at most near chance (seemingly to calculate an exhaustivity inference). Second, we found significant effect of our context manipulation in both populations. This finding suggests that context-sensitivity is crucial to fixing the interpretation of questions at the literal level (supporting theories like Beck & Rullmann 1999; George 2011, Ch2; Asher & Lascarides (1998); van Rooij (2003), 2004), as well as at the expected speaker meaning level (cf. Schulz & van Rooij 2006; van Rooij & Schulz 2006; Spector 2007; Zimmermann 2010). More than that, I sugest that these findings bear against analysis of question meaning that encode only (weak or strong) exhaustivity.

Finally, in Chapter 7 I conclude by discussing some open questions about the relationship between experimental and theoretical semantics (and pragmatics). Do the data I've presented here actually support context-sensitivity at the level of literal meaning as I've argued? Are degraded acceptances of mention-some evidence for an underlying semantic representation, or are they merely reflective of participant permissiveness or tolerance? What do differences between task factors (like dependent measures) reveal about semantic competence, if anything? In attempt to answer these questions, I review similar debates from experimental syntax, attempt to draw analogy

where possible, and importantly describe where the analogy fails and why.

# Chapter 2

# Non-exhaustivity and Question Semantics

**(NON-)EXHAUSTIVITY** refers to an interpretational variability in questions: the number of answers to a question a hearer must provide to *answer* the question, or that an agent must know in order to be ascribed knowledge-wh.[1]

(8) Where can I find coffee?

(9) Dana knows where to find coffee.

Consider a world where the coffee shops include three places, Hidden Grounds, Penstock, and Peets. A NON-EXHAUSTIVE, or MENTION-SOME, answer names some (but not all) of the answers to the question. In contrast, an EXHAUSTIVE, or MENTION-ALL, answer names all the answers to the question.

(10) **NON-EXHAUSTIVE/MENTION-SOME**
   a. Hidden Grounds.
   b. Hidden Grounds and Penstock.

(11) **EXHAUSTIVE/MENTION-ALL**
   a. Hidden Grounds, Penstock, and Peets. — WEAK-EXHAUSTIVE
   b. Hidden Grounds, Penstock, and Peets, and perhaps other places. — INTERMEDIATE EXHAUSTIVE
   c. Only Hidden Grounds, Penstock, and Peets. — STRONG EXHAUSTIVE

At this point in the discussion, it would be reasonable to exclaim, "But wait, how can one ever name *all* the answers?" You're right. Clearly there are more answers to the question than those three coffee shops. This becomes particularly obvious when we consider how- and why-questions. However, semantic theories take it for granted that the domain of answers is restricted in some way.

[1] Or even more generally, for whatever relation to the question denoted by question-embedding verb, the number of answers that an agent must be in that relation to.

In this chapter, we will review the main theoretical accounts of question semantics in order to understand what semanticists say about (non-) exhaustivity. Before addressing different semantic theories, we first look at the data they are supposed to capture. In Section 2.1, I will walk through the main empirical observations about (non-)exhaustivity, focusing on linguistic form factors. In Section 2.2, I will present the main theories of question semantics generally, and in Section 2.4 I will present the main theoretical treatment of non-exhaustivity specifically. The reason that these are separate sections is because dominant semantic theories treat non-exhaustive readings as exceptional, more limited in distribution and availability.

## 2.1 Empirical generalizations, empirical observations

In this section of the chapter, I will present the empirical observations that directly relate to non-exhaustivity and mention-some readings and answers. The main finding of this section is that non-exhaustivity is modulated by both the surface-level linguistic form of the question, and by the discourse goals that either explicitly or (are inferred by the hearer to) implicitly drive the context.

### 2.1.1 Generalization 1: Wh-words

Different wh-questions license different levels of (non-) exhaustivity. This observation was first made explicit by Ginzberg (1995), and shortly there after Asher & Lascarides (1998) noted that this manifests when we look at the data used in support of different semantic theories: theories which argue for an exhaustive semantics typically cite who-questions, while those which argue for a non-exhaustive semantics typically use non-who-questions (often how- and why-questions).

(12) a. Who came to the party? / Dana knows who came to the party. MA/#MS
b. Where can I get coffee? / Dana knows where I can get coffee. MA/MS
c. How do you get to Central Park? / Dana knows how you get to Central Park. #MA/MS

Constructed examples presented in support of MA readings usually feature who-questions,

as in (12a), while examples in support of MS readings feature where-questions, as in (12b)/(12c)—or more generally, non-who-questions. Consider, for example, (13), from Asher and Lascarides (1998):

(13) Dana knows how to get to the treasure.

It seems natural to interpret (13) as true just in case Dana knows at least one way to get to the treasure, and unreasonable that she should know all of the possible ways to get there. What matters is that Dana is able to find a way to the treasure.

Asher and Lascarides suggest that while wh-words might differ in how they resolve (non-) exhaustivity by default (i.e., that who-questions are exhaustive), these preferences may be overridden by contexts that make explicit a questioner's goals and mental state. They argue given the variability of interpretations observed, a unified question semantics should provide a weak (monotonic) meaning (a non-exhaustive one). This may then be strengthened via pragmatics (to an exhaustive reading) given the questioner's plans and cognitive state, rather than the other way around (p. 262, see also Chemla & Singh 2014).

Further differences between wh-words may be observed in how the referential domain of a wh-word is fixed. For example, Ginzburg (1995) notes that who- and where-phrases differ in the granularity of their referential domains. Consider the two contexts and in (14) and (15).

(14) Mary has just stepped off a plane in Helsinki.
   a. Flight Attendant: Do you know where you are?
   b. Mary: Helsinki.

(15) Mary has just gotten out of a taxi in front of her hotel.
   a. Taxi Driver: Do you know where you are?
   b. Mary: Helsinki.

(16) Mary knows where she is.

(16) seems true in (14), but false in (15). According to Ginzburg, where is vague with respect to granularity of location, while who typically only refers to individuals. Though he does not relate this directly to MS/MA, he notes that with where questions, the

questioner's contextually-provided goals determine the level of granularity appropriate.

However, we might create comparable scenarios with who questions which also demonstrate granularity effects. Consider the following fictional scenarios.[2]

(17) Luke Skywalker is talking to Han Solo about his dismay concerning the Galactic Empire's attempts to purge the galaxy of the Jedi. A menacing character dressed in all black with a breathing mask is suddenly revealed.
   a. Han Solo: Do you know who that is?
   b. Luke: Darth Vader.

(18) Luke Skywalker is expressing his despair to Obi-wan Kenobi about his lost opportunity to ever have one final moment to see his father. A menacing character dressed in all black with a breathing mask is suddenly revealed.
   a. Obi-Wan: Do you know who that is?
   b. Luke: Darth Vader.

(19) Luke knows who that is.

Just as with (16), we might argue that the truth of (19) depends on whether it is a response to (17) or (18): (19) appears to be true in (17), but seems false in (18). As in (16), the response in (19) seems out of place, like the person who utters the embedded question is missing or not clued in on something. While we admit that this case is slightly different from Ginzburg's, it serves to demonstrate that for both where and who, the context may determine the level of specificity or granularity with which an embedded question is acceptable.

In a similar vein, we can also cite who-questions that seem to be naturally interpreted on an MS reading, as the examples in (20) and (21) show.

(20) Who's got a light? (Groenedijk and Stokhof, 1984; van Rooij, 2003)

(21) I need a ride. Who's going to the party? (Dayal, 2016)

Both questions are headed by who, and yet both permit an MS answer. If one person steps forward and truthfully offers, "Me," the speaker should be satisfied.

Likewise, Asher and Lascarides discuss another example where we naturally have

[2]The aware reader may note that the following stories seem to reflect a *de re/de dicto* ambiguity. The discussion of these kinds of knowledge claims were the particular focus of Boër & Lycan. Indeed, these two examples parallel's their Superman cases, which we mention later.

a non-exhaustive know-who. Imagine that Jill is a gossip columnist, writing on the celebrities who attended Elton John's party. (22) can be true even if Jill doesn't know any cameramen who were at the party.

(22) Jill knows who attended the post-Oscar party at Elton John's house.

Some might argue that (22) reveals an exhaustive answer when the domain is restricted to the set of celebrities (e.g., Who [of the celebrities/relevant party-goers] attended the post-Oscar party?) (see discussion in George, 2011, Section 6.2). However, if this is the case, it is still unclear how domain restriction alone could be the determining factor and why with rampant domain restriction in natural language (e.g., with quantifiers and definite descriptions), MS readings still seem to be blocked in some cases but not others.

Domain restriction of the set picked out by who also does not seem to readily explain other examples we might create, where an MS answer seems felicitous. Consider the following scenario.[3] Imagine that our friend Mark is incredibly cliquish, and typically only invites philosophers to his parties. I am trying to prove that he's biased, while you are defending him. In fact, Mark had a party just last night, so we have the following dialogue in (23).

(23) a. Me: Who came to the party last night? / Who was invited to the party last night?
b. You: Jill, a linguist.

Note that your response in (23b) is both felicitous and non-exhaustive. When I ask either question in (23a), I may intend a restriction to the set of philosophers (i.e., who, of the philosophers), because of my beliefs about Mark. I may even plausibly intend you to give me an exhaustive answer. However, again, not only is your answer felicitous but it's non-exhaustive as well.

Did the hearer misconstrue the speaker's intended restriction? The set of philosophers would be the natural restriction available from the common ground (cf. von Fintel, 1994), and when we consider my expectations about the answers. However,

[3]Thanks to Caley Howland for bringing this scenario to my attention.

if my point is to prove that all party-goers were philosophers, then it is more likely that I do not intend this set as a restriction. To put this more explicitly, I would not ask instead of (23a), Which philosophers came to the party? with the expicit restriction because the answers to this question would not prove my point. Rather, I need to know everyone who went, not just philosophers. Yet, I do *expect* that the answers to the more general question will be only philosophers. My expectation of the what the domain of answers is is different than the domain of reference I may or may not have intended. In this way, there is no misconstrual of the domain of reference, although my expectations about the answers are different than what they turn out to be.

A third candidate explanation might be possible. You may think that a restriction is to the complement of the contextually available set, (i.e., any non-philosophers). There may be independent reasons to prevent this kind of move. Further, the question remains nonetheless of how this kind of restriction is licensed. No obvious semantic mechanism is present to trigger the restriction.

It is possible to create contexts where the domain is explicitly restricted, and an MA answer is felicitous for a how- or a why-question. Imagine the following scenario (p.c. with N. Theiler and F. Roelofsen):

(24) GUARD
An apartment building has jus hired a new night watch guard. The guard is learning the floor plan. There are three fire exits on each floor: one from the front stairwell, one from the back stairwell, and out the windows.

(25) The guard knows how to exit the building in an emergency.

It seems natural for (25) to be MA. Note how constrained the domain is, and how naturally MA is. Intuitively, MS is infelicitous in the context as we have defined it. A night guard who had only MS knowledge in this scenario would not only be negligent of their job, but potentially endanger the lives of those living in the apartment. It further seems implausible that the guard would not know all the ways to exit, given their small number.

The point is not that an MA context cannot be constructed for a how- or why- question. Rather, the fact that we can do so is further evidence of the role of context. In

(24), world knowledge conspires with the domain of reference to determine what is plausible and necessary for a guard to know about their job. A night guard certainly *should* know all the ways to exit the building, that is part of their job. If there is a fire, they will certainly need that knowledge to direct all the tennats out of the building as quickly as possible.

The drive for exhaustive answers/interpretations arises when the domain is restrictied enough for that to be achieved. While how-/why- questions may allow for MA given those restricted circumstances, in the wild they typically do not because the domain is unclear. Following researchers like Zimmermann (2010), Schulz & van Rooij (2006), it seems that this is pragmatic in nature. Grice's Maxim of Quantity emphasizes the drive to be as informative as is requrired (Quantity 1), while not being more informative than is required (Quantity 2). In the case of questions in context where the domain is explicitly or implicitly restricted, it is easy to be as informative as possible with respect to the entire domain. However, in cases where the domain is not clear, it seems that the drive for exhaustivity goes away, or is at least lessened. Since the underlying goals will determine the relevant standards and threshold of informativity, and since this information is not always forth-coming in a particular context, it is a further a safe strategy to approach exhaustivity as much as possible.

For any question, we can construct a context where discourse goals license non-exhaustivity. It is the contribution of these discourse goals that matters and give rise to interesting interactions with the linguistic features of the (embedded) question.

While levels of granularity and exhaustivity may be different sorts of beasts, they share one salient commonality: in order to establish truth conditional content, precisification of the speaker's referential intent is necessary. A hearer must recruit whatever information is available to them in order to resolve the intended level of granularity/specificity. This could include information conveyed in the linguistic form of her utterance as well as any contextual information that may elucidate the speaker's goals. Aloni (2001, 2005) suggests that all these elements are crucially involved in the pragmatics of fixing the reference domain for wh-phrases. It seems that (non-)exhaustivity

is just another manifestation of that process.

Beck & Rullmann 1999 provide evidence for weak exhaustivity in degree how-questions. Depending on the monotonicity of the question predicate, the question will be resolved by either the maximal or minimal degree. In those cases where the minimal degree resolves the question, Beck & Rullmann argue that weak exhaustivity is required. Further, when a degree question occurs with at least or at most, they argue that a mention-some representation must be semantically available to derive the correct intrepretation (Beck & Rullmann 1999, p.285).

(26) a. Wieviele Leute waren mindestens da?
how-many people were at-least there
'How many people were there at least?'

b. Wieviele Leute waren hoöchstens da?
how-many people were at-most there
'How many people were there at most?'

(27) a. Hans weiss, wieviele Leute mindestens da waren.
Hans knows how-many people at-least there were
'Hans knows how many people were there at least.'

b. Hans weiss, wieviele Leute hoöchstens da waren.
Hans knows how-many people at-most there were
'Hans knows how many people were there at most.'

Finally, it has also been observed that the availability (or rather, unavailability) of MS readings can be conditioned by the use of a D-linked wh-phrase (Pesetsky, 1987). While a singular-marked which phrase (e.g., which child) can give rise to both MA and MS readings (although it is unclear whether they have equivalent availability, see Dayal (2016) and Groenendijk and Stokhof (1984)), Comorovski (1996) observes, as does Spector (2007), that plural-marked which-phrases (e.g., which places) block an MS answer (as opposed to monomorphemic wh-phrases). However, Dayal (2016, p.79) presents the following scenario as evidence against this as an absolute restriction:

(28) Suppose a researcher needs a few people with AB blood type to test a new drug. The study requires her to test some but not necessarily all the patients in the hospital. She has a list of patients but not their blood types. The researcher

asks (28a) of the administrator who has the information mapping patients to blood types:

a. Which (of the) patients can we approach for this test?
b. You can approach Bill and Sue.
c. Or you can approach Jim and Tammy.

The answer in (28b) picks out a sub-group of individuals in the domain of which patients. Dayal notes that it is thus both mention-some, and felicitous even though the wh-word is a plural D-linked, contra Comorovski (1996).

Xiang and Cremers (2016) tested the availability of mention-some readings in modal and non-modal questions, with both monomorphemic and d-linked wh-phrases. They found no effect of wh-word, providing evidence against Comorovski's claim and support for Dayal's. They also found a significant effect of modal for both who and plural which N phrases (e.g., Mary remembered which children/who can lead the dance vs. Mary remembered which children/who have an accessory in common). However, aspects of the experimental design may have led to, or at least influenced, this pattern. Notably, the modal predicate can lead the dance was explicitly included in the lead-in and as part of the visual stimuli prior to appearing in the target statement, while the non-modal predicate have an accessory in common was not. As a result, in the non-modal condition, participants may have had to execute additional inferences to calculate both Mary's perspective and what she remembers, given that this information was not explicitly stated. It is possible that the additional task demands involved in this condition incur processing costs, resulting in the observed response patterns. Moreover, the two predicates have an accessory and lead the dance were not fully crossed for presence/absence of a modal, so a tight comparison between the two conditions cannot be made. Given that these design points leave open questions about the source of the results, we consider it empirically unresolved as to whether and to what extent these factors give rise to an MS reading.

### 2.1.2 Generalization 2: Question-Embedding Verbs

Some theoreticians have supposed that the distributional differences between readings of embedded questions arise from similar semantic restrictions. This discussion has typically revolved around the three levels of exhaustivity (i.e., weak, intermediate, strong), where it is generally claimed that know prefers strong exhaustivity, though allows both strong and weak readings (Heim 1994, Zimmermann 2010). Cremers and Chemla (2016) showed experimentally that know-wh gives rise to the range of exhaustive readings. Further, classic examples in support of the non-exhaustive reading in embedded questions are often presented with know. Recall the examples below:

(29) a. Dana knows where to find an Italian newspaper. Groenendijk & Stokhof (1982), (1984)
b. Dana knows how to get to the buried treasure. Hintikka 1976/Asher & Lascarides (1998)

In contrast to know, some verbs, for example predict, are argued to prefer only weak exhaustive readings (Beck and Rullmann, 1999; Heim, 1994; Klinedinst and Rothschild, 2011; Sharvit, 2002).

George (2011, Ch2) argues that questions are ambiguous between a strong MA and an MS reading due to the presence/absence of an EXH operator. Further, they suggest that some verbs select for a structure without the EXH operator. Thus, know selects for it, while non-factives like predict or emotive factives like be surprised do not. Secondly, George argues that data presented in support of weak exhaustivity actually supports the existence of non-exhaustivity. If these two points are right, then we would expect to see asymmetries in the availability of non-exhaustivity depending on which verb is embedding the question. However, given the naturalness of the two sentences above with know, we may question how robust this effect is.

### 2.1.3 Generalization 3: Existentials and Existential Priority Modals

One of the more robust generalizations is that the presence of a modal auxiliary in the question appears to license mention-some readings and answers. The aspects of the

modal relevant for non-exhaustivity are both its existential force and the contextually-determined conversational background, which provides a goal-oriented interpretation. The conversational background is comprised of a modal base, which picks out the set of worlds where the prejacent $\phi$ in can $\phi$ is satisfied, and an ordering over those worlds, determined by the deontic flavor of can (Kratzer, 1981; 1991). In Portner's (2009) classification, these are the existential priority modals.

The modal need not be overt. Infinitival clauses are also natural on mention-some readings (recall our recently discussed (29a) and (29b)), and naturally paraphrased with modals— Dana knows where we can find an Italian newspaper or Dana knows how we can get to the buried treasure. This comparison highlights the natural relationship between infinitival clauses and modality, and supports Bhatt's (1999) proposal that infinitival clauses contain a covert modal. Bhatt enriched the Kratzerian picture to capture this goal-orientedness in infinitival clauses by contextually restricting the modal base to the worlds where not only $\phi$ is true, but where the agent's actions maximize the likely satisfaction of their goals. Bhatt further notes (fn. 12, pg. 140), that non-exhaustivity is linked to the absence of indicative tense. Following this logic, the fact that examples like (29) contain an embedded infinitival, while (30) has neither a covert nor an overt modal, could explain the perceived difference in MS acceptability between the two.

(30) Dana knows who came to the party.

There is no question that there seems to be an asymmetry when sentences are presented out of context. Can we provide a context where MS is felicitous? The answer is Absolutely. Just recall Asher & Lascarides's example about the Oscar party from (22). If Dana is a gossip columnist, she is certainly not interested in the non-celebrity attendees. Even if she names a single celebrity (e.g., Madonna), (30) is true.

We can also find natural non-exhaustivity in the absence of any modal. Consider (31 from Schulz & van Rooij 2006):

(31) Who has a light?

The natural context we imagine when we hear (31) is one where a smoker has asked it.

A smoker needs only one lighter to light up, thus the goal is non-exhaustive. However, if (31) were shouted at a Rush concert while the band begins to play "2112," the effect would be very different. In this context, the speaker's goal would be to get *everyone* to pull out their lighters.

Some hold that there are different varieties of non-exhaustivity (Dayal, p.c.; Xiang, p.c.). Thus, the non-exhaustivity licensed by the presence of a modal is just one kind of non-exhaustivity, while the readings in questions like (31) might be derived differently, perhpas as a free-choice reading.

### 2.1.4 Generalization 4: Quantifying Question Particles, and Beyond

Various other factors can affect whether a question is interpreted exhaustively or non-exhaustively. Karttunen (1977) discusses phrases like for example which marks non-exhaustivity in root questions, but is infelicitous in embedded questions. Karttunen uses this data to argue that questions do not have a semantically existential reading (contra Hintikka 1976).

(32) a. Who, for example, came to the party
b. ∗Dana knows who for example came to the party.

(33) a. How, for example, do I get to the buried treasure.
b. ∗Dana knows how for example to get to the buried treasure.

Note that for example is also infelicitous in (33b), suggesting that the infelicity of for example in embedded contexts is orthogonal to answerhood.

Further, Karttunen argues that (34) would be true if the question could be semantically existential, but he reports that the sentence is a contradiction.

(34) Dana knows who came to the party but she doesn't know that Fox came.

While for example or speaker-oriented phrases like it may not be embeddable, cross-linguistically, embedded questions allow "non-exhaustivity markers" (cf. Beck & Rullmann, Bade ms.).

(35) NON-EXHAUSTIVITY MARKERS
a. Dutch

Jan wil weten wie er zoal (niet) op het feest waren.
Jan wants know who there ZOAL (not) at the party were
'John wants to know who for example were (not) at the party.'

b. German
Hans will wissen, wer so (?nicht) auf dem Fest war.
Hans wants know who SO (not) at the party was
'John wants to know who for example were (not) at the party.'

At the same time, there are also "exhaustivity markers" cross-linguistically. When these occur in a question, it must be interpreted exhaustively (Beck 1996, Reis 1992, Zimmermann 2007, Beck & Rullmann 1999).

(36) EXHAUSTIVITY MARKERS

a. Dutch (Beck & Rullmann 1999)
Hij weet wie er allemaal op het feest waren.
he knows who there all at the party were
'He knows who all were at the party.'

b. German (Beck 1996, Reis 1992, Zimmermann 2007, Beck & Rullmann 1999)
Er weiss, wer alles auf dem Fest war.
he knows who all on the party was
'He knows who all were at the party.'

c. What all did you get for Christmas?

d. Irish English (McCloskey, 1995)/Some dialects of American English
John knows what all you got for Christmas.

Zimmermann (2010) actually argues that so does not mark non-exhaustivity. However, he argues that the presence of such quantifying quetsion particles (QQPs) cross-linguistically is evidence against semantic exhaustivity: if questions are semantically exhaustive, why would a language encode explicit exhaustifying particles?

As previously mentioned, in degree questions, Beck & Rullmann 1999 argue that the presence of at least or at most indicates that the question must have a Hamblin denotation, essentially a semantic mention-some reading. This holds for both root and embedded questions.

(37) a. Wieviele Leute waren mindestens da?
how-many people were at-least there
'How many people were there at least?'

b. Wieviele Leute waren höchstens da?
how-many people were at-most there
'How many people were there at most?'

(38) a. Hans weiss, wieviele Leute mindestens da waren.
Hans knows how-many people at-least there were
'Hans knows how many people were there at least.'

b. Hans weiss, wieviele Leute höchstens da waren.
Hans knows how-many people at-most there were
'Hans knows how many people were there at most.'

Zimmermann (2010) provides two arguments that the existence of QQPs is problematic for semantic theories on which questions are strongly exhaustive. First, if questions were semantically exhaustive, it would be mysterious that cross-linguistically we would find particles that demand strong exhaustive interpretations of questions in the first place. Second, if questions were strongly exhaustive, their meaning would be equivalent to the meaning of the question with an exhaustive QQP. Yet this is not the case.

### 2.1.5 Generalization 4: Discourse Cues to Interpretation

The discussion in this section has been anticipated throughout earlier discussion. We have suggested that a common theme across the cases we have considered is that any perceived baseline for (non-) exhaustivity may be the result (at least in part) of the context highlighting non-exhaustive discourse goals. Groenendijk & Stokhof noted that sensitivity to "human interests" was key for mention-some. In support of this, they note that (39) does not appear to allow a mention-some reading (see also Dayal 2016).

(39) Who is elected depends on who is running.

It is single representation theories which take such context sensitivity seriously, and attempt to provide formal explanations for it. Crucially, the notions often cited are human interests (Groenendijk & Stokhof 1982, 1984), the questioner's mental state (Boër & Lycan 1976, Ginzburg 1995, Asher & Lascarides 1998), plan (Asher & Lascarides

1998), goal (Ginzburg 1995), and decision problem (van Rooij 2003, 2004).

(40) TREASURE (Asher & Lascarides, 1998)
   a. Dana: How do I get to the buried treasure?
      Fox: You go to the secret island.
   b. Dana doesn't know how to get to the secret island.
   c. Dana does and doesn't know how to get to the buried treasure.

(41) HELSINKI (Ginzburg, 1995)
   a. Taxi Driver at the hotel: Do you know where you are?
      Jill: Helsinki.
   b. Jill doesn't know what the cross-streets are, or the neighborhood she's in.
   c. Jill does and doesn't know where she is.

(42) SUPERMAN (Boër & Lycan, 1976)
   a. The editor at the *Daily Planet* pointing to Clark Kent: Do you know who that is?
      The copy-boy: Clark Kent.
   b. The copy-boy doesn't know that Clark Kent is Superman.
   c. The copy-boy does and doesn't know who Clark Kent is.

These cases illustrate that whether a knowledge-wh ascription is accepted, depends on whether the agent in question has epistemic access to the right level/specificity/granularity of reference. In (40), while B's response may constitute an answer to A's question, it does not resolve it unless A already knows how to get to the secret island. In (41) and (42), the fact that the attitude holder in question lacks additional knowledge which prevents us from unequivocally accepting the knowledge report.

We can use these to illustrate the sensitivity to the speaker's goal/plan/decision problem. In (40), if A's goal is to go get the buried treasure and she is missing the relevant information to lead her to the secret island, then we say that A doesn't know how to get to the buried treasure. However, if her goal is to record facts relevant to the buried treasure, we might be more inclined to accept the knowledge ascription, even if she doesn't know where the secret island is. In (16), if Jill's goal is to walk around Helsinki, then perhaps she doesn't know where she is. If her goal is to attend a conference in the hotel, perhaps she does, and so on.

If (non-) exhaustivity is resolved relative to such contextual parameters, we might take it as a puzzle that questions out of context exhibit "preferences" for exhaustivity

or non-exhaustivity. However, our prior expectations and world knowledge govern inferences that we make about a speaker's intentions in many cases (Degen 2015, Degen & Tanenhaus 2015; Goodman & Stuhlmuller 2013; Degen, Tessler, & Goodman 2015; Zondevan, Meroni, & Gualmini 2008; Tonhauser, Bever, & Degen (in press), and many others), and govern the way we process information much more generally. In the case of questions, the speaker's goal may be more or less obvious to the answerer, thus the answerer will have to make a choice about how to answer to best satisfy the questioner's goals. We interpret the speaker of Who has a light? to want a non-exhaustive answer exactly because the people who typically ask that question are smokers, and they only need a single light to light their cigarette. Similarly, we might interpret the speaker of Who came to the party? to want an exhaustive answer because typically people who ask that question have that goal. Expectations about what goals a speaker might have, given the question that they asked can play a crucial role here of guiding the hearer in her own decision problem. Perhaps, the purported default-ness of exhaustivity falls out from a simple conversational heuristic. In the absence of explicit cues specifying how much information the speaker requires, the hearer provides more information so that the speaker may decide amongst the alternatives given. In a sense, then, the move is to put the ball back in the speaker's court. Nonetheless, these expectations may be easily overridden when goals are made explicit. Thus, the perceived "puzzle" of default preferences can be explained by our prior expectations about the connection between questions and questioner goals.

### 2.1.6 Desiderata for a theory of (non-)exhaustivity

We have just discussed many observations about the distribution of (non-)exhaustivity. It appears that the linguistic form of the question imposes baseline restrictions or requirements on the level of (non-)exhaustivity that the question should be resolved to. Yet, throughout the discussion of those linguistic restrictions and particularly in the last section, I pointed out that we can systematically manipulate context and those baseline restrictions evaporate. Thus, the extent to which a question is exhaustive

should fall out from the prior probability that the speaker goals match that level of exhaustivity.

## 2.2 The semantics of questions in truth-conditional semantic theory

A long tradition of semantic research stemming from mathematics and logic analyses the meaning of a proposition expressed by a declarative in terms of its truth-conditions, evaluated with respect to the world. If one asserts (43), we can assess whether the proposition expressed is true or false, given information we can collect in the world regarding the book(s) Dana read.

(43) Dana read *Ancillary Justice*.

A question, however, does not *prima facie* have truth conditions. What would it mean to look into the world to determine that a question is true? In the sections that follow, I review the classic approaches to question meaning against this backdrop. Note that the first theories we will discuss only capture exhaustivity, and our discussion will reflect this, being somewhat historic at first in virtue of introducing the basic semantic formalism.

Hamblin (1973) proposed that the meaning of a question is the set of its answers, and consequently, that knowing a question is equivalent to knowing what counts as an answer to the question.

(44) a. $[\![$ Which book did Dana read? $]\!]$
b. $\lambda p.\exists x \in \mathsf{book}.p = \lambda w.\mathsf{read}_w(x)(\mathsf{Dana})$

Given Hamblin's proposal, the question in (44a) can be thought of as denoting the set described in (44b). Note that this is not the notation that Hamblin himself used. Rather, this is a Hamblin set described in Karttunen notation (see, e.g., Kratzer & Shimoyama (2002) for Hamblin notation). By substituting in for $x$ each entity in the extension of which book, we can generate a set of propositional answers to the question, and determine the truth or falsity of each one. Note that (44b) prevents responses like something or some book from counting as answers, a desirable outcome.

Hamblin provided a formal mechanism to generate this set of answers called Point-wise Function Application (subsequently referred to as Hamblin Function Application). Point-wise Function Application is an operation that combines two sets: each element of one set with each element of the second set. The output is then collected up into a set. In the case of a question like (44), the first set is a set of entities that the wh-word ranges over, and the second set contains the function $\lambda x.\lambda w.\mathsf{read}_w(x)(\mathsf{Dana})$. The resulting set in this case is a set of propositions, the output of saturating the function denoted by the predicate with each entity in the wh-word's domain. These semantic, or congruent, answers are as identical syntactically and semantically to the question as possible, with the missing information filled in. The formal definition is in (45). Point-wise Function Application takes each element from the set of books, $\beta$, and feeds them one-by-one to the function $\lambda x.\lambda w.\mathsf{read}_w(x)(\mathsf{Dana})$, $\alpha$, and returns the set of propositional answers, $\gamma$.

(45) If $\{\alpha, \beta\}$ is in the set of $\gamma$'s daughter nodes, $[\![\alpha]\!]^w \subseteq \mathrm{D}_{\langle\sigma,\tau\rangle}$ and $[\![\beta]\!]^w \subseteq \mathrm{D}_{\langle\sigma\rangle}$ then $[\![\gamma]\!]^w = \{a(b) \mid a \in [\![\alpha]\!]^w \wedge b \in [\![\beta]\!]^w\}$

Hamblin's proposal in turn gives us a way to assign truth values to sentences in which questions are embedded, as in (46), in which a matrix verb takes an interrogative complement (although he himself did not analyze embedded questions). As a result, we can treat the meaning of the root question and the corresponding embedded question the same. Belnap (1982) later referred to this as the Equivalency Thesis.

(46) a. Fox knows $[\![$ which book Dana read $]\!]$
b. Fox knows $\lambda p.\exists x \in \mathsf{book}.p = \lambda w.\mathsf{read}_w(x)(\mathsf{Dana})$

Directly exporting Hamblin's (1973) semantics for root questions to embedded questions as in (46a) predicts that a sentence like (46a) has the meaning expressed in (46b): for Fox to know which book Dana read, he has to know the possible answers. However, many have had noted that this is not the intuitive meaning for (46a), but rather the set of answers must be restricted to the true answers (cf. Karttunen 1977). Thus, another step is required, and semanticists since Hamblin have grappled with identifying what exactly that step is, and whether the question denotes at base the Hamblin

set, or another more restricted set.

Before identifying what those more restricted sets are, it is worth noting for our discussion of non-exhasutivity, that Hintikka (1976) proposed that questions permit both an existential and a universal reading, and corresponding answers. He thought that (46) had multiple truth conditions, as captured below in (48). (a) (the existential reading) requires that Fox knows of at least one book that Dana read it, while (b) (the universal reading) requires that he know of all of the books that she read, that she read them. The existential reading is our non-exhaustive/mention-some reading, while the universal reading is exhaustive/mention-all.

(47) Fox knows what Dana read.

(48) a. $\exists x$ [Dana read $x \wedge$ Fox knows that Dana read $x$]
b. $\forall x$ [Dana read $x \rightarrow$ Fox knows that Dana read $x$]

Karttunen (1977) took issue with the non-exhaustive reading, pointing out that if (a) were a possible reading of (47), then (49) would not be a contradiction. However, in a situation where *Ancillary Justice* is one of the things that Dana read, it is.

(49) #Fox knows which books Dana read, but he doesn't know that she read *Ancillary Justice*.

Karttunen also argued that question meaning should encode only the possible *true* answers, which captures what was later called **weak exhaustivity**. Consider the pair in (50). The verb tell is not veridical when it embeds a propositional complement as in (50a), but appears to become so when it embeds a question as in (50b). Karttunen argued that this reveals a truth requirement imposed on the set of answers. More recent authors have disagreed with Karttunen (for example, Spector & Égre 2015)

(50) a. Fox told Alex that Dana read *Ancillary Justice*.
b. Fox told Alex which books Dana read.

We can describe the Karttunen denotation as in (51).

(51) $\lambda p.\exists x.p = [\mathsf{book}(x) \wedge \lambda w.\mathsf{read}_w(x)(\mathsf{Dana}) \wedge p(w_0)]$

The formula in (51) denotes the set of propositions $p$, such that for some book $x$, the proposition $p$ is true in the actual world $w$, and $p$ is equal to the proposition that Dana

read $x$.

Karttunen treats wh-words as existential quantifiers: which book denotes $\lambda P.\exists x.[\text{book}(x) \wedge P(x)]$. Interestingly, this move finds empirical support in the fact that in some languages: existential quantifiers and wh-words are homophonous, as with nani in Japanese. The logical form in (51) can be expressed graphically, as in Figure 2.1 below. This figure captures the fact that answers may be overlapping, because a question might have more than one true answer in a world.

Karttunen does not need to employ Hamblin Function Application in question composition because of the basic difference in the wh-phrase denotation. For Karttunen, first the declarative base of the question is shifted to a proto-question, $\lambda p.[[p = \lambda w.\text{read}_w(x)(\text{Dana})] \wedge p(w_0)]$, which combines with the existential quantifier what via a wh-quantification rule, and returns the set of true answers. Despite these compositional differences, Hamblin- and Karttunen-style theories are classified together as propositional set approaches or often as alternative semantics, because the meaning of a question is the set of its propositional answers, or the alternatives. We refer the reader to Chapter 2 of Dayal (2016) for more details about the composition.

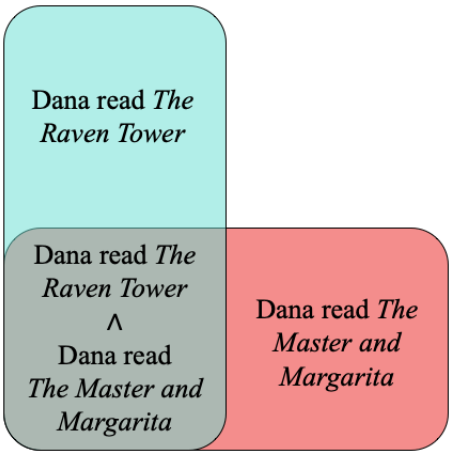


**Figure 2.1:** Graphical representation of Proposition Set Answers.

A Karttunen semantics predicts that a know-wh declarative as in (46) is compatible with Fox's ignorance about the books that Dana *did not* read. In other words, (52a) and (52b) do not entail (52c).

(52) a. Fox knows what Dana read.

b. Dana didn't read *Ulysses*.
c. Fox knows that Dana didn't read *Ulysses*.

In contrast, Groenendijk & Stokhof's (1982, 1984) influential theory analyzes questions as partitions on worlds. A partition semantics predicts that a question meaning licenses a valid inference from (52a) and (52b) to (52c). Partitions deliver mutually exclusive answers: once the true partition is determined, all other partitions are ruled out, thereby yielding *strong exhaustivity*. Thus, to know what Dana read entails knowing all of the books she read, and also knowing what she did not read. Often, this is glossed using only: Fox knows what Dana read ≈ Fox knows that Dana read ONLY *Ancillary Justice*.

Formally, a partition is an equivalence relation (symmetric, reflexive, and transitive) between extensions evaluated with respect to different indices. We can express it for our target sentence as in (53).

(53) a. ⟦ Which books did Dana read? ⟧
b. $\lambda w_i.\lambda w_j.[[\lambda x.\text{book}_{wi}(x) \wedge \text{read}_{wi}(x)(\text{Dana})] = [\lambda x.\text{book}_{wj}(x) \wedge \text{read}_{wj}(x)(\text{Dana})]]$

In partition semantics, a question denotes a function from worlds to a function from worlds to the exhaustive true answer. Intuitively, a partition can be thought of as a filter that chunks possible worlds into mutually exclusive parts, as represented graphically in Figure 2.2. When the partition chunks two worlds together, it treats them as indistinguishable. If Dana read only *Lilith's Brood* in two worlds, a partition semantics will group those worlds in the same partition (here, bottom right). This partition will be different from the one that groups worlds where Dana read *Ancillary Justice* and *Lilith's Brood* (here, top right). Thus, a question's meaning is a single proposition that is the complete true answer, rather than a set of propositions. Groenendijk & Stokhof introduced an operator that delivers a particular cell of the partition, identifying the unique true answer to the question (1984, pp. 299). The operator has a meaning equivalent to only, essentially rendering the answer set mutually exclusive and exhaustive. It forms the basis of the EXH operator which many later theories employ to derive


**Figure 2.2:** Graphical representation of Partitions.

strong exhaustivity (cf. George, 2011), and which many have used in theories of grammatical scalar implicature (e.g., Chierchia, Fox, Spector 2012).

Let us briefly compare Karttunen's and Groenendijk & Stokhof's theories. Consider a world in which Dana read only *Ancillary Justice* and *Lilith's Brood*. Figure 2.1 presented us with a graphical representation of proposition set semantics, while Figure 3 presented us with partitions. Figure 44 does not encode any information about negative answers; all it provides is two overlapping answers: that Dana read *Ancillary Justice* and that Dana read *Lilith's Brood*. One of the key differences between Karttunen's proposition set semantics, which encodes weak exhaustivity, and partition semantics, which encodes strong exhaustivity, is what is encoded in the explicit answers delivered. For Karttunen, the answer Dana read *Lilith's Brood* is neutral with respect to the subject's beliefs about the books that Dana did not read, while partition semantics delivers answers that make determinations about every object in the domain. Thus, is it not possible to know which books Dana read without knowing whether *Ancillary Justice* was read. Since these two foundational semantic proposals, the field has built more advanced proposals to capture both readings, as well as much more complex phenomena that arise with question semantics.

### 2.2.1 Some shortcomings of these classic approaches

These early theories are incredibly influential, however they present some empirical and conceptual issues. First, Karttunen and Groenendijk & Stokhof's theories predict a single interpretation to questions. Since questions seem to allow multiple readings, researchers in the 90's proposed mechanisms to capture both weak and strong exhaustivity. We will introduce these in the next section.

These theories on their own cannot capture non-exhaustivity. Recall that Karttunen outright rejects the reading's existence. In contrast, Groenendijk & Stokhof did acknowledge non-exhaustivity with a now classic example:

(54) a. Where do they sell Italian newspapers in Amsterdam?
b. Who has got a light?
c. Where can I find a pen?

They note that this question has a natural non-exhaustive interpretation, which they called the MENTION-SOME reading. They further observe that (54a) allows both mention-some and mention-all answers, and the matching (54b) likewise allows the two readings. This ambiguity is resolved relative to "human interests": if the questioner is a tourist then the mention-some answer is the most felicitous answer; if the questioner is a seller looking to break into the local newspaper market, then an exhaustive answer is most felicitous. We will discuss the semantics of non-exhaustivity in detail in Section 2.4. It will suffice for the moment to say that neither of these theories provide a semantic account of non-exhaustivity, but appear to endorse a "pragmatic" account without explaining in too much detail what that means.

Even weak exhaustivity is too strong a requirement in general, and the referential domain of the wh-term must necessarily be restricted to some salient subset. Thus, to even derive an exhaustive set, the wh-domain must first be specified. While this may seem reasonable from a logical (perhaps, a-psychological) standpoint, I do not believe that it is a reasonable for a characterization of my semantic competence. Surely, it is true that *sometimes* we can find such a set, but this is exceptional, and often only after an answer has been made (even then the answer will not consistently be exhaustive).

To posit that the meaning of a question is a set of answers, assumes that a speaker who felicitously asks a question in search of information, has at least idea of the answer space, or else does not know the meaning of the question. If I have never been to Paris before, but I want to visit the Centre Pompidou, I might ask How do I get to the Centre Pompidou? without having much of an idea of the different ways to do so. I might know that Paris has the Métro, I might suspect that they have a decent system of buses given that we're in Europe, I certainly know that walking can in general be a viable way to get around if the distances are close, as is surely calling a taxi if the distances are far. But those answers will also depend on (1) my starting point of venture, (2) the starting time of venture. Probably if I'm a tourist on the streets of Paris, I am looking for the Centre right now, and my indexical location is a safely assumed starting point for my interlocutor to give an answer. My ideas about the space of answers will derive from a complex background of experience and world knowledge.

This diffuse body of information possibly will be lingering in the air when I ask my question, but it is not necessarily the case that I can derive a set of answers as predicted by the Karttunen and Groenendijk & Stokhof style theories. It seems that on these theories I cannot know the meaning of my question, which I clearly *do* for that is precisely why I asked it. I can even ascribe knowledge-wh to a Parisian who gives me an answer to the question without my knowing exhaustively the answers. Yet there again it would seem incorrect on these theories to say that I know the meaning of what I am saying.

### 2.2.2 Reconciling weak and strong exhaustivity on classical theories

We have just introduced classical theories of question meaning. These theories capture exhaustivity, but not non-exhaustivity. Karttunen's semantics gives us weak exhaustivity, while Groenendijk & Stokhof's give us strong exhaustivity. Many have proposed type-shifting operators to reconcile the descriptive insights of Hamblin, Karttunen and Groenendijk & Stokhof, and to account for various empirical facts. Early proposals include Heim 1994; Dayal 1994, 1996 for cross-linguistic facts about scope

marking, and presuppositions associated with number marking in questions; Lahiri 1991, 2002 for quantificational variability effects; Beck & Rullmann 1999 for (non-)exhaustivity facts in degree questions.

These authors have argued that in these phenomena, the correct interpretation depends on the availability of a Hamblin set—thus, their operators take as input a Hamblin set ($\lambda p.\exists x.[p = \lambda w.\mathsf{read}_w(x)(\mathsf{Dana})$) and derive from that more restricted answer sets. Dayal's operator (55a) picks out the unique true proposition, and Lahiri's (55b) picks out propositions from the Hamblin set which are also in some set C. C can be determined partly by the lexical semantics of an embedding verb.

(55) a. $[\![\text{ANS}_D\text{Q}]\!] = \iota p[p \in Q \wedge p(w) \wedge \forall p' \in Q[p'(w) \rightarrow p \subseteq p']]$
b. $[\![\text{ANS}_L\text{Q}]\!] = \lambda p.[p \in Q \wedge p \in C]$

Heim's operators are presented in (56). $\text{ANS}_1$ delivers weak exhaustivity, while $\text{ANS}_2$ delivers strong exhaustivity. In contrast to Dayal and Lahiri's operators, Q denotes an intensionalized Karttunen set, $\lambda w.\lambda p.\exists x.[\mathsf{book}(x) \wedge p(w) \wedge p = \lambda w'.\mathsf{read}'_w(x)(\mathsf{Dana}]$.

(56) a. $[\![\text{ANS}_1\text{Q}]\!] = \lambda w. \cap Q(w)$
b. $[\![\text{ANS}_2\text{Q}]\!] = \lambda w_i.\lambda w_j.[\text{ANS}_1(Q)(w_i) = \text{ANS}_1(Q)(w_j)]$

$\text{ANS}_1$ yields the conjunction of the true answers, and $\text{ANS}_2$ yields a partition. Note that $\text{ANS}_2$ is defined in terms of $\text{ANS}_1$; the weak exhaustive meaning is more primitive than the strong exhaustive one. An embedding predicate may select for one operator or the other. Heim suggested that know selects for $\text{ANS}_2$. Beck & Rullmann 1999 modified Heim's two operators, and include a third one which delivers a non-exhaustive meaning to cover their observations about degree questions.

We might say that these theories appeal to a covert ambiguity, because the phonological string associated with a question alone does not distinguish between the multiple abstract semantic representations which correspond to different meanings. Further, it is a lexical ambiguity because the representations differ only in which ANS operator is present, rather than its position. This characterization might be controvertial because ANS operators as type-shifters are not necessarily present at LF. In contrast,

accounts like George (2011, Ch.6), Nicolae 2014, or Xiang (2016) would count as structural ambiguities because they attribute different readings to the structural differences in the underlying representation. (Although, both Nicolae and Xiang use semantic reconstruction to derive the right scope effects, rather than actual LF movement as in George's case.)

George (2011, Chapter 2) derives the two readings not from two different ANS operators, but via the presence or absence of an exhaustivity operator, X, as shown by the two LFs in (57). The Q operator existentially quantifies over the question abstract to derive a Hamblin set (and non-exhaustive readings), while the X operator returns an exhaustified set of propositions.

(57) a. $[\![$ Q [ what Dana read ] $]\!]$ NON-EXHAUSTIVE
$\lambda p_{\langle s,t\rangle}.\exists\beta_e.[p = \lambda w.\text{read}_w(\beta)(\text{Dana})]$

b. $[\![$ Q [ X [ what Dana read ] ] $]\!]$ STRONG EXHAUSTIVE
$\lambda p_{\langle s,t\rangle}.\exists\beta_{\langle e,t\rangle}.[p = \lambda w.[\beta = \lambda x.\text{read}_w(\beta)(\text{Dana})]]$

Similar to Heim, George posits that different embedding verbs may select for the X operator. Unlike Hamblin/Karttunen and other proposition set theories, George treats wh-words as lambda abstractors rather than as existential indefinites. George and Beck & Rullmann represent two perspectives on how to capture the multiple readings associated with questions.

Strong and weak exhaustivity are not the only possibilities. Another is intermediate exhaustivity (Spector 2005, Klinedinst & Rothschild 2011). A weak exhaustive semantics would predict that (47) is true in a situation where Fox knows the true answers, but is either ignorant about the books Dana didn't read, or falsely believes that Dana read a book that she actually did not read. This reading is sometimes described as being *false-answer sensitive*, referring to a more general effect whereby judgements of know-wh reports are rejected when the attitude holder has false beliefs about the false answers.

Klinedinst & Rothschild (2011) propose that an exhaustivity operator may be applied in two different places in the LF of a declarative with an embedded question, as shown by the LFs in (58), where $\alpha$ stands for a question-embedding verb. Questions

have at base a Karttunen-style weak exhaustive semantics.

(58) a. $[_{CP}\ [_{TP}\ s\ [_{VP}\ \alpha\ [_{CP}$ EXH [ what Dana read ]]]]] EMBEDDED
b. $[_{CP}$ EXH $[_{TP}\ s\ [_{VP}\ \alpha\ [_{CP}$ what Dana read ]]]] MATRIX
c. $[_{CP}\ [_{TP}\ s\ [_{VP}\ \alpha\ [_{CP}$ what Dana read ]]]] NONE

The intermediate exhaustive reading is derived via matrix exhaustification, and the strong exhaustive reading via embedded exhaustification. Finally, the weak exhaustive reading is derived when there is no exhaustivity operator present in the LF. This proposal thus has the benefit of capturing Heim's insight that weak exhaustivity is primitive, and strong exhaustivity is derived.

### 2.3 (Non-)Exhaustivity and Embedding Predicates

Given a question-declarative pair such as the one in (59), a key question is, to what extent the answers permitted in the embedded question are linked to or constrained by the matrix verb?

(59) a. What did Dana read?
b. Fox knows what Dana read.

It is well known that verbs have both syntactic subcategorization restrictions and semantic selectional restrictions on their arguments (Grimshaw, 1979). For example, a verb like know can embed either an interrogative or a declarative proposition (60), a verb like wonder can only embed a question (61), and a verb like think can only embed a declarative (62).

(60) a. Fox knows where Dana bought coffee.
b. Fox knows that Dana bought coffee.
(61) a. Fox wondered where Dana bought coffee.
b. $*$ Fox wondered that Dana bought coffee.
(62) a. $*$ Fox thinks where Dana bought coffee.
b. Fox thinks that Dana bought coffee.

Many researchers have attempted to provide a unified explanation of embedding predicates (Karttunen (1977), Groenendijk & Stokhof (1982), Ginzburg (1995), Lahiri (2002),

Egré (2008), Theiler (2014), Romero (2015), Spector & Egré (2015), Uegaki (2015), Theiler, Roelofson, & Aloni (2019), Mayr (2019), Uegaki & Sudo (2019), a.o.). Some researchers connect selectional restrictions to semantic properties of the embedding verb (e.g., factivity or veridicality). Responsive predicates in particular are trouble-some because they allow both declarative and interrogative complements, and their semantic properties do not always appear consistent across complements. For example, tell is non-veridical when it embeds a proposition, but appears to be when it embeds an interrogative. While this point originated with Karttunen (1977), many recent scholars have questioned it (see Spector & Egré (2015)).

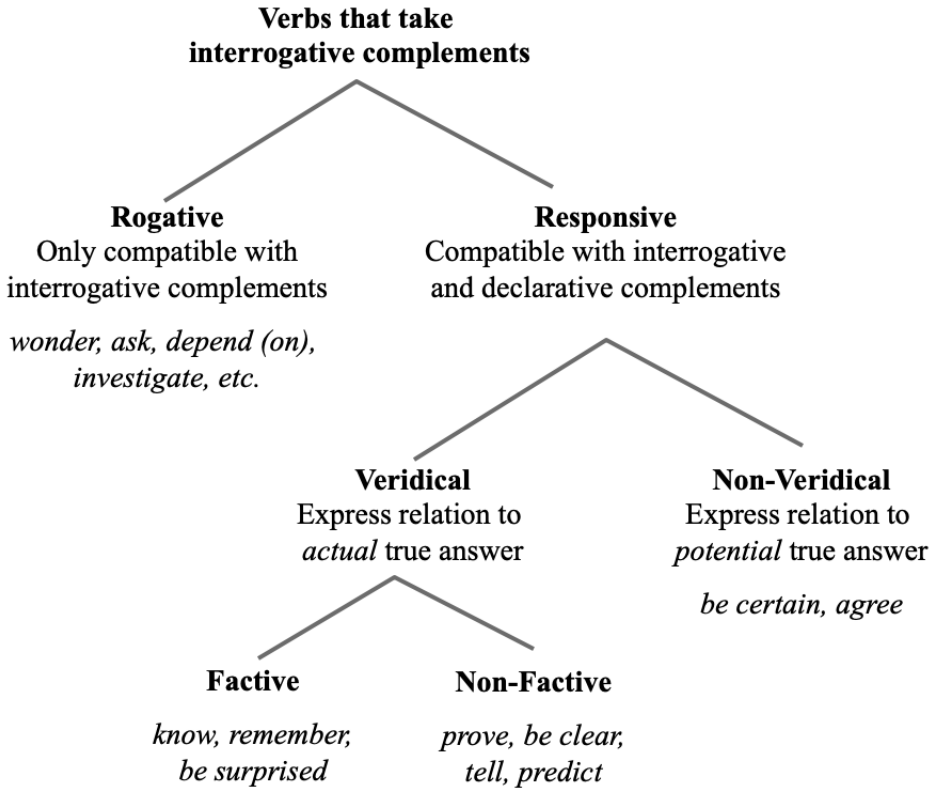


**Figure 2.3:** Lahiri's typology of interrogative-embedding verbs.

One question that arises is how to articulate the lexical entry for these verbs, whether to proliferate entries for each syntactic frame a verb takes. Many treat the declarative-embedding use as basic, and attempt to reduce the interrogative-embedding uses to this basic one. This can be achieved in a number of ways: by separate lexical entries

for each complement (cf. Spector & Egré, 2015; Karttunen, 1977), or by positing operator(s) that type-shift interrogatives to declaratives (Heim, 1994; Dayal, 1996; Beck & Rullmann, 1999; Lahiri, 2002). Still others take a different approach. Uegaki (2015) argues for a reduction in the other direction. George 2011 derives both uses from a common lexical entry. Inquisitive Semantic accounts like Theiler et al. (2018, 2019) attempt to avoid the problem all together, because declaratives and interrogatives have the same semantic type. See Theiler et al. (2018), and Uegaki (2019) for thorough reviews of this issue. However, this solution encounters trouble with verbs like wonder, which do not embed declaratives (and should, if the two clauses have the same semantic type).

Many theoreticians have in turn suggested that the distributional differences in exhaustivity of embedded questions arise from semantic selection restrictions (see George, 2011; Guerzoni & Sharvit, 2007; Heim, 1994; Klinedinst & Rothschild, 2011; Lahiri, 2002; Spector & Egré, 2015; Uegaki 2015; Theiler 2014).

It is commonly claimed that know selects for strong exhaustivity (Groenendijk & Stokhof (1982), (1984); Berman 1991, Heim 1994, George 2011, Schulz & Roeper 2011). However, there are clearly cases where strong exhaustivity—and even weak exhaustivity—is not required for know, as in (63).

(63) a. Dana knows where Fox can get a cup of coffee.
b. Dana knows how to get to Quantico.

Indeed, Hintikka (1976) and Asher & Lascarides (1998) have argued for a non-exhaustive semantics for questions on the basis of know-how-questions, and Beck & Rullmann on the basis of degree questions (know-how much/many).

**Emotive factives** such as be surprised or be happy, are often presented in support of a Karttunen-style semantics. They seem to robustly allow weak exhaustivity, and perhaps even disallow strong exhaustivity (in contrast to know) (Berman 1991; Beck & Rullmann, 1999; Heim, 1994; Klinedinst & Rothschild, 2011; Sharvit, 2002). Lahiri (2002) and Uegaki (2015) (amongst others) have argued that emotive factives do not allow strong exhaustivity because of their monotonicity properties. Monotonicity is

defined for a relation between two sets that either preserves or reverses an ordering. Here, the relevant ordering is entailment. Consider the simple entailments between the sentences in (64). The truth of (64a) entails the truth of (64b). However, the truth of (64a) does not entail the truth of (64c).

(64) a. Cezi is a gray cat.
b. Cezi is a cat.
c. Cezi is a gray tabby cat.

An upward monotone function preserves truth from a subset to a superset (e.g., from gray cat to cat). A downward monotone function preserves truth from superset to subset (e.g., from gray cat to gray tabby cat).[4] Consider what happens when sentences like those in (64) are embedded, as in (65) and (66).

(65) a. It surprised Dana that Cezi is a gray cat.
b. It surprised Dana that Cezi is an cat.
c. It surprised Dana that Cezi is a gray tabby cat.

(66) a. Dana knows that Cezi is a gray cat.
b. Dana knows that Cezi is an cat.
c. Dana knows that Cezi is a gray tabby cat.

Despite the fact that (64a) entails (64b), this entailment is not preserved when these sentences are embedded under surprise, as in (65), but it is when they are when embedded under know. This is because know is upward monotonic on its complement, while surprise is non-monotonic—it is neither upward or downward monotonic on its complement. However, recent literature has suggested that strong exhaustivity may indeed be available with emotive predicates. (Klinedinst & Rothschild 2011, Theiler 2014, Cremers & Chemla 2017; Uegaki & Sudo 2019).

Negative Polarity Items (NPIs) have been argued to be licensed in downward-monotone/entailing environments (Ladusaw, 1979). Thus, their acceptability might present a diagnostic for exhaustivity. Observe the contrast in (67): surprise licenses NPIs with a declarative complement but not with an interrogative complement ((67a)

[4]The relevant notion for NPI licensing is actually Strawson entailment (von Fintel 1999) not classical entailment as presented in (65)-(67). $p$ entails $q$ if and only if every context where $p$ is true, $q$ is also true. But $p$ Strawson entails $q$ iff $p$ classically entails q and all the presuppositions of $p$ and $q$ are met.

v. (67b)). This suggests that the problem is tied to the wh-clause. However, the contrast between (67a) and (67c) suggests that (67a) is not ungrammatical because of the embedded question per se, but rather because of the interaction of (be) surprise(d by) with the embedded question.

(67) a. ∗ Dana is surprised by who has ever been to Paris.
b. Dana is surprised that Fox has ever been to Paris.
c. Dana knows who has ever been to Paris.[5]

Given this pattern, Guerzoni & Sharvit (2007) argue that emotive factives with embedded questions do not license NPIs, because they are weakly exhaustive. Only a strongly exhaustive operator can create a downward monotonic environment that licenses NPIs. Other explanations are given by Nicolae (2013) and Mayr (2013). For Nicolae, the exhaustivity operator that creates a downward entailing environment is optional, therefore explaining why (67a) is ungrammatical with NPIs. See also Schwarz (2017) for arguments against accounts which posit a covert exhaustivity operator.

Klinedinst & Rothschild (2011) argue that **non-factive verbs** (in particular, tell and predict) provide evidence for the intermediate exhaustive reading.[6] Consider (68) in a situation where Frank and Emilio are the only people who sang.

(68) John predicted/told me who sang.

If John predicts/tells me that Frank and Emilio sang, but has no opinions about anyone else, then Klinedinst & Rothschild report (following Spector 2005, 2006) that (68) seems intuitively true. Thus, predict and tell do not appear to require strong exhaustivity. However, if John predicts/tells me that Frank, Emilio, and Ted sang, now (68)

[5] A reviewer suggested that, to the extent that the presupposition that Dana finds out that Cezi is a tabby when she finds out that Cezi is a cat is satisfied, (38a) Strawson entails (38c). If that's right, then these data show that surprise is (Strawson) downward monotonic. To our knowledge, it is not commonly assumed that surprise is associated with such a presupposition. Therefore, we will assume that these data show that surprise is non-monotonic. However, Cremers & Chemla (2017) claim that their experimental study of emotive factives shows that surprise (and forget) did pattern more with downward monotone predicates, so perhaps the issue is not so clear.

[6] Spector & Egré (2015) call these communication verbs, because they sometimes allow for factive readings.

is reported to be false.

Some verbs do not distinguish between true and false answers. For example, the non-factive verbs agree and be certain are argued to permit false answers (Berman 1991; Lahiri 1991, 2002; Beck & Rullmann 1999; Spector 2005; George 2011; Spector & Egre 2015; Theiler et al. 2018), as shown in (69). In (a), Dana and Fox could have the same beliefs about who was elected, but they need not be accurate. The same can be said for (b): Dana could be certain about who attended the party without being correct.

(69) a. Dana and Fox agree on who was elected.
b. Dana is certain (about) who was at the party.

Some have provided analyses of be certain to account for these non-veridicality facts, while maintaining a strongly exhaustive semantics (see Uegaki, 2015; Theiler et al., 2018).

### 2.3.1 Experimental evidence for multiple readings of exhaustivity in embedded questions

Given the various claims about the readings licensed by different embedding verbs, and disagreements about exhaustivity, researchers in recent years have turned to experimental methods in an attempt to understand which readings are available. By recruiting these methods to achieve more robust data, these researchers hope to clarify the theoretical landscape to determine a proper treatment of question semantics.

White & Rawlins (2016, 2018) have elucidated our understanding of attitude verb selectional restrictions by conducting large-scale acceptability judgements on over 1000 English clausal embedding verbs in dozens of different syntactic frames (the "MegaAttitude dataset"). Their computational model of selection encodes systematic mappings from semantic type to syntactic distribution. They trained this model on the acceptability data, and found that it derived selectional patterns consistent with many of the theoretical claims in the literature discussed above. White & Rawlins (2018) tested hypotheses about the relationship between the ability of a verb to embed a question

(responsivity) and veridicality/factivity, and found that neither veridicality nor factivity were predictive of responsivity across the entire set of verbs. However, a correlation emerges when verb frequency is factored in: more frequent verbs show correlations between veridicality and factivity, while less frequent verbs do not. White & Rawlins note that this pattern diverges from a well-known result in the morphological literature, where low-frequency forms exhibit strong correlations with rule-based generalizations.

In an acceptability judgement task, Cremers & Chemla (2016) asked whether sentences such as (70) allow for weak, intermediate, or strong exhaustive readings, in contexts where different readings were made true or false. Their results confirm that know gives rise to strong exhaustive readings, as well as both intermediate and weak exhaustive readings. The verb predict gives rise to all three readings.

(70) John {knew/predicted} which squares were blue

Sensitivity to false answers has been a focus of recent investigations from the semantic perspective (Spector 2005, van Rooij & Schulz 2004, Klinedinst & Rothschild 2011, Theiler et al. 2016, 2018). Phillips & George (2018) examined the effect of false answers on judgements of know reports, and found that participants judge these reports to be more acceptable when the proportion of false to true beliefs that the agent holds is lower, and less acceptable when the proportion is higher. Thus, the phenomenon may not be categorical, but rather gradient.

Cremers & Chemla (2017) tested a range of embedding verbs to examine grammaticality with different complements and the range of exhaustivity permitted. Their main focus was on emotive factives such as be surprised. First, these verbs do not seem to license strong exhaustive inferences (repeated below): recall that the inference from (a) to (b) does not hold in (71) the way it seems to in (72) (Groenendijk & Stokhof (1982), (1984); Berman 1991, Heim 1994).

(71) a. It surprised Dana who came to the party.
b. $\not\Rightarrow$It surprised Dana who didn't come to the party.

(72) a. Dana knows who came to the party.

b. ⇒Dana knows who didn't come to the party.

It has been also observed that whether complements are only possible with embedding verbs that are strongly exhaustive (Nicolae 2013, 2015; Guerzoni & Sharvit 2014). Indeed, emotive factives are generally ungrammatical with whether complements, as shown in (73):

(73) a. Dana knows whether Fox came to the party.
b. ∗It surprised Dana whether Fox came to the party.

Finally, emotive factives appear to be non-monotonic and do not license NPIs (recall discussion of (64)-(67)).

There have been many different explanations proposed as to why emotive factives exhibit these patterns, which attempt to link these two to the verbs' entailment properties (Guerzoni & Sharvit 2007/2014, Guerzoni 2007, Sæbø2007, Abels 2007, Herbschrift 2014, Nicolae 2015, Roelofson et al. 2018). Cremers & Chemla compared monotonicity, the availability of strong exhaustive readings, and the acceptability of whether-clauses, to determine whether these properties were linked. Across all verbs, the selectional properties were consistent with those reported in the literature. However, emotive factives were only found to be degraded with whether-questions, rather than completely ungrammatical. As for monotonicity, generally verbs patterned as predicted by the literature, with the exception of the emotive factives (be happy and surprise). Though these were claimed to be non-monotonic, be happy patterned with upward entailing verbs, while be surprised patterned with downward-entailing verbs. Finally, they found that be surprised licensed strong exhaustive readings, contrary to the predictions from the literature.

Chemla & George (2017) tested agree reports as in (74a)-(74b) in a variety of situations where two agents' beliefs about the colors of letters were aligned completely or partially.

(74) a. John and Mary agree {on/about} which letters are blue.
b. John and Mary don't agree {on/about} which letters are blue.

Participants judged (74a) true and (74b) false when John and Mary's beliefs about

the blue letters matched, regardless of whether they were ignorant of or had false beliefs about the other letters. These results suggest that agree licenses intermediate exhaustivity. Thus, the value of experimental work is to confirm and enrich the theory, and to reveal variability among categories that is not easily captured within existing theoretical proposals.

## 2.4 The semantics of non-exhaustivity

Up until this point, we have focused almost exclusively on exhaustive readings, noting only in passing the availability of non-exhaustive readings. There has been considerable debate about what exactly licenses non-exhaustive answers/interpretations of embedded questions, and whether there is a semantic of pragmatic mechanism which explains it. Take the question in (75) as a starting point.

(75) Where can I find an Italian newspaper.

The most natural answer (it seems) to this question is one that provides a non-exhaustive (or mention-some) answer (Hintikka 1976, Groenendijk & Stokhof (1982), (1984)). However, this does not mean that an exhaustive answer is ungrammatical; it is simply not felicitous or optimal in most discourse contexts. It seems that changing the context or the goals of the speaker posing the question influences which answers are preferred, and this implicates pragmatics.

Many questions naturally admit non-exhaustive answers/interpretations. Some examples from the literature include the ones in (76). Note that non-exhaustivity is felicitous across both root and embedded contexts.

(76)
a. How do I get to the buried treasure?
b. Who's got a light?
c. Who will take me to the party?
d. How many people were there at least/at most?
e. Who are some of the people that came to the party?
f. What is a common Russian name?
g. What is an example of a mythical creature?
h. Fox mostly knows what Dana read.
i. Dana knows where to buy gas for the car.

Often the theoretical treatment of non-exhaustivity rests on the decision whether the non-exhaustive answer/reading is complete or partial (recall the discussion from Section 2.2.1). If complete, then a semantic explanation is desired; if partial, then "pragmatics" is often appealed to as explanation. Considering, then, the questions in (75) and (76), something about them seems to make non-exhaustivity quite natural, while the ones in (77) seem infelicitous when paired with non-exhaustive answers/readings.

(77) a. Who came to the party? / Dana knows who came to the party.
b. Which gas stations are open? / Dana knows which gas stations are open.

Our puzzle is this asymmetry, *Why do some questions appear to license non-exhaustivity, while others do not?* In this section, we will survey both the theoretical claims about non-exhaustivity, and critically analyze the extent to which they can address the observations discussed in the first half of this chapter. Over the next few chapters, I will present empirical support for the view that (non-) exhaustivity is not semantically restricted to certain questions. Rather, that it is a feature of interpretation that is calculated by a hearer, in conjunction with inference about the speaker (Questioner's) goals, the context, and the hearer's prior expectations about what goals naturally pair with what sentences/questions. This view emphasizes the role of context, and is compatible with an underlying semantic ambiguity, or underlying underspecified semantics. Before giving away the game too much, let us now see what has been said about the non-exhaustive reading.

I will use 'question' henceforth to refer to the phonological string containing an interrogative clause—this is the input to the interpretational process. Supposedly in interpretation, syntax and semantics will map that phonological string to certain syntactic and semantic representations, respectively. When it comes to (non-) exhaustivity, our inquiry concerns the inventory and status of the underlying semantic representations. Assuming that, the following question will guide our partition of theories: Is the phonological string associated with an interrogative clause (out of context) associated with multiple possible underlying semantic representations?

The answers to that question are either yes, in which case a theory posits grammatical ambiguity; or no, in which case a theory has different choices for what the single representation is. I partition the theories in this manner because we will see that, while the theories disagree whether there is any semantic cause of non-exhaustivity, they all agree that, indeed require, some sort of necessary "pragmatic licensing". Thus, to truly adjudicate between theories we require an understanding of those necessary pragmatic mechanisms, in so far as they can be differentiated based on the needs of particular theories and the posited underlying semantic representations. Note that I have also left out many theories. So far, I have presented only theories which explicitly treat with weak and strong exhaustivity. For example, Heim (1994) would be considered a multiple representation theory, but only with regards to weak and strong exhaustivity. We might assume then that non-exhaustivity is derived contextually on her theory, but she does not give an explicit analysis.

In the next two sections, I present the different manifestations of these two types of theories in more detail, beginning with single representation theories.

### 2.4.1 Single representation theories

One line of approaches (Groenendijk & Stokhof (1982), (1984); Asher & Lascarides (1998); van Rooij (2003)) argues that there is a single underlying semantic representation, corresponding to the phonological string associated with a question. What exactly this single representation is or encodes differs between theories.

Exhaustive theories (Groenendijk & Stokhof (1982), (1984); Karttunen 1977; Heim 1994[7], George (2011, Ch. 6)) posit an underlying exhaustive semantic representation of questions. For Groenendijk & Stokhof, the single representation is a partition, for Karttunen it is the weak exhaustive answer set; Heim (1994) derives both of these but not a non-exhaustive representation. How to derive non-exhaustivity semantically from weak/strong exhaustivity?

[7]Heim argues for two answerhood operators, one to capture weak and the other strong exhaustivity. I include here because there is no semantic non-exhaustivity for her. I will however discuss her operators later in the section on ambiguity theories.

G&S argue that mention-some is pragmatic because it is licensed when the speaker goals permit non-exhaustivity. While they discuss possible semantic and pragmatic accounts of mention-some, they argue in favor of "pragmatic multi-interpretability," (1984, p. 459). What this means is that, an exhaustive answer/reading is complete answerhood, while the mention-some answer is partial answerhood (1984, p. 530). What kind of pragmatic mechanism delivers partial answerhood? G&S say, "What kind of answer is called for depends on the context in which the interrogative is used," (*ibid*). However, G&S also argue that the data of embedded mention-some is evidence for a semantic account (pp. 533). The reason for this is a methodological assumption about the informational encapsulation of semantic processes from pragmatic ones.

Partiality is not Gricean, because no Gricean maxim would achieve the weakening effect required to go from strong to non-exhaustive. Indeed, Asher & Lascarides (1998) articulate this point. Pragmatic mechanisms typically make stronger but defeasible inferences available, while semantic inferences are supposedly weaker but not-defeasible[8]. Further, if we maintain the methodological assumption that Groenendijk & Stokhof suggest, these theories cannot account for embedded mention-some.

A plausible weakening mechanism might include domain restriction: if the domain of answers is restricted to a small enough set, then the question might appear non-exhaustive, but really exhaust that small subset. Already, to make any exhaustive theory plausible—more generally, to fix the domain of the wh-phrase in the way required by the theory—an adequate account of how the domain of answers is restricted is necessary because the set of true answers alone could be infinite. Similar to cases where a universal quantifier is not intended to literally qauntify over a whole set, but is rather implicitly restricted to some salient/relevant subset.

The reason I emphasize this point is because the manner in which the domain is restricted and determined is not merely pragmatic, but importantly so because it is guided by the hearer's top-down expectations about the context, and importantly expectations about the goals of the discourse. Discourse goals determine crucial aspects

[8]Note there are non-monotonic semantics for modals, cf. Gilles' semantics for modals.

of reference, including referential specificity or granularity as well as how much information is neccessary and sufficient (the issue of (non-)exhaustivity).

Mention-some answers seem to allow multiple compatible answers. One question is whether a domain restriction account intuitively captures the optionality that mention-some answers are felt to encode. Domain restriction is not a pragmatic mechanism, but a semantic one. On standard accounts (cf. von Fintel 1994), the restricted subset must be linguistically available in the context. As George (2011, Ch6) shows, mention-some answers are available in a range of contexts where there is no antecedent sub-domain, nor where one can be plausibly retrieved. This point we also argued with an example like (23).

Other single representation theories posit an underlying existential representation that is pragmatically strengthened (Schulz & van Rooij 2006; Spector 2006, 2007; Zimmermann 2010).

For Ginzburg (1995), a question is truly semantically underspecified for exhaustivity. He models a discourse-level 'resolvedness' relation between a question, a piece of information, and contextual variables like the speaker's goal and mental state. His theory does not provide a compositional semantics which takes these facts into account. As a result, several have critiqued this account for its inability to derive a compositional explanation of embedded non-exhaustivity (cf. Asher & Lascarides (1998)).

Asher & Lascarides (1998) use Segmented Discourse Representation Structures (SDRT) (Asher 1993; Asher & Lascarides 1994), to dynamically model the interaction between discourse, sentence meaning, and compositional semantics. Questions denote sets of answers, including non-exhaustive ones. Their theory is similar to Ginzburg's, in that the main work of determining exhaustivity is a rhetorical relation between a question and a (potentially-resolving) proposition, which holds if the proposition is an element of the set denoted by the question. The relation connects these two utterances in the discourse, in combination with information from the questioner's cognitive state and their plan. Their SDRT formalism allows them to formally implement this relation compositionally, so that it derives the intended effects in embedded contexts—SDRT

representations can be arguments to propositional attitude verbs (Asher 1993). They thus capture Ginzburg's insights in a formal framework which allows multiple sources of information—not purely linguistic information—to interact with truth-conditional semantic composition.

Van Rooij (2003, 2004) essentially includes a covert operator, sensitive to the speaker's decision problem. Decision problems are a formal notion that originate from Bayesian decision theory (Savage 1954). According to van Rooij, the decision problem determines which answers are most useful in a given context and the operator ranks answers according to how relevant and useful they are to solving the problem. However, the utility of a mention-some reading can never be higher than the utility of the mention-all reading (van Rooij (2004), p10). Van Rooij explicitly states that we would expect questions to receive mention-all interpretations generally, or by default, in virtue of this fact that the utility of that answer is higher than the utility of a mention-some answer. There are some cases where the utility of a mention-some answer is equal to that of a mention-all answer, as dictated by the If this is right, hearers should never prefer a non-exhaustive reading over an exhaustive one. At most, both should be equally available.

### 2.4.2 Ambiguity theories

Ambiguity theories propose multiple underlying representations for the same phonological string associated with a question (Beck & Rullmann 1999, Lahiri 2002, George 2011, Nicolae 2014, Fox 2014/2018, Xiang (2016), Inquisitive Semantics). Specific theories achieve this ambiguity in different ways.

#### Answerhood operators

Generally, answerhood operators act as filters on a set of answers. Depending on the semantics of the operator, it may return different answer sets. These type-shifting operators typically take as input a Hamblin set, which is argued to be the base denotation of a question.

While Heim's theory does not explicitly derive non-exhaustive readings (her two ANS operators only account for weak and strong exhaustivity), Beck and Rullmann's (1999) semantics includes two ANS operators for weak and strong exhaustivity, plus a third operator for non-exhaustivity. Below, I present the meaning of the MS operator, which takes a world and an intensionalized Hamblin-set as arguments.

(78) $\text{ANS3}_{\text{BR}}(w)(Q) = \lambda P.[\exists p.[P(w)(p) \wedge Q(w)(p) \wedge p(w)]]$

This operator shifts a wh-clause to a generalized quantifier over propositions.[9] It returns the set of all sets of propositions that contain at least one true proposition from the Hamblin/Karttunen set, true at $w$. In the embedded case, the entire wh-clause must QR to avoid a type-mismatch in object position. The LF is presented in (79).

(79) $[_{\text{VP}}\ [_{\text{CP}}\ \text{ANS3}_{\text{BR}}\ [_{\text{CP}i}\ \text{what Dana read}\ ]]\ [_{\text{IP}}\ \text{Fox knows}\ [_{\text{CP}i}\ t_i\ ]]]$

Lahiri (2002) similarly employs two ANS operators to derive weak and strong exhaustivity. He generalizes Beck & Rullmann's operators to include a contextual variable which relativizes the (maximal) informativeness of an answer to context. Essentially, replacing the $w$ variables with $C$ variables, representing contexts.[10] I present the ANS operator in (80a).

(80) $\text{ANS}_{\text{L}}(C)(Q) = \cap\lambda p.[Q(p) \wedge C(p)]$

He also allows embedded questions to be interpreted either *in situ* or raised. In raised position, he derives quantificational variability effects, which we will not talk about here. For the non-exhaustive reading, an embedded question is type-shifted using the ANS1 operator, and then must be raised to either IP or VP of the matrix clause. There, it is bound by a covert quantifier, whose semantics is similar to the adverb enough (see, Lahiri 2002, p.162). The strength of this covert quantifier is contextually determined. In combination with the contextual restrictor on the ANS operator, these two elements derive different strengths of (non-) exhaustivity. An LF for Fox knows what Dana read

[9] Type $\langle\langle s, \langle\langle s, t\rangle, t\rangle\rangle, t\rangle$.

[10] For Lahiri, 'context' can refer to a larger discourse context where goals are conveyed, or more narrowly to lingusitic context, to capture lexical semantic differences between embedding verbs.

is presented in (81).

(81) $[_{\text{CP}}\ [_{\text{CP}}\ \text{ANS}_{\text{L}}\ [_{\text{CP}i}\ \text{what Dana read}]]\ [_{\text{IP}}\ \text{ENOUGH}_i\ [_{\text{IP}}\ \text{Fox knows}\ [_{\text{CP}i}\ t_i\ ]]]]$

For various compositional reasons, the interrogative CP c-commands the quantificational adverb that binds it. While typically, this would mean that the adverb cannot bind the CP, Lahiri includes a rule called "Adverbial Binding" which permits binding in this case.[11]

**Exhaustivity operators**

George (2011, Ch. 2) argues that the data which has been used in support of weak exhaustive answers in reality provides evidence for a non-exhaustive denotation. In their semantics, a question's base denotation is a Hamblin set, which derives non-exhaustivity.

(82) a. $[\![\ \text{Q}\ [\ \text{what Dana read}\ ]\ ]\!] =$
$\lambda p_{\langle s,t\rangle}.\exists\beta._e[p = \lambda w.\text{read}_w(\beta)(\text{Dana})]$

(83) a. $[\![\ \text{Q}\ [\ \text{X}\ [\ \text{what Dana read}\ ]]\ ]\!]$
$\exists\beta_{\langle e,t\rangle}.[p = \lambda w.\lambda x.[\ \text{read}_w(x)(\text{Dana}) = \beta\ ]]$

George's account is structural, because exhaustivity is derived when an exhaustivity operator is present in the LF, and non-exhaustivity when it is absent.

**Semi-Ambiguity Theories**

The Ambiguity theories discussed so far predict that both exhaustive and non-exhaustive meanings are in principle always grammatically available. We noted at the beginning of this section that there is a purported asymmetry between question forms—not all questions allow non-exhaustive readings. To capture this fact, some theories have attempted to provide a grammatical explanation for mention-some in a subset of questions. I call these *Semi-Ambiguity Theories* because they posit ambiguity in a subset of question forms.

[11] Lahiri uses Bittner's (1994) semantic framework.

The term 'Semi-Ambiguity' may suggest that the following theories predict that only certain questions types are ambiguous. In one sense that is correct, because the ambiguity they posit is grammatically derived in those sub-types of questions. In another sense it is incorrect, because grammatical ambiguity in a sub-set of questions is not necessarily inconsistent with mention-some availability in other questions. Given the theoretical mechanisms posited by these theories to explain the grammatical ambiguities, we might try to infer the predictions for other question types. Yet, it should be acknowledged upfront that these accounts do not exhaustively explain all questions types, and thus all observations of non-exhaustivity.

Three kinds of questions, in particular, have motivated such accounts. Questions with existential quantifiers, questions with existential modals, and embedded questions with infinitival clauses (George 2011, Ch6; Nicolae 2014; Fox 2014; Dayal 2016; Xiang 2016). Compare (84) and (85).

(84) a. Where are **some** of your students from? EXISTENTIAL QUANTIFIERS
Dana knows where **some** of your students are from.

b. Where is **a** gas station? EXISTENTIAL INDEFINITES
Dana knows where **a** gas station is.

c. Where **can** I find an Italian newspaper? EXISTENTIAL PRIORITY MODALS
Dana knows where I **can** find an Italian newspaper.

d. Dana knows where **to find** an Italian newspaper. INFINITIVAL CLAUSES

(85) Who came to the party?
Dana knows who came to the party.

(84a) contains the existential quantifier, (84b) an existential indefinite, (84c) contains the existential priority modal, can, and (84d) contains an infinitival embedded clause. These are often contrasted with questions like (85), argued not to allow a MS interpretation.[12] In this section, we review four such grammatical accounts. The first two, George (2011, Ch.6) and Nicolae 2014, deal mostly with existential quantifiers as in (84b), and suggest tentative treatments of (84a) in light of clear problems extending those analyses. The remaining two, Fox (2014) and Xiang (2016) deal specifically with

[12]We will see in the next section that this is not quite true. When the goals of the questioner are explicitly taken into account, we will see that the appropriateness of a given level of (non-) exhaustivity is determined relative to those goals.

existential modals as in (84a).

George (2011, Ch6) offers an alternative account for deriving non-exhaustivity to the one presented from their Ch.2. Unlike the earlier account, here they assume that the exhaustivity operator is obligatory. They account for (84) via quantifier raising (QR, May 1985). Mention-some is derived when the existential undergoes Quantifier Raising, and syntactically scopes over the exhaustivity operator.

(86) Where is there a pharmacy?
   a. [ Q [ [a pharmacy]$_i$ [ X [ where is $x_i$ ]]]]
   b. $\lambda p_{\langle s,t\rangle}.\exists\beta(p = \lambda w.\exists x.[[\text{pharmacy}_w(x) \wedge \lambda z.\text{location}_w(z)(x)] = \beta\ ])$

Essentially these truth conditions read, 'There is some set $\beta$ that specifies a proposition There is some pharmacy $x$ such that the set contains all locations that are locations of $x$.' They suggest that these are intuitive truth conditions for a mention-some reading. Further, George suggests that this scope approach may be extended to other elements with existential quantificational force, including modals like can and indefinites. However, Nicolae and Fox both object to this account on the grounds that modals do not QR (Nicolae 2013; Fox 2014).

Nicolae's (2013) solution employs a slightly different scopal approach. When an existential quantifier appears in a question, and there is no maximally informative true answer, a repair strategy is necessitated to reinterpret the question correctly. In this case, the existential can be reinterpreted as a complex quantifier bearing a +WH feature. Then, it undergoes two cycles of movement (wh-movement followed by a type-mismatch driven QR), resulting in quantification over a family of questions. In order to derive the right meaning, Nicolae invokes a quantifying-in operation, which allows selection of a subset of questions.

This approach cannot straightforwardly apply to modals because they cannot bear WH features or QR (Nicolae 2014, Fox 2014). In order to capture the modal data, some auxiliary assumptions are made. First, that there is a covert distributivity operator EACH sister to the wh-trace (following Fox 2014). Second, in the way that quantifiers above were allowed to be reinterpreted as complex objects, she suggests that wh-phrases are also allowed to do so. Thus, in the way that the existential quantifier

moves twice, the wh-phrase does here.

(87) Dana knows who can chair the committee.
  a. [ WHO$_i$ [ Dana knows [$_{\text{CP}}$ ID$_k$ [[$_j$ $\exists$ $t_i$ EACH ][[$_{\text{C}'}$ C $t_k$ ][ can $t_j$ chair ]]]]]]
  b. WHO = [ $\exists$ [ $\exists$ people ] ]
  c. ID$(Q)$ = $\lambda w.\exists p \in Q \wedge p = \iota p[p(w) \wedge \forall p' \in Q[p(w') \rightarrow p \subseteq p']]$

The presupposition of a unique maximal answer is contained in the definition of this ID operator. While her solution for existential quantifiers involve syntactic movement of the quantifier, her solution for modals does not. There, it is the wh-phrase that is reinterpreted to a higher type (that of a quantifier), and QRs.

Nicolae notes several issues with this solution (see, Nicolae 2013, p182). In particular, it seems to overgenerate especially in light of the fact that not all wh-phrases allow such re-interpretation (for example, d-linked wh-words, Pesetsky 1987). Further, the analysis relies on a secondary movement above an attitude predicate. Thus, in root questions she must assume that there is a covert attitude predicate.

Fox (2014/2018) derives non-exhaustivity through the modal's interaction with a distributivity operator with a meaning similar to each. First, he modifies Dayal's (1996) ANS operator which picks out the unique maximal element in the Hamblin set—the element that entails all other elements.

(88) ANS$_{\text{FOX}}(Q)(w)$ = $\lambda p.w \in p \in Q \wedge \forall q[w \in q \in Q \rightarrow q \not\subset p]$

This is a function from the Hamblin set (and the world of evaluation) to a set of propositions. Those propositions are the set of *maximally informative* true answers to Q—those which are not asymmetrically entailed by other propositions in Q. An exhaustive reading is derived when ANS$_{\text{FOX}}$ returns a singleton set, and the non-exhaustive reading is derived otherwise.

The cases where the set is non-singleton are those cases where the question contains a modal. Thus, modal questions are ambiguous between the two LFs below:

(89) a. [ ANS$_{\text{FOX}}$ [ who$_i$ [$_{\text{IP}}$ can [ [ $X_i$ EACH ] [ $\lambda x$.chair$(x)$ ]]]]] EXH
  b. [ ANS$_{\text{FOX}}$ [ who$_i$ [$_{\text{IP}}$ [ $X_i$ EACH ] [ [ $\lambda x$.can chair$(x)$ ]]]]] NON-EXH

In these two LFs, the top shows the exhaustive reading, while the bottom shows the

non-exhaustive reading. When the distributivity operator scopes over the modal, we derive a mention-some denotation that has multiple maximal elements.

Xiang achieves a scope effect without syntactic movement of the modal. First, the wh-word is raised to a higher type by a type-shifting operator. Unlike Nicolae, this is not a repair strategy triggered by presupposition failure, but a mandatory component of the analysis of questions. The complex wh then QRs to a position either above or below the modal. The resulting possible LFs are presented in (90a) and (90b).

(90) a. $[_{\text{CP}}$ [ SHIFT who ][ $\lambda X_{\langle et,t\rangle}$ $[_{\text{IP}}$ can [ $X_{\langle et,t\rangle}$ [ $\lambda x$.EXH chair$(x)$ ]]]]] NON-EXH
b. $[_{\text{CP}}$ [ SHIFT who ][ $\lambda X_{\langle et,t\rangle}$ $[_{\text{IP}}$ $X_{\langle et,t\rangle}$ [ $\lambda x$.can EXH chair$(x)$ ]]]]] EXH

The type-lifted wh-trace out-scopes the modal to derive the exhaustive reading; while it is out-scoped by the modal to derive the non-exhaustive reading. Finally, an ANS operator takes as its arguments, a world of evaluation and the meaning derived from the LFs above, and returns a set of maximally informative true answers.

Xiang argues that this reading is a very specific kind of mention-some called 'mention-one'. The reason that, on a mention-some reading of the modal question, mention-some is mention-one, is because the scopal relation makes the individual answers maximally informative because the mention-all answer (the conjunction of mention-one answers) is contradictory. This technical term refers to answers that name either a singular *or plural* individual, depending on the question predicate. Further, these generally do not need to be prosodically marked for ignorance, nor do they give rise to exhaustivity inferences. For example chair the committee is a predicate that can take either a singular individuals or a plural individual as an argument. A given context will determine whether singular/plural individuals are truthfully permitted to chair the committee, and an ANS operator filters in only true answers to the question denotation. Thus, in a context where only one chair is allowed, plural individuals will be false answers, and thus not part of the question denotation.

Modal questions are ambiguous between three denotations: a mention-all denotation (where the maximally informative answer is the conjunction of the true single answers), a mention-some denotation (where the true individual answers are each

maximally informative), or a disjunctive mention-all denotation (where the maximally informative answer is the disjunction of all the true mention-some answers). The disjunctive mention-all answer involves an additional covert operator, whose meaning is equivalent to the Mandarin exhaustification particle dou.

This theory is better understood from the standpoint of the speaker/hearer dynamic. When a speaker asks a modal question, the hearer must determine which interpretation to assign to the question. The hearer must make a decision to provide an answer, based on the interpretation she assigns to the question. The answer that the hearer gives might provide some clues to how she interpreted the question.

When an answerer gives an answer like "A and B" (with a falling intonation), this answer is homophonous between the different technical notions of answer that Xiang employs. Underlyingly, this answer could be semantically either a conjunction of two (singular) individual answers ($\Diamond$[Only $A \wedge$ Only $B$]), or itself constitute a plural individual answer ($\Diamond$ Only $A \oplus B$). On a mention-some denotation of the question, in a context where plural individuals are false or not permitted by the question predicate, then "A and B" is a contradiction because of the covert exhaustivity operator. More accurately then, the answer "A and B" is interpreted as "Possibly only A and only B".

It is important to note that, if the answerer gives a non-exhaustive answer like "A and B" and plural individual answers are not contextually true or possible arguments to the predicate, according to Xiang, this indicates that the answerer has interpreted the question on its mention-all reading. Because the answer is partial, the answerer must signal her ignorance about the mention-all answer by prosodically marking it. Otherwise, if she does not do so, then the answer will be interpreted exhaustively.

Let's review the data she includes to support these prediction (Xiang, 2016, pp. 40-41). First, she argues that mentioning a non-exhaustive answer like "A and B" leads to an exhaustivity inference in the absence of "ignorance-marking" prosody (e.g., a rising intonation marked by ↑):

(91) Context: only Dana, Mary and Sue can chair the committee, and there can be only one chair.
Who can chair the committee?

a. Dana. ↓ no exhaustivity inference
b. Dana and Mary. ↑
c. # Dana and Mary. ↓ # because exhaustivity inference
d. Dana or Mary. ↑
e. # Dana or Mary. ↓ # because exhaustivity inference.

Imagine only one person is needed to chair, but we are considering three people. According to Xiang, "A and B" and "A or B" with falling intonation as in (91c) and (91e) are infelicitous because they induce an exhaustivity inference which is not satisfied by the context. For these answers to be felicitous, they must associate with a rising intonation, as demonstrated in (91b) and (91c).[13] In contrast, a mention-one answer as in (91a) is felicitous with a falling intonation.

Xiang presents two scenarios to show that mention-intermediate is also infelicitous in embedded questions *even when the discourse goals explicitly license it*.

(92) Dana knows who can chair the committee.
a. 'For some individual $x$ such that $x$ can chair, Dana knows that $x$ can chair the committee'
b. 'For all individuals $x$ such that $x$ can chair, Dana knows that $x$ can chair the committee'
c. # 'For some three individuals $xyz$ such that $xyz$ each can chair, Dana knows that $xyz$ can each chair the committee'

(93) Norvin says to us, "On my exam, you'll have to name...multiple *wh*-fronting."
a. one language that has True
b. all languages that have True?
c. three languages that have False
Test: "Norvin said that we'll have to know where we can find multiple *wh*-fronting"

According to Xiang, (92c) and (93c) are infelicitous. Interestingly, (93) is quite similar to an example Lahiri gives in *support* of non-exhaustive answers (Lahiri 2002, pp. 162):

(94) a. What are the possible structures for the following sentence? (Give at least three.)
b. I can tell you what the possible structures for the following sentence are.

[13] Note that this is the terminology that Xiang uses, but to simply identify an intonation as 'rising' is underinformative as there are many different types of rising intonation. See, e.g., Hirschberg 1985.

Here, Lahiri argues, I need not give you all structures, but just enough to satisfy you. Dayal (2016, p. 80) also offers the following examples in support of mention-intermediate:

(95) I need two people to help me move my things.
  a. Who can I ask?
  b. You could ask Dana and Fox, or Walter and Alex.

(96) I need two people for my project on Spanish bilingualism.
  a. Who can speak Spanish fluently?
  b. Dana and Fox can. So can Walter and Alex.

Many native English speakers accept these mention-intermediate answers as felicitous. Xiang (pc) suggests two possible reasons for why this might be. First, mention-intermediate answers entail mention-one answers. Thus, a speaker might accept mention-intermediate because it satisfies mention-one. Example (93) is intended to show that, when you embed a know-wh under the weak necessity modal have to, mention-intermediate is semantically blocked. Additionally, the predicate passed the exam does not take plural individual arguments, so a mention-intermediate answer would not be permitted as satisfying the mention-one denotation.

However, exhaustivity inferences are incredibly common in declarative utterances, in the absence of any nearby question in the discourse (Roberts 1996/2012; Levinson 2000; Sperber & Wilson 1986; Schulz & van Rooij 2006, Zimmermann 2010). Thus, the exhaustivity interpretation in these answers plausibly falls out from that phenomenon (as argued by Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010). Further, given that the context is constructed to require a a single committee chair, we would expect that a mention-intermediate answer be infelicitous. If we changed the goals, to allow require two people to chair the committee, then at least (91c) is definitely felicitous. This illustrates the fact that mention-some is mention-one *only if the discourse goals are mention-one*.

I'd like to suggest other possible explanations. If there is infelicity here, perhaps it can be attributed to an attempt to quantify over an exact number using who when there is a better option available, which three people. This explanation would not require a

specific interpretation of mention-some as mention-one, but rather would illustrate competition between alternative utterances in a given context. Imagine a case with an existential quantifier as in (97), where Dana ate three cookies. Would we say that this is infelicitous if the speaker intended to quantify over those three cookies, as is attempted in (92c)?

(97) Fox knows Dana ate some cookies.
   a. For some cookie $x$ Fox knows that Dana ate $x$
   b. For all cookies $x$, Fox knows that Dana ate $x$
   c. For some three cookies $xyz$, Fox knows that Dana ate $xyz$

(97) seems perfectly true if Dana ate an intermediately-exhaustive number of cookies. It's not straightforwardly clear to me that (92c) is any more or less infelicitous than (97c) or (93c). We certainly would not want to say that this suggests that some requires quantification over a singleton.

### 2.4.3 Summary and Discussion

In the previous section, I have presented the extant accounts of non-exhaustivity. Here, I now highlight the predictions that they make about the availability and distribution of the reading. These theoretical accounts all share one property: when you consider how an ambiguity between mention-some and mention-all is resolved, whether grammatical or not, some mechanism is needed to *license the correct grammatical interaction*. For example, any scope effect requires a pragmatic mechanism to either license the particular scopal interpretation or not. This point is illustrated by the fact that a question with a modal or existential may receive either an exhaustive or a non-exhaustive interpretation. Which interpretation it receives will depend on whether the context provides the right kind of information to license that interpretation. Thus, whether non-exhaustivity is grammatically available under the right circumstances, pragmatically derived under the right circumstances, or disambiguated under the right circumstances, those circumstances are given by pragmatics.

## 2.5 The pragmatics of questions

It is important to mention that some notion of context-sensitivity is necessary to account for the interpretational variability in (non-)exhaustivity we have observed, no matter how it's derived in an underlying semantic theory. A given context will determine which interpretation a question receives (or, which answer to a root question is most appropriate). We can attempt to identify the predictions that particular semantic theories make. However, most semantic theories do not themselves discuss in detail the pragmatic mechanisms that their theories would require. Thus, much of the following discussion will involve abduction from what the theories do say, and what seems plausible prerequisite assumptions in light of the evidence discussed for/against mention-some.

There are three strategies for accounting for this context-sensitivity. Each of these three strategies has been explicitly or implicitly proposed by semanticists accounting for (non-)exhaustivity in questions. Unfortunately, it is often difficult to tease apart these accounts on the basis of acceptability judgement behavior, because they predict acceptability given appropriate contextual licensing.

### 2.5.1 Ambiguity resolution

The first strategy posits context-sensitivity in virtue of an underlying semantic ambiguity. The context-sensitivity here is akin to the context sensitivity seen with bank (place by the river/financial institution/a sloped inward tilt of a vehicle along a curve...) and the girl saw the pirate with the telescope(the girl saw [the pirate] [with the telescope]/the girl saw [the pirate [with the telescope]]). These strings are ambiguous between two different interpretations, and the context will determine which interpretation the speaker intended. The hearer must resolve this ambiguity in order to fix literal meaning.

Semantic theories on which questions give rise to an ambiguity (cf. Beck & Rullmann 1999; George 2011, Ch. 2) require such mechanism for disambiguation. Modal theories (George 2011, Ch 6; Fox 2014/2018; Nicolae 2015; Xiang (2016)) fall under this

category because they posit ambiguity in modal (or existential) questions. However, on modal theories, non-modal questions are not ambiguous, so they would not be context-sensitive in the way an ambiguous expression would be.

There's another kind of ambiguity. An utterance might be ambiguous between a literal meaning, and a non-literal, or speaker meaning. We might think that this kind of ambiguity resolution recruits non-linguistic mechanisms which involve, perhaps, general abductive inferences about the context given the speaker's utterance. This kind of ambiguity might be considered a different beast than the ambiguity described above, because it is not located at the grammatical level. We will discuss this in two sections.

### 2.5.2 Free variable resolution (precisification)

Many semantic theories posit free variables whose values resolve to certain aspects of the context. Call this free variable context-sensitivity (cf. Kaplan 1989, Stanley 2000; Stanley & Szabó 2000). Note that Kaplan distinguished between pure indexicals, which have a fixed conventional content (the character) that determines a context-independent truth-value, and true demonstratives, whose content is context-sensitive and does not determine a context-independent truth-value. We will group these two notions together because their differences do not matter for the present purposes.[14]

A candidate example of this would be universal quantifiers, whose quantificational domains are implicitly restricted by context: everyone is asleep is not understood to mean everyone in the world is asleep, but something more like every contextually relevant person is asleep.

This kind of semantic theory fits well within a Stalnakerian theory of the context,

[14]Where the distinction might be useful is in thinking about a theory like van Rooij's, where a decision problem must be contextually fixed to determine the utility of answers. This might be considered conventional, and thus a pure indexical in Kaplan's sense. In contrast, a contextual variable that constrains the quantificational domain of the wh-word (cf. Aloni 2001), or the set of alternatives (cf. Chierchia, Fox, & Spector 2012) might reasonably be considered true demonstratives because they need context to specify the salient sets. But this is debatable.

because variables can be formally treated as pronouns anaphoric to preceding expression in the common ground. However, one problem noted by Cresswell (1973) was that, to reduce all instances of context sensitivity to pronominal indexation in this way, would require a proliferation of variables that would give rise to combinatorial explosion. Further, on a Stalnakerian theory, all the propositions necessary to fixing the values of these variables must be available prior to interpretation (1973, p. 111, as noted in Stanley & Szabó 2000, footnote 1). This gives rise to a 'frame problem'.

Semantic theories of questions have used free variable in different ways: for precisification of an underspecified semantic representation (Ginzburg (1995); van Rooij (2003), 2004), or strengthening of an existential representation due to free variables (Asher & Lascarides (1998)). Both theory types will predict that both mention-some and mention-all interpretations will be context-dependent and parametric to specific aspects of the context. This strategy allows for context-dependence in embedded questions, for what Asher & Lascarides term "pragmatically rich representations...[that] can be used as arguments of propositional attitudes," (pp. 266). "Pragmatics" here refers to the involvement of beliefs, mental states, goals/plans in fixing the semantic content of an expression.

Van Rooij's decision-theoretic semantics delivers exhaustivity parameterized to an agent's decision problem, which determines the expected utility of answers. An answer reduces uncertainty, and will have a higher utility the more uncertainty it reduces. As a mention-all answer will typically reduce the most uncertainty, given that it covers the space of possibilities, these answers/interpretations will have the highest expected utility. A question is interpreted as mention-some when the expected utility of a mention-some answer is equal to the expected utility of a mention-all answer (with additional weighting given that a mention-some answer is less effortful than a mention-all answer). Given that this formalized utility is to track the acceptability of interpretations of embedded questions, we would expect that we will not

find mention-some interpretations accepted at a higher rate than mention-all interpretations. We can call this van Rooij's Utility Hypothesis.[15]

### 2.5.3 Pragmatic vs. Semantic Strengthening

The final pragmatic strategy posits that the additional interpretation is derived via a **strengthening** inference. There are two ways this strategy can operate. We take as an example phenomenon, (scalar) implicature (Grice 1967, 1989; Horn 1972; Gazdar 1979). A scalar implicature is an inference from a weaker meaning to a stronger one, as with utterances with the existential quantifier some. When a speaker utters something like, I ate some of the cookies, it is logically compatible with a stronger universal meaning, that the speaker ate all of the cookies. However, often the hearer will infer that, when the speaker utters the existential claim and not the universal claim, it is because (the speaker believes that) the stronger claim is false.

The mechanism which strengthens the existential claim (negating the universal claim) could be semantic, as in the covert exhaustivity operators posited by researchers like Chierchia, Fox, Spector for grammatical scalar implicature. In contrast, the traditional explanation is that the strengthened meaning falls out from a rational cooperative inference that the hearer makes on the basis of "what was said" (i.e., the existential claim). This is the Gricean explanation, and it is an extra-linguistic inference on the basis of a semantic representation. Given what we know from the psycholinguistics literature, we would expect this kind of inference to be context-sensitive as well (Schulz & van Rooij 2006; Degen 2015; Degen & Goodman 2014; Breheny et al. 2006; Grodner et al. 2010; Breheny et al. 2013).

Some researchers align question semantics/pragmatics in this way (Schulz & van Rooij 2006; Spector 2006, 2007; Zimmermann 2010; and for clefts: Geiss et al. 2018;

[15]It is known from behavioral economics that many constraints actually go into calculating the expected utilities (cf. March & Simon, 1958; Lindblom, 1959; McGee 1991; Feldman 2006). Thus, might think that van Rooij's theory is providing a useful simplification of decision theory using only the informational content of an answer, but that ultimately his theory should be refined once the additional constraints of making decisions are identified.

Destruel & DeVaugh-Geiss 2018)[16]). Further, pragmatic explanations of exhaustivity can easily appeal to independent factors to explain embedded exhaustivity, like the semantics of the embedding predicate (e.g., with know, as argued in Heim 1994; discussion in Zimmermann 2010).

### 2.5.4 Semantic vs. Pragmatic Weakening

It is seems taken for granted that questions are, by-and-large, at least weakly exhaustive, and recent theories posit mandatory covert exhaustivity operators (cf., George 2011, Ch 6; Nicolae 2013, Fox 2014; Xiang 2016) whose meanings closely resemble the operator posited by proponents of grammatical scalar implicature (cf. Chierchia, Fox, Spector). If questions are indeed underlyingly (weakly or strongly) exhaustive, they would need to be *weakened* even further to derive non-exhaustivity. What kind of weakening mechanism would be available? I will discuss three separate possibilities: tolerance, domain restriction, and a Gricean inference.

Essentially, one could argue that question meaning appears non-exhaustive, but if the wh-domain is restricted to a small enough subset, you derive the appearance of non-exhaustivity. Though it is highly contested how this subset is identified, a classic proposal is that it must be linguistically available in the preceding discourse (in the Common Ground, von Fintel 1994). However, we can present several different examples where this criterion does not hold, suggesting that if anything the subdomain is fixed relative to the speaker's intention (recall the example (23), discussion from George 2011, Ch 6). Let us briefly review some of those cases against a domain restriction account.

George 2011 concludes that the strategy is not sufficient to cover many basic cases. Consider (98) from George 2011, pp. 211-212, and (99), an elaboration from van Rooij (2003).

[16]Given our characterization of context-dependency, Asher & Lascarides (1998) might be included as this kind of a theory, because they posit underlying existential representation that is strengthened based on the resolution of free variables. I do not discuss them here because they do not employ an exhaustivity operator to derive stronger degrees of exhaustivity.

(98) Context: Professor Worth is an outspoken critic of the mayor.
   a. A: Who criticized the mayor's plans for renovating the high school?
   b. B: Professor Worth did.
   c. A: Thanks, now I know who criticized the mayor's plans for renovating the high school.

(99) Context: A is at a party and has forgotten their lighter. In fact, most people at the party have lighters.
   a. A: Who has a light?
   b. B: Pam over there.
   c. A: Thanks, now I know who has a light.

It is highly unlikely that no one besides Professor Worth would criticize the mayor's plans, such that (98b) could be considered an exhaustive answer, and (99b) is explicitly not an exhaustive answer. Yet, in these contexts the knowledge reports are felicitous on a mention-some reading.

The point comes out in force with data such as (100).

(100) a. How do I get to the train station?
   b. Why did the Roman Empire collapse?

Similar to cases where a universal quantifier is not intended to literally qauntify over a whole set, but is rather implicitly restricted to some salient/relevant subset.

Alternatively, it has been suggested that a hearer may *tolerate* a non-exhaustive interpretation, though one is not semantically available (Xiang, p.c.). Hearers might accept a non-exhaustive interpretation especially if the context makes that interpretation true, but the interpretation really is not acceptable. We can draw an analogy to false-answer sensitivity, where some participants accept embedded wh reports even when the agent holds false beliefs about the answers to the question (cf. Spector 2006, 2007; Kleindinst & Rothschild 2011; Phillips & George 2018; Xiang (2016)). Note that in that case, hearers *reject* these false-belief cases, and (unsurprisingly) especially when test sentences are know-wh reports. Perhaps other kinds of permissiveness might be possible explanations, in a "pragmatic slack" or "loose speak" kind of way (Lasersohn 1999). This kind of permissiveness occurs when a speaker's utterance is true enough for the current conversational purposes. For example, a speaker might say that Dana arrived at three o'clock, even if she didn't arrive *exactly* at three. A hearer who hears

Dana arrived at three o'clock would not typically react with surprise or anger if they found out Dana arrived at 3:01, because three o'clock is close enough to 3:01. Lasersohn's theory concerns how speakers often fail to *strictly speaking* utter truths, rather than how hearers are always *strictly speaking* accepting truths. We will see that hearers do not reject mention-some at the rate that they reject the false answer scenarios (even when the agent knows the weak exhaustive answer).

We can dismiss a Gricean explanation because in general, the Gricean maxims from which a hearer draws additional inferences, affect a defeasible *strengthening* of semantic content, but not a *weakening* of it. And yet, if we examine the data presented and discussions against semantic mention-some, it would seem that some theories in fact depend on certain auxilary assumptions about the division of labor between semantics and pragmatics, as assumed by Grice. And yet, some theories seem to assume that mention-some is pragmatic in a Gricean way. In some cases the assumption is explicit (e.g., Groenendijk & Stokhof (1982), (1984)), while in other places is underlies the data, an argument I will call the *unembeddability of pragmatic phenomena*.

A Gricean inference is assumed to take as input a fully truth-conditional proposition: what the speaker means/implicates is a function of what the speaker said (the proposition expressed by the literal meaning) and the alternative utterances the speaker could have made.

The hearer reasons from the literal meaning, and (her beliefs about) the speaker's intended meaning, by using the principles of rational cooperative communication as embodied in the four maxims (Quantity, Quality, Relevance, and Manner). This view, combined with an early view about cognitive architecture from sentence processing (Frazier & Fodor 1978; Forster 1979) and philosophy of mind (Fodor 1983), developed into a hypothesis about the interface between language and mind. Namely that linguistic/semantic processes are *informationally encapsulated* from domain-general rational processes. As such, inferences involving rational reasoning of the Gricean sort cannot interfere in semantic processes occur in the scope of semantic operators or embedded clauses generally. See, e.g., Chierchia, Fox, & Spector, 2012; Chemla & Singh

2014; Degen & Tanenhaus 2015; Degen & Tanenhaus 2019 for in-depth discussion.

This view has been brought into question by discourse representation theory (Kamp 1981; Kamp & Reyle 1993; and segmented DRT, Asher & Lascarides 2003, Simons 2011, 2017), by the emergence of grammatical theories of scalar implicature (Chierchia, Fox, Spector 2012), and by psycholinguistic findings which cause doubt to informational encapsulation views of cognitive architecture (e.g., Degen & Tanenhaus 2019; Elman, Hare, McRae 2004; MacDonald, Pearlmutter & Seidenberg 1994; Seigenberg & MacDonald 1999; Tanenhaus & Truswell 1995; McRae & Matsuki 2004, a.o.).

The question of the embeddability of mention-some is a double-edged sword: if mention-some is unembeddable, this is used to argue that it is pragmatic and not semantic; if mention-some is embeddable, this is used against pragmatic theories in support of semantic theories. Either argument here assumes the general unembeddability of pragmatics as discussed above. We find this in Karttunen (1977), footnote 4; Groenendijk & Stokhof (1984), footnote 14, pp. 558 and discussion on pp. 533; Beck & Rullmann 1999 pp. 286; and in Xiang (2016) pp. 44-45 (reiterating Karttunen's insight). In illustration of this kind of data, consider example (101a) from Karttunen (1977), p.9 footnote 4:

(101) a. A: Who, for instance, came to the party last night?
b. B: Dana.

(102) Fox knows who (∗for instance) came to the party last night.

The root question in (101a) permits a mention-some answer as in (101b). Karttunen notes that phrases such as for instance or for example are "conventional devices for indicating that exhaustiveness is not desired," (*ibid*). He further observes that they are not permitted in embedded questions, as in (102).

## 2.6 Conclusion

In this chapter, we have reviewed theoretical, and experimental approaches to the meaning of questions. We first introduced the challenges a truth conditional approach to questions encounters, then turned to discuss the history of formal semantic accounts

of the meaning of questions. We then reviewed the different ways these theories handle a variety of question types, and the particular problem that non-exhaustivity poses to formal accounts. Finally, we turned to focus on the elusive non-exhaustive reading.

Our review of the linguistic and discourse factors licensing non-exhaustivity has revealed significant variability in the availability of non-exhaustive readings. We saw joint constraints from both discourse and linguistic form. We reviewed three linguistic factors that will play a crucial role in the following chapters: the embedding matrix verb, the wh-word heading the question, and the presence/absence of a modal element. We also examined in detail how, for each linguistic constraint posed, we can construct a context to make a non-exhaustive reading felicitous by manipulating the discourse goals.

We are left asking how systematic and robust are the linguistic constraints on non-exhaustivity, and to what extent context exerts an influence? That is, to what extent is (non-) exhaustivity derived from or licensed by the linguistic form of the question, and to what extent can context override the influence of these linguistic cues? We conducted two sets of experiments to answer these questions. The following chapters address these questions.

# Chapter 3

# Experiments 1 and 2: Establishing the bounds of non-exhaustivity

There are two main questions I address in this chapter: to what extent is mention-some restricted across question forms, and to what extent is it licensed by contextual goals, in spite of question form.[1] We will see that, while some question forms appear to correlate highly with mention-some over others when contextual goals are not explicitly manipulated (Experiment 1), when goals are manipulated explicitly, participants choose the reading/answer that best matches those goals regardless of question form.

A question, root or embedded, is exhaustive/mention-all (MA), if it permits a mention-all answer to the root question, or reading of the embedded question. A question is non-exhaustive/mention-some (MS), if it permits a mention-some answer, or reading. The question is whether, and which questions allow both, and under what conditions. The previous chapter surveyed the literature on (non-)exhaustivity in questions, and revealed disagreements about the distribution of the mention-some reading and its proper treatment. On most semantic theories, a question's denotation is mention-all. Thus, the proper answer to a root question will always be the exhaustive list answer, and any embedded question report (for example, Dana knows where to find coffee) will be true only if Dana knows the exhaustive answer. The acceptability of the mention-some answer/interpretation, in both root and embedded questions is a puzzle if a question's denotation is always exhaustive.

In surveying the literature, two observations stood out about the distribution of mention-some. On the one hand, questions appear to exhibit *baseline interpretations*

[1]The bulk of this chapter comes from Moyer, M & K. Syrett. (2019). (Non-)exhaustivity in embedded questions: contextual, lexical, and structural factors. *Proceedings of the 23rd Meeting of Sinn und Bedeutung*.

(in the words of Asher & Lascarides) for exhaustivity or non-exhaustivity. When we say that a question has a baseline interpretation for exhaustivity(/non-exhaustivity), we just mean that without explicit linguistic context provided, the question is interpreted exhaustively(/non-exhaustively). Importantly, we will see that these baseline interpretations are defeasible: when contexts are made explicit, interpretations shift according to the context.

Recall a representative sample of question forms (103):

(103) a. **Who** came to the party? MA/#MS
Dana **knows who** came to the party.

b. **Where can** I get coffee? MA/MS
Dana **knows where** I **can** get coffee.
Dana knows where **to get** coffee.

c. **How** do you get to Central Park? #MA/MS
Dana **knows how** you get to Central Park.

(103a) is often presented as a mention-all question, (103c) as a mention-some question, while (103b) as allowing either. Who-questions are default mention-all, how-questions default mention-some, and where-questions perhaps do not have a default. The facts are more fine-grained than this, however. Modulation of these baselines is found in not just the **wh-word** heading the question, but also in the presence of a **modal** element (or a non-finite clause in an embedded question), and from the **matrix verb** embedding a question.

These emerge when the question is presented without any explicit linguistic context. They are thus defeasible. The second observation is that despite these baseline interpretations, the felicity of a mention-some answer to a root question *and* the acceptability of a mention-some interpretation of an embedded question are also modulated by particular aspects of the discourse context: the questioner's **goal** and in some cases, her **mental state**. For example, the root question in (103b) is interpreted as mention-some when asked by a tourist on the street, but as mention-all when asked by a journalist reviewing the local cafés. The difference between these two contexts lies in the questioner's goal: the tourist most likely does not need an exhaustive list to satisfy her goal of drinking a cup of coffee, while the journalist would need one to conduct

a complete review of the local establishments. Importantly, these judgements transfer to the embedded question: Dana is judged to have knowledge if her mention-some knowledge satisfies the contextual goals.

What determines a given question's baseline interpretation of (non-)exhaustivity out of context? To what extent can a given context override those baseline interpretations? Semantic theories of question meaning disagree in their answers to these questions. In this chapter, I present two sets of experiments that quantify both the linguistic and contextual discourse factors that modulate the availability of mention-some interpretations in embedded wh-questions. In Section 1, I present the hypotheses and predictions of semantic theories. In Section 2, I present Experiment 1, which establishes a baseline to identify how generalizable the mention-some interpretation actually is (data that has, up to this point, been missing from discussions). To preview the findings, I show that the linguistic factors we have isolated do indeed modulate acceptability, and crucially, that while the presence of a modal boosts acceptability of MS readings, the absence of a modal does not yield categorical rejection of MS interpretations. In Section 3, I present Experiment 2, in which we manipulate the context and show that (non-)exhaustivity in wh-questions is influenced by the speaker's goals in asking the question, despite question form. In attempt to understand the nature of *default preferences*, in Section 4, I present a more in-depth analysis of the effects of each story presented to participants from Experiments 1 and 2. The point of this discussion is to understand what world knowledge particular scenarios may import, outside of the manipulated contextual factors. I suggest that the idea that questions exhibit baseline interpretations out of context is due to the fact that questions are never really interpreted out of context. Rather, the hearer imports a context from their prior experience and world knowledge about the likely goals of a speaker asking the question, or situation surrounding an utterance of an embedded question. Finally, I conclude in Section 5.

## 3.1 Hypotheses and Predictions

We noted four factors that potentially influence the acceptability of the mention-some reading. Three form factors (the wh-word, matrix embedding verbs, and the presence/absence of an existential modal) and at least one contextual factor (contextual goals, and possibly an agent's mental state).

### 3.1.1 Accounting for Linguistic Form

The first hypothesis is that the linguistic form of the question semantically determines whether the question is mention-some or mention-all. This is the main hypothesis tested in Experiment 1. More specifically, we can generate predictions based on the observed linguistic factors. First, following the observations of Ginzburg (1995) and Asher & Lascarides (1998), the wh-word hypothesis predicts that who-questions will have a lower baseline acceptability on mention-some readings than where-questions. This prediction is not grounded in any mechanism proposed by a semantic theory, but rather from general observations in the literature. Second, the Matrix Verb Hypothesis predicts differences between different question embedding verbs. This prediction is grounded in the observation that know-wh seems to require strong exhaustivity in its embedded question complement. One example theoretical proposal is George 2011, who suggests that question embedding predicates may select for a complement that contains an exhaustivity operator, know being one such predicate. Differences between know, a verb thought to select for (strong) mention-all, and a verb which does not select for (strong) mention-all (we will use predict, Beck & Rullmann 1999, Klinedinst & Rothschild 2011) thus may provide evidence for some kind of lexical semantic or semantic selectional restriction. Note that we expect these factors to interact, given the observation that some know-where and know-how reports felicitously allow mention-some, as in (103b) and (103c) (though we do not systematically examine how-questions in these studies).

The Modal Hypothesis predicts that mention-some will be more acceptable in modal

questions than in non-modal questions. We would also expect a similar difference between non-finite and finite (non-modal) clauses, assuming Bhatt's (1999) semantics which encodes covert priority modality in infinitival clauses. As stated, this hypothesis is fairly weak, merely predicting *differences* between modal(/non-finite) and non-modal questions. This hypothesis is compatible with different explanations for why we would find these interpretational differences. Modal semantic theories (George 2011, Ch 6; Nicolae 2014; Fox 2014; Xiang (2016)) claim that the (existential priority) modal interacts scopally with some other element in the question (see the previous chapter for the summaries of particular theoretical accounts).

It is tempting to think that these theories make the strong prediction that non-modal questions do not give rise to mention-some, because no modal is available for the scope interaction that these theories posit. This is especially natural considering that the data supporting these theories involve asymmetries in acceptability between modal and non-modal questions (cf. Xiang & Cremers 2017). However, Xiang (p.c.) states that these theories make no claims about mention-some in non-modal questions, and thus do not make this prediction.

A mere asymmetry in acceptability could mean many different things, and does not necessarily provide support for or against an underlying grammatical difference because of the modal, without a coherent notion of the contextual modulation of ambiguity, or of tolerance in the case of an interpretation that is not supported by the presence of an underlying representation. In the comparison between modal and non-modal questions, it is important to consider the effect size, and how much the degraded reading is degraded (cf. Gibson, Piantadosi, & Fedorenko, 2011). If non-modal questions are degraded *from ceiling* with respect to modal questions, the interpretation of the results would be different than if non-modal questions are degraded *from floor*. In order words, if we see the predicted asymmetry, do participants accept non-modal mention-some more often than not, or less-often-than-not? The first option is not obviously consistent with a theoretically meaningful distinction, while the second is perhaps more obviously so.

Xiang (2016)'s account of mention-some imposes a very particular kind of non-exhaustivity on the question: mention-some specifies exactly one *option*. This single option can be either a singular or a plural individual, depending on whether the question predicate allows plural individual arguments and whether the context makes the plural or singular individual option true (the ANS operator filters in only true answers in the question denotation). Only these single options, on a mention-some reading of a question, permit non-exhaustivity (they do not give rise to an exclusive interpretation). On a mention-all interpretation of the question, all non-exhaustive answers are partial answers. For example: on the mention-some reading of Dana knows where we can find coffee, a mention-one reading is true if and only if Dana knows one place to find coffee (and possibly more). Since the predicate does not typically take plural individual arguments, an interpretation on which Dana knows two (but not all) cafés is not semantically available; it is partial (or, "mention-few"). It is not clear that this hypothesis predicts an asymmetry in acceptability between the two non-exhaustive interpretations of a statement like Dana knows where we can find coffee, because the "mention-two" interpretation entails the mention-one interpretation: if Dana knows two places where we can find coffee, then she knows one (Xiang, p.c.). Thus, while a context may make our "mention-two" reading true, participant acceptance of Dana knows where we can find coffee on that interpretation would not indicate that there is an underlying semantic representation of that reading, but fall out from the entailment facts. To preview our results, Experiment 2 shows that participants do not distinguish between single or plural non-exhaustive answers, but it is an open question what these data say about the theory.

The role that a modal would play if not grammatical as predicted by modal theories, then, is a disambiguating signal which differentially updates the probability distribution of one meaning over the other. Another way to say this is that the modal makes a non-exhaustive meaning more likely/salient. This is compatible with a semantic explanation provided by modal theories, but it is also compatible with a wider

range of explanations. We may then think of the modal as a defeasible *cue to interpretation*. This reinterpretation accords with thinking in the sentence processing literature (for example, in Elman, Hare & McRae 2004), and recently, in constraint-based accounts of pragmatic processing (Degen & Tanenhaus 2019). These psycholinguists note that cues to interpretation may have an additive effect when they co-occur. If this is right, then perhaps we will see interactions between multiple factors, suggesting that the question form factors are better understood in this light as cues to interpretation. This hypothesis, while it predicts the importance of a modal, is consistent with semantic theories that posit across-the-board ambiguity or underspecification.

The Null Hypothesis, then is that there will be no differences in the acceptability between different question forms. This hypothesis is compatible with many different kinds of underlying semantics theories, including ones that posit across-the-board ambiguity for all values of (non-)exhaustivity in embedded questions (i.e., Beck & Rullmann 1999; George 2011, Ch.2) and those which posit semantic underspecification for (non-)exhaustivity (i.e., Ginzburg (1995); Asher & Lascarides (1998); van Rooij (2003), 2004). Theoretically, these two different kinds of semantics posit very different kinds of representations—recall that in the last chapter these were presented as separate approaches to non-exhaustivity. I group them together here because they make the same predictions about the three question form factors.

### 3.1.2 Accounting for Context-Sensitivity

No matter the underlying semantic representation, all theories must allow some context-sensitivity, some mechanism to explain why a question's interpretation shifts. In Section 2.5 of Chapter 2, I explicated several different ways that an expression could exhibit context sensitivity. We will not be able to give definitive evidence for one or the other kind of strategy here.

If mention-some is merely tolerated because of partiality, but not semantically available in non-modal questions (cf. modal theories, George 2001, Ch. 6; Fox 2014, Nicolae 2014; Xiang (2016)), then we might see a higher proportion of rejections than

acceptances. However, if mention-some is available semantically in these cases, but merely dispreferred for lack of contextual licensing (as Dayal (2016) notes pp. 71-82), the we might see more acceptances than rejections. Either way, it will be tricky making inferences from response rates around the chance mark.

## 3.2 Experiment 1: How generalizable is the mention-some reading?

In this experiment, we test the hypothesis that the Linguistic Form of the embedded question determines whether a mention-some interpretation is acceptable. We focus on the contribution of the three form factors discussed above, and isolate each one by using a fully-crossed factorial design.

### 3.2.1 Design and Materials

The experiment had a 2x2x2x4 design with three QUESTION FORM factors MATRIX VERB (know, predict), WH-WORD (who, where), and FINITENESS of embedded clause (+FIN, –FIN), and ANSWER TYPE: Mention-All (MA), Mention-Some (MS), Mention-All+False Report (MA+FR), and False Report (FR). These four ANSWER TYPES differ in how many answers a character provides in the specific trial. FRs were included as a control. FINITENESS was the only factor manipulated between subjects, in order to make sure that there was no influence of this factor within a participant's experimental session. Contexts were minimally changed across this factor to satisfy the felicity conditions of finite/non-finite clauses. An example of a trial with know, where, –FIN and an MS ANSWER appears in (104). The dependent measure was a binary yes/no judgement.

(104) The places that serve cappuccinos around the neighborhood are A, B, C, and D. E and F do not. Mary usually gets her cappuccino at D. Jane is going to be in the neighborhood tomorrow. She loves cappuccinos, and texts Mary to ask where to get a cappuccino.

Mary responds,

a. "D." MENTION-SOME (MS)
b. "A, B, C, and D." MENTION-ALL (MA)

c. "A, B, C, D, and E." MENTION-ALL+FALSE REPORT (MA+FR)
d. "E and F." FALSE REPORT (FR)

Jane reports, "Mary knows where to find cappuccinos."
Is Jane right?

There were eight total sentence frames, for every possible combination of the three factors pertaining to question form. The target sentences featured eight different embedded verbs following the wh-word to allow for generalization across predicates within the embedded clause:

(105) *Embedded verbs*

a. who: find, view, store, locate, hide, bury, sell, display

b. where: recruit, interview, invite, ask, call, hire, select, contact

These manipulations yielded a total of 64 unique sentence tokens. Stimuli were assigned to four lists in a pseudo-randomized Latin square fashion. In addition to the 64 unique test items, there were 10 root question filler sentences of the form *Which of the following X is not Y?*, with four possible answers listed. Filler questions served as comprehension and attention checks and addressed common world-knowledge based category membership, for example, *Which of the creatures is not a mammal?*

### 3.2.2 Participants

232 undergraduates enrolled in introductory-level courses were recruited from the Rutgers University Linguistics and Cognitive Science subject pool. 14 participants were removed from final analysis for non-native speaker status. The experiment was designed and administered using Qualtrics survey software. Each participant was run in a quiet laboratory setting, seated at an iMac. Participants were asked to read a series of brief contexts, and after each one, respond to a question corresponding to a preceding statement. Each context was comprised of 3-4 sentences, and ended with a question. A person then delivered an answer to the question, corresponding to one of the Answer types manipulated. Participants chose either *yes* or *no* in response to the prompt (e.g., *Is Jane right?*).

### 3.2.3 Predictions

As mention-all answers are uniformly predicted to be available for all questions (excepting how-questions, not tested in these experiments), the Mention-All MA ANSWER condition will serve as a True Control. We predict uniform acceptability in this condition, with no effects of any linguistic form factor. Indeed, this prediction has been supported by formal and informal surveys from Cremers & Chemla (2016) (predict and know), and Klinedinst & Rothschild (2011) (predict), respectively. Similarly, as false answers are uniformly predicted to be unacceptable, the FR ANSWER condition will serve as a False Control. Klinedinst & Rothschild (2011) note that predict allows for a non-veridical reading, so it is possible that predict targets will be accepted slightly higher than know targets in this condition.

Though not of main interest to the current inquiry, the Mention-All+False Report (MA+FR) Condition tests the acceptability of mixed true and false reports. Results from such conditions often play a role in the debate over strengths of exhaustive denotations. Experimental results from Cremers & Chemla (2016, 2017) and Phillips & George (2018) suggest that these conditions will receive acceptability degraded relative to the MA condition, but higher than the FR condition.

The critical condition is the Mention-Some (MS ANSWER) context, because it presents a true mention-some reading. The Strong Modal Hypothesis predicts that -FIN target embedded question reports will be accepted in this condition, but +FIN targets will be rejected because a mention-some reading is only available in embedded modal questions. The Weak Modal Hypothesis only predicts differences between the two conditions, where -FIN targets will be more acceptable than +FIN targets. Thus, we will be able to test the prediction that a modal is necessary for mention-some. The wh-word hypothesis predicts that where targets will be more acceptable than who targets, and the matrix verb hypothesis predicts that know-wh targets will be less acceptable than predict-wh targets.

### 3.2.4 Results

Experiment 1 results are presented in Figure 3.1. Each graph corresponds to a factor tested, across the four ANSWER TYPE conditions. All analyses conducted were non-

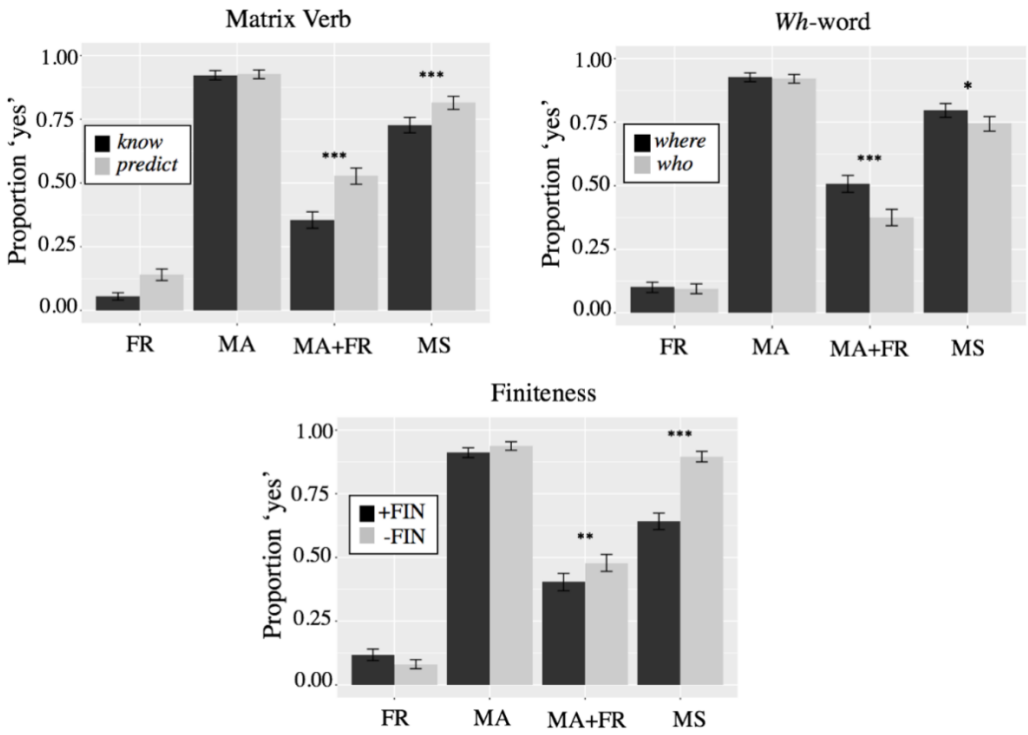


**Figure 3.1:** Experiment 1 results given three form factors and four answer types.

parametric Kruskal-Wallis tests. We found overall main effects of VERB ($\chi^2(1) = 53.714$, $p<0.0001$), WH-WORD $\chi^2(1) = 9.71$, $p < 0.005$), FINITENESS ($\chi^2(1) = 43.567$, $p<0.0001$), and ANSWER ($\chi^2(3) = 823.42$, $p < 0.0001$). Breaking down each factor per ANSWER, all effects are significant for both MS and MA+FR: VERB (MS: $\chi^2(1) = 18.892$, $p < 0.0001$; MA+FR: $\chi^2(1) = 51.731$, $p < 0.0001$); WH-WORD (MS: $\chi^2(1) = 6.61$, $p < 0.05$; MA+FR: $\chi^2(1) = 30.219$, $p < 0.0001$); and FINITENESS (MS: $\chi^2(1) = 156.7$, $p < 0.0001$; MA+FR: $\chi^2(1) = 9.513$, $p < 0.005$). We also found a significant interaction between VERB and FINITENESS ($\chi^2(3) = 118.66$, $p < 0.0001$). We then zoomed in on the critical MS ANSWER condition, as shown in Figure 3.2. Here too, all factors were significant: VERB ($\chi^2(1) = 18.892$, $p < 0.0001$), WH-WORD ($\chi^2(1) = 6.61$, $p < 0.05$), and FINITENESS ($\chi^2(1) = 156.7$, $p < 0.0001$).

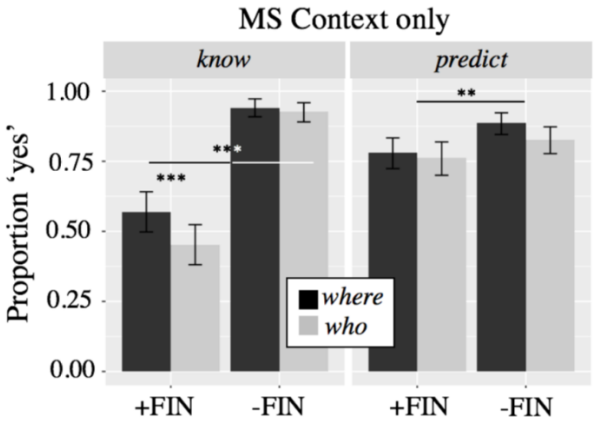


**Figure 3.2:** Experiment 1 results for MENTION-SOME given three question form factors.

### 3.2.5 Discussion

The results of Experiment 1 confirm aspects of the Question Form Hypothesis, because linguistic form factors significantly affect the acceptability of mention-some. However, these question form modulations did not reveal strict and categorical acceptance or rejection of mention-some, but rather revealed relative differences of greater or lesser acceptability. First, a significant effect of WH-WORD revealed that where questions were accepted more than who questions on mention-some readings. This confirms the wh-word hypothesis, which was based on observations from Ginzburg (1995) and Asher & Lascarides (1998) that these two types of wh-questions are not on the same footing with respect to (non-)exhaustivity. Second, a significant effect of MATRIX VERB revealed that know questions were accepted less than predict questions on a mention-some reading. This confirms the matrix verb hypothesis, and further supports the intuition that the two verbs impose differing restrictions on the semantic properties of their complements. However, these results do not reveal the locus of these restrictions, whether in the complement itself or in the lexical semantics of the verb.

The FINITENESS manipulation tested the predictions of the modal hypothesis. We indeed found a significant effect of FINITENESS, which supported the modal hypothesis. While the presence of the -FIN embedded clause significantly boosted MS acceptability in comparison to +FIN clauses, the acceptance rate in +FIN clauses was far from 0%—indeed, it was over 50%. Thus, a modal is not necessary for the acceptability of

mention-some, but significantly improves it.

Further, an interaction between MATRIX VERB and FINITENESS revealed that the effect of FINITENESS was driven by know-wh. While targets with both verbs showed a significant effect of FINITENESS, this effect was much larger with know-wh. This further supports the idea that some kind of restriction is imposed by know but perhaps not by predict. If mention-some were truly unavailable in +FIN clauses, we would expect equal or more rejection in predict-wh. These results could then be taken as providing initial evidence in support of accounts based on across-the-board ambiguity (Beck & Rullmann 1999; Lahiri 2002; George 2011, Ch.2) or underspecification (Ginzburg (1995); Asher & Lascarides (1998); van Rooij (2003), 2004), both of which predict mention-some to be just as available as mention-all, modulo verb restrictions and contextual licensing.

We have captured an asymmetry in acceptability between modal and non-modal questions, supporting modal theories. The theoretical import of 50% acceptability of mention-some in non-modal questions is in question. On the one hand, if this is interpreted as a low number, it could be argued to support the lack of a grammatically available mention-some reading (supporting modal theories). On the other hand, if the number is interpreted as high, it could be taken to support a semantics that allows grammatical mention-some in non-modal questions, but requires more explicit contextual support.

In the first case, the explanation for why participants *accept* at all might appeal to a notion of *tolerance* (Xiang, p.c.). Mention-all is consistently accepted at a high rate because without a modal, the mention-all answer is the answer that asymmetrically entails all other answers (using the Fox/Xiang notion of answer informativity)

In the second case, the explanation for why participants *reject* at all, might appeal to contextual factors which are required (Ginzburg (1995); Asher & Lascarides (1998); Beck & Rullmann 1999; van Rooij (2003), 2004; George 2001, Ch. 2; noted in Dayal (2016), 71-82). Mention-all is consistently highly rated, perhaps because, as van Rooij (2003) notes, this kind of answer/interpretation is the most (Shannon) informative,

and that fact could be explained in the semantics (van Rooij's semantics) or as a general principle of rational communication invoking maximizing the conveyance of information (Grice's Quantity Maxim, compatible with the semantic theories previously mentioned).

Both these explanations end up appealing to context. Note that participant tolerance of mention-some has a different behavioral signature than tolerance to false answers: the latter are rejected near floor, a clear case of a bad reading with a tasking of participant tolerance.

The comparison to know-wh responses in mixed true and false conditions (MA+FR) is salient. There we saw participants reject targets more often than not. This fact is often taken to license the inference that know-wh does not permit false answers, and the little participant acceptances are called *tolerance*. Thus, more rejection than acceptance in this instance licenses an inference on the part of the experimenter or theoretician about the grammar. Why can we not make a similar inference in the MS condition? Perhaps because the acceptance in non-modal know-wh is at 50%

Now the comparison to predict is informative with respect to this question. Many acknowledge a non-factive and a veridical reading of predict-wh (Kleindinst & Rothschild 2011; Spector 2006; Spector & Egré 2015). We found that participants accepted MA+FR conditions—which are true on a non-factive reading—around 50%. There is a true (grammatical) reading of the question, but it did not yield near-ceiling responses because there was another *false* (grammatical) reading of the question. From this result, we would not reject an underlying non-factive reading, but rather suggest that participants differed on which reading they based their responses on. Participant preference are the cause of 50% response rates. Mention-some interpretations were different from these. Mention-some interpretations were accepted much *more* than 50%. Can we now then make the inference that there is a grammatical mention-some in non-modal questions?

A further explanation for the degraded acceptability of non-modal mention-some is that test items did not provide enough contextual support for the mention-some

reading (Dayal (2016): 71-82). If, as Asher & Lascarides (1998) note, questions exhibit baseline interpretations for exhaustivity or non-exhaustivity, then we would expect that questions may fall one way or the other based on those preferences without context to direct interpretation (echoing Dayal (2016): 71-82). Since Experiment 1 did not manipulate context in this way, this explanation is quite plausible especially when coupled with hearer preferences as discussed above. The mention-some reading is context-dependent no matter how you look at it. For *all* theoretical accounts, context must disambiguate or precisify the underlying representation. While it is difficult to tease these two theories apart, it is possible to investigate whether and to what extent context does drive interpretation.

### 3.3 Experiment 2

In Experiment 2, we investigate the role of the questioner's goals by manipulating contextual information. We operationalize a notion of "what's-at-stake" to track contextual goals that license mention-some and mention-all. A HIGH STAKES context is one where human health or lives are at risk, while a LOW STAKES context is one without any such life-threatening issue (for instance, choosing a good diner or hair salon). By design, in our HIGH STAKES contexts, the goal is to save human lives, which we expect to be an exhaustive goal (to save *all* the human lives). Thus, we expect a mention-all answer to be the most informative. In contrast, our LOW STAKES contexts (in which no human lives are at risk) present goals where multiple answers are possible, thus an mention-some answer not only suffices, but may be preferred.

We note two things about our notion of STAKES. First, it is not isomorphic to exhaustivity. It is possible in principle to have a HIGH STAKES context where the questioner's goals are non-exhaustive, and a LOW STAKES context where the goals are exhaustive. For example, this difference might arise under constraint of time pressure, or where a HIGH STAKES goal may only be satisfied by a single person, etc. Nonetheless, one might assume that in most cases, when the stakes are HIGH, one values above all an answer that is not only true, but thorough. Indeed, we take care to design our

contexts so this is systematically the case. Second, we recognize that what counts as HIGH or LOW STAKES is also context-dependent. Nonetheless, this approach at least gives us a first look at the contribution of one way in which contexts could manifest speaker goals, and influence (non-)exhaustivity in answer reports.

In addition to this goal-oriented manipulation of context, we introduce a more fine-grained manipulation of ANSWER TYPE. We manipulate ANSWER TYPE in two ways. First, we include both singleton (mention-one, mo) and intermediately non-exhaustive (mention-some, MS) answers. If the acceptability of non-exhaustivity is parametric on a notion like informativity relative to a goal, then intermediate non-exhaustivity should be more acceptable than singleton non-exhaustivity because simply, the former will carry more (Shannon) information than the latter. This result might be taken to show support for a theory like van Rooij's (2003, 2004) decision-theoretic semantics. Additionally, Xiang's (2016) semantics assigns a special grammatical status to these singleton mention-one answers for modal-questions on their mention-some interpretation. Recall the discussion from Chapter 2 Section 2.4.2. Intermediate non-exhaustive answers will not be in the question denotation on its mention-some reading (these are contradictory), and will be considered partial answers on the (conjunctive) mention-all reading.

In an acceptability judgement task, will we find participants differentiating degrees of *non*-exhaustivity? Xiang (p.c.) notes that, since intermediate non-exhaustivity entails singleton non-exhaustivity (mention-one), a task where truth is under question will not be able to differentiate these two interpretations; an observation of acceptability of intermediate non-exhaustivity could thus appear in virtue of the fact that this *interpretation* (verified by a context) entails a grammatical *reading*. For the moment, we keep this point in mind and include the manipulation to establish an empirical base for these issues.

We also manipulated the level of informativity of mo/MS answers, relative to some ranking. To give an intuitive example, recall the tourist asking, Where can I find coffee? Given her likely goals, a hearer should mention a *nearby* or *local* coffee shop, rather

than one on the opposite side of town. While still an answer, the *nearby* coffee shop is more relevant, or *informative* to the tourist than the shop across town.

Note that the notion of INFORMATIVITY we introduce here does not map directly on to the logical notion employed by semanticists, and encoded in an answerhood operator that picks out the maximally informative answer (Fox 2014). Recall (88). There, a maximally informative answer is one that is not asymmetrically entailed by another answer. Essentially, the answer set denoted by a question may or may not have multiple such elements—in an exhaustive question, there will only be one, while a mention-some question allows for more than one.

This notion of informativity treats all non-exhaustive answers the same as long as they have the same number of answers: for a question like there *Where can I find coffee?* is no distinction (entailment) between an answer naming *Peet's* from an answer naming *Stumptown*. In our intuitive example, if Peet's is closer than Stumptown, we consider *Peet's* to be a more informative answer than *Stumptown*.

Thus, the exhaustive answer carries all the information (available in the context) by naming all the answers, while a non-exhaustive maximally informative answer will name one or two answers which carry a high amount of information, and a non-exhaustive minimally informative answer provides one or two answers which carry a minimal amount of information. This will become clearer when we present our stimuli in Section 3.1.

### 3.3.1 Design and Materials

As in Experiment 1, in Experiment 2 we manipulated FINITENESS in target sentences as a between-subjects factor. The –FIN target condition contained three within-subjects factors: ANSWER TYPE (MENTION-ALL (MA), MENTION-SOME (MS), MENTION-ONE (MO), and FALSE REPORT (FR)), INFORMATIVITY (MAX, MIN, for MO/MS ANSWERS), and STAKES (HIGH, LOW). The +FIN target condition had two within-subjects factors (ANSWER TYPE and INFORMATIVITY). We did not manipulate STAKES in the +FIN condition, and only targeted LOW STAKES for the following reason: we predicted that in a

HIGH STAKES condition, an MA answer would be favored. Given that +FIN embedded questions already favor MA answers, we would predict to see little to no change. The question is whether a LOW STAKES context can pull answers away from MA towards MO/MS. For both conditions, we included both MENTION-SOME ANSWER TYPES, where two answers were given, and MENTION-ONE ANSWER TYPES, where a singleton answer was given.

Each context featured a main topic, a main character conducting a search form some contextually-relevant information, and a set of ranked entities relevant to the topic. The main character in search of the information posed a wh-question to a group of individuals, who then each provided an answer connected to the ranking. The participant's task was to evaluate the knowledge of these individuals, based on their answers and the given context.

The ANSWER TYPE and INFORMATIVITY manipulations yielded 6 possible answers in the set of answers, which were randomized so that the same answers did not always appear together, and so that participants would see different answers for each story. Thus, there were six answer type permutations. At any given time, only three ANSWER TYPES were randomly displayed by an algorithm, in order to reduce the cognitive load on the experimental participants, and to ensure that it was not the case that the same types were always pitted against each other (thereby forcing certain comparisons and reducing the probability of a response bias from surfacing on every trial). An example of a HIGH STAKES and a LOW STAKES trial type follow.

(106) HIGH STAKES
Scientists have discovered a new strain of a dangerous virus that has contaminated oysters in the Mid-Atlantic. The Center for Disease Control is trying to prevent as much contamination as possible by tracking down all the oysters. In this area, luckily only 6 restaurants usually buy oysters from the contaminated area: Restaurant A ordered 10 crates, Restaurant B ordered 8, Restaurant C ordered 5, Restaurant D ordered 2, Restaurant E ordered 1, Restaurant F ordered 0.

The supervisor for this county asks his inspectors, "Where should we check for contaminated oysters?"

Inspector A says, "Restaurant A, B, C, D and E." MA

Inspector B says, "Restaurant A." MO-MAX
Inspector C says, "Restaurants D and E." MS-MIN

**Who knows where to look for oysters? (Choose all that apply.)**

(107) LOW STAKES
Johanna is new to Minneapolis and wants to try local coffee shops. The Ultimate Coffee Guide 2018 ranks cafes on a ten-point scale, where ten is the highest number of points. Minneapolis has the following ranking for coffee roasteries: Café A has 10 stars, Café B has 8, Café C has 5, Café D has 2, Café E has 1, Café F has 0.

Johanna asks three of her classmates originally from the city, "Where should I go for coffee?"

Classmate A says, "Cafés A and B." MS-MAX
Classmate B says, "Cafés E." MO-MIN
Classmate C says, "Café F." FR

**Who knows where to go for coffee? (Choose all that apply.)**

At the end of each trial, participants were instructed to answer the question about the individuals' knowledge by choosing all that apply. There was also a *None of the above* option. This multiple-choice question allowed participants to choose more than one answer, allowing us to determine if multiple answer types were permitted in a given scenario.

### 3.3.2 Participants

318 native speakers of English participated. The study was constructed and administered using Alex Drummond's Ibex Farm platform. Participants were recruited online through Amazon Mechanical Turk. IP addresses were restricted to US only, and further questions were included to ascertain native speaker status. 6 participants were removed for browser incompatibility issues, and 6 participants were removed for non-native English speaker status.

### 3.3.3 Predictions

This study tests the extent to which the interpretation of embedded questions is driven by contextual demands over and above any restrictions imposed by the linguistic form of the question. We thus test the Context-Sensitivity Hypothesis detailed in Section 3.1.2. Showing that interpretation is context-sensitive will not tell us whether the underlying semantic representation is underspecified (Ginzburg (1995), Asher & Lascarides (1998); van Rooij (2003), 2004) or ambiguous (Beck & Rullmann 1999; Lahiri 2002; George 2011, Ch.2), but it may reveal the necessity or predominance of context over linguistic form. If interpretation is sensitive to context as predicted by the Context-Dependence Hypothesis, MS/MO answers will be accepted in the LOW STAKES condition, and more than in the HIGH STAKES condition. These theories do not predict variability to any question form manipulation.

Regarding the manipulation of answer INFORMATIVITY, we expect MAX-INFORMATIVE MS/MO to be more acceptable than MIN-INFORMATIVE ones, because by design maximally-informative answers are better question resolvers than minimally-informative ones. Additionally, we test both singleton and intermediate non-exhaustivity by manipulating MO and MS answers.

On van Rooij's (2003, 2004) semantics, mention-some answers will never be more useful than mention-all answers in any context, from the sheer fact that a mention-all answer contains more information than a mention-some context. While a mention-some answer may be *equal* to a mention-all answer in utility, it will never exceed the utility of a mention-all answer. This will track the judgements in acceptability of these answers. Van Rooij's Utility Hypothesis thus predicts that both (MAX-INFORMATIVE) MO and MS will be accepted less than MA answers, even in the LOW STAKES condition.

While only mention-one answers are grammatically avaialable in modal questions in low-stakes contexts on some modal theories (Xiang (2016)), we might expect that intermediate non-exhaustive answers to be dispreferred because there is no semantic representation to support them. However, Xiang (p.c.) predicts that these will be acceptable because they entail the grammatical mention-one reading. In contrast, van

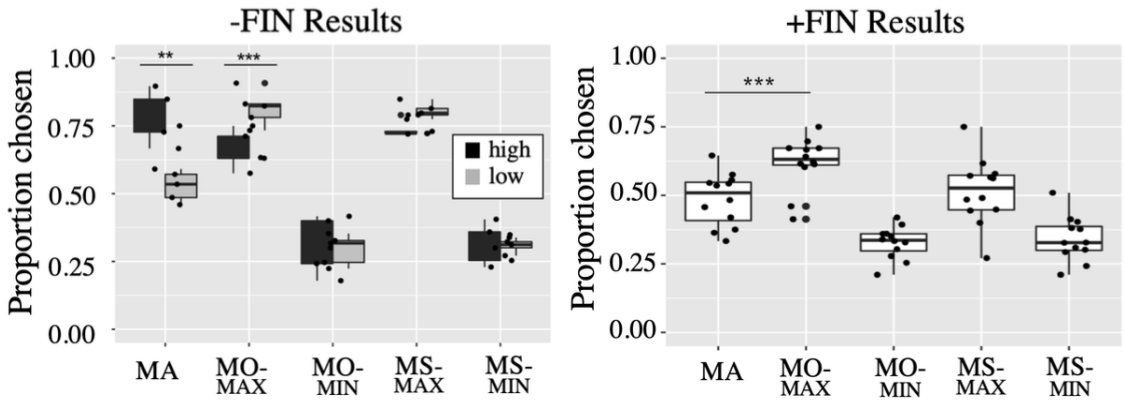


**Figure 3.3:** Experiment 2 Results.

Rooij's semantics would predict that intermediate non-exhaustive (MS) answers will be preferred to singleton (MO) answers because they are more informative.

### 3.3.4 Results

Experiment 2 results are in Figure 3.3. We find an overall effects of ANSWER TYPE ($\chi^2(2) = 14.626$, $p < 0.001$), FINITENESS ($\chi^2(1) = 19.547$, $p < 0.0001$), INFORMATIVITY ($\chi^2(1) = 664.8$, $p < 0.0001$) across the entire sample. We also find an interaction between ANSWER TYPE and INFORMATIVITY ($\chi^2(3) = 667.2$, $p < 0.0001$), which reveals that participants' choice of MO/MS answers was significantly affected by whether they were MAX or MIN informative: in particular, MAX answers were accepted significantly more than MIN answers for both MO ($p < 0.0001$) and MS ($p < 0.0001$). However, overall MAX MS/MO answers did not differ from each other ($p$=0.1).

There was also an interaction between ANSWER TYPE and TENSE ($\chi^2(5) = 52.362$, $p < 0.0001$), but this should not be surprising since participants in the -FIN condition saw both HIGH and LOW STAKES scenarios, while participants in the +FIN condition saw only LOW STAKES scenarios.

Beginning with the -FIN condition, we also find effects of ANSWER TYPE ($\chi^2(2) = 16.675$, $p < 0.001$), INFORMATIVITY ($\chi^2(2) = 16.675$, $p < 0.001$), but surprisingly there was no main effect of STAKES alone ($\chi^2(1) = 0.84772$, $p$=0.4). STAKES was only significant in interaction with other variables (2-way with ANSWER TYPE $\chi^2(5) = 49.003$,

$p$ <0.0001; 3-way with ANSWER TYPE and INFORMATIVITY $\chi^2(7) = 569.47$, $p$ <0.0001). This reveals the following: MO-MAX answers are significantly more acceptable than MA answers in LOW STAKES scenarios ($p$ <0.0001), but MA answers are so in HIGH STAKES ($p$ <0.001). The effect of stakes almost disappears in MS MAX answers ($p$=0.05). Further, we see that in HIGH STAKES, no comparison between MA and MS MAX revealed significant differences, but in LOW STAKES, participants did differentiate them, choosing MS MAX significantly more than MA ($p$ <0.0001). Incidentally, in LOW STAKES, choice of MS-MAX was not significantly different from choice of MO-MAX ($p$=0.7), but was in HIGH STAKES ($p$ <0.05). MS-MAX answers pattered with MO-MAX in LOW STAKES, and MA in HIGH STAKES.

In the +FIN condition, we find significant effects of ANSWER TYPE ($\chi^2(3) = 6.5063$, p<0.05), and INFORMATIVITY ($\chi^2(1) = 157.43$, p<0.0001), and an interaction between the two ($\chi^2(3) = 170.44$, p<0.0001). As in the -FIN Condition, we see significant differences between MO-MAX and MA ($\chi^2(1) = 18.814$, $p$<0.0001). Unlike in the -FIN condition, the difference between MS-MAX and MA was not significant ($\chi^2(1) = 0.72216$, $p$=0.4), but the difference between MS-MAX and MO-MAX was ($\chi^2(1) = 16.372$, $p$<0.0001).

In the LOW STAKES condition, we see a significant effect of FINITENESS ($\chi^2(1) = 25.004$, $p$<0.0001), a two-way interaction with ANSWER TYPE ($\chi^2(5) = 32.64$, $p$<0.0001), and a 3-way interaction with ANSWER TYPE and INFORMATIVITY ($\chi^2(9) = 549.5$, $p$<0.0001), revealing that participants choice of answers depended on both whether the answer was MAX or MIN INFORMATIVE and whether the participant was in the +FIN or -FIN condition.

We removed one trial from the +FIN Condition from our analysis because it turned out to be HIGH STAKES, and we were only interested in LOW STAKES for this condition. We discuss this more in Section 4, where we provide an in-depth post hoc analysis of trial effects to gain insight into the contribution of world knowledge.

### 3.3.5 Discussion

The results reveal the following. In the -FIN condition (left graph), MA answers were preferred over MS and MO answers in HIGH STAKES contexts. As expected, mention-all answers were preferred when contextual goals are exhaustive. However, acceptance of MS-MAX answers did not significantly differ from MA answers. As the informativity of a mention-some answer increased, the answer contains more information relevant to the contextual goals. These two points suggest that participants calculate a threshold whereby answers that approach exhaustivity are as acceptable as exhaustive answers.

In the LOW STAKES -FIN condition, MS-MAX and MO-MAX answers were both preferred over MA answers. Even in the +FIN condition, MA answers did not consistently win out: here, too, participants accept a maximally informative MS or MO answer more than an MA answer. These results bear against van Rooij's Utility Hypothesis. While mention-one/mention-some answers may at most equal the utility of mention-some answers, according to van Rooij (2003, 2004), participants judged MO-MAX/MS-MAX significantly *more*—not equally—acceptable than MA. Additional notions are required to capture the behavior of hearers when they resolve the exhaustivity of wh-questions.

Why might this be the case? On the one hand, van Rooij predicts that ma answers should always be preffered because they are the most informative. It's possible with an cost function that penalizes utterance length (or a notion of representational complexity where MA interpretations are more complex than MS one), one could derive higher utility for mention-some answers over mention-all answers because the latter are penalized. Note however, that in a sense this may reveal the complexity of defining the constrains on decision problems, known from the literature in behavioral economics.

At the same time, the context-dependence of exhaustive answers seen here contradicts context-independent semantics that is exhaustive, and the idea of defeasible baseline interpretations is vindicated (cf. Asher & Lascarides (1998)). While exhaustivity may be a default inference, perhaps for reasons of maximizing informativity in a semantic or Gricean manner, we can see that contexts can also render mention-all readings dispreferred to mention-some (Asher & Lascarides (1998), Schulz & van

Rooij 2006, Spector 2007, Zimmermann 2010). Perhaps then, this should be taken as evidence for a non-exhaustive underlying semantics, especially given the difficulty in establishing strong exhaustive answers to how- and why questions, and the fact that we find this effect across modal and non-modal questions.

Further, there is no significant difference between MO and MS (on matched INFORMATIVITY), meaning that participants treat mention-some akin to mention-one. While we don't see across the board rejection of mention-some in the +FIN condition, we see a higher proportion of acceptance of MS-MAX/MO-MAX in -FIN condition than in the +FIN condition. The presence of a modal boosts the acceptability of mention-some readings, but the absence of one does not block them. This is compatible with modal theories

Participants' choice of MS-MAX answers is critical in this experiment, and I argue reveals that participants calculate a threshold of exhaustivity as acceptable, given contextual demands. In LOW STAKES, MS-MAX is as good as MO-MAX, and in HIGH STAKES, it is as good as MA. The result mirrors that of Phillips & George (2018), where the number of false beliefs an agent has proportionally determines the acceptability of a know-wh report: the more false beliefs, the more reject and the less false beliefs the less rejection. We reveal the effect for both answer informativity and answer exhaustivity, parameterized to context. Contextual goals determine the threshold which specifies how many answers are enough.

Experiment 1 showed us that the linguistic form of the question affects the availability of a mention-some reading: we saw significant effects of WH-WORD, MATRIX VERB, and MODAL. However, these effects were not on the magnitude of unacceptable and acceptable. With the modal particularly, we noted that the presence of one was not necessary for participants to access mention-some, but that the lack of one led to degraded responses (but, not categorically so). Given the Context-Sensitive Hypothesis, we predicted that given proper contextual support, non-modal mention-some would receive more judgements of acceptability. In Experiment 2, we found that +FIN (non-modal) questions received degraded judgements for all answer types, relative to

-FIN (modal) questions. Yet, participants' preference for mention-some answers over mention-all answers in LOW STAKES was still significant.

These findings demonstrate that the discourse context provides central information that is relevant to resolving a given question, including which answers are most informative and how much information is needed to resolve exhaustivity and the questioner's goal. This effect of context was observed regardless of finiteness in the embedded clause, and therefore did not depend on the presence of a modal element to license either an MO or MS answer. Finally, our results show no degradation in (maximally informative) non-exhaustive *non-singleton* answers. This result held even when participants could explicitly compare MS and MO answers. If such MS answers can signal ignorance, our empirical results show that they do neither consistently nor obligatorily do so.

These results are compatible with two approaches to the interpretational variability seen in embedded question interpretation: a semantics that encodes across-the-board ambiguity with respect to (non-)exhaustivity and a semantics that is underspecified for it. Whatever mechanism(s) are involved in this process of integrating contextual information, more information about the behavioral signatures of precisification and disambiguation would be incredibly illuminating to resolve such a theoretical stalemate. However, perhaps we have accrued some evidence for underspecification in the threshold-like behavior of Experiment 2. One could argue that the economy of mechanisms posited to capture a threshold-like result would be greater in an underspecified semantics than in a semantics which delivered a representation for each reading. Given that parsimony is often a theoretical boon, an underspecified semantics then may be more parsimonious than a semantics that posits ambiguity.

It is always a possible move to reject the connection between semantic and structural features posited by semanticists and the behavioral measures used in this experiment. If one rejects the proposition that participant judgments of non-modal mention-some reveal that the semantics makes such meaning available for that structure, then why would we accept such a measure as telling us anything about semantics more

generally? To make this move is to relinquish the methodology of acceptability judgements as a reliable indicator of semantic representations.

I'd like to suggest that linguistic form factors are cues that help a hearer determine how to resolve (non-)exhaustivity in a WH-question. The story (perhaps a familiar one) goes as follows. Speakers often utter expressions whose meanings are underspecified. A hearer who is uncertain about the meaning of the speaker's utterance, must recruit a wide range of information, about which the hearer might also be uncertain, to understand the speaker's meaning. In the case of a question, many elements of meaning are underspecified by the pronunciation of the question itself, including (non-)exhaustivity, granularity, whether the reference should be construed *de re* or *de dicto*. In this respect, questions are no different from other elements of language like quantifiers and descriptions. A hearer who is asked a question, must determine how to answer by also determining the speaker's goal in asking the question. In resolving the goal behind the question, the hearer determines how exhaustively to answer. A speaker, for her part, can make the job easier for the hearer by signaling that her goal is exhaustive or non-exhaustive—by providing *cues to her goal*.

Recall that languages may encode (non-)exhaustivity in the form of particles which are non-exhaustive or exhaustive. In English, a speaker can signal non-exhaustivity by using some or any, exhaustivity by using all, and even to signal a restricted domain by using phrases like local or nearby. Likewise, we saw earlier that other languages also allow such particles, like German so and alles. An existential modal like can would also be such a cue.

Further recall Asher & Lascarides' (1998) suggestion that questions may exhibit default preferences for exhaustivity or non-exhaustivity (the following discussion is summarized from Asher & Lascarides (1998), pp.269). They derive this defaultness from a general principle of interpretation, whereby one picks the interpretation that logially entails all others (Dalrymple et al. 1998's Strongest Interpretation Principle). With one caveat: when we evaluate a knowledge claim, the interpretation must be compatible with the current *cognitive task*. For know-who-questions, the cognitive task

is often compatible with the strongest interpretation, an exhaustive one. In most cases, it is reasonable that the attitude holder has exhaustive knowledge. If Dana and I are discussing a party had amongst our mutual friends, it's reasonable that Dana's knowledge be exhaustive in (108) because we are interested in our mutual friends which is plausibly a small set of people.

(108) Dana knows who attended the party.

(109) Dana knows who attended Elton John's post-Oscar party.

However, if Dana is a celebrity gossip columnist, then her knowledge need not be exhaustive for (109) to be true. She need not know whether non-celebrities came to the party. So, the cognitive states of the speakers constrain exhaustivity. With know-how-questions, and even know-where questions, Asher & Lascarides note, this effect is much greater given the vast number of possible ways to get to the treasure, or to be (precisely) located.

(110) Dana knows how to get to the treasure.

(111) Dana knows where she is.

Given the context-dependence of (non-)exhaustivity, when we interpret a root or embedded question out of context, we naturally retrieve a context against which to evaluate the question for (non-)exhaustivity and fill in the missing information about contextual goals. Whether we judge a question exhaustive or non-exhaustive will depend on aspects of the context we have retrieved. This point was brought out by examples like the following:

(112) Who has a light?

(112) is naturally construed as non-exhaustive because our world knowledge of smokers needs tells us that the person who asks (112) only needs one light. Of course, as social attitudes change, we would expect this context to become less available as smoking becomes less common. This is how world knowledge works. Indeed, when I presented this example to our lab manager Taylor Martinez (who is about ten years

my junior), she immediately thought of flashlights or lighting a candle. This is still a non-exhaustive goal (you only need one), but the ready context for her was different from mine (and the common one in the literature), because of our different exposures to smoking. When I was growing up, smoking was still permitted in many restaurants and public buildings. Between the years of 2004 and 2007, many states banned smoking indoors, such that in the intermittent years, it became rare to see smoking indoors. The fact that our differing experience led us to retrieve different contexts in which such a question would be uttered exemplifies the point I am making. No context was explicitly provided to give us a goal, yet our experience and expectations provide us with the missing information. The most likely goal for her was different than the most likely goal for me.

What cues us in reconstructing the relevant missing information? In part, the information contained in the predicate have a light. Our knowledge of the meaning of that predicate includes several different scenarios given the polysemy of light. These resolutions may include different goals, which derive from our experience and knowledge of the world and how speakers use questions to convey their goals.

In the next chapter, I'll present a corpus analysis and related studies which attempt to quantify some of the factors which could influence how we retrieve contexts. First, we can examine the effects of particular trials could reveal how prior world knowledge associated with the contexts and embedded predicates used in our experimental stimuli might lead to participants' acceptance or rejection of mention-some.

## 3.4 Default Preferences: Qualitative analyses of context, goals, and world knowledge

The analysis presented in this section attempts to provide a more in-depth look at variability in mention-some due to the effects of the particular contexts presented in Experiments 1 and 2. This analysis investigates the contributions stemming from the

constructed contexts and/or from the embedded predicates used in experimental targets. It will deepen our understanding of the way in which context influences interpretation in embedded questions.

### 3.4.1 Effects of Trial in Experiment 1

Experiment 1 tested 8 different embedded verbs per wh-word. Participants saw each embedded verb twice because the MATRIX VERB comparison (know vs. predict) was within-subjects. These verbs are presented again in (113), and represent the spread of topics of the different scenarios in which embedded question reports were evaluated. These different scenarios were coded as STORY.

(113) a. who: recruit, interview, invite, ask, call, hire, select, contact

b. where: find, view, store, locate, hide, bury, sell, display

Figure 3.4 presents participant responses in each different scenario, for the critical MS condition. A Kruskal-Wallace test reveals no significant effect of STORY across the entire sample ($\chi^2(15) = 22.034$, $p = 0.1$), including the other answer factors. However, in the MS ANSWER Condition alone, we do see a significant effect of STORY ($\chi^2(15) = 22.034$, $p = 0.01$), and an interaction with wh-word ($\chi^2(15) = 29.609$, $p = 0.01$). The interaction reveals that the effect of STORY is driven by where-questions ($\chi^2(7) = 15.09$, $p = 0.03$).

The where-question scenarios reveal this variability due to STORY (the bottom graph). Contrast WHERE-HIDE and WHERE-VIEW with WHERE-SELL. The first two show a small effect of FINITENESS, while the latter shows a much larger effect. We also see the interaction with MATRIX VERB more in WHERE-LOCATE and WHERE-FIND.

In contrast, who-questions (the top graph) show a consistent effect of FINITENESS and the interaction with MATRIX VERB, where finite know-questions are rejected more on mention-some than finite predict-questions.

Consider WHERE-VIEW and WHERE-HIDE, in which we see little difference between finite and non-finite clauses. In our design, participants saw two tokens of each embedded verb (thus two difference stories about, e.g., hiding) because the MATRIX VERB

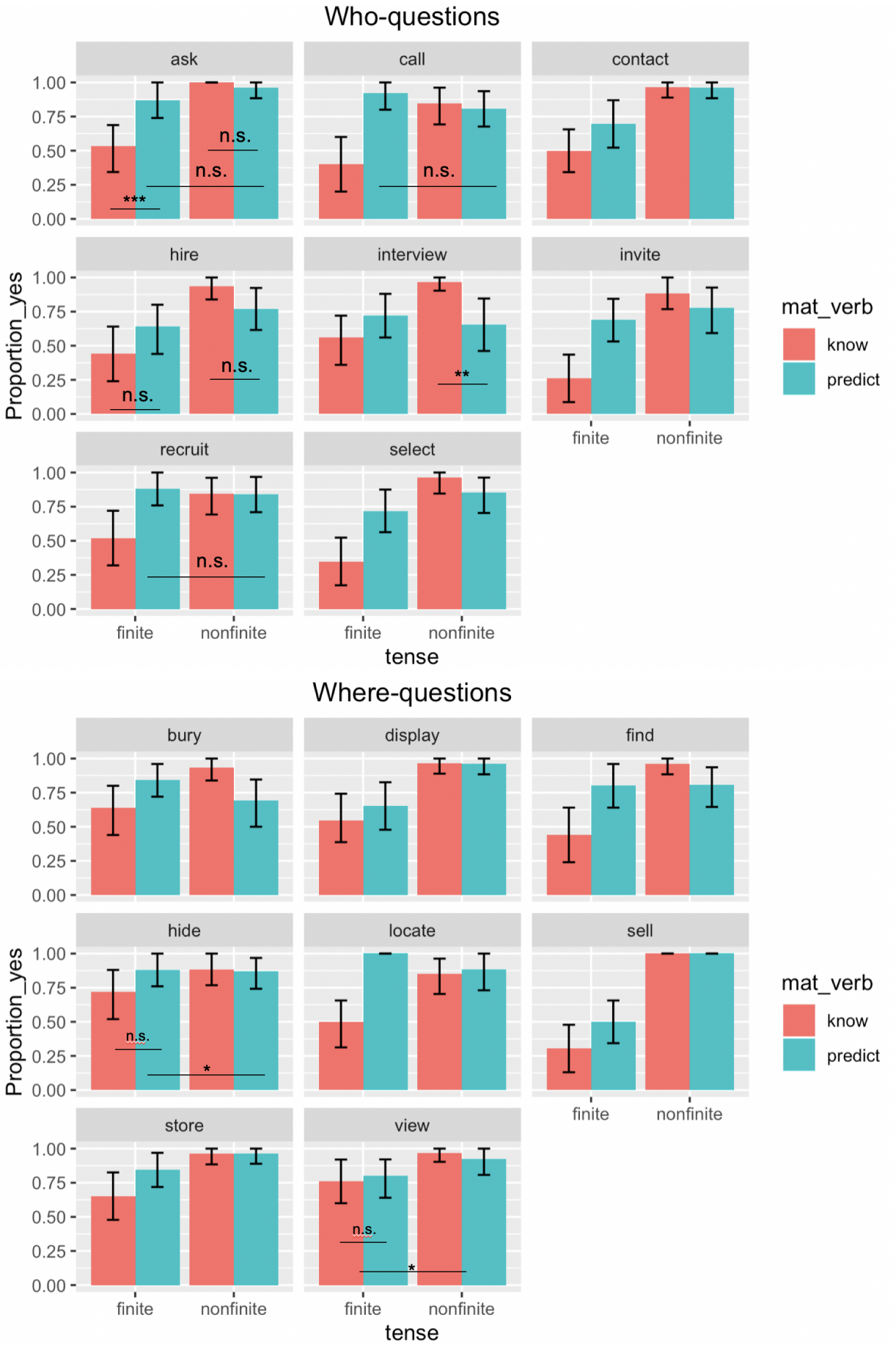


**Figure 3.4:** Graphs of participant responses per STORY, in the MS ANSWER condition only.

comparison (know vs. predict) was within-subjects. While embedded verb was a fully crossed factor, individual scenarios were not because this was not a factor of main interest, and because adding STORY as an additional factor in the Latin square would have expanded the study greatly. Thus, one scenario always appeared with know-wh, while the other always appeared with predict-wh.

The stimuli for those trials are presented below. Let us review where-view first.

(114) **KNOW-WHERE-VIEW - FINITE**
Zack and Joe are going on vacation to California and want to see the Hollywood sign.

Zack's mom asks his dad, "Where do you think they will view the sign?"
His dad responds, "I think Beachwood Canyon."

Actually, they go to Beachwood Canyon, Lake Hollywood Park, and the Griffith Observatory, but not Downtown or Runyon Canyon Park.

Mom reports, "Dad knows where they viewed the Hollywood Sign."

Is Mom right?

(115) **PREDICT-WHERE-VIEW - FINITE**
Mary has a special bucket list item she wishes to check off her list, and first on the list is to view the Northern Lights. She and her friend Abby are going on a road trip, from Maine to Alaska, and stopping in New York, and Ohio before going up to Minnesota and then Canada.

Mary's mom asks her dad, "Where do you think they will view the Northern Lights?"

Her dad responds, "I think Alaska."

In fact, they did view the Lights in Minnesota, Canada and Alaska, but not in New York and Ohio because it was too cloudy.

Mom reports, "Dad predicted where they viewed the Northern Lights."

Is Mom right?

Note that the stories establish the mention-all answer—Zack and Joe see the Hollywood sign in three places (but not two others) and Mary sees the Northern Lights in three places (and not in two others). Yet, a mention-some answer is accepted 75% of

the time or higher.

What about these two scenarios is licensing mention-some? In both, the main character's goal is to view some sight. Of course, a viewing event may be satisfied by several repeat instances of the same viewing. I can look out my window several times a day and view the street corner, and each separate time I do so counts as a true case of viewing the street corner. Nothing here provides insight. Rather, it is the characters' particular goals in viewing sights *they have never before seen* that licenses mention-some interpretations. This is one common goal driving tourism and the creation of bucket lists. With such a goal, often a single event of viewing (probably the first one) satiates the goal. Of course, one could create a bucket list item to "see the Northern Lights as many times as possible", or be a tourist who ventures to see the Hollywood sign from every angle and perspective. These goals would then lend themselves to mention-all interpretations. However, I expect that they are less-typical goals for a tourist or one creating a bucket list.

(116) **PREDICT-WHERE-HIDE - FINITE**
Davey eats too much chocolate. He went to Costco the other day and bought a bulk package of Ferrer Roché. He asks his friend Mackenzie to hide the chocolate for him.

Ruth asks Davey, "Where do you think she will hide the chocolate?"
Davey says, "I think in the linen closet."

In fact, Mackenzie hid the chocolate in the dresser drawers, in the linen closet and inside Davey's suitcase, but not in the freezer.

Ruth reports, "Davey predicted where she hid the chocolate."

Is Ruth right?

(117) **KNOW-WHERE-HIDE - FINITE**
Chris is always taking Alice's snacks so she wants to hide them from him. Alice and Barbara conspire to hide her food.

Chris asks Geoff, "Where do you think they will hide the snacks?"

Geoff says, "I think in the mattress."

Actually, they hid the snacks in the mattress, and in the couch, but not in the

shoebox in the closet.

Chris reports, "Geoff knows where they hid the snacks."

Is Chris right?

In these scenarios, a reasonable mention-some goal is less forth-coming than in the previous ones. I'd like to suggest the following. When one is searching for hidden food, the natural reason behind the search is to eat the hidden food. Similar to the VIEWING scenarios, this goal is satisfied by a single event of finding hidden food, and most likely the first. Thus, knowing or predicting a mention-some answer here is sufficient because finding *some* hidden food is enough for the purposes of eating the food. This could especially be the case if participants interpret the event of searching depicted in the scenarios as occuring *right now*: the searcher wants food now, so the first place to find the hidden treats satisfies that craving. Of course, one might have an insatiable craving whose satiation requires all the hidden treats.

In both cases, these facts do not fall out of the question's meaning, but from the world knowledge associated with typical touristic goals/bucket lists or the (subjective) satisfaction conditions of a food craving.

Compare those two sets above to SELL, where we see a large effect of FINITENESS.

(118) **KNOW-SELL - FINITE**
Julia has just created a new perfume, and is deciding where to sell it. Her friend Paula is helping her market it. Julia's brother and sister are discussing the situation.

Her brother asks her sister, "Where do you think they will sell the perfume?" Her sister says, "I think the local boutique."

Actually, Julia ends up selling it at the local boutique, the pop-up and Perfumerie. But not at Macy's or Lord and Taylor.

Julia's brother reports, "My sister knows where they sold the perfume."

Is the brother right?

(119) **PREDICT-SELL - FINITE**
Dante and Bea are trying to sell a collection of ugly Christmas sweaters. They are trying to get permits to sell at the mall, at the local boutique, at a pop-up in

town, and online. Their friends, Felicia and Gabe are discussing whether they will be successful.

Gabe asks Felicia, "Where do you think they will sell the sweaters?"
Felicia says, "I think online."

As it turns out, they are unable to get a permit to sell at the mall, but are able to get permits to sell at the other places.

Gabe reports, "Felicia predicted where they sold the sweaters."

Is Gabe right?

In these two FINITE scenarios, participants accepted (118) 30% of the time and (119) 48% of the time on average. These responses are neither different from each other nor from chance.

What might the relevant goal here be? A natural reason to sell something is to make a profit. One makes more profit, the more one sells one's product(s). This is particularly clear in scenario (118), where the main character has created a new product. Its being sold at many places will increase the chances of it being a successful product that makes profit for the seller.

Note that even though this might be a likely goal for a selling scenario, participants accepted non-finite targets at ceiling. The non-finite scenarios are presented in (120) and (121):

(120) **KNOW-SELL - NON-FINITE**
Julia has just created a new perfume, and has permits to sell it at the Perfumerie, Macy's, Lord and Taylor, at a pop-up in town, and at the local boutique.

Actually, the best selling locations are the local boutique, the pop-up and Perfumerie. Macy's and Lord and Taylor sell only two bottles.

Her friend Paula asks her where to sell the perfume.
Julia says, "The local boutique."

Paula reports, "Julia knows where to sell the perfume."

Is Paula right?

(121) **PREDICT-SELL - NON-FINITE**
Danny is trying to sell his collection of ugly Christmas sweaters. He has permits to sell at the mall, at the local boutique, at a pop-up in town, and online.

He asks his friend Felicia to guess which places will be successful.
Felicia says, "The pop-up in town."

As it turns out, the online store and mall location did poorly, but the local boutique and the pop-up did very well.

Danny reports, "Felicia predicted where to sell the sweaters."

Is Danny right?

Why were NON-FINITE scenarios particularly compatible with mention-some? I present some possible explanations.

The first point concerns the felicity conditions of non-finite clauses. While we attempted to make minimal changes between the FINITE and NON-FINITE conditions, there were some changes that were necessary to make a finite or non-finite embedded question felicitous. In the FINITE stories, answers were factual—the events encoded by the embedded predicate happened. Thus, the questioner was asking about past events that were not ordered in any way. In contrast, the covert deontic modality encoded in the non-finite question (cf. Bhatt 1999; Kratzer 1989, 1991) led answers in this condition to take on a flavor of optimality or "bestness". While we tried to counteract this layer of meaning by explicitly stating that several answers were "best" (as can be seen in both (120) and (121)), perhaps the superlative meaning licensed a mention-some answer because typically a superlative picks out a single item. Further, its possible that this logic was reinforced by the fact that another character is making an assertion about the answerer's knowledge. So participants may have inferred that the answerer had knowledge of the single best answer to satisfy the superlative, thus explaining why they asserted a mention-some answer.

If this is how participants interpreted these non-finite scenarios, we might expect that mention-all answers would be degraded relative to mention-some answers because a mention-all answer does not satisfy this superlative meaning. If we take a look

at responses for these two stories across all ANSWER conditions in Figure 3.5, we do see that MA answers are accepted slightly less than MS answers, but not significantly so in either predict or know targets.

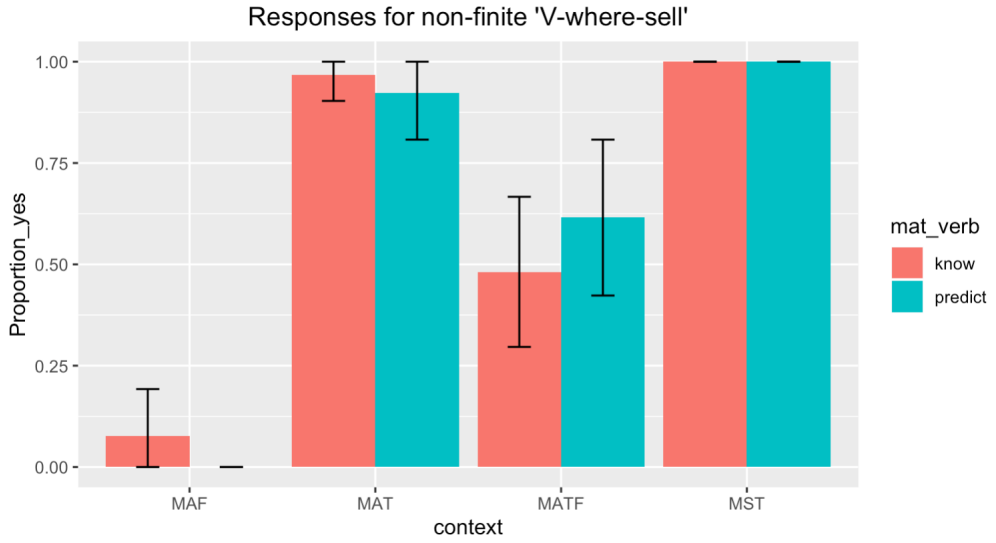


**Figure 3.5:** Graph of responses for v-where-sell in all ANSWER conditions.

There is another possibility. If someone is selling something, you would expect them to diligently research where to sell their product. It's possible that participants are thus interpreting the characters as having exhaustive knowledge in virtue of this fact even though they give a mention-some answer. This explanation would apply to (120) because the seller is the person whose knowledge we are evaluating, but it is not the case for (121).

### 3.4.2 Effects of Story in Experiment 2

A Kruskal-Wallace test reveals significant effects of STORY in the NON-FINITE ($\chi^2(15) = 22.034$, $p = 0.01$) but not the FINITE conditions ($\chi^2(4) = 5.9394$, $p$=0.2), as well as an interaction between STORY and ANSWER TYPE in the NON-FINITE Condition ($\chi^2(15) = 22.034$, $p = 0.01$). This is all good news, and unsurprising because we manipulated STAKES by creating stories that were HIGH and LOW STAKES. Thus, the effect due to particular stories is driven by this manipulation. Given that we did not manipulate HIGH STAKES in the FINITE Condition, we should not expect to see an effect of STORY

there, if the individual stories were consistent enough.

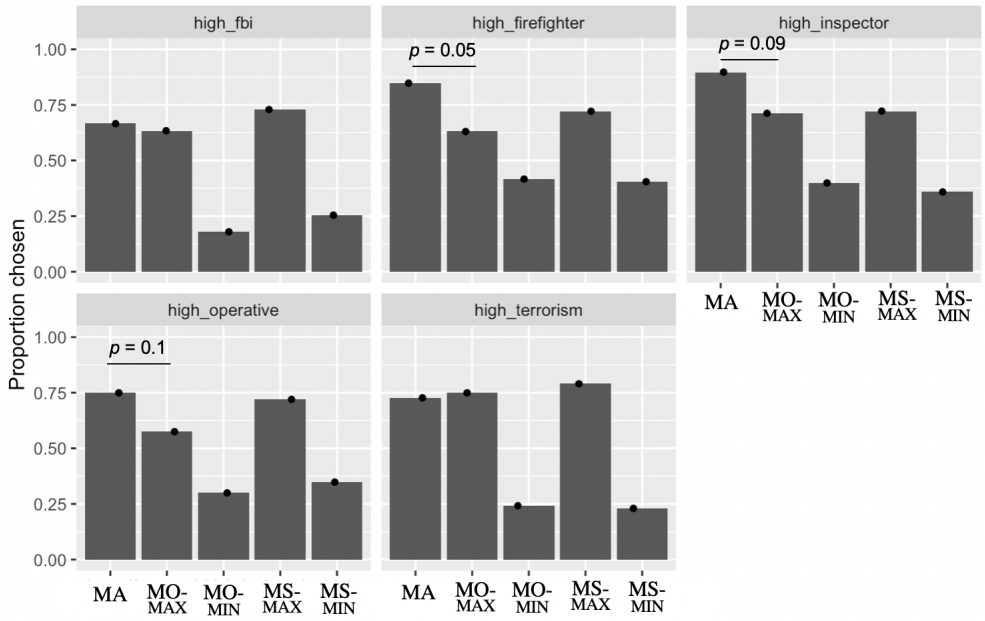


**Figure 3.6:** Results of Experiment 2, HIGH STAKES NON-FINITE Condition, broken down by STORY.

Let's first focus on HIGH STAKES, presented in Figure (3.6). Across the entire sample for this condition, the difference between MA and MO-MAX answers is significant ($\chi^2(1) = 6.0274$, $p$=0.01), but the difference between MA and MS-MAX answers is not ($\chi^2(1) = 0.381$, $p$=0.5). Further, while the graph appears to present at least trends towards significant differences for most trials, a Test of Equal or Given Proportions reveals that MA and MO-MAX/MS-MAX answers did not significantly differ in any individual trial. The exception was the FIREFIGHTER story, where only the MA/MO-MAX comparison was significantly different ($\chi^2(1) = 3.9894$, $p$=0.05).

The STAKES manipulation was designed to elicit differing responses between MA and mo-max(/MS-MAX) answers: the HIGH STAKES condition aimed to present exhaustive goals to favor MA, while the LOW STAKES condition aimed to present non-exhaustive goals to favor MO-MAX/MS-MAX. As Figure (3.6) shows, qualitatively it appears that participants did not distinguish between these two answer types in FBI and TERRORISM; indeed, there does not appear to be any trend toward a significant difference here. Further, in both trials, it even appears that MS-MAX answers are accepted

at a slightly greater proportion than MA answers. Let's examine these two stories more closely.

(122) **FBI**
A serial killer is on the loose, and has kidnapped several people. FBI forensic analysts have determined a ranking for the most likely places where the victims might be found (where a higher number indicates more likely):

Place F is ranked at the top as the most likely place where the killer is operating.
Place A is next,
then Place B,
Place C,
Place E, and finally
Place D is ranked last, as the least likely place.

The local state police detective working the case asks her three chief investigators, "Where should we search for the kidnapped people?"

(123) **TERRORISM**
The Capitol Police have just gotten word of a possible terrorist attack targeting specific stores in the area. An independent consultancy firm has calculated the risk of certain stores to be targeted. The firm has not yet released the study, but has ranked the stores from high risk to low risk:

Store E is most at risk, with a .5 probability,
Store B is next with .4 probability,
Store A has a .25 probability,
Store D, a .1 probability,
Store F, a .05 probability, and
Store C is least at risk, with a 0 probability.

The Chief asks his three top advisors, "Where should we set up extra surveillance?"

In real life, our goals are often constrained by practical factors like time-sensitivity, manpower, and other resources. Though our stories did not manipulate these directly, it's possible that participants imputed such constraints into the scenarios. Consider (122). Successfully saving kidnapped individuals may well rest on fast action (is there not a greatly diminished expectation to find the victim alive after 72 hours?). Thus, if certain locations are more likely to lead to results, then it would make sense that those places should be prioritized by an efficient search party. An exhaustive search, while

more thorough, would waste time. This could explain why MS-MAX answers were preferred over both MA and MO-MAX in FBI. Similarly, in (123), if time or resources are a constraint, then perhaps the best action is to move on the most certain locations.[2]

Such variability was not observed in LOW STAKES NON-FINITE Condition (top graph in Figure 3.7). We see participants' robust preference for both MO-MAX ($\chi^2(1) = 49.223$, $p<0.0001$) and MS-MAX ($\chi^2(1) = 43.81$, $p<0.0001$) over MA across the entire LOW STAKES sample.

There appears to be more variability in the FINITE Condition (bottom graph in Figure 3.7). In the analysis presented above, we removed the DENTIST story because we determined that it was actually a HIGH STAKES scenario and we were only interested in LOW STAKES for the +FIN condition. However, we do not find an effect of STORY in this condition. Like the NON-FINITE Condition, we see significant differences between MO-MAX and MA ($\chi^2(1) = 18.814$, $p<0.0001$), but not between MS-MAX and MA ($\chi^2(1) = 0.72216$, $p=0.4$). Indeed, MS-MAX and MO-MAX were also significantly different ($\chi^2(1) = 16.372$, $p<0.0001$).

Consider the DENTIST scenario, presented below.

(124) DENTIST
Dee is looking for a new dentist. She asks various family members for their opinions. One of her aunts mentions that her cousin Joe has had chronic dental issues for the past few years. As a result, he has tried several local dentists and has ranked them according to affordability and competence:

Dentist A is ranked first,
Dentist B is the next,
Dentist C is third,
Dentist D is fourth,
Dentist E is fifth, and
Dentist F is last.

Dee asks, "Where did Joe go for his dental work?"

We originally considered this to be LOW STAKES because there was no (intended)

[2]Note however, that in a sense this may reveal the complexity of defining the constrains on decision problems, known from the literature in economics and behavioral economics, and suggest that ultimately an account that involves calculating utility must be much more complex than deciding merely on informational content of an answer.

emergency conveyed by the scenario, and certainly no lives were apparently at risk. Our operationalized definition of STAKES made the very coarse-grained distinction between human lives at risk (=high stakes), or not (=low stakes). We did acknowledge that by this definition, a scenario may be exhaustive and low stakes or non-exhaustive and high stakes. I believe this scenario nicely illustrates these possibilities. While no lives are at risk, it's possible that this scenario would warrant a more exhaustive response, perhaps because bad dentistry *could* be risk in this way. If this were the case, we would expect to see high acceptance of MA answers, which we do not see. We leave this mystery, and move on to the discussion.

### 3.4.3 Discussion

The discussion in this section explores the possibility that partiticpants impute additional information into these experimental scenarios, in order to resolve (non-)exhaustivity. We saw variation in both experiments suggesting that some participants may have imposed additional constraints on the stories, like time-sensitivity or resource maximization (experiment 2), or used their general expectations about scenario-specific speaker goals (experiment 1). In both cases, this information is not only not present in the designed scenarios, but also not part of the semantic content of the target questions.

## 3.5 General Discussion

Across two experiments, we have presented empirical data that reveal the following. First, MS answers are not as constrained as has been previously assumed. They are acceptable when associated with either infinitival embedded questions or with finite embedded questions. The presence of a modal element facilitates licensing of a non-exhaustive answer, but is not a necessary precondition for it. Second, the type of embedding verb and the wh-word both play a role in answer acceptability, highlighting the role of lexical semantics. Third, contextual information about the questioner's goals in the discourse context play a key role in the resolution of exhaustivity. Finally, the quality of a mention-some answer matters: those that are more informative are

valued more than those that are not. Thus, in this work, we have identified a set of surface-level and contextual cues that the speaker can manipulate and that the listener can recruit to arrive at an intended interpretation. Together, these points bear against semantic theories which assign only an exhaustive meaning to questions.

I suggested in the previous chapter that world knowledge and hearer's prior expectations play a role in the resolution of exhaustivity. Given the context-dependence of (non-)exhaustivity, when we interpret a root or embedded question out of context, we retrieve a context against which to evaluate the question for (non-)exhaustivity. Whether we judge a question exhaustive or non-exhaustive will depend on aspects of the context we have retrieved. How we retrieve a context, and what context we do retrieve will naturally depend on several different factors that involve our experience and knowledge of the world. Amongst these, may include the baseline frequency of the construction, the relative co-occurence of the cues in the question (assuming that indeed linguistic factors are cues to interpretation), the conditional probability of (non-)exhaustivity given those cues, and the conditional probability of certain contextual goals given those cues.

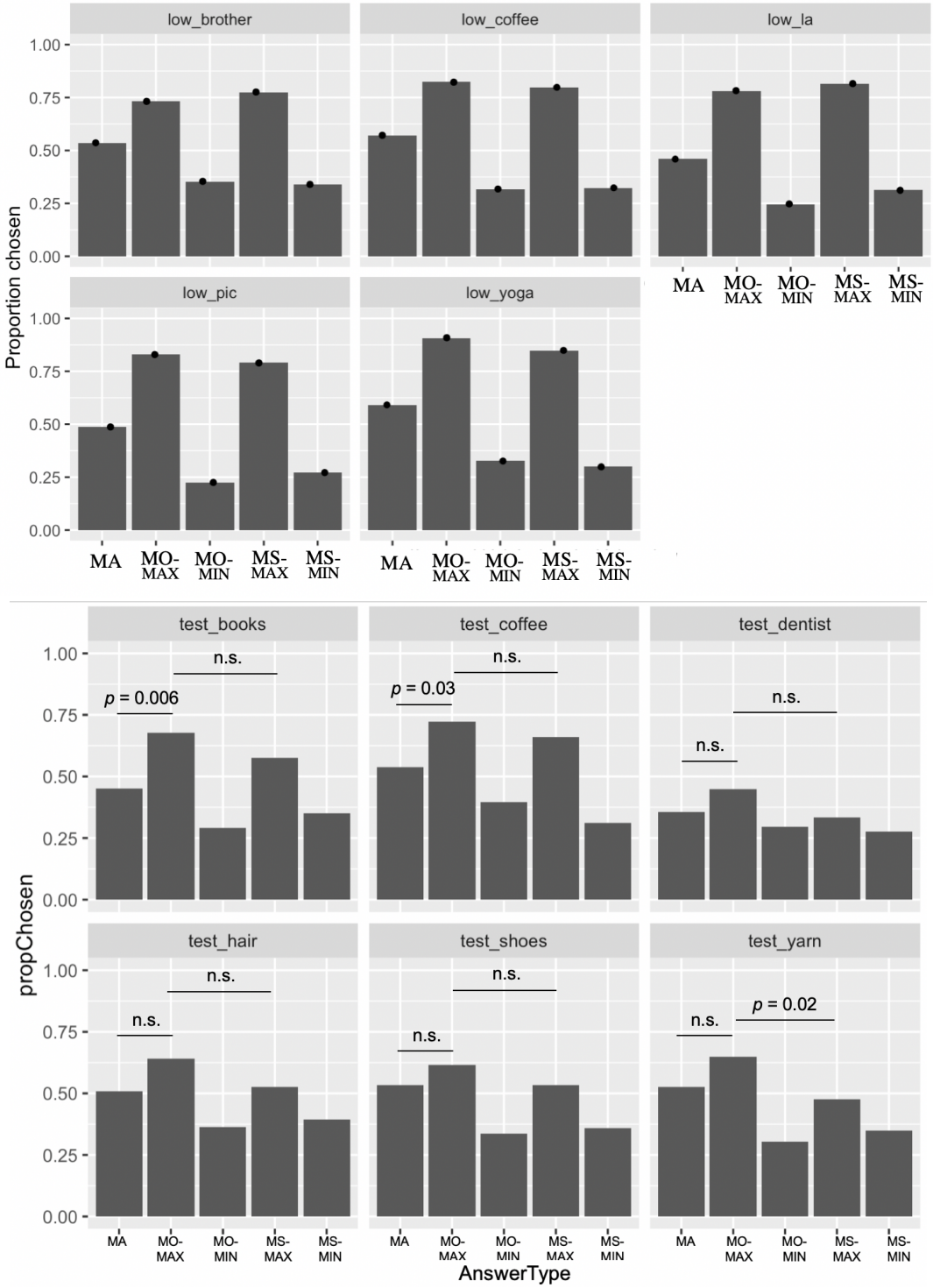


**Figure 3.7:** Results of Experiment 2, LOW STAKES, NON-FINITE (top) and FINITE (bottom) Conditions, broken down by STORY.

# Chapter 4

# Experiments 3a and 3b: Conditional probability of (non-)exhaustivity

Up to this point, we have focused on testing predictions made by semantic theories about the acceptability of mention-some readings of embedded questions across contexts and question forms. We found that while a modal question boosts the acceptability of mention-some, non-modal questions can also be non-exhaustive. However, these were degraded relative to modal questions. We argued that if contextual goals were explicitly manipulated to be non-exhaustive, that non-modal questions would be felicitous on mention-some interpretations. Dayal (2016: 71-82) has argued similarly that non-modal mention some needs explicit contextual licensing. In our second experiment, we found that both mention-some and mention-all acceptability was conditioned upon discourse goals. Not only can non-modal questions be felicitously mention-some, but exhaustive readings of questions are not default, they too are context-dependent (Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010). We additionally found that independent factors, like the matrix embedding verb imposed interpretational constraints on the embedded question: know-wh questions were more exhaustive than predict-wh questions. This supports insights from, e.g., Heim (1994), Zimmermann (2010), George 2011 that know imposes exhaustivity on its wh-complement.

I further argued that the responses in Experiment 2 revealed that participants were calculating something akin to "mention-enough", and thus best supported accounts of semantic underspecification (Asher & Lascarides (1998); Ginzburg (1995); van Rooij (2003), 2004). One could argue that the data are also consistent with semantic ambiguity (Hintikka 1976; Beck & Rullmann 1999; Lahiri 2002; George 2011, Ch. 2). However,

given that participants appeared to access a range of intermediate non-exhaustivity in Experiment 2, an underspecified semantics simply explains with a single underlying representation. In contrast, an ambiguity semantics would potentially need many representations. From a cognitive perspective, it is an open question whether one kind of underlying representation is better than the other. Some have argued that underspecification accounts are safeguards against combinatory explosion (Poesio 1996; Ebert 2005), while others have made the case that ambiguity actually facilitates communication (Piantadosi, Tily, & Gibson 2012).

In fact, ambiguity and underspecification theories are equivalent from the point of view of the semantics-pragmatics divide, and simple experimental methods may not suffice to empirically distinguish between them. Both accounts will show interpretational variation based on context, and those contextual features we think modulate interpretation; both accounts require a pragmatic mechanism deployed by the hearer to either precisify or disambiguate the speaker's utterance. Indeed, in any case where we might postulate ambiguity, we might equally postulate underspecification via contextual variables instead, and vice-versa. To empirically choose between these two semantic theories, we must then examine the behavioral signature of disambiguation and precisification, and determine which one best matches (non-)exhaustivity resolution. This is no small task, and one that we will not answer here.

Precisification can be defined as a process by which a hearer takes an underspecified semantic representation, and refines it through the inclusion of contextual information (Pinkal 1996; Egg 2010). This might mean that the representation receives an extension, or truth-value (van Deemter 2010), but the question of truth-value is not our concern here. Similarly, disambiguation can be defined as a process by which a hearer determines which of multiple representations, a speaker intended, through the inclusion of contextual information (cf. Jurafsky 1996).

Both are processes whereby the *hearer* determines the speaker's meaning by integrating multiple sources of information together: the proprietary linguistic information (the semantic representation, or the classical "literal meaning" of the utterance), with non-linguistic information (e.g., context, world-knowledge, prior expectations). Only, in the case of questions, the literal meaning alone does not determine (non-)exhaustivity. Rather, the question form itself can provide defeasible *cues* to the speaker's meaning (to the degree of (non-)exhaustivity intended).

This conception of (non-)exhaustivity provides explanations for judgements in the theoretical literature. Linguists are also hearers, and the process of determining truth conditions (or answerhood conditions) inherently involves the interpretation of a question in a context. How a hearer does this may depend on several different factors, as we've been discussing.

For one, it has been standard in the literature to discuss question interpretation out of context. Given the necessity of context to interpretation here, hearers must impute a context where none is explicitly (linguistically) provided. Thus, the perceived omnipresence of mention-all could be the result of a 'safe bet" heuristic strategy a hearer deploys. Perhaps, in absence of sufficient information regarding the speaker's goals, a hearer might interpret or answer more exhaustively or not as a quicker way to discharge their responsibility as hearer or answerer. Indeed, exhaustivity inferences arise more generally than in questions (Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010; Geiss et al. 2018; Destruel & DeVaugh-Geiss 2018).

In other cases, the baseline likelihood of a particular resolution of (non-)exhaustivity might differ with different questions. Beyond the form factors we have so far been discussing, the world knowledge associated with a particular question may lead a hearer to have prior expectations about the relevant speaker goals associated with it. Recall the case of Who's got a light?, where without any explicit linguistic context, we assume the speaker is a smoker and has a non-exhaustive goal.

Thus, the probability that a particular resolution is drawn may depend on not only the linguistic form of the question asked (both the syntax and semantics of the words

plus the world knowledge associated with them), but the context, and a range of subjective prior beliefs that the hearer has. It is a combination of both bottom-up linguistic information and top-down beliefs and expectations. This view is consistent with current thinking, not only amongst psycholinguists about the nature of language comprehension (Degen & Tanenhaus 2014, 2016, 2018; Elman, Hare, McRae 2004; MacDonald, Pearlmutter & Seidenberg 1994; Seigenberg & MacDonald 1999; Tanenhaus & Truswell 1995; McRae & Matsuki 2004, a.o.), speech perception (Kleinschmidt, Weatherholtz & Jaefer 2018), acquisition of syntax (Pearl 2006), and word learning (Frank & Goodman 2014, Frank, Goodman, & Tenenbaum 2006); but also amongst cognitive scientists more generally (visual perception: Feldman & Singh 2006, Feldman 2016; inductive reasoning: Pearl 2000; concept learning: Tenenbaum 1999).

Probabilistic models have been fruitfully applied to the study of language in discourse, including coreference (Kehler, et al. 2008; Kehler & Rohde 2018), syntactic ambiguity (Jurafsky 1996; Rohde 2008), scalar implicature (Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Degen 2015; ), specification of thresholds in gradable adjectives (Lassiter & Goodman 2013), and non-literal language use (metaphor: Kao et al., 2014, hyperbole: Kao et al., 2014; irony: Kao & Goodman 2015), amongst many others.

In this chapter, I aim to dig into the probabilistic relationship between surface-level linguistic cues and (non-)exhaustivity by answering the question: What is the probability of interpretation conditioned on these surface-level cues? We thus provide empirical grounding for the Cue Hypothesis discussed in Chapter 3. I conduct two answer rating tasks. The first task manipulates the presence/absence of a modal (can) and the wh-word heading the question (who, where, and how). The second study looks more closely at the relationship between contextual goals and the presence/absence of an existential modal, using the notions of "high" and "low stakes" contexts introduced in Experiment 2.

## 4.1 Experiment 3a: Interpretation Conditioned on Linguistic Form

Experiments 1 and 2 provide some evidence that the linguistic form of the question and contextually provided discourse goals determine the acceptability of (non-)exhaustivity in a question. However, here we explicitly test the Cue Hypothesis, that "exhaustive cues" will lead to higher ratings of exhaustive (MA) answers, and "non-exhaustive cues" will lead to higher ratings of non-exhaustive (MO/MS) answers. The next two studies attempt to lend empirical support to this hypothesis.

### 4.1.1 Design and Materials

Data and materials for this study can be found at https://github.com/mcmoyer11/Conditional_probability . This study manipulated one within-subject factor, WH (who, where, how). We included three between -subjects factors: MODAL (MODAL, NOMODAL), ANSWER (MO, MS, MA), and TASK (ACCEPT, LIKELY). Thus, this study was a 2x3x2 factorial design, with 12 unique trial types. In addition to these test trials, there were 32 filler items. Fillers involved either judgments of pronominal co-reference, or judgments of the naturalness of a picture of an animal.

Example test trials are presented in (125) and (126). As in Experiment 1, participants read a short scenario in which a character asks another character a root question, for example, Where did Fido hide his toys? In contrast to Experiments 1 and 2, here no answer is explicitly given. Rather, it is stated that the Questioner concludes, based on the answer, that the answerer knows-wh. At test, the participant is then presented with an answer, and asked how likely or how acceptable that answer is, given the story. Here, answers were manipulated for (non-)exhaustivity: MA, MS, and MO.

(125) NOMODAL
Fido the dog buried his toys in the backyard last week. He hid them so well that now he cannot seem to find any. A neighbor, Jill, comes over to help Mary, Fido's owner, find the toys in the yard.

Fido hid his toys behind the shed, next to the large tree, under the swing set, and under the deck. No toys were in the middle of the yard, on the side of the house, or near the mailbox.

Jill asks Mary, "Where did Fido hide his toys?"

Based on Mary's answer, Jill concludes, "Mary knows where Fido hid his toys."

a. How **acceptable** is it for Mary to give an answer like, ACCEPT TASK
b. How **likely** is it that Mary gave an answer like, LIKELY TASK
   i. "Behind the shed"? MO
   ii. "Behind the shed, and next to the large tree"? MS
   iii. "Behind the shed, next to the large tree, under the swing set, and under the deck"? MA

(126) **MODAL**

Fido the dog buried his toys in the backyard last week. He hid them so well that now he cannot seem to find any. A neighbor, Jill, comes over to help Mary, Fido's owner, find the toys in the yard.

Fido can hide his toys behind the shed, next to the large tree, under the swing set, and under the deck. But not in the middle of the yard, on the side of the house, or near the mailbox.

Mary asks Jill, "Where can Fido hide his toys?"

Based on Jill's answer, Mary concludes, "Jill knows where Fido can hide his toys."

a. How **acceptable** is it for Mary to give an answer like, ACCEPT TASK
b. How **likely** is it that Mary gave an answer like, LIKELY TASK
   i. "He can bury them behind the shed"? MO
   ii. "He can bury them behind the shed, and next to the large tree"? MS
   iii. "He can bury them behind the shed, next to the large tree, under the swing set, and under the deck"? MA

Note that, in other work we explicitly compared definite descriptions which impose a maximality requirement, with base nominal expressions that do not (Moyer, Husnain, Syrett 2019). The hypothesis was that the former would grammatical block mention-some readings if maximality was violated but the latter would not. We found that, when contextual goals are non-exhaustive, participants accepted mention-some readings for both noun types (no significant differences between the two noun conditions), even when maximality was violated. Malamud (2011)'s semantics of definites gives a van Rooij-style decision-theoretic analysis of maximal and minimal readings of definites which are parametric on context, and provides a nice explanation of those results. However, the interaction between definiteness/maximality and (non-)exhaustivity in

questions is an open issue.

In many of our stimuli, we use possessive nouns without fear of introducing a confound due to maximality, given the previous results just discussed. At the same time, if participants are accessing a reading where maximality violations render the target sentence false, it could lower the acceptability of mention-some. We will keep this point in the back of the mind.

### 4.1.2 Participants

238 participants were gathered and run on this experiment through Amazon Mechanical Turk. Participants were restricted to those with U.S. IP addresses, who had a HIT completion rate of 99% or higher, and who had completed more than 1,000 HITs. These additional restrictions were to ensure that the participants were of a high quality and would take the task more seriously. The study was designed and administered through Qualtrics survey software (Provo, UT).

### 4.1.3 Predictions

This study tests the hypothesis that the linguistic factors we have been calling "cues to (non-) exhaustivity" are indeed cues for the hearer to resolve (non-)exhaustivity. If this is true, then generally speaking, exhaustive cue conditions will lead to higher ratings of MA answers, and lower ratings of MS/MO answers, while non-exhaustive cue conditions will lead to higher ratings of MS/MO answers and lower ratings of MA answers.

Specifically then, we may think of exhaustive cues as being finite (non-modal), and who-questions; and we may think of non-exhaustive cues as being modal, and how-questions. Further, as understood in the processing literature, we would expect that cue co-occurrences can affect the strength of an interpretation. Note that we are not measuring interpretation strength directly in this experiment. To approximate strength, we might look at the variance in responses. Higher variance in ratings or an answer across participants could indicate that those cues do not reliably cue to that

answer; likewise, low variance could indicate that the link between the answer and a cue is strong.

We are also implicitly testing the form theories discussed in the previous chapter. Modal theories predict asymmetries between MODAL and NON-MODAL questions for MS/MO answer condition. Further the manipulation of MS and MO introduced in Experiment 2 allows us to test the acceptability of degrees of non-exhaustivity.

### 4.1.4 Results

A Wilcoxon-Mann-Whitney test reveals a significant effect of TASK: $W$ = 1200000, $p$ <0.0001, and an interaction between TASK and ANSWER ($\chi^2(5)$ = 217.54, $p$ <0.0001). This is confirmed by a comparison between the two distributions shown in Figure 4.1.

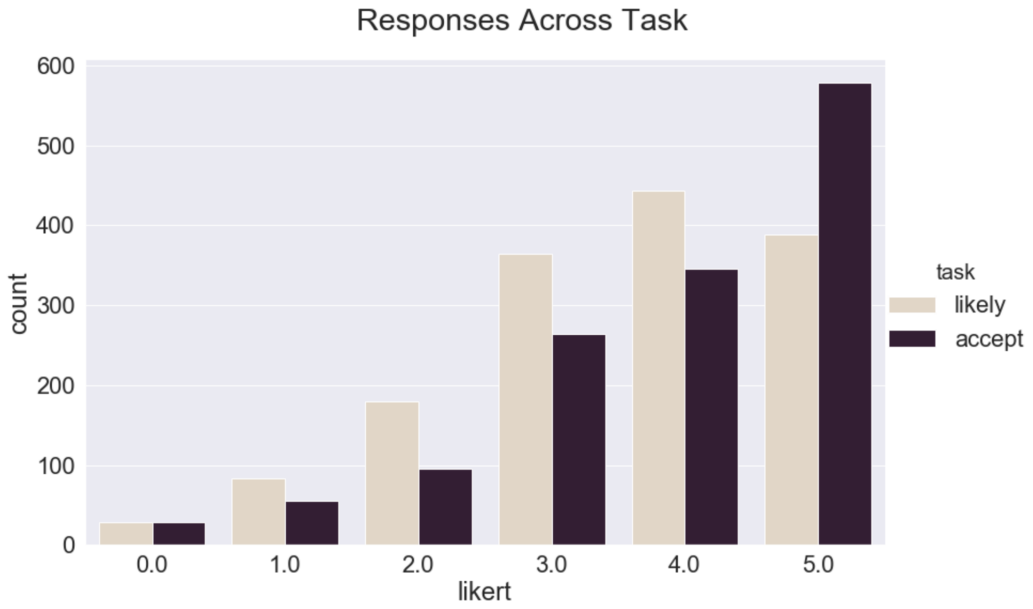


**Figure 4.1:** Comparison of Likert response distribution over the two tasks. The LIKELY task has 1488 observations, while the Accept Task has 1367.

Given these differences in task, we will analyze the two datasets separately. In both tasks, there were main effects of ANSWER (LIKELY: $\chi^2(2)$ = 17.261, $p$ <0.0002; ACCEPT: $\chi^2(2)$ = 129.52, $p$ <0.0001), and MODAL (LIKELY: $\chi^2(1)$ = 43.691, $p$ <0.0001; ACCEPT: $\chi^2(1)$ = 24.87, $p$ <0.0001). WH-WORD was significant in the LIKELY task ($\chi^2(2)$ = 19.206, $p$ <0.0002) but not the ACCEPT task ($\chi^2(2)$ = 1.6813, $p$=0.4). Finally, in both tasks we

found a two-way interaction between ANSWER and MODAL (LIKELY: $\chi^2(5) = 73.137$, $p < 0.0001$; ACCEPT: $\chi^2(5) = 208.55$, $p < 0.0001$) and a three-way interaction between ANSWER x MODAL x WH (LIKELY: $\chi^2(17) = 107.39$, $p < 0.0001$; ACCEPT: $\chi^2(17) = 222.11$, $p < 0.0001$).

While there was no overall significant difference between MS and MO answers ($\chi^2(1) = 1.5726$, $p = 0.2$), there were differences in each task (LIKELY: $\chi^2(1) = 7.8094$, $p < 0.005$; ACCEPT: $\chi^2(1) = 25.278$, $p < 0.0001$). There was an overall significant difference between MS and MA answers ($\chi^2(1) = 77.884$, $p < 0.0001$), driven by the ACCEPT task (ACCEPT: $\chi^2(1) = 132.28$, $p < 0.0001$) but not the LIKELY task ($\chi^2(1) = 1.4519$, $p$=0.2). TASK was a between-subjects measure, so no participant saw both dependent variable, and thus there would not have been interference between measures.

The following graphs present the breakdown of each factor across the three ANSWER conditions. First, Figure 4.2 shows the MO condition. Here, we can see that responses are overall on the high end: median ratings are all 3 or above. First, NOMODAL condition receives significantly lower ratings than the MODAL condition in both tasks (LIKELY: $\chi^2(1) = 42.889$, $p < 0.0001$; ACCEPT: $\chi^2(1) = 70.108$, $p < 0.0001$). Second, the median rating for NON-MODAL WHO (M=3) in the ACCEPT task is lower than for the other WH-WORDS (M=4). This difference is not significant ($p$=0.09 between WHO and WHERE). Finally, the median responses are about one point lower in the LIKELY task than in the ACCEPT task (MODAL: $\chi^2(1) = 53.221$, $p < 0.0001$; NOMODAL: $\chi^2(1) = 15.105$, $p$=0.0001).

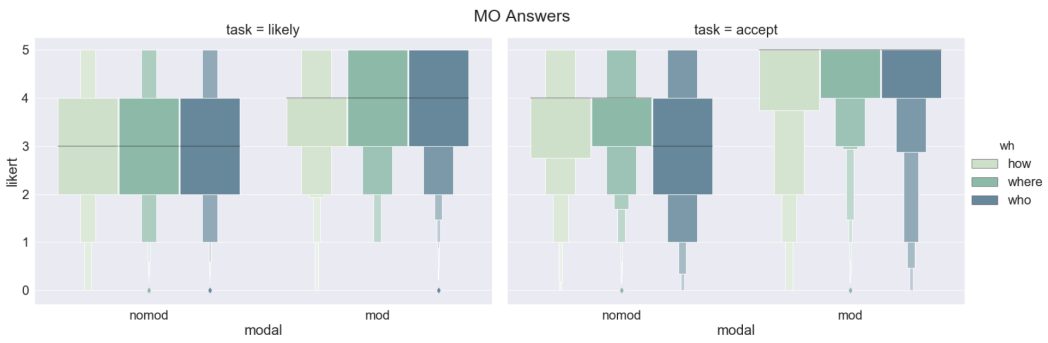


**Figure 4.2:** Responses split by MODAL, WH and TASK, for MO answers.

Figure 4.3 shows the MS condition. Just observing the graphs reveals that there do

not appear to be many differences between factors, and ratings are generally high ('3' or above). Again, we see that NOMODAL conditions are degraded compared to MODAL conditions, but this is significant only in the LIKELY task ($\chi^2(1) = 7.1023$, $p$=0.008) and not in the ACCEPT task ($\chi^2(1) = 3.0243$, $p$=0.08).

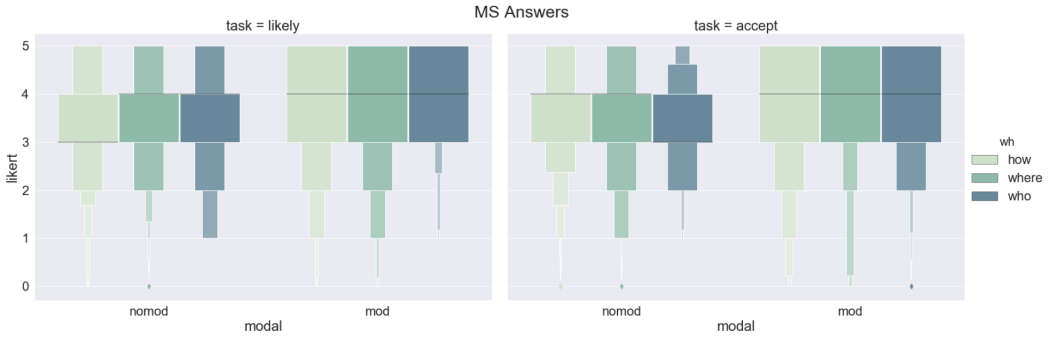


**Figure 4.3:** Responses split by MODAL, WH and TASK, for MS answers.

Finally, Figure 4.4 shows the MA condition. Again, the graphs reveal little overall differences, even between MODAL and NOMODAL conditions. However, NOMODAL condition received significantly lower ratings than the MODAL condition in the LIKELY task ($\chi^2(1) = 7.8079$, $p$=0.005) but not in the ACCEPT task ($\chi^2(1) = 3.6394$, $p$=0.06). We also see significant effects of WH in the LIKELY task, across both MODAL ($\chi^2(2) = 7.3343$, $p$=0.03) and NOMODAL ($\chi^2(2) = 17.093$, $p$=0.0001). In both cases, the effect is driven by a difference between WHO and HOW (MODAL: $p$=0.03; NOMODAL: $p$=0.0002), particularly in the NOMODAL condition.

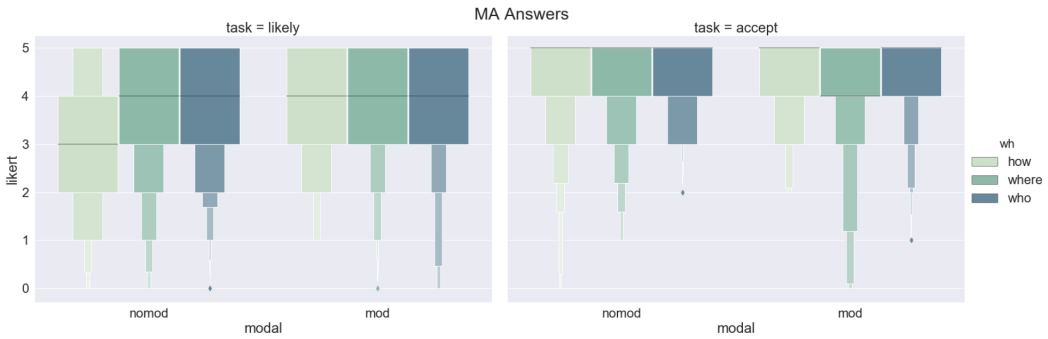


**Figure 4.4:** Responses split by MODAL, WH and TASK, for MA answers.

### Analyzing Variance

One way to test variance in Likert scale data would be to test for homogeneity of variance, or homoscedasticity. If there is less variance in participant responses to certain conditions, it could indicate that those conditions more consistently provide a cue to interpretation than conditions with greater variance in participant responses. More variance could indicate either that participants are not consistently responding based on that condition, or could perhaps indicate that there are different underlying response patterns.

Given the non-parametric nature of the data, we can perform a Fligner-Killeen test for homogeneity of group variances based on ranks. The null hypothesis is that the variances in the different groups are the same. We hypothesized that, if the variance in cue co-occurrence conditions was less than in non-co-occurrence conditions, this could indicate a stronger relationship between the cues and a given answer. Figure 4.5 plots the variance for each condition. First, the difference in variance in ratings

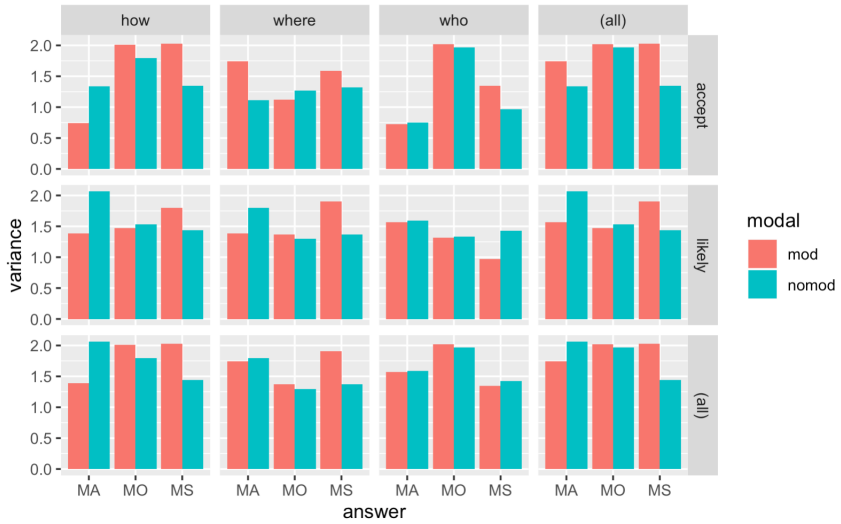


**Figure 4.5:** Variance in each condition.

for either MO or MS answers between MODAL conditions for was not significant (MO: $\chi^2(1) = 2.3765$, $p = 0.1$; MS: $\chi^2(1) = 0.276$, $p = 0.6$), meaning (according to the above logic) that participants were neither more nor less consistent in their ratings for modal

vs. non-modal questions. In contrast, the difference in variance in responses for wh-word was significant for MO answers ($\chi^2(2) = 9.2457$, $p > 0.01$), but not MS answers ($\chi^2(2) = 3.9014$, $p = 0.1$).

In particular, we hypothesized that two such conditions would be of interest: NONMODAL who-questions, and MODAL how-questions. For neither who- nor how- questions, was there a significant difference between MODAL conditions in the variance of responses to MS/MO answers (who-questions: $\chi^2(1) = 0.395$, $p$=0.5; how-questions: $\chi^2(1) = 0.048$, $p$=0.5). This means that we cannot reject the null hypothesis that these MODAL and NOMODAL groups have the same variance.

Interestingly, the variance between MODAL conditions was significantly different for MA answers overall ($\chi^2(1) = 6.561$, $p$=0.01), but driven by how-questions ($\chi^2(1) = 8.3967$, $p <$0.005). A look at histograms from the two conditions reveals that responses in the MODAL conditions are lower than responses in the NOMODAL condition. A closer look at these conditions reveals lower means in the NOMODAL condition ($\overline{x}$ = 3.7 vs. $\overline{x}$ = 3.9), but equal medians (M=4).

Here, we actually see TASK differences which make the MODAL difference come out. Overall there were lower medians and means in the LIKELY than in ACCEPT (M = 3, vs. M = 4; $\overline{x}$ = 3.3 vs. $\overline{x}$ = 4.4), suggesting that an MA answer is an acceptable, but unlikely response to a how-question. But this was driven by the NOMODAL condition (M = 3; $\overline{x}$ = 3), suggesting that an MA answer to a non-modal how-question is less likely. This is consistent with the intuition expressed by researchers like Asher & Lascarides (1998) that exhaustive answers to HOW-questions are somehow exceptional.

## 4.2 Discussion

This study showed that the Likert rating of an answer was indeed sensitive to the surface-level cues in the question's linguistic form. We saw that MS and MO answers were degraded when the question did not contain a modal, that MO answers were degraded especially for NOMODAL who-questions (in the ACCEPT task only), and that MA answers were degraded for NONMODAL how-questions. Note that these last two

results lightly replicate the observations from the literature that who-questions really disprefer non-exhaustive answers, while how-questions disprefer exhaustive ones.

Degraded median ratings were nonetheless high across the board; none dipped below a 3. The manipulated factors did not render any answers completely unacceptable or unlikely. Instead, the degraded ratings suggest that these form factors have the potential to render an answer more or less optimal, not completely ungrammatical or unacceptable. This is consistent with the hypothesis that form factors are defeasible cues to interpretation.

The results from this experiment replicates findings from Experiments 1 and 2, and thus provides support for the hypothesis that the form of a question may provide cues to the most optimal answer. In particular, we find support for a Weak, but not a Strong, Modal Hypothesis, and that the kind of wh-question (who- or how-question) may also affect what kind of answer is optimal.

What about the Cue Hypothesis? We have found some support for it given that our dependent measures were somewhat sensitive to cues. But the strength of the relationship between cues and interpretation is still open. The analysis of homoscedasticity revealed that the variance in ratings of MS/MO answers did not significantly differ between levels for our cues of interest. In particular, we did not find differences between modal and non-modal who- and how-questions. However, we did find significantly different variance in ratings for MA answer, driven by MODALITY in how-questions.

Just because there were no differences in variances for MS/MO answers, does not mean the Cue Hypothesis is ruled out. Rather, this finding is consistent with each level of a factor of interest providing a cue (e.g., both the presence and the absence of modality), rather than only one level (i.e., the presence of modality but not its absence) providing a cue for one reading. The cues go both ways.

We saw some task variation. The effects of the wh-word were modulated by TASK: MA answers were rated slightly less *likely* answers to nonmodal how-questions than other to other wh-questions (but not less *acceptable*), and MO answers were rated less *acceptable* answers to nonmodal who-questions than to other wh-questions (but not less

*likely*). These TASK effects could suggest that there is sensitivity to a particular kind of contextual evaluation. While different from the kind of contextual manipulation from Experiment 2, we can see these task effects as illustrating the context sensitivity of these judgements (cf. Degen & Goodman 2014; Degen & Tanenhaus 2019; Roberts 2017).

One possible explanation for the generally high ratings in this experiment is experimental confound. The test prompt asked participants responded to evaluate *an answer like* __, rather than to evaluate a particular kind of answer. They could have interpreted *an answer like* to mean *an answer from the provided list* or *an answer in the form of* __, rather than as the intended mention-some/mention-all difference. Experiment 3B addresses this potential confound, and additionally manipulates context operationalized again as 'stakes', as in Experiment 2.

## 4.3 Experiment 3b: Interpretation Conditioned on Form and Goal

In this follow-up experiment, we remove a potential confound from Experiment 4a by asking the participant to judge a particular answer, rather than *an answer like...*. Additionally, we focus only on the relationship between modality and context on the acceptability of mention-some/mention-one answers.

### 4.3.1 Design and Materials

Data and materials for this study can be found at https://github.com/mcmoyer11/Answer_rating . This study manipulated within-subjects STAKES (HIGH, LOW) and ANSWER (MA, MF, MS, MO) and MODALITY (MODAL, NOMODAL) between subjects. Stories were the same as in Experiment 2, with slight modifications. An example HIGH STAKES scenario is provided in (127).

(127) HIGH STAKES Scientists have discovered a new strain of a dangerous virus that has contaminated oysters in the Mid-Atlantic. The Center for Disease Control is trying to prevent as much contamination as possible by tracking down the oysters which were sold to restaurants.

The CDC supervisor is tasked with tracking down the oysters. With this goal in mind, she asks her task force,

a. "Where can we locate the contaminated oysters?" MODAL
b. "Where are the contaminated oysters located?" NOMODAL

In fact, the oysters were delivered to Restaurants A, B, C, and D, but not Restaurant E.

Please rate the acceptability of the following answer to the Supervisor's question.

a. "Restaurants A, B, C, and D." MA ANSWER
b. "Restaurant A." MO ANSWER
c. "Restaurants A and B." MS ANSWER
d. "Restaurant E." MF ANSWER

### 4.3.2 Participants

263 participants were gathered and run on this experiment through Amazon Mechanical Turk. Participants were restricted to those with U.S. IP addresses, who had a HIT completion rate of 99% or higher, and who had completed more than 1,000 HITs. These additional restrictions were to ensure that the participants were of a high quality and would take the task more seriously. The study was designed and administered through Qualtrics survey software (Provo, UT).

### 4.3.3 Predictions

The predictions are generally the same as in previous experiments. The Context Sensitivity Hypothesis predicts a significant effect of STAKES: that ratings of MS and MO answers will be higher in LOW than in HIGH STAKES conditions. The Strong Modal Hypothesis predicts that MS/MO answers will receive low ratings—as low as for ungrammatical structures—in NOMODAL question conditions. The Weak Modal Hypothesis only predicts that these NOMODAL conditions will receive lower ratings *relative* to MODAL conditions, but not necessarily that these low ratings will be as low as for ungrammatical structures.

We include a MENTION-SOME/MENTION-ONE manipulation nonetheless in order

to establish empirically the facts about the acceptability of these answer types. Given that MO answers are grammatical and not pragmatic according to this hypothesis, we might expect that MO answers will be more available than MS answers, and thus receive higher ratings in the MODAL conditions compared to MS answers. However, as previously noted, MS answers entail grammatical MO answers, thus we might expect to find no difference between these two non-exhaustive answer types.

Finally, the Utility Hypothesis of van Rooij holds that MS/MO answers will never be more acceptable than MA answers. The results of Experiment 2 bore against this: in LOW STAKES conditions, actually (maximally informative) MO/MS answers were judged *more acceptable* than MA answers. Thus, van Rooij's hypothesis predicts merely that MA answers will always be rated high, while the others rated lower or as high, where contextually appropriate. Concretely then we would expect that LOW STAKES scenarios would see MS/MO answers potentially rated as high as MA answers, and MS answers rated higher than MO answers.

### 4.3.4 Results

Figure 4.6 presents the total results for test items. We find main effects of STAKES ($\chi^2(1) = 86.79$, $p < 0.0001$) and ANSWER ($\chi^2(1) = 49.775$, $p < 0.0001$), but not of MODAL ($\chi^2(1) = 2.3662$, $p = 0.4$). We also see an interaction between STAKES and ANSWER ($\chi^2(3) = 130.41$, $p < 0.0001$).

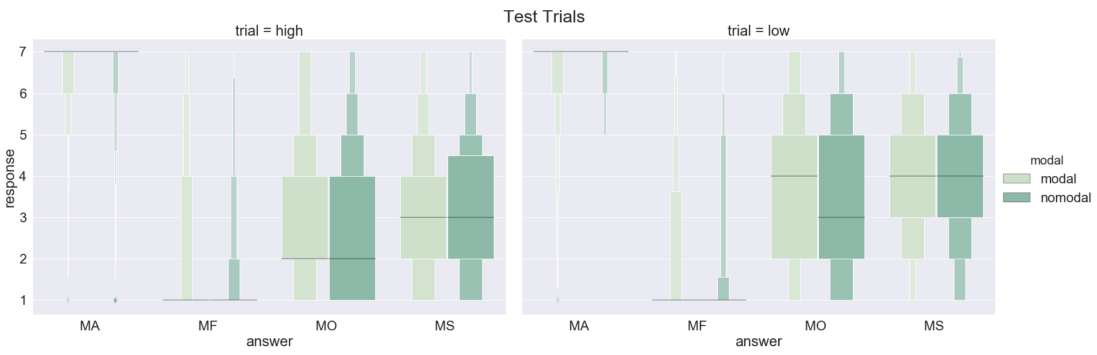


**Figure 4.6:** Responses for test trials, split by MODAL and STAKES for each ANSWER type.

No matter how the data was divided, the MODAL vs. NOMODAL comparison was

not significant in this experiment. This was further confirmed by a modal comparison: the probit regression model with MODAL as a predictor did not significantly improve the model fit over the model without it ($p$ = 0.06).

The difference between MS and MO answers was significant: a MO answer was never rated higher than an MS answer—which in turn, was never rated higher than an MA answer. Despite the fact that MS answers entail MO answers, the former are rated higher than the latter. This supports the notion that the acceptability of (non-)exhaustivity is parametric on how much information sufficiently resolves contextual goals, consistent with theories like Ginzburg (1995), Asher & Lascarides (1998), van Rooij (2003, 2004), Schulz & van Rooij (2006), Spector (2007), Zimmermann (2010).

The STAKES manipulation was significant for both MS ($\chi^2(1)$ = 46.927, $p$ <0.0001) and MO answers ($\chi^2(1)$ = 37.013, $p$ <0.0001). Further, in both HIGH and LOW STAKES, MO and MS answers were significantly different ($\chi^2(1)$ = 21.514, $p$ <0.0001; LOW: $\chi^2(1)$ = 23.511, $p$ <0.0001). The next two figures present histograms of responses for MO and MS answers, respectively.

The distribution of ratings of MO answers in HIGH STAKES conditions (Figure 4.7) is skewed to the lower end of the rating scale. In LOW STAKES conditions, ratings shift more centrally (more so for MODAL conditions than for NOMODAL ones), but are still pretty evenly spread. Fligner-Killeen tests of homoscedasticity reveal that there were no significant differences in variance between levels of any factor for MO answers.

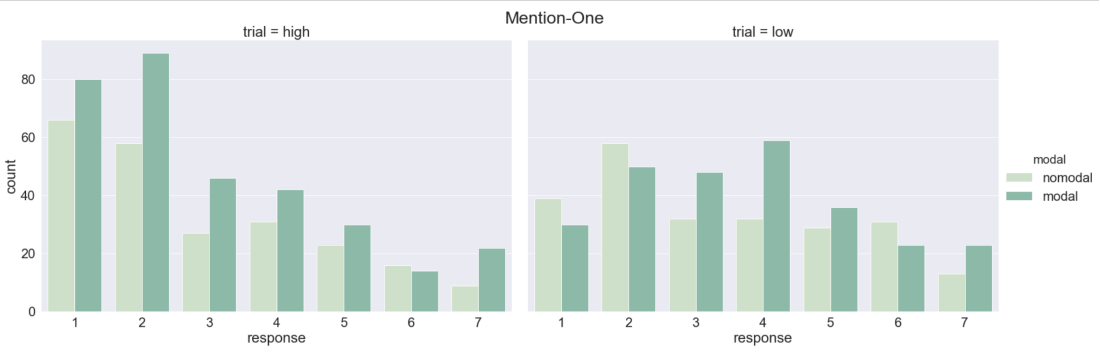


**Figure 4.7:** Histogram of responses for MO ANSWER trials, split by STAKES and MODAL.

The ratings of MS answers in HIGH STAKES conditions are pretty evenly distributed

in the lower half of the scale, in contrast to the more sharply left distribution seen with MO answers. In LOW STAKES, we see a much more dense central distribution, with less ratings in the lower range of the scale. Fligner-Killeen tests of homoscedasticity reveal that the difference in variance was significant between MODAL levels ($\chi^2(1)$ = 8.51, $p$=0.004), and in interaction with STAKES ($\chi^2(3)$ = 8.61, $p$=0.04), but not STAKES alone.

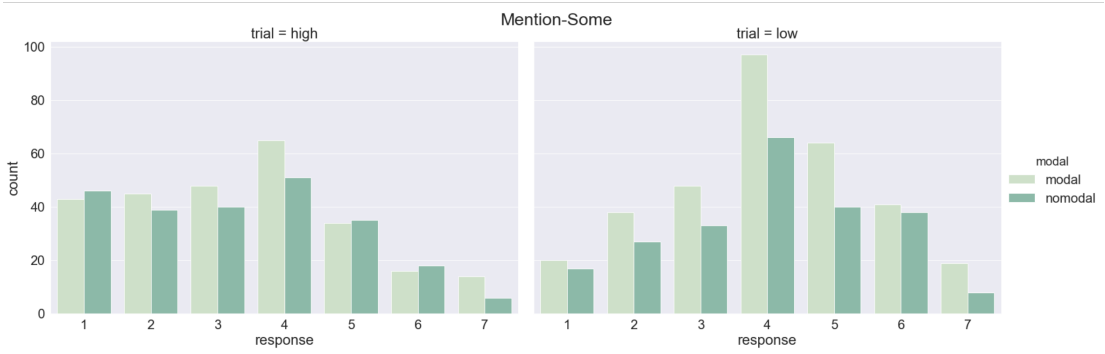


**Figure 4.8:** Histogram of responses for MS ANSWER trials, split by STAKES and MODAL.

Finally, Figure 4.9 plots the variance for each condition. We see that the control conditions, MA and MF answers, show low response variance. In contrast, both MS and MO conditions reveal much higher variance in ratings, and more so for MO answers than for MS answers ($\chi^2(1)$ = 17.57, $p$ <0.0001). Note that MS answers received higher ratings than MO answers, and there was less variance in participant responses to MS answers. Further, note that neither the presence of a modal or low stakes goals did anything to quell the high variance in participant responses.

### 4.3.5 Discussion

The results of this experiment are different from the previous experiments. First, while we found a significant effect of STAKES here, we did not see either MS or MO answers receiving *higher* ratings in LOW STAKES trials than MA answers as we did in Experiment 2. In fact, MA answers were always rated highest of all the answers, regardless of STAKES. While Experiment 2 provided evidence *against* a simple version of the van

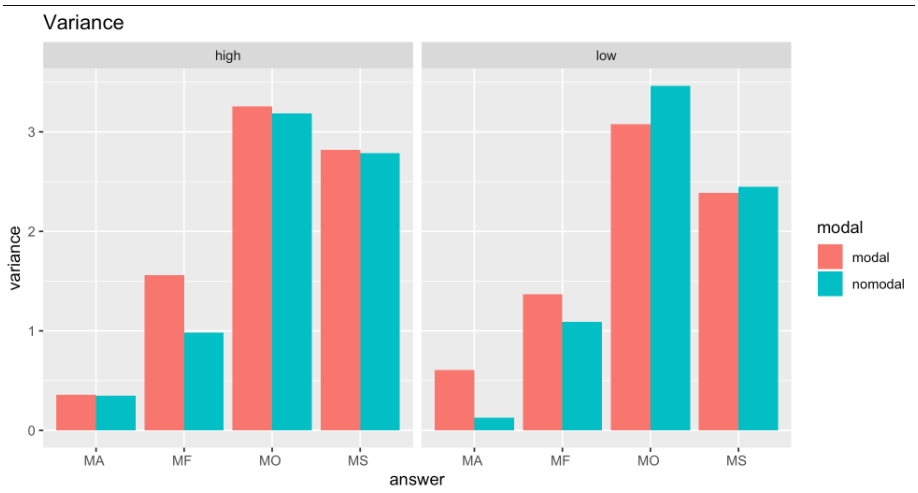


**Figure 4.9:** Variance in ratings

Rooij Utility Hypothesis (one that merely takes into account the informational content of an answer, and not additional constraints like time-sensitivity), this experiment provides evidence *in favor* of it as well as other theories which take seriously the contextual specification of (non-) exhaustivity. In contrast to Experiment 2, it appears that here participants value transmitting as much information as possible, regardless of contextual considerations, regardless even of the linguistic form of the question.

Why might this be the case? In this task, we did not look at embedded questions. It's possible that in root question-answer exchanges, participants deployed a general principle strategy to maximize the information conveyed in their answers, consistent with a Gicean Maxim of Quantity. That non-exhaustive mention-some and mention-one answers are rated lower across the board is consistent with this explanation. The fact that we did find significant effect of STAKES suggests perhaps competition between the general principle just described, and sensitivity to particular contexts which might require more cognitive resources. We know that often speaker and hearers deploy efficient but fallible heuristic principles when making judgements (Kahneman & Tversky 1979, a.o.). The general principle described is consistent with some theories of pragmatic exhaustivity inferences (Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010; Geiss et al. 2018; Destruel & DeVaugh-Geiss 2018), as well as semantically

derived ones (van Rooij (2003), 2004). Again, the extent to which we would find similar effect with why- and how-questions would be particularly informative as to the plausibility of semantically derived exhaustivity, given the implausibility of defining a set of exhaustive answers for those question types.

The STAKES manipulation did lead to significant differences between MS and MO answers. Means and medians were highest in LOW STAKES MS ANSWER conditions, and lowest in HIGH STAKES MO ANSWER conditions. Further, there was higher variance in participant responses to MO than to MS answers in LOW STAKES, while there was more variance in responses to MS than to MO answers in HIGH STAKES. Essentially, in HIGH STAKES ratings to MO answers were more clearly *un*acceptable, while in LOW STAKES answers to MS answers were more clearly acceptable.

Further, in no condition were MO answers were rated higher than MS answers, even in MODAL conditions. Further, there was much more variance in participant responses to MO answers than to MS answers in the LOW STAKES context. It would seem that if MO answers are grammatically licensed, they would give rise to less variance, rather than the opposite. On the other hand, if participants even in the LOW STAKES condition interpreted modal questions as mention-all (on the semantics of Xiang (2016)), then this result is expected.

However, in this experiment we found no evidence for the claim that modal questions boost the acceptability of MS/MO: MODAL conditions neither boosted ratings, nor reduced variance in responses. It is possible that participants ignored the question form, and only evaluated whether the answer provided as much information as possible relative to the goal. Given the close relation between goals and questions as suggested by some researchers (Question-under-discussion, Roberts 1996/2012), perhaps the presence of an explicit goal made an explicit question redundant or unnecessary. Thus, participants could evaluate answers relative to the explicit goal without considering the explicit question asked.

## 4.4 General Discussion

Experiments 3a and 3b aimed to provide a concrete link between the occurrence and co-occurrence of linguistic form cues and the acceptability(/likelihood) of (non-) exhaustivity in an answer. In both cases, it did not appear that participants were strongly influenced by any form factor to reject (or, give a low rating) MS/MO or MA answers. While we did find significant effects of form factors in Experiment 3a—both MODAL and WH—we found none in Experiment 3b, even with the addition of the STAKES manipulation. Further, the effects in Experiment 3a were not large enough to render MS or MO answers completely ungrammatical or unacceptable—these were rated high across the board. In contrast, in Experiment 3b, participants always preferred ma answers above MS/MO ones.

We hypothesized that, since discourse goals were explicitly stated in Experiment 3b, but left implicit in Experiment 3a, the linguistic form of the question would be more crucial to providing cues to the speaker's goal in the latter task than in the former. For this reason, it might not be surprising that we do not find effects of linguistic form in Experiment 3b: the cues were not necessary because the context was sufficiently informative.

The data bare against aligning any special grammatical status to either exhaustive or non-exhaustive answers/interpretations. The differences found between the LIKELY and ACCEPT tasks of Experiment 3a, and between Experiments 3a and 3b themselves reveal the extent to which contextual factors (broadly construed) govern answer (non-)exhaustivity. For hearers, it appears that contextual demands weigh heavier in the resolution of (non-)exhaustivity than does the linguistic form of the question. At the same time, whatever the mechanism which integrates contextual and linguistic information, it is sensitive *at times* to the fine-grained distinctions in the linguistic signal. This relationship is found in many other phenomena at the semantics/pragmatics interface as well (Hobbs 1979; Asher & Lascarides 2001; Asher & Lascarides (1998); Kehler & Rohde 2018; Kehler et al 2008; Schloeder & Lascarides 2020).

Questions are underspecified on the surface for (non-)exhaustivity. The hearer

must resolve (non-)exhaustivity by reasoning about the linguistic signal, the context, and the speaker's intended message. We have suggested that there is a probabilistic relationship between these differing sources of information and that grammatical differences in the linguistic signal should be read as cues to the speaker's intended meaning. In quantifying that underlying probabilistic relationship, it is important to understand the baseline frequency and co-occurrence of these linguistic cues, the context, and their relationship to (non-)exhaustivity. In the next chapter, we present a corpus study of naturalistic speaker productions to lay that groundwork. We venture further into the mind of the speaker by manipulating context using slightly modified stimuli from Experiment 3b, to test their cue productions given explicit discourse goals.

# Chapter 5

# Corpus Study and Experiment 5 (Production): Examining speaker production of cues to (non-)exhaustivity

In the picture of (non-)exhaustivity in questions that we've put forth, the degrees of (non-)exhaustivity that a hearer accesses is probabilistically linked to both the linguistic form of the question, and to contextual factors relevant to determining the speaker's goal. In the last chapter, we presented two experiments which show that hearers are differentially sensitive to these cues in resolving (non-)exhaustivity: when the discourse goals are underspecified in the context, hearers can recruit linguistic information in the question form to resolve (non-)exhaustivity (Experiment 3a); when discourse goals are explicit, the information in the question form becomes unnecessary (Experiment 3b).

We have suggested that the speaker can manipulate the probability that a hearer would draw a particular resolution by providing information that makes plain her goals or her intended meaning. She could do this by explicitly stating her goals, or by uttering a question whose linguistic form cues the hearer to her goal.

What questions are speakers producing naturalistically? Are they producing questions which convey their desire for exhaustive or non-exhaustive answers? In this chapter, I aim to get inside the relationship between goals and linguistic form by looking at the question forms that speakers produce naturalistically, and when given specific contexts and goals; and by determining the conditional probability of interpretation given those cues by seeing how hearers rate answers given those factors.

Constraint-based accounts of pragmatic processing make crucial connections between cue frequency and the speed and robustness of interpretation. They assume (supported by evidence from sentence processing) that utterance comprehension is

probabilistic, context-driven, and constraint-based (Degen & Tanenhaus 2014, 2016, 2018; Elman, Hare, McRae 2004; MacDonald, Pearlmutter & Seidenberg 1994; Seigenberg & MacDonald 1999; Tanenhaus & Truswell 1995; McRae & Matsuki 2004, a.o.). Quantifying these features for (non-)exhaustivity in questions is crucial to understanding the question interpretation and production. Additionally, quantifying the probabilistic information in the input is important for understanding what language learners are exposed to, and what they can exploit in the process of learning (see, for example, Syrett 2007, Dudley 2017).

This chapter includes two studies. The first looks at naturalistic productions of questions via a corpus study, and the second looks at question production when discourse goals are explicitly manipulated using the notions of 'high' and 'low' stakes as in Experiment 2 and 3b. Before progressing to the studies, we review the cues of interest in Section 5.1.

## 5.1 Cues of interest

The general prediction is that cues linked to robust interpretations will be more frequent, and more frequently co-occurring with other such cues (Degen & Tanenhaus 2015; 2019; Elman, Hare, McRae 2004; MacDonald, Pearlmutter & Seidenberg 1994; Seigenberg & MacDonald 1999; Tanenhaus & Truswell 1995; McRae & Matsuki 2004, a.o.). The main factors of interest are the ones that have been discussed at length throughout this dissertation so far: matrix verbs, wh-words, and modals/non-finite clause types. These cues are interesting in how they frequently they occur, as well as co-occur with each other. Let us briefly review the reasons why these are factors of interest, which should be familiar by now.

### Modality and Non-Finiteness

Our first factor of interest is modal/non-finite clauses, the presence of which seem to greatly boost the acceptability of non-exhaustive answers/readings (George 2011, Ch.6; Nicolae 2015; Fox 2014, 2018; Xiang (2016)). Examples are provided below to

remind the reader.

(128) Modal/Non-Finite Questions
  a. Where can I find coffee?
  b. Dana knows where we can find coffee
  c. Dana knows where to find coffee

(129) Non-Modal/Finite Questions
  a. Who came to the party?
  b. Dana knows who came to the party

If the presence of such modality (covert or overt) is grammatically necessary for mention-some (as argued by these semantic accounts), then we might expect that matrix verbs which do not encode strong exhaustivity—i.e., surprise and predict—will have a higher co-occurrence with these clause types than with finite (non-modal) clause types. In fact, we might reasonably expect that these verbs would co-occur with modals and non-finite clause types since these are the ones that permit non-strong-exhaustive readings.

Up to this point, we have treated modality as something of a homogenous class, and have focused only on the overt modal can, following the semantics literature. However, we know that modal meaning varies on two dimensions: flavor and quantificational force (Kratzer 1989, 1992; Portner 2009). Some examples are presented below.

(130) Force
  a. Existential: can, could, might, may, possible
  b. Universal: should, must, necessary, ought

(131) Flavor
  a. Deontic: interpreted with respect to laws (You should wear a seat belt) or obligations (You ought to call your mom)
  b. Bouletic: interpreted with respect to ability (You can have one), or desire (You should try this cake)
  c. Teleological: interpreted with respect to goals (You can take the subway (to get to Central Park))
  d. Epistemic: interpreted with respect to a body of knowledge (It might be raining, The keys must be in the cabinet)

What aspect(s) of modality are relevant to non-exhaustivity? Discussion in the semantics literature are not explicit about the answer. George 2011, Fox (2014), and Xiang

(2016) have focused on can's existential force as a possibility modal. The advantage is that it renders the modal facts a natural class with mention-some questions that have existential quantifiers. This would predict that other possibility modals like the epistemic may and might should also permit non-exhaustive readings.

In contrast, Dayal (2016) discusses can as a priority modal, following Portner's (2009) classification. The category of "priority modality" cuts across modal force to refer to modal flavor: priority modals refer to necessary or possible ways of achieving a goal or priority set by context. It seems uncontroversial that existential priority modals (can and could) give rise to non-exhaustive readings. But we may ask two questions: Can other existential (non-priority) modals give rise to mention-some readings? and, Can universal priority modals give rise to mention-some readings?

To answer the first question, consider examples (132)-(134) modified slightly from Dayal (2016), Section 2.3.

(132) **Context**: Fox and Dana see a light in the office. Often, Walter, Alex, and Pat are in the office working late.
   a. Fox: Who might be in at this time?
   b. Dana: # Walter may be in. Or Alex.

(133) **Context**: Fox needs help. Often, Walter, Alex, and Pat are in the office working late.
   a. Fox: Who might be in at this time?
   b. Dana: Walter may be in. Or Alex.

(134) Dana knows who might be in at this time

While a mention-some answer to a root epistemic modal question is infelicitous in (132), a mention-some answer is felicitous in (133). In the latter, the explicitly provided goal licenses a mention-some answer. Dayal is less certain whether an embedded epistemic modal as in (134) is felicitous on a mention-some reading, even with contextual support.

Another way to pull apart the first question, is to see whether the non-priority aspect of can allows mention-some.[1] This would mean accessing an inherent-ability reading of can. Compare (135) to (136).

[1]Thanks again to Lydia Newkirk for this suggestion.

(135) **Deontic Context**: in order to be eligible for coronavirus testing in New Jersey, a person must meet two criterion. First, they must know that they've come into contact with someone who has tested positive. Second, they must be manifesting the following symptoms: fever, dry cough, exhaustion, shortness of breath.
   a. Who can get tested for coronavirus?
   b. Dana can. Or Walter.

(136) **Inherent-Ability Context**: Fox is curious to know whether anyone here speaks French.
   a. Who here can speak French?
   b. Dana can. Or Walter.

(135) seem perfectly acceptable, and indeed resembles the example from Dayal (2016) which showed that plural marking in wh-words does not block mention-some. Dayal notes that the disjuction unambiguously signals mention-some. As for (136), a brief poll of native English speakers reveals that a mention-all answer seems preferable unless an explicit non-exhaustive goal is provided (echoing Dayal, 2016). While this could suggest a grammatical restriction on mention-some, it is also compatible with default mention-all interpretation with a non-grammatical etiology, as suggested in previous chapters. Note that if specified that we needed a French translator, then (136b) is unequivocally felicitous.

To answer the second question, let us consider the following brief exegesis of Bhatt (1999), Section 4.3.4. Bhatt argues that infinitival questions carry a covert modal, and discusses the various ways modal flavor and force of this covert modal are realized and constrained. He notes that in some cases, it is natural to paraphrase infinitival question with the universal priority modal should, and in other cases with an existential priority modal (can or could). Now, Bhatt notes that those paraphrases with should are often non-exhaustive, but distinguishes this from mention-some (pp. 156-158). Consider (137a) and (137b), which are judged true in the given context, according to Bhatt.

(137) **Context**: The only way to become popular is by talking to Magnus, Herb, and Penna, or by talking to Daniel, Stefan-Árni, and Baldur. Didda knows that she can become popular by talking to Magnus, Herb, and Penna, but *doesn't know* she can become so by talking to Daniel, Stefan-Árni, and Baldur.
   a. Didda knows who to talk to at the party.
   b. Didda knows who she should talk to at the party.

Bhatt says explicitly that (137b) allows a non-exhaustive interpretation consistent with the context provided, but in a footnote he distinguishes this non-exhastivity from a mention-some. A mention-some answer, he says, is any member of the question denotation that is a subset of the exhaustive answer. Thus, a mention-some answer for (137) could be one where Didda knows that talking to Magnus will make her popular. Bhatt asserts that this reading of (137a)/(137b) is false. The reason that the non-exhaustive interpretation in (137) is not mention-some, is that the answer Didda should talk to Magnus, Herb, and Penna is not a member of the question denotation, according to Bhatt. Intuitively, it seems that this is a perfectly acceptable mention-some answer

The goal of the current work is to understand the general phenomenon of non-exhaustivity, including mention-some readings—we do not distinguish mention-some from non-exhaustive readings. Regardless of whether we call this reading mention-some or non-exhaustive, should is clearly not an existential modal, and allows for a non-exhaustive reading. According to Rubinstein (2012), while strong necessity modals like must and have to impose a necessity according to all mutually-agreed-upon priorities, weak necessity modals like should also include consideration of priorities which may not be mutual (like personal preferences). In contrast, Rullmann et al. (2008) and von Fintel & Iatridou (2008) argue for a domain restriction account of weak necessity modals, in which they are weak in virtue of a small restricted domain. In either case, it seems there would be room for non-exhaustivity, in virtue of these non-overlapping priorities, or an extremely restricted quantificational domain. In the latter case, the domain restricted non-exhaustive reading is indeed different from an existential non-exhaustive reading. We discussed in Chapter 2 reasons for why domain restriction does not seem the the best way to capture non-exhaustivity.

Cross-linguistically, modals are lexically realized in a variety of ways, and much of modal meaning (even in English) must be resolved contextually. Many languages, like English, realize modal flavor in a variety of ways, with distinct morpho-syntactic repercussions (St'át'imcets (Lillooet Salish): Matthewson et al. 2007, Rullmann et al. 2008, Davis et al. 2009; Javanese (Austronesian): Vander Klok, 2008, 2012; Blackfoot

(Algonquin): Reis Silva 2009, 2013; Kwakwala (Wakashan): Menzies 2012; Nez Perce (Penutial): Deal 2011; Nsyilxcen (Okanagan Salish): Menzies 2012). Even so, a given modal flavor will still depend on relevant facts and opinions specified by context (cf. Abusch 2012).[2]

In contrast, English, like many other languages, lexically realizes modal force. Gitskan, Nez Perce, and Nsyilxcen (Okanagan Salish) are examples of languages that lack duals, but have a single modal which is felicitous in both possibility and necessity contexts. The answer of how exactly to analyze these modals is open, but the two logical approaches parallel the general issue in analyzing questions: one could argue that the modals are existential in force and then strengthened in context (Deal 2011 for Nez Perce and Peterson 2010 for Gitksan), or one could argue that the modals have universal force and must be weakened (via domain restriction) in context (Rullmann et al. 2008, von Fintel & Iatridou 2008).

For the learner, these facts make modality a somewhat unpredictable cue to mention-some. As a cue, then, we would expect to see differences in the learning trajectory based on the language-particular facts, how much the relevant meaning is lexicalized or left to contextual specification.

Before moving on to the next factor of interest, it is worth noting that there are several ways that non-finite modality interacts with them, discussed in Bhatt (1999), Section 4.2.2. Rather than placing those discussions here, I will address them as they arise in their respective sections.

### Matrix Verbs

First, we are interested in the particular matrix verbs know, surprise, and predict because they have factored into arguments about (non-)exhaustivity. To summarize the claims, know is often said to favor strong exhaustivity,[3] if not to require it (cf. Groenendijk &

[2] I am grateful to Lydia Newkirk for references of modality cross-linguistically.

[3] As a reminder, on a strong exhaustive reading of an embedded question report like Dana knows who came to the party is true iff Dana knows for each person, whether they came to the party. Equivalently, iff she knows that the people who came are the only people who came. Dana must know not only who came, but that no one else did. In contrast, weak exhaustivity leaves room for Dana to know nothing

Stokhof (1982), (1984); Berman 1991; Heim 1994, Beck & Rullmann 1999; George 2011, Schulz & Roeper 2011). While Heim's (1994) theory made it possible to assign either weak or strong exhaustive meanings to embedded questions, it was acknowledged that know-wh (on the basis of know-who questions) are interpreted as strongly exhaustive more often than not. In contrast, emotive factives like surprise and non-factives like predict have been crucial to arguments against strong exhaustivity (Berman 1991; Heim 1994; Beck & Rullmann 1999; Sharvit 2002; Guerzoni & Sharvit 2007; George 2011; Klinedinst & Rothschild 2011; Uegaki 2015). These are said to prefer/require weak exhaustivity, or even non-exhaustivity (on George's semantics). While no verbs are said to prefer mention-some *per se*, it is then a puzzle that know-how and know-where questions felicitously permit non-exhaustivity.

Bhatt (1999), Section 4.2.2.2, discusses how matrix verbs may interact with the modal force in non-finite questions. In particular, he notes that non-finite know-questions most naturally have paraphrases with the possibility modal could rather than with should. We might think that this is somewhat contrasting with the general sense of know as an exhaustive cue. Here, we might think that non-finite know-wh would be better called a non-exhaustive cue given the way the modal force is resolved.

**Wh-Words**

Second, there are asymmetries in the baseline preferences for (non-)exhaustivity exhibited by different wh-questions (Ginzburg (1995); Asher & Lascarides (1998)). In brief, how-questions prefer non-exhaustivity, while who-questions prefer exhaustivity. Asher & Lascarides (1998) note that this fact emerges when we examine the data used in support of these different readings: those who argue for a non-exhaustive semantics typically use data from how-questions and why-questions (Hintikka 1976, and Asher & Lascarides (1998)), while those who typically argue for a weak or strong exhaustive semantics support their theories with who-questions (Karttunen 1977; Groenendijk & Stokhof (1982), (1984); Heim 1994).

---

about non-party-goers.

Additionally, we will include degree questions in the current analysis (e.g., How many eggs does the recipe need?). Beck & Rullmann 1999 provide evidence for weaker degrees of exhaustivity in how-questions. Depending on the monotonicity of the question predicate, the question will be resolved by either the maximal or minimal degree. In those cases where the minimal degree resolves the question, weak exhaustivity is required. Further, when a degree question occurs with at least or at most, they argue that a mention-some meaning must be available to derive the correct interpretation.

Note that the wh-type interacts with CLAUSETYPE in the following way. First, Bhatt (1999) Section 4.2.2 discusses that non-finite who questions are better paraphrased with the universal modal should, while non-finite where and how questions with the existential modal could. While the discussion from the modal section above would suggest that should paraphrases are not incompatible with non-exhaustivity, we might think that non-finite who questions would not be a strong cue in virtue of allowing multiple readings, while non-finite how and where questions would be stronger cues to non-exhaustivity.

Secondly, in Section 4.3.6, Bhatt discusses the interaction between non-finite clauses with degree questions. He claims that non-finite degree questions only have should paraphrases:

(138) a. Penna knows how many people to invite to the party.
b. The NATO spokesman knows how much to say about the bombing of civilian facilities.
c. Olafur knows how detailed to make his presentation.

In contrast to the earlier claim about non-finite how-questions, we might consider non-finite degree questions specifically to lean more towards cueing exhaustivity, or not being a strong cue at all, given the variation in (non-)exhaustivity found with should paraphrases (see previous discussion in the section on modals).

In additional to tracking degree questions generally, we will specifically look at degree questions co-occurring with at least and at most, and degree questions occurring with non-finite clauses. The first co-occurrence combination would constitute a non-exhaustive cue, while the second we might call an exhaustive cue.

**Miscellaenous Cues**

Finally, there are several other words which may provide cues to exhaustive and non-exhaustive goals. The term "cue word" should be understood as any word that might serve to signal (non-)exhaustivity to the hearer.

(139) **Exhaustive cue words**
  a. all: Where all can we find coffee? / Where are all the places to find coffee?
  b. every: Who was everyone at the party?

(140) **Non-exhaustive cue words**
  a. i. some: Where are some places to find coffee?
     ii. any: Where can we find any sugar?
     iii. at least/at most: How many eggs can you eat at most/at least?
  b. best, local, favorite, near/nearby, close, suggest, good, common

The words in (139) and (140a) have been discussed often in the literature because cross-linguistically we find such words which impose additional quantificational constraints on the question (Beck & Rullmann 1999, Zimmermann 2007, Zimmermann 2010, Bade, ms). These are called Quantifying Question Particles. Typically, discussion of such cues begins and ends with all/some because of the sense that the presence of one or the other in a question is sufficient to license (non-)exhaustivity regardless of any other property of the question. Similarly, any is suggestive of a free choice question (Dayal 2016). Another point made by Beck & Rullmann 1999 is that when degree questions occur with at least and at most, they argue that the only interpretation which derives the correct truth conditions is the mention-some interpretation.

The words in (140b) have been little discussed, and not as a group. These words all signal a restriction on the referential domain of the wh-word. For example, our classic case of the tourist looking for a cup of coffee might ask a question using any of the words in (140b) to make her goals clear to the speaker (e.g., Where is a nearby/local/close coffee shop?). We might think that such domain restrictors might signal to the hearer that a non-exhaustive answer is acceptable.

## 5.2 A corpus analysis

### 5.2.1 Methods

The following data and code may be found at https://github.com/rangat/whAnalysis for the initial parsing scripts, and https://github.com/mcmoyer11/Questions-Corpus-Analysisfor Jupyter Notebooks for spot-checking and looking through the data. For this study, we used four corpuses. Three of those come from the Natural Language Toolkit (NLTK) free corpora package: Australian Broadcasting Corpus (ABC), Reuters, and the Penn Treebank. We did not use the Brown corpus because the language was somewhat archaic. Additionally, the bulk of our data come from the British National Corpus (BNC). Table 5.1 shows the breakdown of number of sentences used per corpus. Each corpus

| Corpus | Number of sentences |
|---|---|
| British National Corpus | 419,075 |
| Australian Broadcasting Company | 2410 |
| Reuters | 1247 |
| Penn Treebank | 251 |

**Figure 5.1:** Corpuses used in the study, with number of sentences from each corpus.

was tagged for part-of-speech and sentence tokenized using the Natural Language Toolkit (NLTK) POS tagger (Bird, Loper, & Klein 2009). We extracted all sentences occurring with either a who, where, or how. These were stored as a .json object. Finally, we applied a set of ordered heuristics to code for type of question (QUESTTYPE) and type of clause (CLAUSETYPE). The general strategy for tagging involved a search for matching patterns (linear orders) of POS tags, or a combination of POS tags and sets of strings (tokens of particular words). The next section discusses these pattern matching heuristics in more detail.

#### Heuristics

In the discussion below, I group the heuristics into two categories based on whether they tagged sentences to be excluded or to be included in the analysis. The "exclusionary" heuristics tagged RELATIVE CLAUSES, AMBIGUOUS sentences (those whose

QUESTTYPE status could not be further determined), and FRAGMENTS (essentially, those sentences without any verb after the WH, like Who?). The "inclusionary" heuristics tagged ROOT QUESTIONS and EMBEDDED QUESTIONS, based on typical patterns associated with these two kinds of constructions, or in some cases the lack of patterns associated with the excluded categories.

For most categories, we had a strict and a weak heuristic. The strict heuristic required a pattern match involving a larger sequence of POS tags, while the weaker heuristic involved one with less POS tags. This was necessary because of the variety of patterns associated with each category, and the nature of the pattern-matching strategy. Often the strict pattern match would miss some acceptable sentences for that category, but the weaker one was too permissive and would tag undesired sentences as the category. The heuristics applied sequentially throughout the entire dataset, removing sentences as they were tagged, allowing more permissive heuristics to safely apply. Thus, by combining weak and strict heuristics, as well as ordering them, we could be confident that the heuristics were accurate more often than not. To further maximize accuracy for analysis, each heuristic was spot-checked after initial parsing of the dataset, and the dataset modified directly if errors were found. These spot-checking scripts are located at https://github.com/mcmoyer11/Questions-Corpus-Analysis.

### Exclusionary QUESTTYPE Heuristics

We will discuss the heuristics one-by one, grouped by shared output category tag. The number in parenthesis represents the order that the heuristic was applied to the data. The first set of heuristics in (141), all tagged for RELATIVE CLAUSE.

(141) **RELATIVE CLAUSE**

| | | |
|---|---|---|
| a. | VB_{RC}_WH | STRICT (4) |
| b. | {RC}_WH | WEAK (5) |

{RC} = the set of NLTK POS tags for RC heads

Relative clauses are interrogatives embedded under DPs, and are thus not themselves

the complement of matrix verbs. We can approximate relative clause syntax by looking at whether DPs occur between matrix verbs and each wh-phrase. This is essentially what the two heuristics above do. Since all words were tagged for POS before any heuristic applied, DPs were determined by the following POS tags: 'NN', 'NNS', 'NNP', 'NNPS', 'DT', 'JJ', 'PDT', 'POS', 'PRP', 'PRP$', 'CD'. There were 330027 sentences tagged as RELATIVE CLAUSE.

The first (141a) is stricter, and checks whether a sentence matches the pattern of VB_{RC}_WH. This strict heuristic captured the bulk of relative clauses (80%). The weaker heuristic in (141b) merely labeled sentences matching relative clauses appearing sentence-initial, where there was no verb preceding the wh-word. (141b) captured the remaining 20% of relative clauses. Example sentences are provided below in (142) and (143).

(142) **STRICT RELATIVE CLAUSE HEURISTIC**
It recommended that those who have to move away from home to attend university should be automatically eligible for youth allowance. (#151)

(143) **WEAK RELATIVE CLAUSE HEURISTIC**
The Docking family, who are based on the fringes of Darwin' s rural area at Berry Springs, are making the most of drier early build - up weather (#194)

Notice in the first sentence there is a matrix verb, but none in the second. Without the WEAK RELATIVE CLAUSE heuristic, the second sentence would have not been labeled as a relative clause and even may have been labeled as something else.

The two heuristics in (144) and (145) tagged sentences that were FRAGMENTS (those without any verbs) or AMBIGUOUS (those whose status could not be easily determined).

(144) **FRAGMENT**
¬VB (1)

(145) **AMBIGUOUS**
ELSE (9)

There were 8662 sentences tagged as FRAGMENT, and 17810 tagged as AMBIGUOUS. Examples of sentences caught by these two heuristics are presented below.

(146) **Sentences tagged as FRAGMENT.**

a. How about a Sainsbury's one? (#251956)
b. Ace Barton, that's who! (# 269791)
c. But there's one I thought how stupid. (# 296171)

(147) **Sentences tagged as AMBIGUOUS.**

a. Where no evidence of infection can be found, the complaint is sometimes called prostatodynia ('prostate pain').(#5039)
How it must have hurt him. (# 12562)
b. How nice to run into you. (#16544)
How naïve of her to let Roman de Sciorto's powerful charm override her normal caution! (#24470)
c. Where aircraft get tampered with and fuel caps get left off and possibly fuel tanks contaminated. (#1342)
How those resources are apportioned is a matter for the council. (#5633)

Spot-checking revealed that there were several cases that were legitimate instances of root questions with contracted verbs. These were separated and re-labeled appropriately to be added to the analysis. As for the AMBIGUOUS cases, there were also several legitimate root questions missed by ROOT QUESTION heuristics. These were sentences occurring with a final ? and an additional punctuation mark. These were separated and relabeled. All remaining cases of AMBIGUOUS questions appear to be legitimately ambiguous, or at least not directly of interest to the current analysis (like sentential subject interrogatives).

### QUESTTYPE Heuristics (Inclusionary)

(148a) and (148b) introduce the heuristics used to label sentences that were to be included in analyses. These heuristics labeled sentences as EMBEDDED and ROOT QUESTIONS. Let us begin with ROOT QUESTIONS.

(148) **ROOT QUESTIONS**

a. WH_{AUX}_{RC}_VB (6)
$\wedge \neg$WH_{RC}_{AUX}
$\wedge \neg$ {AUX}_VB_{RC}
b. ?_# (7)

{RC} = the set of NLTK POS tags for RC heads
{AUX} = the set of (non-modal) auxiliary verbs

The first rule for ROOT QUESTIONS checks for Subject-Aux Inversion, and it labelled 6867 sentences as root questions. The second heuristic searched for strings ending in

?, and labelled 23219 sentences. This latter heuristic failed to catch sentences where a ? was followed by another punctuation mark, as mentioned above.

(149) **Caught by (148a)**
How did you do it? (# 4587)
Who was the burglar breaking in through an entryphone? (#8222)

(150) **Caught by (148b)**
How much old research is based on fraud? (# 1888)
Where do we learn how to behave in the intimacy of marriage? (# 16119)
Who played Hilts, the cooler king, in The Great Escape? (# 17717)

There are three EMBEDDED QUESTION heuristics, each below. The first tagged 34975 sentences, the second 5130, and the third 333.

(151) **EMBEDDED QUESTIONS**
a. {RC}_VB_¬{RC}_WH (2)
b. VB_WH ∧¬ S-AUX-INV (3)
c. VB_WH (8)
{RC} = the set of NLTK POS tags for Relative Clause heads

These heuristics decrease in strictness. Example sentences are provided below, respectively. The first one tags sentences which match a pattern like DP_VB_¬DP_WH, which requires a matrix subject, a matrix verb, and no DP-ish object intervening between the matrix verb and the WH. The second looks for sequences of V_WH where there's no subject-auxiliary inversion after the wh-word (implemented by Rule 148a). Finally, the third heuristic is the weakest, looking for strings that matched a pattern where a verb preceds the wh-word.

(152) **Caught by (151a)**
He questioned who would monitor and pay for the proposed body. (#395)
The authority cannot confirm how much meat headed for the domestic market has been wrongly labelled. (# 1781)

(153) **Caught by (151b)**
And see how wide it can be and also in some respects how remote it can be. (# 52256)
And didn't know where to find them, leave them alone and they will come home wagging their tails behind them (# 71802)

(154) **Caught by (151c)**
Reflecting how Laura herself invariably shied away from mention of her gifts, I was left with my thoughts. (# 51436)
Explain how you might seek evidence to evaluate these hypotheses. (# 51737)

The third heuristic mislabeled sentences like 'How could I refuse?' he said. (#3782) as EMBEDDED QUESTIONS most likely because of the additional punctuation marks which caused a lot of trouble. During spot-checking, these were caught and correctly re-labelled as ROOT QUESTIONS.

In general, the parsing script did a thorough job in its initial parsing of the dataset for QUESTTYPE, and greatly reduced the number of hours which would have been needed to code the data. Systematic errors revealed during spot-checking were easily rectified by modifying the .json file directly. The numbers reported in the Results are from the post-spot-checking dataset.

#### CLAUSETYPE Heuristics

The CLAUSETYPE heuristics are presented below. Note, the heuristics are the same for both ROOT and EMBEDDED QUESTIONS, except that the NON-FINITE CLAUSETYPE tag technically only applied to EMBEDDED QUESTIONS. There actually were some instances of NON-FINITE ROOT QUESTIONS, which will be briefly discussed in the Results section.

(155)

| | | |
|---|---|---|
| a. | WH_{MOD} | MODAL (1) |
| b. | WH_to_VB | NON-FINITE (2) |
| c. | ALL ELSE | FINITE (3) |

{MOD} = the set of modal auxiliary verbs

The MODAL CLAUSETYPE tag searches for patterns matching a modal auxiliary following the wh-word, the NON-FINITE CLAUSETYPE searched for patterns where a to occurs between the wh-word and the verb. Finally, everything else was tagged as FINITE CLAUSETYPE. Note that while modal questions are tensed finite clauses, here we use the FINITE CLAUSETYPE tag to refer specifically to *non-modal* finite clauses.

### 5.2.2 Results: CLAUSETYPE and Wh-Word Overall

In the following sections, I will discuss the results from this corpus search with a focus on the factors of interest we have been discussing. While there are many interesting

avenues to explore in the dataset, I adopt this plan for the sake of space. In such cases, I include more exhaustive data in the Appendix.

Analyses use Pearson's $\chi^2$ tests of proportion where observations are large enough to do so. Figure 5.2 presents the overall distribution of QUESTTYPE in the final dataset.

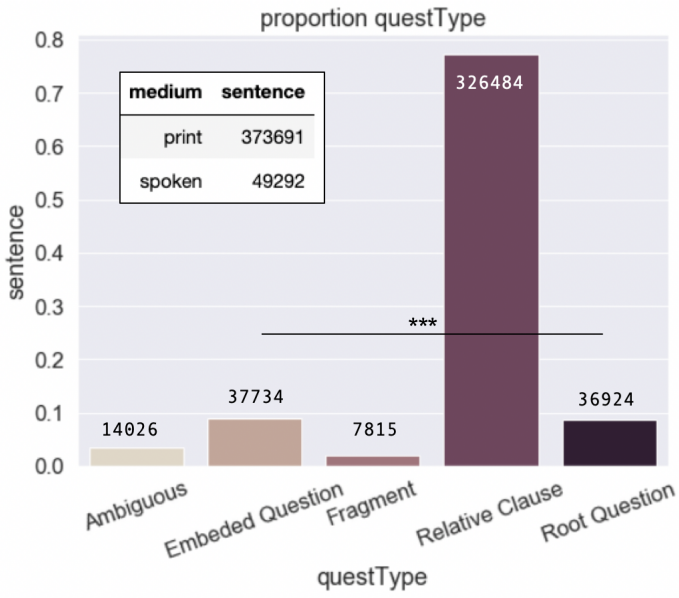


**Figure 5.2:** Distribution of QUESTTYPE over entire final dataset.

There were 422,983 wh-clauses in the entire dataset. We can see that the overwhelming majority are RELATIVE CLAUSES, about 77%. Only about 8.9% of wh-clauses are classified as EMBEDDED, and only about 8.7% are classified as ROOT. About 3% were tagged as AMBIGUOUS, and a little under 2% as FRAGMENT.

The next several graphs take a closer look at the distribution of factors of interest in EMBEDDED and ROOT questions alone. Figure 5.3 presents the distribution of CLAUSETYPE across the three types of wh-questions for both ROOT and EMBEDDED. First, FINITE clauses are produced with the highest frequency across the board ($\chi^2(1) = 57576$, $p < 0.0001$). Second, how-questions are also produced significantly more frequently across the board ($\chi^2(1) = 5758.8$, $p < 0.0001$), and significantly more so than other wh-types in each type of clause.

Of these HOW-questions, about 29% are degree questions— they occur with an adjective in between how and the first verb (12625 observations). Overall, there is

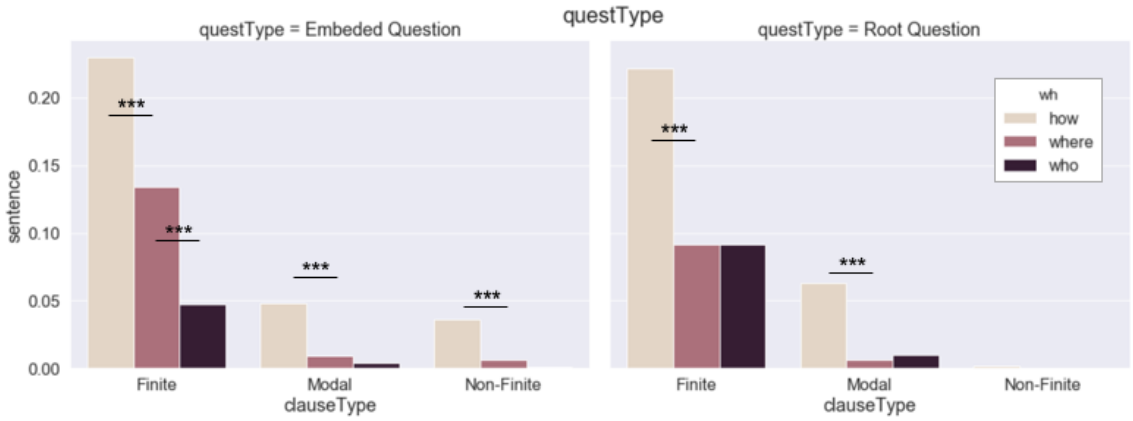


**Figure 5.3:** Distribution of CLAUSETYPES across WH and ROOT and EMBEDDED QUESTTYPE.

nothing outstanding about the distribution of factors in these that differs from the rest of how-questions. In particular, the total of undifferentiated how-questions is about 19% constituted by MODAL CLAUSETYPE; when degree questions are removed, actually the proportion rises 3 points to about 21% MODAL CLAUSETYPE ($\chi^2(1) = 57.358$, $p > 0.0001$).

The remaining results will be discussed in the following order. First, Section 5.2.2 will discuss the results for MODAL CLAUSETYPE, Section 5.2.2 will discuss matrix verbs, with a focus on know, predict, and surprise. Know is often said to be (strongly) exhaustive, while predict and surprise to be (weakly) exhaustive or even non-exhaustive. Thus, comparing the distribution of cues and their co-occurrence in questions with these verbs may be informative. Finally, Section 5.2.2 looks at the distribution of additional words which may provide cues to (non-)exhaustivity. These are the words identified in Section 5.1.

#### Results: MODAL CLAUSETYPE

Figure 5.4 presents the distribution of modal auxiliaries across ROOT and EMBEDDED QUESTIONS, and across the three wh-types. What we see is that can is the most frequent, followed by would and could. Recall that can and could are existential priority modals and license mention-some readings. In comparison, epistemic and universal modals are much less frequent (except for would).

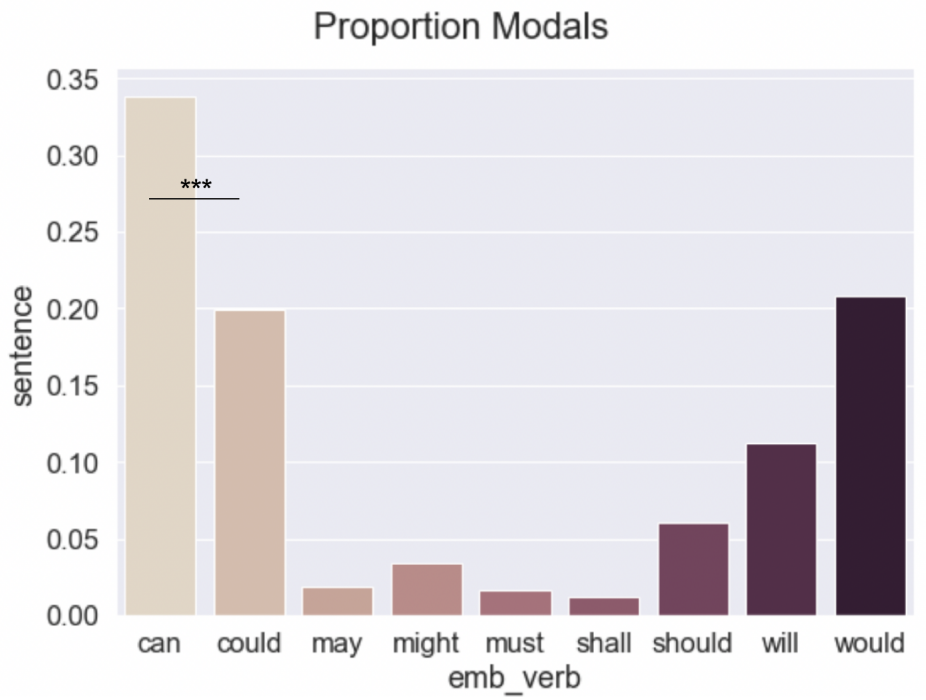


**Figure 5.4:** Distribution of modal auxiliaries.

Following the overall trend in the dataset, 79% of modal questions are modal HOW-questions ($\chi^2(1) = 6845.9$, $p < 0.0001$). Figure 5.5 presents the proportion each modal auxiliary occurs relative to QUESTTYPE and wh-type. Here we see that, across both ROOT and EMBEDDED questions, can/could occur in highest proportion in how-questions, while would, a necessity modal, occurs in the highest proportion in who-questions. When we compare the distribution of other factors in can/could-questions

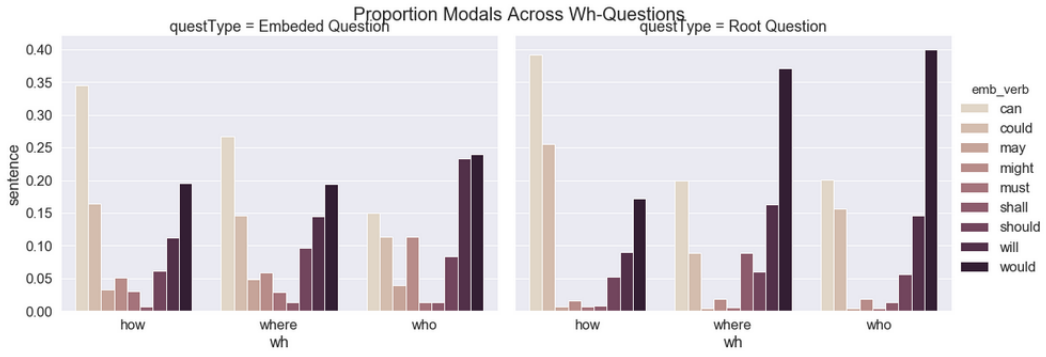


**Figure 5.5:** Distribution of modals across questType and Wh in corpus.

to will/would, we find that in both cases, how-questions are the most frequent wh-type.

Where they differ is in the frequency of matrix verbs. For can/could, surprisingly, know is not the overall most frequent (11%, at 242 observations), but see is (13%, at 292 observations). In fact, this difference is significant ($\chi^2(1) = 5.357$, $p$=0.02). In contrast, know is the most frequent matrix verb for willwould-questions (14%, 208 observations). In fact, know is significantly higher frequency in will/would questions than in can/could questions ($\chi^2(1) = 6.7771$, $p < 0.001$).

**Results: Matrix Verbs**

Overall, know is the most frequently occurring verb (20% of all embedded questions) followed by be (11%). Figure 5.6 presents the matrix verbs which occur more than 300 times. Note that 's has not been lemmatized because it is an acceptable contraction for both have and be. These frequencies are not consistent when we look across different

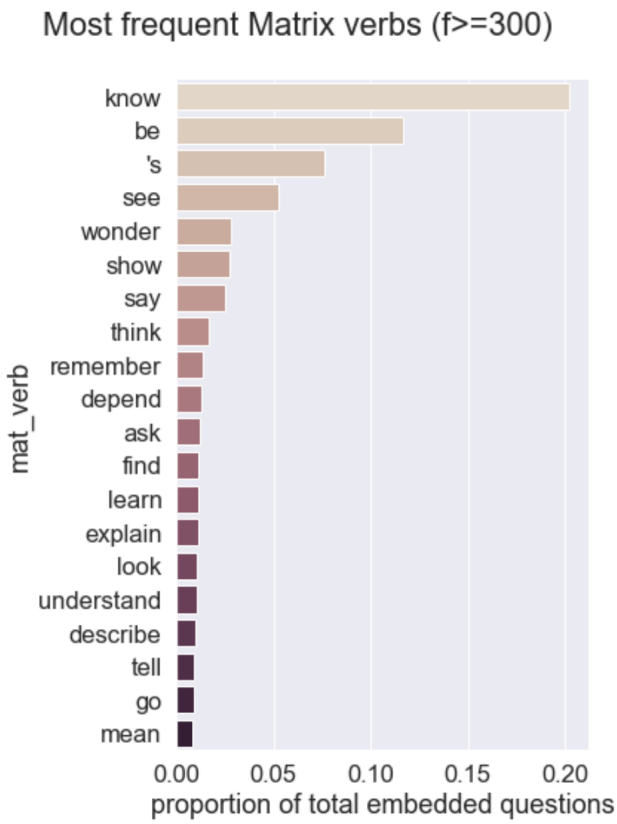


**Figure 5.6:** Distribution of MATRIX VERBS across all embedded questions.

wh-type and CLAUSETYPE. Figures 5.7 -5.9 presents the most frequent matrix verbs

broken down by these two factors. The number in parentheses is the frequency cutoff used for graphing purposes. There were a small number of high frequency words, and a large number of small frequency words in the corpus. We see here some slight differences based on wh-type and CLAUSETYPE, which I will discuss in turn.

For FINITE clauses (Figure 5.7), know outpaces other matrix verbs in frequency for who-questions, and a little less so for how-questions. For where-questions, we see that be-where is almost as frequent as know-where. For the moment, it is hard to make many generalizations about the meanings of these questions, since we have little information about the verbs inside the question, except that they are finite and lack modality. In fact, the most frequent embedded verbs are light verbs like be, have, do, go, which do not carry much semantic content.

For NON-FINITE clauses (Figure 5.8), know again outpaces every other matrix verb, with be coming in at a distant second except for how-questions. For how-questions, we see that learn is the second most frequent matrix verb.

Finally, for MODAL clauses (Figure 5.9) we see a different result. While know is still most frequent for how- and who-questions, be is the most frequent matrix verb for where-questions. The second most frequent for who-questions is be, for how-questions is see, and for where-questions is know.

Besides know, we can take a look at two other verbs that might be of interest theoretically. In particular, since surprise and predict are argued to be weak or even non-exhaustive, we might include them as non-exhaustive cues, in contrast to know, which is often argued to be strongly exhaustive. Note that these verbs have so far been absent from the graphs. They are incredibly infrequent: surprise occurs 63 times (about 0.17%) and predict occurs only 22 times (about 0.06%). Let us compare know with surprise and predict.

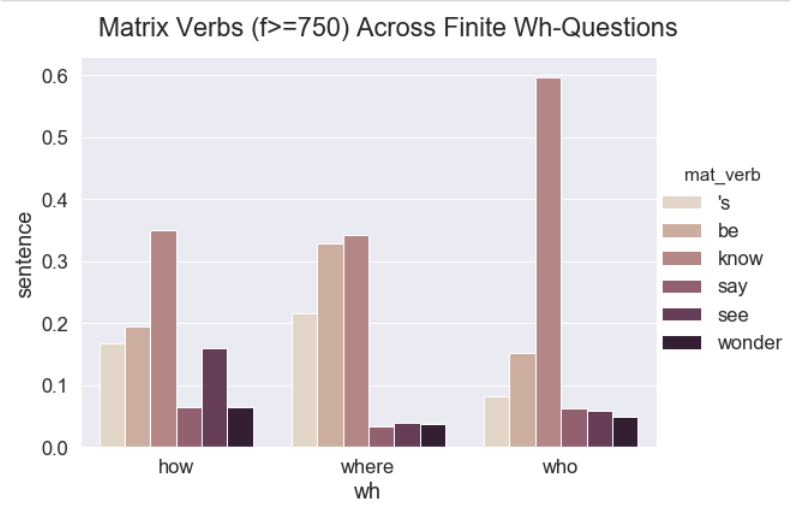


**Figure 5.7:** Distribution of MATRIX across FINITE wh-questions.

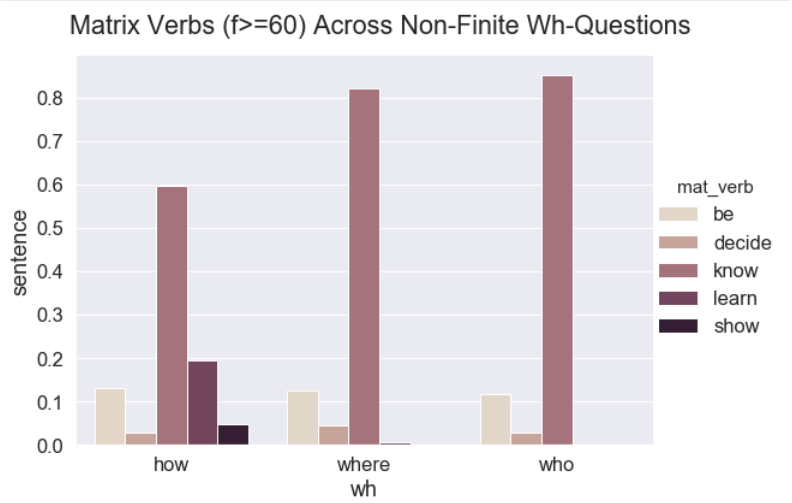


**Figure 5.8:** Distribution of MATRIX verbs across NON-FINITE wh-questions.

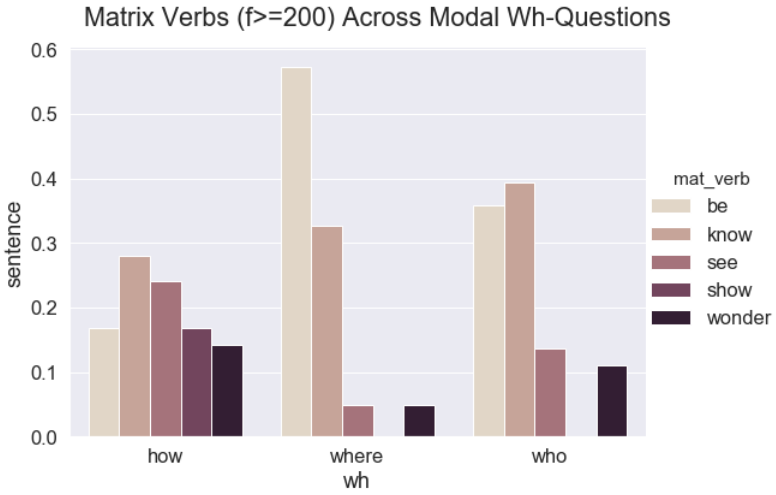


**Figure 5.9:** Distribution of MATRIX verbs across MODAL wh-questions.

**know-wh**

Let us first look at know in Figure 5.10. Know is the most frequent matrix verb in the

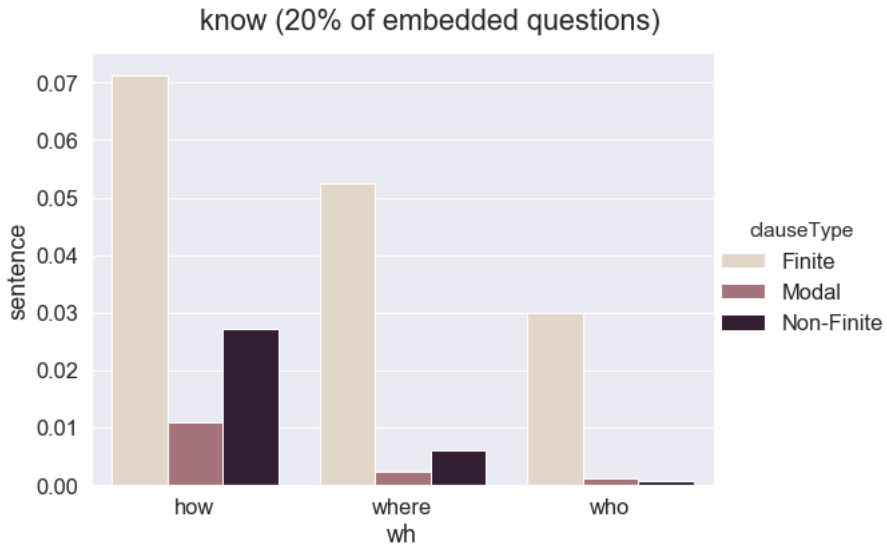


**Figure 5.10:** Distribution of factors in know-wh questions.

corpus, and it occurs most frequently with finite clauses across all three wh-types. If we return to the idea that non-exhaustivity is somehow correlated with modality and non-finiteness, while exhaustivity with finiteness, it would seem that we have know occurring overwhelmingly with an exhaustive cue. Couple this with the fact that know is also an exhaustive cue, and that it is the most frequent matrix verb, then we have a predominance of several cues to exhaustivity in embedded questions.

Interestingly, the most frequent wh-type that know embeds is not one that supposedly cues exhaustivity (i.e., a who-question), but one that supposedly does the opposite (how-, and even where-questions). Putting FINITE clauses aside, know-how questions account for the majority frequency of non-finite and modal clauses, while know-who questions barely occur with these CLAUSETYPES. So what we see is that, if there were correlations between frequency/co-occurrence of these form cues and (non-)exhaustivity, then it would pattern in exactly the way suggested by the previous literature: while know-wh might be default exhaustive (given the overwhelming frequency of finiteness), if any know-wh question would be non-exhaustive it would be know-how questions, and then know-where before know-who.

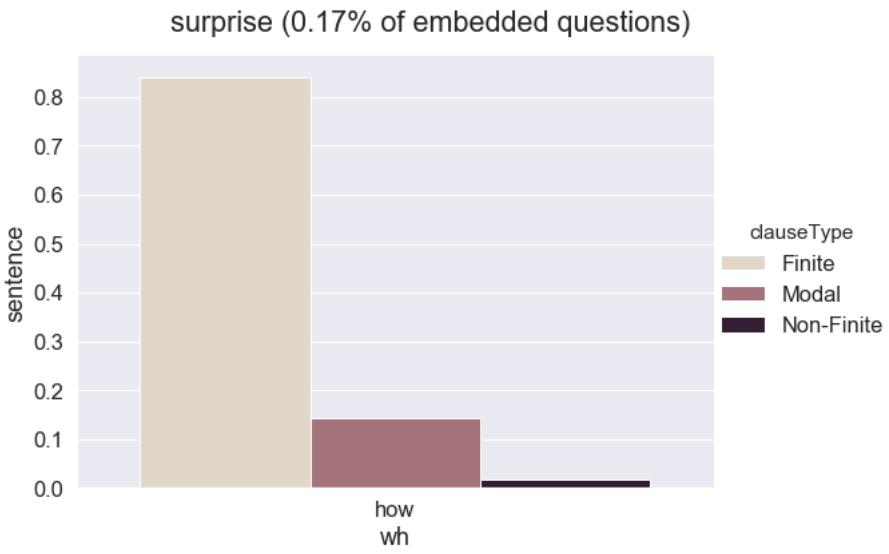


**Figure 5.11:** Distribution of factors in surprise-wh questions.

**surprise-wh**

Perhaps unsurprisingly given the general patterns in the dataset, the dominating CLAUSE-TYPE for surprise-wh questions is FINITE, and these are all how-questions. There are 6 observations of can/could-questions embedded under surprise (9.5%, out of 63 total observations of surprise-wh, all how-questions), 1 NON-FINITE clause (1.5%, out of 63 total).

A quick Google search for surprised who yields 919,000 results, 498,000 for surprised where, and around 11,000,000 for surprised how. The first couple results for surprised who reference, not the embedded construction, but a fictional character named "Surprised Who" from the movie *How the Grinch Stole Christmas*. Subsequent references to that bigram reveal constructions like Who surprised who? before cases with embedded questions. The infrequency of surprise who and surprise where constructions is further confirmed by Google's predictive search algorithm. None of the predicted searches reveal true embedded interrogative constructions. Screenshots of these Google searches are provided in the Appendix.

(156) presents some examples from the corpus. The majority of surprise-wh are degree questions, and bare how-questions are the exception.

(156) a. It was surprising how strong she was. (#310030)
b. You will be surprised by how much more often you dream than you think you do. (#182725)
c. You might be surprised at how things turn out. (#280088)
d. I was pleasantly surprised at how they performed as more experienced opponents (#323688)

Beck & Rullmann 1999 argue that degree questions are resolved by the maximal or minimal degree based on the monotonicity of the embedding predicate. If we do a more thorough analysis of surprise with degree questions, we should look for the monotonicity properties of the predicate to see whether there are more instances of the upper or lower bound degree interpretation. Given that degree questions allow for strong exhaustivity, and that Cremers & Chemla experimental investigation of exhaustivity in emotive factives found that surprise-wh indeed allowed for strong exhaustive readings despite the common intuition in the literature, we might then conclude that surprise provides conflicting evidence for (non-)exhaustivity.

**predict-wh**

While predict-wh constructions are even less frequent than surprise-wh constructions overall, they actually occur overwhelmingly with MODAL clauses across wh-questions, and how-questions over other wh-types. Google searches here confirm the pattern in our dataset: for predict-who 916,000 hits, for predict-where 974,000 hits, and predict-how 4,570,000 hits. The predictive search algorithm also confirms these all as true embedded interrogative constructions for all three wh-types.

(157) presents examples from the dataset.

(157) a. No-one can predict how events will unfold. (#104605)
b. These are solved to predict how it will behave under a set of prescribed conditions. (#200170)
c. No-one can predict how long a person will live. (#77615)
d. In order to predict how the universe should have started off, one needs laws that hold at the beginning of time. (#308944)

77% of predict questions are modal, but these occur with the future-oriented necessity modal will. There are 0 observations of either can or a NON-FINITE clause embedded

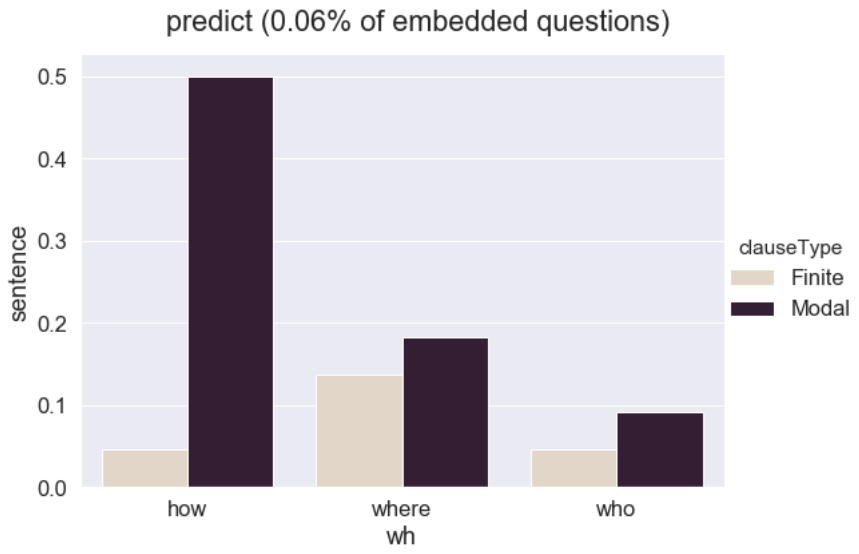


**Figure 5.12:** Distribution of factors in predict-wh questions.

under predict, and only one occurrence with could. If non-exhaustivity is only grammatically derived in the presence of existential modality, then we would not expect predict-wh to provide a stable cue to non-exhaustivity. However, recall from Experiment 1 that participants judged both predict-where and predict-who questions to be true on mention-some readings, regardless of finiteness in the embedded clause.

**Results: Additional Cue Words**

Finally, we can look at the distribution of general cue words. First, note that we find cue words in both ROOT and EMBEDDED questions. This is consistent with the observation about German and Dutch from Beck & Rullmann 1999 and with (Irish) English from McCloskey (1995) that cues can occur in embedded questions. Of the words we identified, about 26% of ROOT and EMBEDDED questions contain one of these.

Of these cues, all, any, and some are most frequent in the corpus, as well as most frequently discussed in the literature. All occurs in 10% of questions overall, any in 7%, and some in only 2%. In fact, all occurs significantly more than any ($\chi^2(1) = 447.72$, $p < 0.0001$). This is perhaps surprising for two reasons. First, if questions are semantically exhaustive, it would be unnecessary to include an explicit exhaustivity marker like all, especially in embedded questions. We would instead expect to see significantly

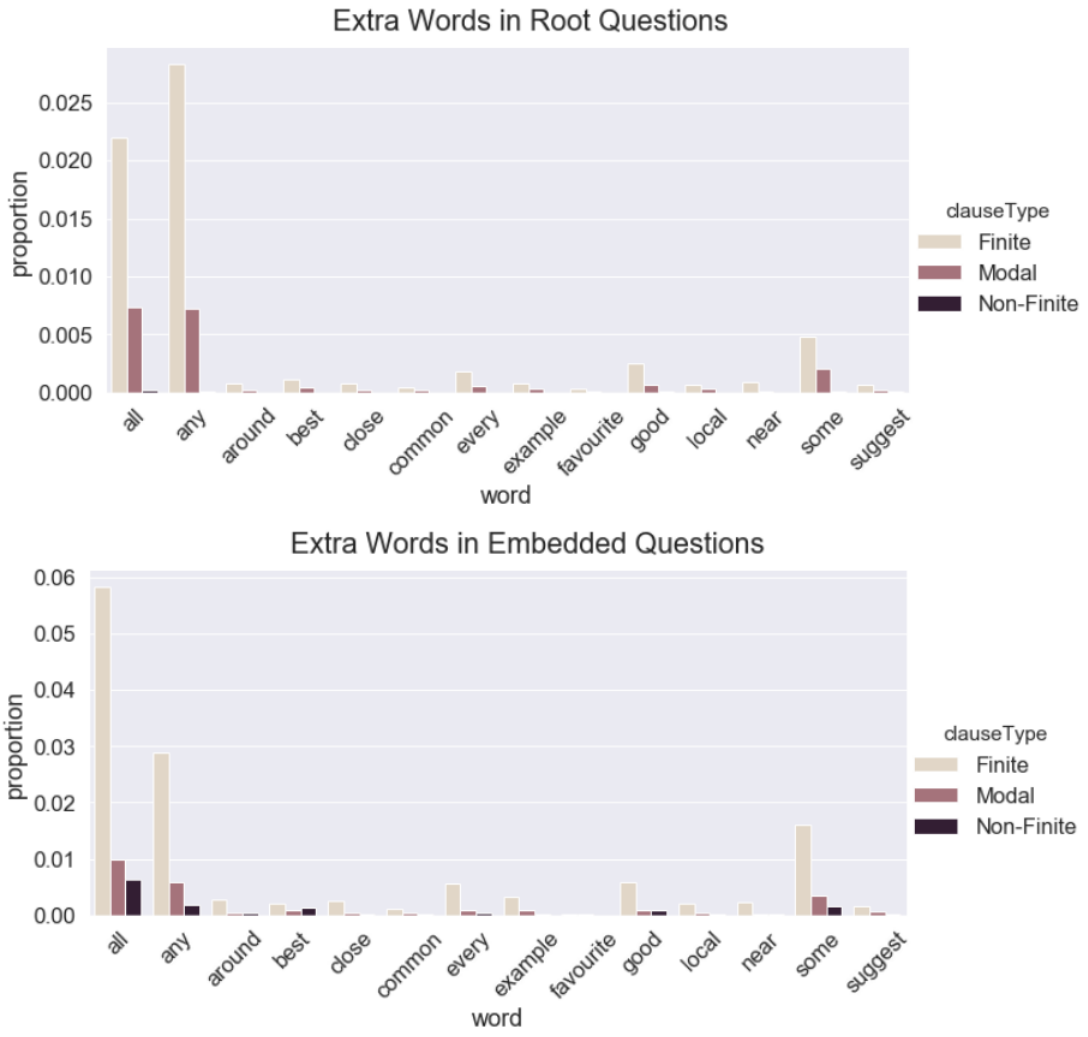


**Figure 5.13:** Distribution of other potential cue words across QUESTTYPE.

more occurrences of any or some to have a weakening effect. Even adding these two together, all still occurs significantly more frequently ($\chi^2(1) = 4.45$, $p < 0.04$), though the effect is smaller. There may be further structural factors that interact when any is present in a questions.

Figure 5.14 shows that these cues are produced in highest proportion in finite clauses. This would make sense if modal and non-finite clauses already cue non-exhaustivity, but FINITENESS alone does not clearly convey exhaustivity or non-exhaustivity (recall the initial discussion of CLAUSETYPE and wh-type). Figure 5.15 plots these three words across questType and wh-type, collapsing across clauseType. Note that root how-questions are an exception to the high-frequency of all. Here, we see that any occurs

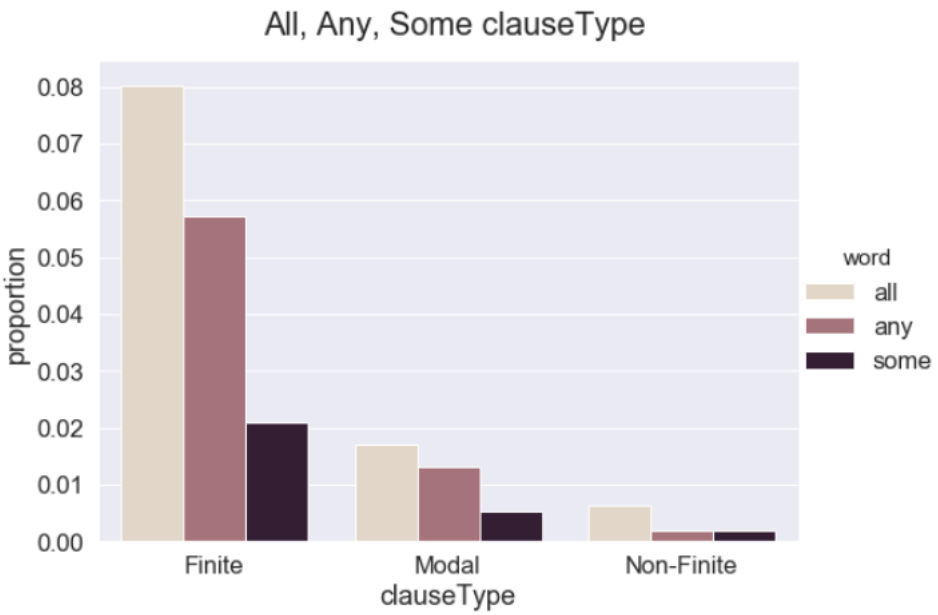


**Figure 5.14:** Distribution of all, any, and some across CLAUSETYPE.

significantly more frequently than all ($\chi^2(1) = 318.72$, $p < 0.0001$). Similar reasoning leads us to conclude that, if questions (or how-questions) are semantically specified for non-exhaustivity, we would not expect a non-exhaustive cue to be the most frequent. Since we see both exhaustive and non-exhaustive cues, these data collude with the previous experiments to provide evidence that questions are ambiguous or under-specified for (non-)exhaustivity.

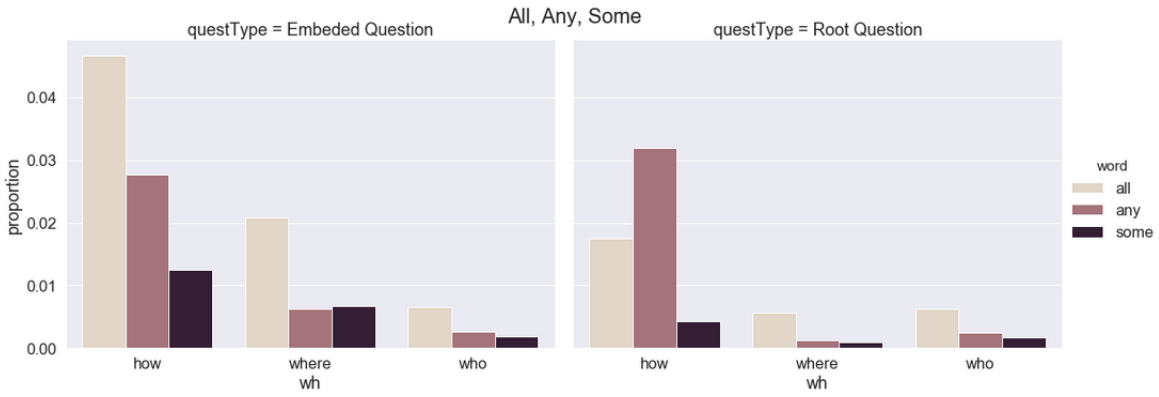


**Figure 5.15:** Distribution of ass, any, and some across QUESTTYPE and wh.

### 5.2.3 Discussion

This corpus study found conflicting evidence for the relationship between frequency and (non-)exhaustivity in the linguistic cues we identified. In virtue of finite (non-modal) clauses being the most frequent clause type, and as a cue to exhaustivity, we might conclude that there is more evidence for exhaustivity. This is consistent with the prevalent sense amongst semanticists that most questions have a (weak or strong) exhaustive reading at least, but not necessarily a non-exhaustive one. At the same time, as how-questions are the most frequent question type, and as a cue to non-exhaustivity, we might conclude that there is more evidence for non-exhaustivity. This is inconsistent with the sense in the literature about the distribution of non-exhaustivity. Further, if the finiteness correlation holds true, we would expect that how-questions be interpreted exhaustively. This is counter to the intuition from Hintikka (1976) and Asher & Lascarides (1998).

The picture is further complicated when we take matrix verb frequency into account: while know is the most frequent matrix verb, it most frequently embeds (finite) how-questions. Further, recall that 25% of questions occur with other words that might cue (non-)exhaustivity in questions. In particular, we saw both exhaustive (e.g., all) and non-exhaustive (e.g., any, some) cue words predominantly occurring in finite clauses, suggesting that finiteness alone is not a homogenous cue to exhaustivity.

Recall the discussion of modality and non-finiteness. Already, we know that modal flavor and force is not only context sensitive generally, but can be influenced by the wh-word and the embedding predicate (in embedded questions). (Non-)exhaustivity resolution is a matter not only of presence/absence of linguistic structures and their combinations, but of the context as well. This is further confirmed by Experiments 3a and 3b, where participants were not overly sensitive to form factors to guide their judgements of answer acceptability.

There are two avenues of research that the current corpus analysis has not directly addressed, but which would be important to establishing a holistic picture of (non-)exhaustivity resolution. First, we did not present an in depth look at other aspects

of the question predicate besides the CLAUSETYPE. In fact, reliably the most frequent verbs in the question nucleus were light verbs like be, have, do, get etc., which carry little semantic content (see Appendix for graphs). Thus, the bulk of semantic content of the question is not provided by the verb itself, but the other content words in the question. Remember the question Who has a light?. In this case, the expectation that the speaker is a smoker with a non-exhaustive goal was triggered not by the verb has alone, but by the DP a light (or by the meaning of the whole predicate to have a light).

Ideally, we would want to look at how this information in the question predicate, beyond/in combination with the verb, is able to predict whether the question is interpreted (non-)exhaustively. In a trial machine learning simulation with the Reddit Corpus (Henderson et al., 2019), we found that this information can reliably predict the wh-word that will head the question, but the structural information we have identified as cues to (non-)exhaustivity alone could not. The question is whether the information in the predicate can predict (non-)exhaustivity in the interpretation. This requires in-depth coding of answers as being either exhaustive or non-exhaustive. This is not trivial.

Other recent advances in natural language processing have developed the tools for analyzing the predictive power of sentences, and collections of sentences (contexts). In particular, Google's Bidirectional Encoder Representations from Transformers (BERT) are particularly useful for sentence to sentence prediction, while Facebook's Robust Bidirectional Encoder Representations from Transformers (RoBERTa) has expanded on Google's AI to allow multi-sentence (context)-to-sentence predictions. If we think that context reliably provides additional information to precisify question interpretation, then we would expect that RoBERTa would be a useful tool for answering that question. The next phase of the corpus research would be to understand the contexts surrounding questions, and would use RoBERTa to predict questions from the preceding context.

Finally, another avenue of research would look in-depth at the relationship between question form, (non-)exhaustivity, and the verbs that embed them. Recent work

in computational linguistics by While & Rawlins (2018, 2019) has looked at lexicon-level selectional properties of complement-taking verbs, and linked them to inferential properties they impose on their sentential complements. Using similar machine learning techniques (topic modeling with latent Dirichlet allocation), may be fruitful in this case to discover the relationship between the structural cues to exhaustivity and the selectional properties of verbs on a much larger scale.

In the next task, we measure question production in a more constrained task that allows us to evaluate the influence of exhaustive and non-exhaustive contextual goals (using the concepts of 'high' and 'low' stakes from Experiment 2), and constrain the range of additional content in the question predicate.

## 5.3 Experiment 4: Production Study

In the corpus study, we got a sense of what questions speakers are producing, and the frequency and co-occurrence of cues. However, the corpus study did not look at the broader context in which questions occurred. Here, we explicitly manipulate this, to understand the relationship between discourse goals and question production. Presumably, speakers with exhaustive goals will utter questions that maximally convey those exhaustive goals, while speakers with non-exhaustive goals will utter questions that maximally convey their non-exhaustive goals. Since questions are underspecified for (non-)exhaustivity, this would manifest in terms of producing questions with the relevant set of linguistic cues, rather than producing questions without those cues. By looking at the question forms that are produced for a given goal, we can understand the extent to which speakers use their knowledge of language to signal their goals through question construction.

This experiment will recruit the notions of 'high' and 'low stakes' used in Experiment 2 and 3b, and the experimental stimuli from Experiment 3b in order to test the effect of context on question production. Following suit with the corpus study, we will code for the factors of interest discussed in Section 5.1.

In addition to manipulating contexts and goals directly, this study also aims to

understand the relevant space of possible questions, with an eye towards future construction of a computational model of question asking and answering.

### 5.3.1 Design and Materials

All materials and data are available at https://github.com/mcmoyer11/Question_production. The study was designed and administered online through Qualtrics survey software (Provo, UT). This study manipulated within-subjects STAKES (HIGH, LOW) between subjects. The stories for this study were similar to the 'high' and 'low' stakes stories from Experiment 2, except they were standardized in their structure in the following ways. Every story involved a search for something. The first paragraph included two sentences to set up the search, and introduced the key topic word that would be asked about. The second paragraph then introduced an explicit goal, and stated that a question had been asked. The sentence always had the form, "With this goal in mind, X asks Y something relevant to the locations of Z." Two examples are provided in (158) and (159). The participant then responded to the question, "What question do you think X asked Y?" by typing the question that they think was asked.

(158) **HIGH STAKES: OYSTERS**
Scientists have discovered a new strain of a dangerous virus that has contaminated oysters in the Mid-Atlantic. The Center for Disease Control is trying to prevent as much contamination as possible by tracking down the oysters which were sold to restaurants.

The CDC supervisor is tasked with tracking down the oysters. With this goal in mind, she asks her task force something relevant to the locations of the contaminated oysters.

**What question do you think the supervisor asked her task force?**

(159) **LOW STAKES: MUSEUMS**
Mark is visiting New York City for the first time. He has heard the city has great museums.

He wants to see museums on his trip. With this goal in mind, he asks the waiter at a restaurant something relevant to the locations of the museums.

**What question do you think Mark asked the waiter?**

In the Instructions and Training, participants were directed to answer in a relevant way, that addressed the goal and question given in the scenario.

### 5.3.2 Participants

56 Rutgers undergraduates enrolled in introductory-level courses were recruited from the Rutgers University Linguistics and Cognitive Science subject pool. 4 participants were removed from final analysis for reporting non-native English speaker status.

### 5.3.3 Predictions

If a questioner is trying to be as unambiguous as possible as to her underlying goal when she's asking a question, then we might reasonably expect her to produce a question that maximally conveys her goals. We will call this the Ambiguity-Averse Speaker Hypothesis, because it predicts that a speaker will avoid utterances that are ambiguous/underspecified in their meaning, by producing questions that maximize cues to contextually specified goals. We believe that this hypothesis is a reasonable reading of implementations of speakers in computational pragmatic models, following Franke & Bergen 2020. These speakers make each conversational move so as to reduce uncertainty in the true state of the world. Questions are uttered so as to elicit answers which will (so the speaker believes) also lead to a reduction in uncertainty. As such, questions which may correspond to multiple goals (and thus elicit too wide a range of answers) should be ranked lower in utility than questions which correspond to a single goal, or a narrower range of goals.

On assumption that our HIGH STAKES stories encode exhaustive goals, and LOW STAKES stories encode non-exhaustive goals, we predict that speakers who are maximizing the utility of their utterances in question production should produce questions with exhaustive cues in HIGH STAKES contexts, and produce questions non-exhaustive cues in LOW STAKES contexts.

On the other hand, recall that in Experiment 3b, participants did not significantly respond to question form manipulations (the presence/absence of a modal). While

this cue in previous experiments has been found to be a strong indicator of non-exhaustivity, the result was not replicated in Experiment 3b. We hypothesized that since the experimental scenarios make the questioner's goal explicit, it is possible that contexts provided sufficient information as to determine the acceptability of an answer. This would illustrate the give-and-take between context and linguistic form in conveying the speaker's goal: where one cue is sufficient to reduce uncertainty about the speaker's goal, the other becomes unnecessary. If this is the case, then we might not expect to see participants maximizing cues in the questions that they produce since the stories already make goals explicit.

### 5.3.4 Coding and Analysis

The raw production data file was cleaned up, it was converted to .json so the parsing script from the corpus study could be run on the dataset to facilitate coding. See Section 5.2.1 for information on the parsing script. As with the corpus study, the data were spot-checked for accuracy and additional coding, in particular cases where responses were coded as AMBIGUOUS or as FRAGMENT. Given that the task was narrowly defined, responses tagged with these were re-tagged as other appropriate QUESTTYPES.

Since this study targeted production of root questions, non-finite clauses and matrix verbs are not of particular interest. Indeed, since non-finite clauses are ungrammatical in root questions (without being embedded in a matrix verb), we should not see them at all.[4] There were some instances of EMBEDDED QUESTIONS and RELATIVE CLAUSES that occurred in yes-no questions, like Do any of you have an idea of where they are? (#123) or Do we know the location of where the hostages are being held? (#132).

We are interested in how the distribution of cues varies with STAKES–whether "non-exhaustive" cues are produced more in LOW STAKES, and "exhaustive cues" are produced more in HIGH STAKES. We take a non-exhaustive cue to be anything in the

[4]There were actually some instances of non-finite root questions found in the corpus study. These are constructions like, How about a quick kiss to seal the pact? (#9698) or "How then to escape the falsity?" he wrote (#123517), which seem more like rhetorical questions and not true information-seeking questions.

question's form that might suggest to the answerer that a mention-some(/mention-one) answer is preferred given the questioner's goals, and we take an exhaustive cue to be anything in the question's form that would do so for a mention-all answer.

Based on the theoretical literature, and the previous experiments in this work, we are interested particularly in the production of modal questions (and of can/could specifically). We are also interested in the production of plural and singular marked d-linked wh-questions. Recall the discussion in Chapter 2. While it has been suggested by some that such plural marked questions should block singleton mention-some (mention-one) answers (e.g., Comorovski 1996; Xiang (2016)), others have observed that this may be overridden by context (Dayal 2016). Further, Xiang & Cremers (2017) failed to find a difference between plural-marked d-linked WH-questions and monomorphemic WH-questions in the acceptability of a mention-some reading of embedded questions, despite predicting that this difference would be significant.

### 5.3.5 Results

Results were analyzed using $\chi^2$-tests of proportion. Figure 5.16 presents the overall distribution of CLAUSETYPEs and QUESTTYPEs in the data set. The majority of wh-questions produced are root questions, but some participants produced yes-no questions with embedded wh-questions. These latter are tagged as EMBEDDED and RELATIVE CLAUSE (15 total).

Figure 5.16 presents the distribution of matrix verbs across sentences labeled as EMBEDDED QUESTIONS and RELATIVE CLAUSES. These 15 sentences are below:

(160) a. Do any of you have an idea of where they are? (#123)

(161) a. Do we know how many tenants are home? (#94)
b. Do we know the location of where the hostages are being held? (#132)

(162) a. Do you know where to find coffee shops? (#168)
b. Do you know which restaurant serves the best italian food around here? (#79)
c. Do you know where the best museums are located around here? (#48)
d. Do you know where the best coffee shops that sell espresso are? (#164)
e. Do you know where the closest bike shop is? (#347)
f. Do you know where any woodworking stores are around here? (#283)

g. Do you know where any local woodworking shops are around here? (#286)
h. Do you know where i can find a bike shop? (#355)

(163) a. Can you find where they are being held? (#137)
b. Can you point me to where some woodworking shops are? (#297)
c. Can you tell me what yoga studios are in the area? (#181)
d. Could you tell me what museums are near this restaurant? (#35)

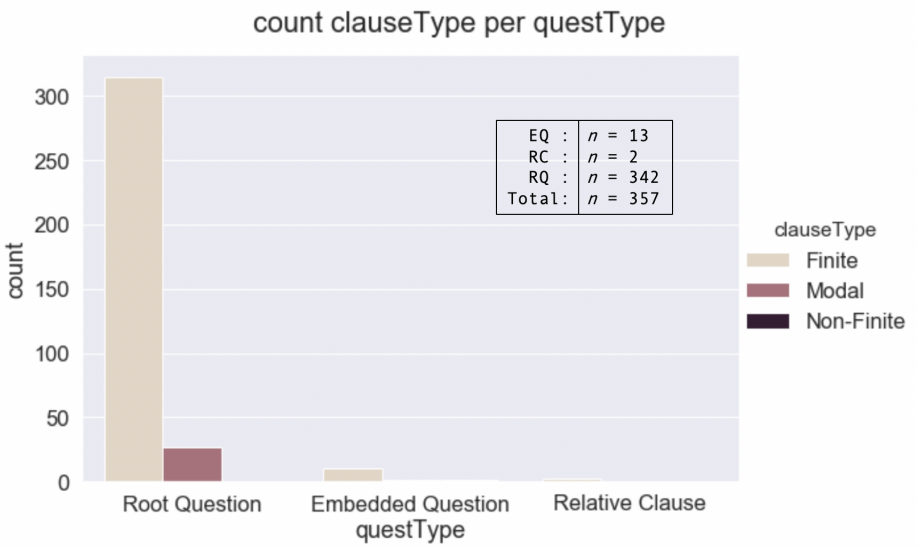


**Figure 5.16:** Distribution of CLAUSETYPE per QUESTTYPE.

As we can see, 10/15 matrix verbs are know. Surprisingly, these are produced more in LOW STAKES rather than HIGH STAKES, as shown in Figure 5.17. Unfortunately, there aren't enough observations to conduct any significance testing. There are only 4 observations in HIGH STAKES conditions, and the rest occur in LOW STAKES.

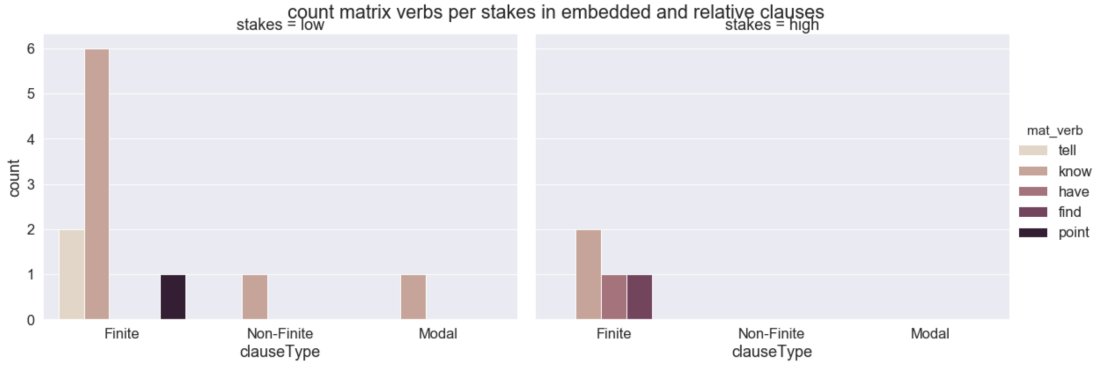


**Figure 5.17:** Distribution of CLAUSETYPE per QUESTTYPE.

Figure 5.18 presents the proportion of CLAUSETYPE levels per STAKES. While the

differences are not significant, we do see that FINITE CLAUSETYPEs are produced in a higher proportion in HIGH than in LOW STAKES contexts, while MODAL CLAUSETYPES are produced in a higher proportion in LOW than in HIGH STAKES contexts. The only significant result is the difference between FINITE and MODAL CLAUSETYPE across STAKES scenarios (HIGH: $\chi^2(1) = 283.29$, $p < 0.0001$; LOW: $\chi^2(1) = 222.09$, $p < 0.0001$).

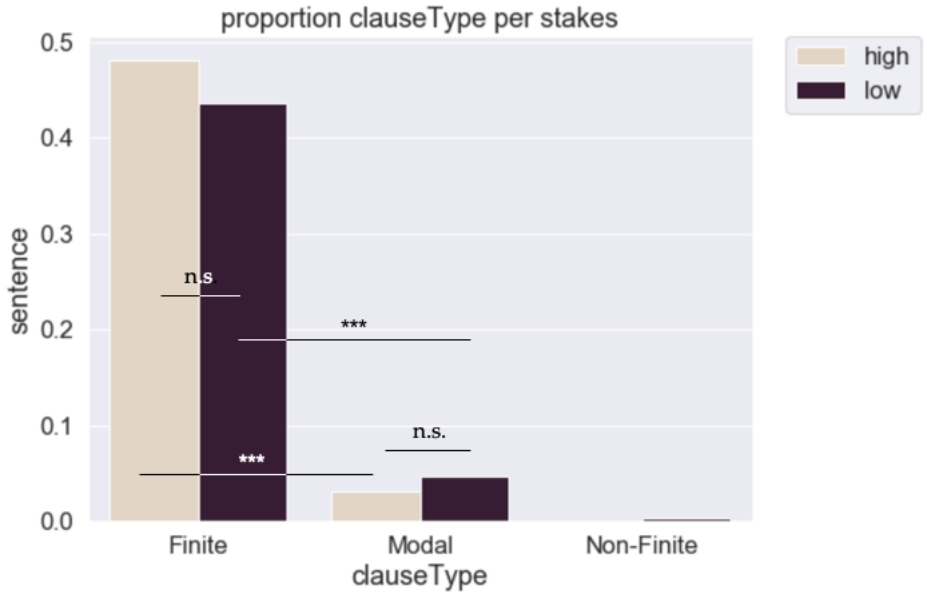


**Figure 5.18:** Distribution of CLAUSETYPE.

The scenarios were created to target where-questions, but it would also be felicitous to produce a question with a d-linked WH-word (Pesetsky 1987), such as which N. Figure 5.19 graphs the distribution of wh-words between HIGH and LOW STAKES. Focusing on what, where, and which, we see that what and where are produced significantly more in LOW STAKES, while which is produced significantly more in HIGH STAKES.

D-linked and plural-marked wh-phrases have been suggested to block mention-some/mention-one (Comorovski 1996; Xiang (2016)), while singular d-linked wh-phrases are said to collapse mention-all and mention-some (Srivastav 1991; Dayal 2017). At the same time, Dayal (2016) has shown that context can override this. If we reinterpret Comorvski's observation as another interpretive default, we might expect plural-marked d-linked wh-phrases to be produced in HIGH STAKE, while singular-marked ones to be produced more in LOW STAKES. Figure 5.20 presents the distribution of these d-linked

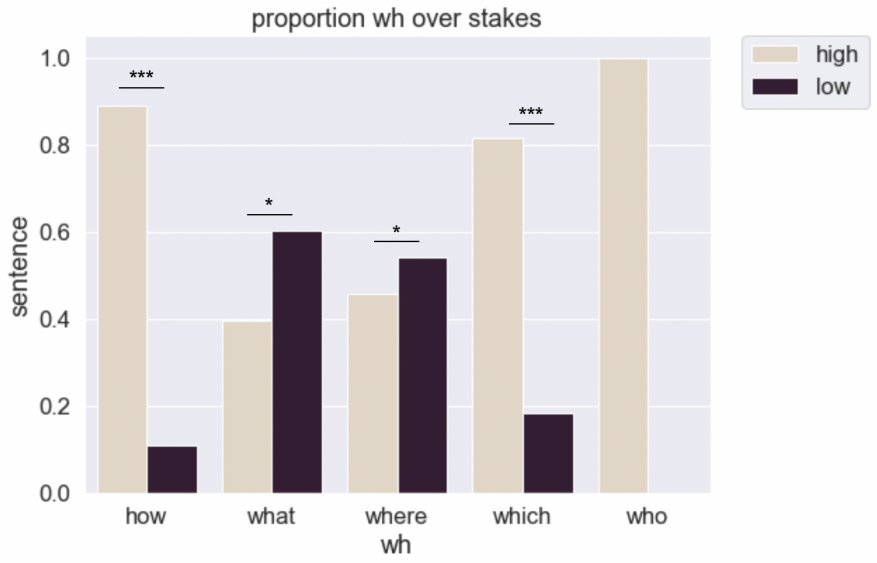


**Figure 5.19:** Distribution of WH.

phrases. Note two things: first, d-linking is overall produced significantly more in HIGH than LOW STAKES ($\chi^2(1) = 5.1579$, $p$=0.02); however, there are no productions of singular d-marking even in LOW STAKES. This suggests that number marking in the wh-phrase is orthogonal to (non-)exhaustivity (as suggested by Dayal 2017).

Finally, Figure 5.21 presents the distribution of other words that we might expect to be cues to (non-)exhaustive goals. These cue words are infrequent—the most frequent is best, occurring 49 times, only 13% of the total dataset. The cues identified as exhaustive (all, every–the latter which wasn't produced at all) do occur only in HIGH STAKES STORIES, while the cues identified as non-exhaustive occur a majority in LOW STAKES scenarios. There are some which appear in both: close, can, local, near, and around; however, these all occur in significantly higher proportion in LOW STAKES CONTEXTS.

### 5.3.6 Discussion

From this data, it would appear that participants were not overly concerned with producing questions that were maximally informative in the manner we predicted them to be. We found no significant differences due to STAKES in the CLAUSETYPES produced. Similarly to the corpus study, finite (non-modal) questions are the overwhelming majority of clauses. D-linked phrases were produced significantly more in HIGH STAKES

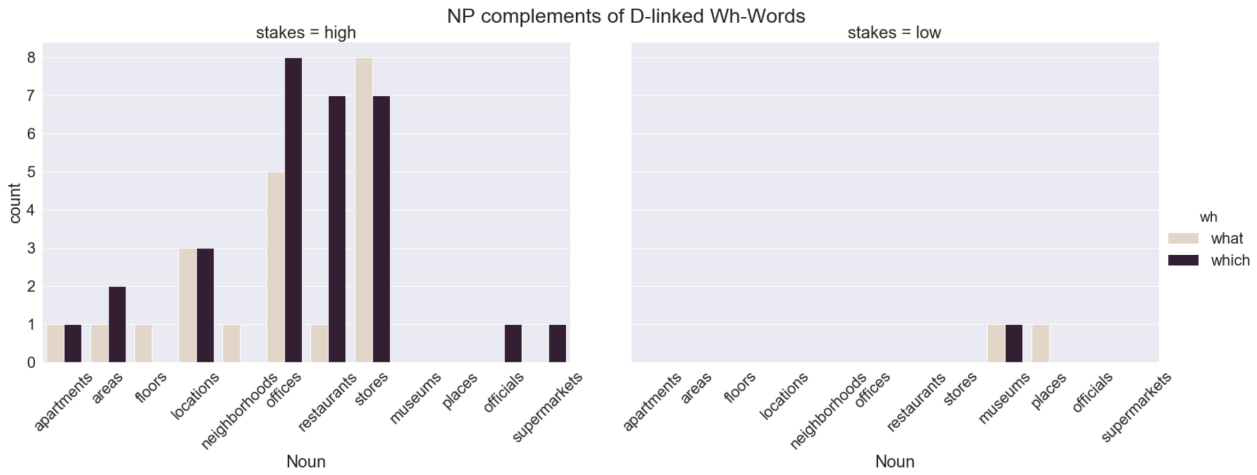


**Figure 5.20:** Distribution of number-marking in d-linked wh-phrases.

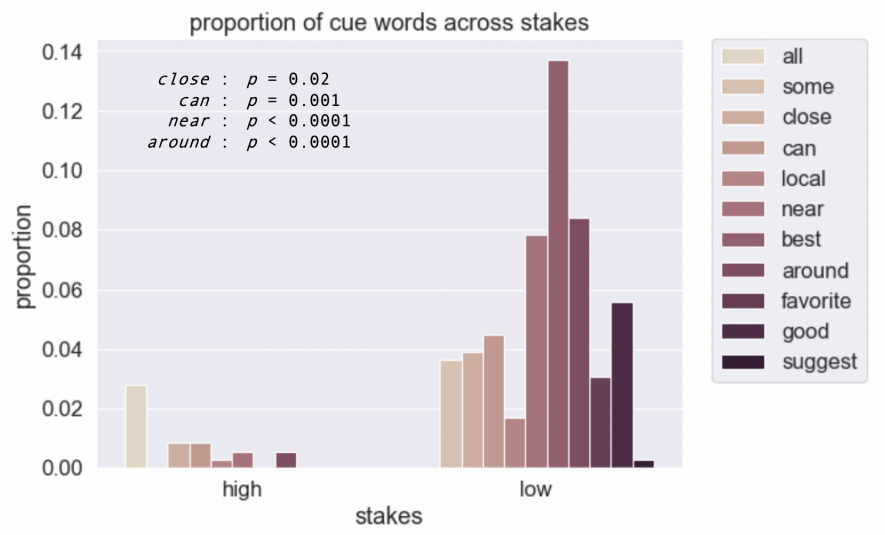


**Figure 5.21:** Distribution of other cue words.

than in low stakes, yet they were not absent from LOW STAKES contexts.

The stories in the Production task were more or less identical to those of Experiment 3b, except without an explicit question or answers. In the answer rating task of Experiment 3b, we found no effect of question form on the ratings. We hypothesized that, since the context provided an explicit goal, participants did not need to rely on cues in the question form to determine the appropriateness of an answer. In the Production task, we find a somewhat similar effect: speakers failed to produce modal or non-modal questions significantly differently between stakes, and only 13% of questions included additional words that might signal (non-)exhaustivity to the hearer.

The reason for this finding is consistent with the explanation we gave for the result in Experiment 3b: perhaps, given the overspecification in the context, participants did not need to produce questions that maximally convey (non-)exhaustive goals. This explanation could be consistent with the standard assumption that mindreading is essential to the pragmatic equation: the speaker might realize that the hearer shares the same contextual information that she does, and thus need not produce a question that makes up for underspecified contextual information. Further, if there is a cost associated with producing a question with more words, then we could easily explain why speakers in this task choose to produce ambiguous (but shorter) questions.

An alternative explanation could be that speakers really are not as concerned with their hearers' epistemic state as we originally thought. Their ambiguous question production here would not fall out from reasoning about shared contextual information, but merely from a *principle of least effort* drive to produce utterances that are easiest for them. This suggestion could be compatible with the above explanation if this drive falls out from some kind of cost to utterance production. We will discuss these possibilities more in the next section.

## 5.4 General Discussion

The corpus study aimed at quantifying speaker's naturalistic production of surface-level form cues, and found that frequency and co-occurrence of cues provides conflicting evidence for (non-) exhaustivity. While how-questions, a non-exhaustive cue, are the most frequent question, FINITE (non-modal) clauses are also the most frequent, and the most frequently co-occurring with how-questions.

The production study aimed at quantifying the extent to which form cues would be produced given the contextual STAKES manipulation. We hypothesized that, if a contextual goal was exhaustive, then speakers would produce questions with more exhaustive cues; and if a contextual goal was non-exhaustive, then speakers would produce questions with more non-exhaustive cues. Instead, we found that participants did not seem to produce a significant amount of cues in the questions they produced.

However, when they did produce cues, those cues aligned nicely with the STAKES manipulation: cues we deemed exhaustive were produced significantly more in the HIGH STAKES contexts, while cues we deemed non-exhaustive were produced more in LOW STAKES contexts.

The goals of this study were to understand the relationship between frequency and co-occurrence of these factors, and judgements of exhaustivity as described by a probabilistic theory of pragmatics (Degen 2015; Degen & Tanenhaus 2019), as well as to understand what kind of probabilistic information would be available to the language learner trying to learn how to resolve (non-)exhaustivity.

What do our results say about these points? Given the tenuous link between these cues and interpretation that we saw in Experiments 3a and 3b, we might not think that they (especially modality) are as important as assumed to (non-)exhaustivity resolution. Indeed, from both the corpus and the production study, we did not find that speakers were making much effort to produce questions that might maximally convey (non-)exhaustivity. While we did not analyze contexts in the corpus study, we manipulated contextual goals in the production study and found that these cues did not differ significantly between contexts, except for the production of complex wh-words.

We suggested in the discussion section of the corpus study that world knowledge associated with the question predicate beyond finiteness in the verb was important to resolving (non-)exhaustivity. The production task highly controlled for the possible predicates that would be felicitous in the contexts provided, by the very fact that we constructed contexts that would be exhaustive or non-exhaustive. Participants relied heavily on finite clause types across both HIGH and LOW STAKES.

In Experiment 3b, we did find a significant interaction between ANSWER and STAKES but not ANSWER and MODALITY. Together, this suggests again that (non-)exhaustivity is not categorically determined by the linguistic factors at hand—in particular, mention-some meanings do not require modality, nor do they seem best conveyed by only modal questions.

Since we did not analyze the contextual information surrounding the question

in the corpus study, nor directly collect (non-)exhaustivity ratings from those questions/contexts, we cannot identify the link between question predicate and context. However, if the production data is representative of a larger pattern, then it is likely that informative contexts in the corpus would explain why we do not see higher production of linguistic cues.

In either case, it would appear that the learner's task in acquiring question interpretation is no less obvious than it might be to the hearer. Rather, it is a complex matter of tracking cues whose interpretations may shift given both the context, and the other cues that they co-occur with.

Assuming as we are now that questions are not semantically specified for (non-)exhaustivity, there are two possible explanations for why speakers would not produce utterances that would make the hearer's task easier. One possibility is that the speaker, engaged in rational inference about whether the hearer's epistemic state would allow her to access the speaker's intended meaning, determines that the contextual information makes the appropriate affordances so that the speaker need not exert more effort than necessary in producing utterances with additional cue words.

It is assumed that central to communication and meaning are speaker intentions (Grice 1957, 1968, 1969, 1989; Lewis 1969; Grosz & Signer 1986; Sperber & Wilson 1986; Thomason 1990; Roberts 1996/2012), and discourse is a game in which interlocutors make moves with their utterances (Wittgenstein 1953; Lewis 1969; Hintikka 1972, 1981), with the ultimate goal of sharing and discovering truths about the world (Grice 1975; Stalnaker 1978). Communication is a cooperative rational act involving the recognition of interlocutor intentions, or theory of mind (Grice 1989; Sperber & Wilson 1986, 2002).

Are speakers really this concerned with their hearers as Gricean pragmatics suggests? Do speakers actually make their utterances maximally unambiguous for their hearers' benefit? Another possibility is that speakers are not concerned with their hearer's mental states, first and foremost, and thus not so concerned with avoiding ambiguous, vague, or semantically underspecified utterances.

One piece of evidence that might suggest that speakers aren't overly concerned with avoiding ambiguity is the case of pronoun production. Pronoun meaning is highly context-sensitive, and highly ambiguous even in a particular context (see Kehler & Rohde 2018; Kehler et al. 2008). A speaker who wished to avoid any ambiguity might avoid producing any pronouns at all. Cross-linguistically, we find pro-drop languages where the interpretation of argument structure is also highly ambiguous and contextually resolved. Yet, pronouns and pro-drop are not only incredibly common in everyday speech, but learned early (Moyer, Harrigan, Hacquard, & Lidz 2014; Oshima-Takane 1985, 1986; Strayer 1977; Shipley & Shipley 1969; Macnamara 1980; Chiat 1981; Clark 1978; Sharpless 1974; Loveland 1984, a.o.)

Another potentially informative case study comes from the psycholinguistic literature on garden path sentences. Garden path sentences are not only ambiguous but can cause comprehension and processing difficulties for hearers because they involve revision of an initial incorrect parse. Empirically, speakers do not avoid producing these difficult utterances (Arnold et al. 2004; Ferreira & Dell 2000; Ferreira & Hudson 2011; Ferreira & Schotter 2013; Jaeger 2010, 2011). Ferreira (2019) suggests that if speakers prioritized unambiguous, unequivocal communication for the sake of their hearers, they would not produce garden path sentences at all. The reason, he offers, is that it is more effortful to avoid producing disruptive utterances.

Clark & Wilkes-Gibbs (1986) found that in discourse, speakers might initially present provisional vague description which they subsequently refine in collaboration with their hearer. Similarly, a study on lexical choice by Brennan & Clark (1996) found that speakers craft the specificity of their utterances based on the common ground between different speakers. Rather than always providing a label with the most specific information, they may produce vague or ambiguous labels if those labels suffice in a given context. In both studies, speakers were concerned with minimizing their own effort.

If a speaker is guided by a principle of *least effort*, it is possible that they would not be guided by reasoning about hearers in cases where such reasoning would be effortful for them. There is some evidence to suggest that it is not always easy for

hearers to use their theory of mind to interpret speaker's intentions (Lin, Keysar, & Epley 2010), and that speakers can sometimes be less clear communicators when there is high information overlap between speaker and hearer (Wu & Keysar 2006).

In the next chapter, we return again to the hearer, with two goals. The first is to understand whether a hearer's predisposition for being more or less literal in the processing of other pragmatic phenomenon can predict their responses in resolving (non-)exhaustivity. The second goal is to understand more the give-and-take between the role of context and linguistic form in driving (non-)exhaustivity resolution. Here, without explicit context, we predict that participants would resort to using the linguistic form of the question to guide their resolutions. With explicit contexts, we predict that participants would not use the linguistic form of the question to guide (non-)exhaustivity resolution.

# Chapter 6

# Experiment 5: Diagnosing (non-)exhaustivity through independently hearer preferences

In Chapter 4, we tested the probability that a hearer would interpret a question on a particular resolution of (non-) exhaustivity given (a) the linguistic form factors in the question, and (b) the discourse goals made apparent in the context. We found that when discourse goals were not explicitly stated in a context, participants rated (non-) exhaustivity conditioned on linguistic form factors (although, these effects were on the scale of more-or-less acceptable, rather than acceptable or not). We also found that the rating was sensitive to whether the dependent variable measured the likelihood or the acceptability of (non-) exhaustivity. However, when discourse goals were explicitly stated in the context, the effect of linguistic form disappeared: participants rated (non-) exhaustivity on the basis of discourse goal, but not linguistic form.

Thus, the bulk of intuitions about (non-)exhaustivity in questions are really intuitions about the likely but defeasbile goals behind a speaker who utters a root question, or a sentence with an embedded question. As a result, we find the theoretical split between theoreticians arguing for weak/strong exhaustivity on the basis of who-questions, and those arguing for non-exhaustivity on the basis of how-/why-questions because different questions encode different expectations about goals. But this inherent context-dependence is not always explicitly acknowledged because those prior expectations can be very strong, and interact with more general principles concerning informative communication.

Speakers themselves do not reliably produce cues to indicate their goal. Thus the so-called cues are not totally indicators of goals. Indeed, research on speech production and audience design suggests that speakers are not first and foremost concerned

with facilitating understanding in their hearers or in avoiding ambiguous utterances, especially when this requires effort (Clark & Wilkes-Gibbs 1989; Clark & Brennan 1996; Arnold et al. 2004; Ferreira & Dell 2000; Wu & Keysar 2006; Lin, Keysar, & Epley 2010; Ferreira & Hudson 2011; Ferreira & Schotter 2013; Jaeger 2010, 2011; Ferreira 2019). In the question production study of Chapter 5, which used the same experimental scenarios as the second experiment from Chapter 4, we found that more often than not, speakers did not produce the questions that would be most informative with respect to discourse goals.

In this chapter, we seek to understand the hearer better. We have hypothesized that, when the context is sufficiently informative with respect to speaker goals, hearers need not rely on the linguistic form of the question to guide how they resolve (non-) exhaustivity. Effects of linguistic form should appear when the context is underinformative. In Section 6.1, we introduce the specific goals of the studies in this chapter, and the hypotheses that we will be testing. In brief, we will use two independent measures of hearer preferences to compare against hearer preferences in resolving (non-) exhaustivity in questions. To measure hearer preferences, we conduct two sentence verification tasks which replicate Bott & Noveck (2004), a study about scalar implicature (Section 6.2), and Chemla & Bott (2013), a study looking at presupposition processing (Section 6.3).

Bott & Noveck (2004) found that, in ambiguous sentences like Some cats are mammals, processing the logical meaning (consistent with a stonger some and possibly all interpretation) was faster than processing the pragmatic meaning (consistent with the some but not all interpretation). The stronger meaning, or *scalar implicature*, is an example of a Gricean inference typically assumed to be extralinguistic (Grice, 1975; Horn 1972; Gazdar 1979). Though modern theories provide ways of deriving the stronger meaning with a grammatical operator (cf. Chierchia, Fox, Spector 2012), the application of this operator is still governed by general principles.

Chemla & Bott (2013) used the methods of Bott & Noveck (2004) to study the processing of presupposition projection in ambiguous sentences like Zoologists don't know

that cats are insects. In this study, the true local reading (consistent with Zoologists don't know that cats are insects because cats are not insects) is processed slower than the false global reading (consistent with Cats are insects but Zoologists don't know this). These two measures were taken before participation in a third task looking at (non-) exhaustivity in questions (Section 6.4).

To preview our results, we find that participants who responded logically/literally on the Bott & Noveck task, were significantly more likely to accept mention-some readings of questions, while participants who accessed scalar implicature readings on that task were significantly more likely to reject mention-some readings of questions. At the same time, we found that both kinds of responders were significantly sensitive to our contextual manipulation. I argue that these results suggest that (1) non-exhaustivity should be represented in the semantics, and (2) that context-sensitivity must therefore be incorporated at the level of literal meaning. We found no correlation with local/global accommodation as measured in the Chemla & Bott task. Finally, we found no significant differences between modal and non-modal questions in any condition.

## 6.1 Goals and Hypotheses

The current study introduces two main manipulations. First, we introduce three different goal manipulations with minimal other changes to experimental factors. This manipulation tests the extent to which hearers are sensitive to the give-and-take between goal-relevant information contained in the linguistic form of the question, versus as contained in the context. This aspect of the motivation is discussed in Section 6.1.1.

Second, we take two independent measurements of hearer preferences in other semantic/pragmatic phenomena. These two measurements serve several purposes. Using a measure of participant literal-ness vs. pragmatic-ness as a reference point, we hope to uncover more support for/against different semantic theories of questions.

I discuss this in Section 6.1.2. In combination with the goal manipulation, we further hope to determine whether hearer sensitivity to context is a property of a literal hearer (and potentially a semantic matter), or of a pragmatic hearer (and potentially a matter of general rational pragmatic reasoning à la Grice). In particular, we discuss the predictions of a model of questions and answers within the Rational Speech Acts Framework (Goodman & Fran 2012; Hawkins et al. 2015; Hawkins & Goodman 2019), which posits hearers who reason differentially about the question asked, the space of possible questions (question alternatives), and the context. I disucss these last two points in Section (6.1.3).

### 6.1.1 The give-and-take between context and linguistic form

Based on the findings discussed in Chapter 4, we hypothesized that linguistic form cues will significantly affect evaluation of (non-) exhaustivity only when the context does not provide sufficient information for a hearer to infer a discourse goal, and thus resolve (non-) exhaustivity. In lieu of such contextual information, a hearer may use the linguistic form of the question as a cue to the discourse or the speaker's goal. We suggested that this fell out from the fact that speakers do not reliably produce questions with linguistic cues to their goals (the corpus and production studies of Chapter 5). Perhaps this is because the shared information in the context makes linguistic cues unnecessary. Indeed, work by Clark & Wilkes-Gibbs (1986) and Brennan & Clark (1996) found that initial expressions in discourse may be vague or ambiguous, but are often refined as discourse unrolls. At the same time, work by Keysar, Lin and colleagues seems to show that both speakers and hearers can be egocentric in communication, relying on information that might not be (perceptually) available to their interlocutor (Keyser et al 2000), or avoiding using theory-of-mind reasoning when performing cognitively demanding tasks (Keysar et al., 2003, Lin et al. 2010).

The Context-Over-Form Hypothesis states that when the context is informative with respect to a goal, linguistic form factors will not drive the acceptability of (non-)

exhaustivity.[1] We test this hypothesis in an experiment that pits contexts which differentially exposit (non-) exhaustivity in discourse goals against question form (presence/absence of the existential priority modal can in the question, and two matrix predicates know and say). We include three goal manipulations: the lack of an explicit linguistically provided goal (NO GOAL condition), an explicitly exhaustive goal (MA GOAL condition), and an explicitly non-exhaustive goal (MS GOAL condition). If our hypothesis is correct, then in the absence of explicit goals (the NO GOAL condition), the Context-over-Form Hypothesis predicts significant effects of question form (MODALITY and/or MATRIX VERB). The reason for this is that, when the context is underinformative about the speaker's goal, a hearer will use the speaker's utterance as a cue to that goal.

In this experiment, we use a slightly different kind of stimuli than in the previous experiments. Rather than introducing separate contexts using short paragraphs, we have one short introductory context in the beginning of each GOAL condition (between-subjects), that sets up the goal. This one aspect changes between conditions, but all other factors remain the same. The general paradigm is similar to the card-playing game paradigms used in Cremers & Chemla 2014, 2017 and Phillips & George 2016. In these experiments, there was a set of instructions that introduced the experiment: a group of friends getting together to play a game of cards. On each trial, participants were presented with a visual display that represented the set of answers to the question, and evaluated a target sentence with an embedded question.

Note that we have also argued throughout this dissertation that hearers impute goals in underinformative contexts. These goals derive from hearer expectations and world knowledge about the uttered question, statement, or situation. Let us consider the typical goals of card games, which is our current experimental context. Intuitively, it seems that the goals of card games are inherently mention-all, because the manner in which one wins a game of cards involves accumulating points. For example, gin rummy, cribbage, and war, the player who accumulates *the most* points/cards is the

[1] This will only hold true for those questions that are indeed ambiguous.

winner. This is not true for all card games. In Crazy Eights, the first to get rid of *all* their cards wins. Other games involve a more complex system of penalties. In Black Jack, the goal is to reach exactly 21 points. In Hearts, players are penalized for collecting points/cards, unless one "shoots the moon" and collects *all* the points cards (the complete hearts plus the Queen of Spades). When a player shoots the moon, she gets 0 points, and everyone else 26. In Spades and Bridge (similar to Black Jack), the goal is to win a targeted number of tricks (a number that players bet on in the beginning of each round). In Black Jack, Spades, and Bridge, missing the target amount can lead to the loss of points or money.

Despite these internal differences, most any game is won by the player who meets the criterion for winning internally, the greatest number of times. This is not an exhaustive discussion of possible card games, but illustrates that often these kinds of contexts typically have either immediate or ultimate goals of winning *as much as possible*. Thus given the general experimental task (playing a game of cards), we allow for the possibility that participants have a higher prior expectation that answers will be exhaustive given the typical goals associated with card games, and thus will answer more exhaustively than not in this task. This is further supported by results from previous experiments (Cremers & Chemla 2015, 2017; Phillips & George 2018), where participants did indeed access weak exhaustive readings, but these responses were rated both lower than ceiling, and lower than strong exhaustive readings. The amount of deviation from ceiling/strong exhaustive readings depended on several factors, including the matrix embedding verb, and whether a false belief was combined with the weak exhaustive reading.

### 6.1.2 Approaching a literal meaning for questions

Since the question of the literal meaning of questions is unresolved, we can use independent measures of hearer preferences to refine our understanding about the different readings of questions. The second goal is to understand how hearers' performance on other tasks tracks with (non-) exhaustivity resolution. Can we predict how a hearer

will resolve a question based on how they interpret other phenomena? Answering this question involves identifying interpretive preferences: is a hearer more "literal" or "pragmatic"? How do these different hearers choose to resolve (non-) exhaustivity?

By using an independent measure of hearer preferences, we can achieve two points: establish the relationship between different hearer types and inferences about goals in our goal manipulation, and get a sense of how different levels of (non-) exhaustivity track with different hearer types. We use two different measures of hearer preferences in this study. The first, which might more closely track a hearer's preference for more or less literal/pragmatic, is performance on a task testing existential statements that give rise to scalar implicatures (Bott & Noveck, 2004). In the case of scalar implicature, there is a clear logical/literal aspect of meaning (existential statements are logically consistent with stronger universal statements), but researchers dispute where the strengthened meaning derives from. The second measure is the extent to which a hearer interprets a presupposition on a local or global accommodation reading (Chemla & Bott, 2013). The relation between a literal/pragmatic meaning and the local and global accommodation readings is less clear, but presuppositions are generally assumed to introduce an at-issue/not at-issue divide.

By aligning a questions task with a task that independently measures a hearer's preference for more logical/literal interpretations as in the Bott & Noveck (2004) task, we can also test semantic theories which posit different literal meanings for questions. If literal hearers reliably interpret questions on a (weak or strong) exhaustive reading (Karttunen 1977; Groenendijk & Stokhof, 1982, 1984; Heim 1994), then we perhaps have evidence to support theories where the literal meaning of questions is (weakly or strongly) exhaustive. If we find that literal hearers reliably interpret questions ambiguously by rating mention-some conditions highly, then we have evidence in support semantic (or underspecified) non-exhaustivity. If we find that literal hearers are sensitive to the presence/absence of a modal in the question, rating mention-some conditions higher for modal questions (when contextually licensed by the MS GOAL condition),

then we have evidence supporting semantic ambiguity in modal questions. An aspect of this goal will be determining whether/the extent to which literal meaning is context-sensitive. I discuss that more in the next section.

### 6.1.3 Goal sensitivity: an aspect of literal meaning or rational pragmatic inference?

Theoretical discussions of (non-) exhaustivity in questions disagree about the division of labor between semantics and pragmatics: whether and which readings/answers are provided by the semantics (reflecting a literal meaning) and which are derived via some pragmatic mechanism (and thus not reflective of a literal meaning). No matter the underlying semantic representation, interpretation will vary with context and that fact requires an explanation that appeals to a context-sensitive mechanism of some kind. We should be careful to distinguish semantic ambiguity that requires disambiguation, from an inference derived via a fully-propositional Gricean inference. It is commonly thought that first kind of theory posits disambiguation (or precisification) that is necessary to establishing the literal meaning of a question. In contrast, the second kind of theory posits a mechanism outside of the literal meaning.

The Rational Speech Acts (RSA) framework of Frank & Goodman (2012), Goodman & Stuhlmüller (2012) is a cognitive hierarchy model (Camerer, Ho & Chung 2004) that has been successfully recruited to model cases of interpretational variability at the interface of semantics and pragmatics, such as scalar implicature (Goodman & Stuhlmüller 2012; Potts et al., 2016, Bergen et al., 2016), vagueness (Lassiter & Goodman 2013), metaphor (Kao et al. 2014), hyperbole (Kao et al. 2014), as well as certain aspects of the Question-Answer exchange (Hawkins et al. 2015). It treats interpretation as recursive Bayesian inference, where a probabilistic hearer reasons about the speaker's meaning in the face of uncertainty about it. The key achievement of RSA is to provide a probabilistic, quantitative computational formalization of the Gricean program that is flexible enough to allow for incorporating novel cues to meaning and

for testing whether these cues are useful for explaining interpretation. In RSA, a pragmatic listener reasons about the speaker's meaning, given the utility of the observed and unobserved utterances for achieving certain goals.

RSA has several attractive features that make it particularly amenable to application in wh-question interpretation. First, the notion of uncertainty is critical in the case of (non-) exhaustivity in questions because on the surface, questions under-represent intended (non-) exhaustivity. Second, RSA can be used to test the effect of context and alternative utterances on (non-) exhaustivity resolution, incorporating the linguistic and contextual cues discussed throughout this dissertation. Third, RSA can be used as a hypothesis-testing tool, to implement multiple theories and select the best one via model comparison. In the model, literal listeners represent the output of semantic computation. Therefore the literal listener can be modified to reflect a hypothesis about what the semantics outputs for a given question.

The question-answer model of Hawkins et al., (2015) and as expanded on in Hawkins & Goodman (2019) posits three different hearer models. While those authors were not interested primarily in (non-) exhaustivity, I will attempt to extrapolate relevant predictions as I introduce each of their hearer models. Note that to determine whether these predictions are exactly correct, we would need to construct the model and referential game in which to test the model behavior. Further, the model behavior and predictions will depend on several other factors, particularly the space of alternative questions (so the hearer can engage in counterfactual reasoning about the speaker's utterance in the typical Gricean manner), the parameter value for how optimally the speaker model chooses the utterances based on their informativity (often implemented as a soft-max function with a "rationality parameter" $\alpha$), and the implementation of a cost function (which might penalize certain kinds of utterances). These parameters must necessarily be constrained and idealized. They often assume a bounded hypothesis spaces, though in natural language it is unclear that the space is bounded in the way assumed by these models. Further, in natural language the space of possible utterances are not always obvious or given, again as assumed in these models. Thus,

while we might consider only some answer types to be semantic, and thus true answer alterantives, in reality a hearer might respond to a question in any number of ways.

Finally, the baseline hearer model, which conceptually represents the literal meaning of an utterance, would then affect how the more sophisticated hearer models work. While the space of utterance alternatives and the determination of literal meaning might be a domain of linguistics proper, the other model parameters constitute aspects of how RSA implements Gricean maxims. At the moment, we have not taken those steps, but seek to establish a better empirical understanding of what a future model would need to cover.

The $A_0$ hearer only gives answers that reduce the most uncertainty about the full state of the world, regardless of whether the answer is relevant to the question asked by the speaker. Thus, this hearer would answer in the same way for any question. If we take a question, Who has a heart to play?, this model would give whichever answer is most informative about the world. Perhaps that answer would name all the players' cards (even the non-hearts), or technically even answer a different question (depending on the domain of questions). The $A_1$ hearer (the literal answerer), responds directly based on the question that the speaker asked (essentially interpreting the speaker's goal to be identical to the question asked). Where $A_0$ perhaps failed to answer about the players with hearts, the $A_1$ would do so. The $A_1$ hearer would further answer exhaustively because in Hawkins' model, (1) question meanings are modeled essentially as partitions (cf. Groenendijk & Stokhof (1982), (1984)) and (2) the value of an answer depends on how much uncertainty it reduces (cf. van Rooij (2003), 2004, Shannon 1948's information entropy). This makes exhaustive answers generally preferable, both because they reflect the more literal meaning of a question, and because they reduce the most uncertainty in virtue of being exhaustive. Hawkins model would thus predict that literal answerers should answer exhaustively

The $A_2$ hearer (pragmatic answerer) reasons about the question asked, but also about the speaker's possible (private) goals. Unlike, the $A_1$ hearer, this hearer does not

assume that the speaker's goal is identical to the question asked. This hearer model is sensitive to both the context in which a question is asked, as well as possible alternative questions. Let us consider two examples used in the paper, modeling results from Clark (1979). First, Clark (1979) found that when the question (166) was preceded by context sentence like (164), liquor store merchants were more likely to specify the exact price rather than respond "literally" with yes/no, than when the question was preceded by (165).

(164) I want to buy some bourbon
(165) I've got $5 to spend
(166) Does a 5th of Jim Beam cost more than $5?

When the $A_2$ hearer interprets the context sentence, it potentially shifts their prior probability distribution over speaker goals, so when it interprets the question, the most informative answer may no longer be the literal answer. Hawkins et al. found that the first context sentence (164) does not change $A_2$'s priors on the speaker's goal (which might be uniform), so $A_2$ responds with the most informative answer (the exact price). In contrast, the second context sentence (165) does update $A_2$'s priors on the speaker's goal, and so biases the hearer to the literal yes/no answer.

Second, between the questions Do you accept Master card? and Do you accept credit cards?, Clark found that restaurant-owners were more likely to give a literal yes/no answer to the first question, but give an exhaustive list of credit cards to the second question. The literal answer is most informative with respect to the first question, but given the alternative questions the speaker could have asked, for the $A_2$ model, the exhaustive answer is more informative than the literal answer. As the space of questions is essentially determined by the specificity of noun phrases, the $A_2$ answerer essentially calculates a manner implicature from the question uttered (as in Bergen et al., 2016): a questioner with a specific goal about Master Cards would not likely ask about credit cards in general, on Hawkins' and colleague's understanding.

Lascarides (p.c.) points out that goals can be much more complex. She gives the example where a speaker might have ordered preference. If Mastercard is ranked

at the top, followed by Visa, and followed by American Express, the speaker might still ask Do you take Mastercard? Hearers can further anticipate this by responding in diverse ways, like No, but we take Visa.

Turning back to exhaustivity vs. non-exhaustivity, this is where the predictions of Hawkins' model becomes murky. In theory, this answerer should reason about possible goals given the possible questions that the speaker *could have asked*, and any information in the context that might shift the prior probability of goals. Thus, we would predict that $A_2$ would respond to questions differently based on a context manipulation, and based on the linguistic form of the question (the question alternatives).

While it might seem straightforward to implement the Hawkins model for (non-) exhaustivity resolution, there are several additional challenges to a straightforward extension. Most importantly, it is not immediately clear what the set of alternative questions and the set of goals should be. For one, in Hawkins' model, goals are specified as questions (QUDs) and the set of alternative question utterances are a subset of these QUDs. When the pragmatic speaker model chooses amongst alternative questions, given a goal, it will choose the utterance that addresses the intended goal as best possible, to maximize transparent communicative intent. This poses a difficulty for modeling questions as ambiguous or underspecified for (non-) exhaustivity at the level of the grammar (modeled by the literal listener), because the question alternatives are a subset of the goals. The model will always choose the goal that is closest in meaning to the utterance. Thus, if a question like Where can I find coffee? is ambiguous between a non-exhaustive QUD (modeled as Where are some places to find coffee? or an exhaustive QUD (modeled as Where are all the places to find coffee?, the pragmatic speaker model will never choose the ambiguous question, but rather the question that unambiguously signals one goal or the other. This is just one example of the type of technical and conceptual challenge involved in appropriately modeling the phenomenon. The production study of Chapter 5 gives insight into what the set of alternative questions might be for the set of goals used in that study, but we leave this for future research.

The Hawkins model assumes an exhaustive literal semantics for questions, and for that model only the $A_2$ pragmatic hearer reasons about contexts and goals. Thus, the Hawkins model in its current formulation would predict that only pragmatic hearers would be sensitive to the goal manipulation, and that literal hearers would only respond exhaustively. However, this and other model parameters could be manipulated in future research to test different semantic theories. The results will thus be informative for building a model of question interpretation. In lieu of generating the different models, we can use an independent measure of hearers in attempt to establish which kind of hearer responds on the basis of the literal meaning of the question (and is thus potentially an $A_1$ hearer), then we might be able to diagnose the literal meaning of the question, and determine if or the extent to which the literal meaning varies with contextual goal manipulation.

To do this, we use two different linguistic phenomena as proxies for establishing hearer preferences. The two phenomena, scalar implicature and presupposition, both involve ambiguity which cuts across different issues in semantics and pragmatics. The ambiguity in sentences with scalar terms, like some, are often thought to align with a semantics/pragmatics divide (although, *pace* grammatical accounts like Chierchia, Fox, Spector 2012). In contrast, the ambiguity that arises between global and local interpretations of embedded presuppositions is not necessarily aligned with a semantics/pragmatics divide, but rather at-issue/non-at-issue content.

## 6.2 Experiment 5a: Replication of Bott & Noveck (2003)

Statements made with the quantifier some are logically compatible with stronger statements with the universal quantifier all. This is demonstrated in (167): a speaker who says (167a) can felicitously follow it up with (167b).

(167) a. I ate some of the cookes.
b. In fact, I(/the speaker) ate all of the cookies.
c. In fact, I(/the speaker) did not eat all of the cookies.

Imagine the child, who, having eaten all the cookies their parent has just baked, utters

(167a) in fear of the repercussions of the stronger (true) statement. Likely, the child says (167a) knowing that their parent will draw an inference that since the child did not say the stronger statement, I ate all the cookies, but could have done so, it must be the case that (the child believes that) the stronger statement is false (Grice 1967, 1989; Horn 1972, 1989; Gazdar 1979). This subspecies of Quantity Implicature is called a *scalar implicature* (SI) because the quantifiers all and some form a scale based on logical strength.

Scalar implicature is equally the darling and *l'enfant terrible* of experimental pragmatics. By far, the most work in the field has centered around the study of this phenomenon, and the disputes over its correct treatment are unresolved. In an influential study, Bott & Noveck (2004) (henceforth B&N) took on the task of testing theories of scalar implicature by translating linguistic theories into theories of pragmatic *processing*.

B&N discuss two theories of SI. According to Relevance Theory (Sperber & Wilson 1986), hearers are guided by a trade-off between effort and pay-off—they compute as much as they can for as little effort possible. In the case of scalar implicature, hearers first compute the literal/logical meaning of the utterance with some (i.e., the meaning consistent with the stronger universal claim, as in (167b)). Only if the context necessitates it, does the hearer then compute the SI (consistent with (167c)) because this interpretation requires an additional processing step. On this view, SIs are not always computed. In contrast, some hold that SIs are always computed (cf. Levinson 2000; Chierchia 2004; Chierchia, Crain, Guasti, Gualmini & Meroni 2001). Levinson (2000) argued that pragmatic inferences are cognitive heuristics, computed fast and automatically. For Levinson, the SI is computed first, and if the context necessitates a literal meaning, the SI is cancelled. Thus, a hearer who hears a sentence like (167a) will automatically infer the stronger interpretation (167c). This strong interpretation may be subsequently cancelled if necessary.

Note that these theories were not theories of processing. However, B&N propose

the following translations into processing theories. The Relevance theoretic view suggests that the literal/logical meaning of a scalar is processed first, and the SI meaning processed second. Thus, this derivational ordering should bear out in terms of processing time: the first interpretation processed should be faster than the second. This processing view is often referred to as a *Logical-First* hypothesis. In contrast, the Default view suggests the opposite. Since the stronger SI meaning is computed automatically by default, it should be processed faster than the weaker logical meaning. The logical meaning is only computed after the SI meaning is cancelled (because the context does not support it), and thus should reflect a longer processing time. This view is often called the *Default-First* hypothesis.

Bott & Noveck's study was the first to propose a link between theories of scalar implicature and processing. They found that logical readings were processed faster *in this task* than SI readings. Research since has replicated their result in some experimental tasks (Breheny et al. 2006; Huang & Snedeker 2009). At the same time, other research has also shown that SI calculation is highly context-dependent, sensitive to dependent measure, and that with proper contextual support, these SIs can be processed fast (Degen 2015; Degen & Tanenhaus 2014; Degen & Goodman, 2014; Sperber & Wilson 1995; Breheny et al. 2006; Grodner et al. 2010; Breheny et al. 2013). Further, in other domains (such as indirect requests, and metaphors) the literal-first result is not replicated at all (e.g., Gibbs 1979, 1983; Ortony et al. 1978).

B&N's study, as well as the current replication, presents experimental stimuli alone without any additional linguistic context. Our purposes in using this study is to get a sense of hearer's predispositions for interpreting sentences, and in particular to understand how a logical/literal hearer would differ from a pragmatic hearer in resolving (non-) exhaustivity, and the extent to which those two hearers differ in sensitivity to features of context.

### 6.2.1 Methodology of Bott & Noveck (2004)

Consider (168), and its two possible interpretations in (169). Unlike the stronger alternative (167c) to (167a) which might be true in some contexts, (169b) is false.

(168) Some cats are mammals.

(169) a. Some *and possibly all* cats are mammals. LOGICAL READING (TRUE)
b. Some *but not all* cats are mammals. PRAGMATIC READING (FALSE)

The Logical-First Hypothesis predicts that the logical reading of (168) should be computed faster than the pragmatic reading. Thus, participants who access the logical reading should accept it, while participants who access the pragmatic reading should reject it. Further, the reaction times on the acceptances should be faster than the reaction times on the rejections, to reflect the fact that the logical reading is computed before the pragmatic reading. In contrast, the Default-First Hypothesis predicts that the pragmatic reading is computed faster than the logical reading. Thus, participants who reject (168) should have faster reaction times that participants who accept it. Thus, by measuring whether participants accept or reject sentences like (168), and how long they take to respond, B&N test hypotheses about the processing of scalar implicature.

Using a sentence-verification task, Bott & Noveck presented participants with different tokens of the six sentence types in Table 6.1. Participants saw 9 different tokens of these six sentence types, for a total of 54 trials. Sentence tokens were randomly drawn from 6 different zoological categories, each with 9 exemplars. Each word in the sentence appeared consecutively on the screen for a duration of 200ms. Across

| SENT TYPE | EXAMPLE | PREDICTED RESPONSE |
|---|---|---|
| T1 | Some cats are mammals | T/F |
| T2 | Some mammals are cats | T |
| T3 | Some cats are insects | F |
| T4 | All cats are mammals | T |
| T5 | All mammals are cats | F |
| T6 | All cats are insects | F |

**Table 6.1:** SENTENCE TYPES from Bott & Noveck 2004.

four experiments, B&N consistently found that participants who access the pragmatic

reading take significantly longer to respond than participants who access the logical reading, consistent with the Logical-First Hypothesis.

### 6.2.2 Design and Materials of B&N Replication

All participants saw three experimental blocks. At the beginning of each block, they were instructed and trained for the following study, and at the end of each block, they were instructed to take a break before beginning the next block. The first block presented was counterbalanced between Bott & Noveck 2004 stimuli (scalar implicature, henceforth BN TASK) or Chemla & Bott 2013 (presupposition, CB TASK). The third block was always (non-) exhaustivity in questions, QUESTIONS TASK.

For the BN TASK, we used a 6x3 factorial design, crossing BN SENTENCE TYPE (T1-T6 in Table 6.1 above) with VERB (NONE, know, say) to yield 18 unique sentence types. Participants saw two different tokens of each unique sentence type for 36 trials total.

Table 6.1 presents the six SENTENCE TYPES from B&N. We included the VERB manipulation to keep the stimuli consistent across the three experimental blocks given that both the CB and QUESTIONS stimuli introduced embeddings, as well as to maintain consistency with Experiment 2 from Bott & Noveck (which included a lead up to each sentence, "Mary says that the following sentence is true".

As with the original B&N, the critical test trials were SENTENCE TYPE T1 in both embedded and unembedded form, because these sentences are false on the strengthened scalar implicature reading (i.e., some, but not all, cats are mammals), but true on the weaker logical reading (i.e., some, and possibly all, cats are mammals). Thus, looking at whether participants 'Agree' or 'Disagree' with these trials, we can determine whether they access a logical or a pragmatic reading of the statement.

We followed the methodology of B&N's Experiment 3, which presented stimuli without any explicit instructions for participants to interpret T1 in one way or the other. We use their response of 'Agree' or 'Disagree' as a proxy for their interpretation. Thus on the basis of PROPORTION LOGICAL RESPONSE ('Agree' responses) to these trials, we will test for correlation with responses on the QUESTIONS TASK and the CB

TASK.

Sentences were generated from 6 different taxonomic categories, each of which contained six exemplars from each of these categories. Each participant saw all six exemplars from every category, but never in the same unique sentence frame: stimuli were pseudorandomized in a Latin-square design with 18 lists.

Sentences were presented to participants automatically one word at a time. Each word was displayed on the screen for a duration of 200ms, following the methodology described in Bott & Noveck (2004).

### 6.2.3 Participants

240 participants were recruited through two platforms: the Rutgers Linguistic and Cognitive Science undergraduate subject pool, and online through Mechanical Turk. 3 participants (undergraduate pool) participated twice, so the data from their second participation was removed. 11 participants were removed for reporting a native language other than English. 3 of those reported two native languages, one of which was English. We removed them from analysis to be conservative. Since this information is the same across all experiments reported in this chapter, we present this section once here.

### 6.2.4 Predictions

Results from the original study, Experiment 3 are presented in Table 6.2. For the critical SENTENCE TYPE T1, responses of 'True' reflect a logical interpretation, while responses of 'False' reflect an SI interpretation. Participants responded 'True' about 40% of the time, and 'False' about 60%, and participants who accepted did so significantly faster than participants who rejected the target (2700 vs. 3300 ms). Thus, this task supports the Logical-First processing hypothesis.

As we use the same kind of task, we expect to replicate the result finding that rejections of T1 take longer than acceptances.

| SENT TYPE | EXAMPLE | MEAN PROP TRUE | MEAN RT (MSEC) |
|---|---|---|---|
| T1 | Some cats are mammals | 0.407 (true) | 2700 |
| T1 | Some cats are mammals | 0.596 (false) | 3300 |
| T2 | Some mammals are cats | 0.887 | 2600 |
| T3 | Some cats are insects | 0.073 | 2700 |
| T4 | All cats are mammals | 0.871 | 2900 |
| T5 | All mammals are cats | 0.031 | 2600 |
| T6 | All cats are insects | 0.083 | 2400 |

**Table 6.2:** Results from Bott & Noveck 2004.

### 6.2.5 Data Analysis

Statistical models ranged from ANOVA where assumptions were met to be consistent with procedures from Bott & Noveck and Chemla & Bott, non-parametric Kruskal-Wallis tests where ANOVA assumptions were not met, as well as logistic regression models. Regression models were computed using RStudio (RStudio Team, 2019) and the `Ordinal` package (Christiansen, 2019), using cumulative link models specified with a logit link for binomial data with the function `clmm`(). Models were also fitted with random effects for Subject and Items. Model comparisons were conducted using Likelihood ratio tests with the R function `anova`().

Following the procedure outlined in Bott & Noveck (2004), responses longer than 6s and faster than 200ms after the final word was presented were removed. For them, this removed roughly 6% of the data from each experiment. Here, this removed roughly 8% of the data. Additionally, all "error trials" were removed for the reaction time analysis. "Error trials" included any trial where a participant responded incorrectly on a control trial (i.e., 'Agree' responses on the false control trials T3, T5, and T6; or 'Disagree' responses on true control trials T2 and T4). For B&N this removed roughly an additional 10-15% of the data depending on the Experiment. Here, removing error trials removed an additional 8% of the data.

Bott & Noveck also removed error responses for the critical test T1 trials in Experiments 1 and 2. We cannot follow this procedure because both responses are justifiable

in this experiment as we did not explicitly instruct the participants to interpret T1 sentences in one way or the other.

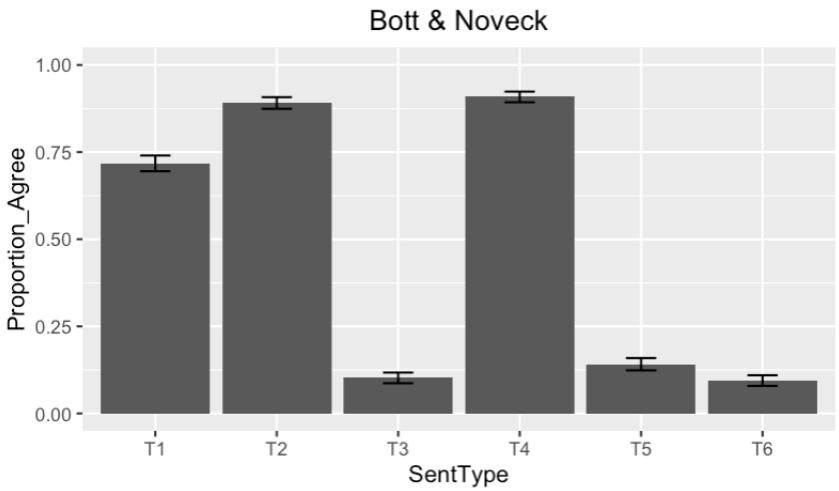


**Figure 6.1:** Proportion 'Agree' for Bott & Noveck Replication.

### 6.2.6 Results

Participants scored high in aggregate on control items, about 90% 'Accept' for True controls (T2 and T4), and 10% or under for false controls (T3, T5, and T6). Participants agreed to critical T1 trials a little under 75%.

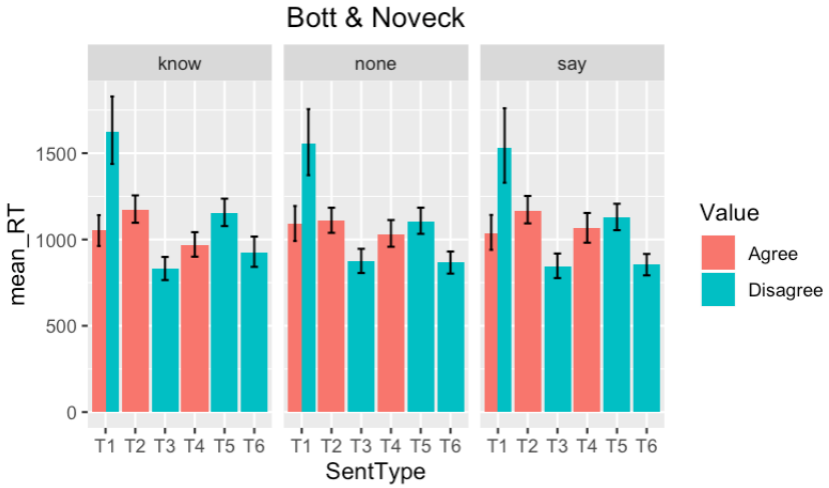


**Figure 6.2:** Reaction Time results for Bott & Noveck Replication.

Figure 6.2 presents reaction times for all SENTENCE TYPEs. Following the analyses reported in Bott & Noveck (2004), we conducted ANOVAs to analyze reaction time

data. Since we did not manipulate the instructions given to the participant in order to test whether they accessed a logical or pragmatic interpretation of some, we instead use their response of 'Agree' or 'Disagree' as a proxy for the interpretation that they accessed. Overall, there was a main effect of PROPORTION 'AGREE' on REACTION TIME (RT) $F(1,7467) = 13.68$, $p < 0.001$) and SENTENCE TYPE ($F(5,7463) = 40.44$, $p < 0.0001$).

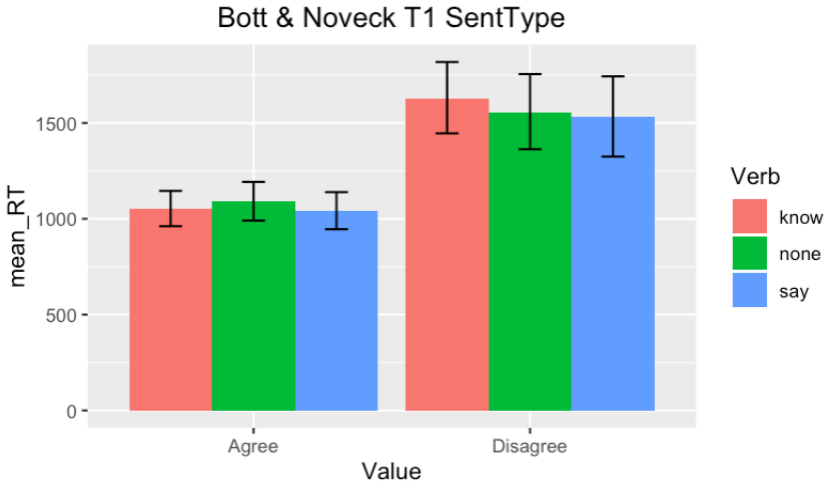


**Figure 6.3:** Reaction Time results for SENTENCE TYPE T1 for Bott & Noveck Replication.

Figure 6.3 plots the reaction time response for the critical trials (SENTENCE TYPE T1). Looking just at these, we find a Main Effect of PROPORTION 'AGREE' on REACTION TIME ($F(1,1316) = 74.52$, $p < 0.0001$) revealing that 'Agree' responses were significantly faster than 'Disagree' responses. Note, we found no main effect of VERB, nor any interaction with another factor.

### 6.2.7 Discussion

As predicted, we replicated the finding from B&N that acceptance of sentences like Some cats are mammals are faster than rejections, suggesting that in this task, and with these kinds of test items, logical interpretations are faster to compute than SI interpretations. Unlike the original study, where participants accepted critical trials only at about 40%, we found that participants overall accepted these sentences much more (just under 75%). It's possible that sample differences could account for this: the replication was conducted in English, and the majority of subjects were recruited through

Mechanical Turk; while the original study was conducted in French, and the subjects were all undergraduates recruited from the Université de Lyon 2. Another reason for the difference could be that B&N had participants perform a truth-value judgement in Experiment 3, while the replication had participants agree or disagree.

## 6.3 Experiment 5b: Replication of Chemla & Bott

The second independent measure of hearer interpretive preference involves presuppositions. Sometimes, the content of an utterance is taken for granted as true, or *presupposed*. Consider:

(170) The king of France is bald.
(171) Dana knows that the king of France is bald.

The definite description in (170) presupposes that there is a king of France (Frege 1892; Strawson 1950; Russell 1905's existence condition). Similarly, the factive verb know in (171) presupposes that its propositional complement is true (Karttunen 1971). The presupposed content is distinguished from the asserted content: (170) asserts that the king is bald, and (171) asserts that Dana has a certain belief about the king. Sometimes the asserted content is called *at-issue*, while the presupposed content is called *not at-issue*. At-issueness is the question of whether the information is backgrounded, or potentially up for discussion in the conversation. Whether this distinction aligns with a semantics/pragmatics divide is controversial.

Importantly, presupposed content behaves differently from asserted content in certain embedded environments. In particular, when embedded under negation, asserted content is negated while presupposed content is not. Presuppositions are said to *project* from out of these environments. Consider (172).

(172) a. The king of France is not bald.
b. Dana doesn't know that the king of France is bald.

(172a) denies the asserted content (that the king is bald), but not the presupposed existential claim (that there is a king); (172b) denies that Dana has a particular belief about the king, but not that the king is bald.

What happens when the complement of a factive verb is false? Consider (173).

(173) a. #Dana knows that cats are insects.
b. Dana doesn't know that cats are insects.

While (173a) is infelicitous, there is a reading on which (173b) is true (and felicitous):

(174) Dana doesn't know that cats are insects, because they're not.

On this reading, it appears that the presupposition does not project out of negation, but is targeted by it, just as with asserted content. Thus, the speaker who utters (173b) may not be presupposing that the complement is true, but rather asserting that it is false. When the presupposition is associated below the negation (in the scope of negation), it is called *local accommodation*. When the presupposition is allowed to scope over negation, it is called *global accommodation*.

Chemla & Bott (2013) (henceforth C&B), following the experimental paradigm and logic of B&N, discuss two natural processing hypotheses. A *Global-First* Processing Hypothesis, posits that the global reading is computed first (and thus faster), and the local reading is computed second (and thus slower), and a *Local-First* Processing Hypothesis posits that the local reading is computed first (and thus faster), while the global reading is computed second (and thus slower).

Further, C&B suggest that these two processing theories could be compatible with competing theories of how global and local accommodation readings arise. A semantic account, like Heim (1983), treats asserted and presupposed content as different components of meaning. The negation operator (as with other operators which allow presuppositions to project) only targets asserted content, allowing the presupposed content to project. When the presupposed content is false as in (173b), it is inconsistent with the common ground (or the speaker's beliefs). Thus, the more likely reading is the one where the presupposition is in the scope of negation. Note that, C&B call this a semantic theory because both readings are derived by semantics. This kind of theory, C&B suggest, is compatible with a Global-First Processing Hypothesis.

In contrast, on a pragmatic account, (e.g., Simons 2001; Abusch 2010; Schlenker

2008, 2009; Abrusán 2011), the negation operator targets both asserted and presupposed content equally. The local reading of (173b), where both asserted and presupposed content are negated, is considered to be the literal meaning of (173b). The global reading is derived via a pragmatic inference (a manner implicature on Schlenker's view). On Schlenker's theory, the local reading is provided by the semantics, while the global reading is provided by a Gricean pragmatic inference. C&B suggest that this kind of theory is compatible with a Local-First Processing Hypothesis.

C&B found that participants took longer to respond to local readings than global ones, providing support, they argue, for semantic theories. However, other research has confirmed that global readings can be fast, but also cognitively effortful (i.e., slow) if the context does not support the presupposition (Schwarz 2007; Tiemann et al. 2011; Schwarz & Tiemann 2017; Romoli et al. 2015; Schwarz 2015b, Schwarz 2014). Processing time is thus a function of the presupposition trigger, how embedded it is, and contextual support.

### 6.3.1 Methodology of Chemla & Bott (2013)

C&B used a sentence-verification task to present participants with different tokens of the five SENTENCE TYPES below in Table 6.3. Critical test sentences were SENTENCE TYPE T1, which involved a negated factive verb (realize in Experiment 1, know in Experiment 2) and false complement. The remaining SENTENCE TYPES were included as controls.

| SENT TYPE | EXAMPLE | PREDICTED RESPONSE |
|---|---|---|
| T1 | Zoologists do not realize/know that cats are insects | T/F |
| T2 | Zoologists do not realize/know that cats are mammals. | F |
| T3 | Geographers do not realize/know that cats are mammals. | T |
| T4 | Zoologists were told that cats are mammals | T |
| T5 | Zoologists were told that cats are reptiles | F |

**Table 6.3:** SENTENCE TYPES from Chemla & Bott (2013).

Similar to B&N's use of taxonomic relations, C&B generated their stimuli from a

list of 60 place exemplars across four geographical supercategory, and 60 animal exemplars across six zoological supercategory. Participants saw 120 items (not including training). Sentences were presented one word at a time, and each word flashed for 200ms.

The experiment was presented in a cover story about aliens. The aliens had invaded Earth, and different alien specialists were trying to learn about the Earth in only their specialty subject. So, alien geography specialists would have learned about Earth geography, but not zoology, while alien zoology specialists would lean about Earth zoology but not geography. This permitted unambiguous true/false control sentences.

### 6.3.2 Design and Materials of C&B Replication

Table 6.4 presents the nine SENTENCE TYPES used in the CB TASK replication. As with the original task, the critical test trials were of SENTENCE TYPE T1. These trials are true on a local reading (i.e., Cats are not insects and Zoologists don't know this), and false on a global reading (i.e., Cats are insects and Zoologists don't know this).

| SENT TYPE | EXAMPLE | PREDICTED RESPONSE |
|---|---|---|
| T1 | Zoologists don't know that cats are insects | T/F |
| T2 | Zoologists know that cats are insects | F |
| T3 | Zoologists don't know that cats are mammals | F |
| T4 | Zoologists know that cats are mammals | T |
| T5 | Babies don't know that cats are insects | T/F |
| T6 | Babies don't know that cats are mammals | T |
| T7 | Zoologists don't say that cats are insects | T |
| T8 | Babies say that cats are insects | F |
| T9 | Babies say that cats are mammals | F |

**Table 6.4:** Stimuli from Chemla & Bott replication.

Unlike the original study, we did not couch the experiment in a cover story about aliens visiting and learning about the Earth. We opted for no cover story to simplify the experiment, and used agents whose knowledge base would be clear: Zoologists should know the taxonomic facts about animals, while babies should not, given that

they lack sufficient world knowledge. This further allowed us to keep the experimental design as similar to the BN TASK as possible.

### 6.3.3 Predictions

Results from C&B are presented in Table 6.5. For the critical SENTENCE TYPE T1, responses of 'True' reflect a local interpretation, while responses of 'False' reflect a global interpretation. Participants responded 'True' a little under 40% of the time, and 'False' about 60% of the time, and 'True' responses were significantly longer than 'False' responses. The results from this task thus support the Global-First Processing Hypothesis.

| SENT TYPE | EXAMPLE | MEAN PROP TRUE | MEAN RT (SEC) |
|---|---|---|---|
| T1 | Zoolog. do not realize that cats are insects | 0.38 (true) | 3.5 |
| | Zoolog. do not know that cats are insects | 0.36 (true) | 2.75 |
| T1 | Zoolog. do not realize that cats are insects | 0.62 (false) | 2.5 |
| | Zoolog. do not know that cats are insects | 0.64 (false) | 2.25 |
| T2 | Zoolog. do not realize that cats are mammals | 0.12 | 2.0 |
| | Zoolog. do not know that cats are mammals | 0.14 | 1.9 |
| T3 | Geog. do not realize that cats are mammals | 0.85 | 1.85 |
| | Geog. do not know that cats are mammals | 0.83 | 1.75 |
| T4 | Zoolog. were told that cats are mammals | 0.93 | 1.5 |
| | | 0.91 | 1.4 |
| T5 | Zoolog. were told that cats are reptiles | 0.11 | 1.75 |
| | | 0.11 | 1.75 |

**Table 6.5:** Results from Chemla & Bott (2013).

While we use slightly different control sentences, and do not couch our replication in a cover story about aliens, our test stimuli are more or less the same as in the original. We thus predict to replicate the original result for test sentences.

### 6.3.4 Data Analysis

Statistical models ranged from ANOVA where assumptions were met to be consistent with procedures from Chemla & Bott, non-parametric Kruskal-Wallis tests where ANOVA assumptions were not met, as well as logistic regression models. Regression

models were computed using RStudio (RStudio Team, 2019) and the `Ordinal` package (Christiansen, 2019), using cumulative link models specified with a logit link for binomial data with the function `clmm()`. Models were also fitted with random effects for Subject and Items. Model comparisons were conducted using Likelihood ratio tests with the R function `anova()`.

In their original study, C&B removed responses longer than 10s, and did not remove error trials as B&N did. However, three participants were removed for scoring under 75% correct on control items. We followed this strategy and removed participants who scored under 75% correct on controls. This removed about 11% of the data. Removing then responses longer than 10s excluded an additional 1.8% of the data.

### 6.3.5 Results

We repeat the SENTENCE TYPES below in Table 6.6. Critical test sentences are SENTENCE TYPE T1. We will compare against SENTENCE TYPE T5 and SENTENCE TYPE T7, which minimally differ based on plausibility that the subject would have knowledge (T5) and factivity in the matrix verb (T7).

| SENT TYPE | EXAMPLE |
|---|---|
| T1 | Zoologists don't know that cats are insects |
| T2 | Zoologists know that cats are insects |
| T3 | Zoologists don't know that cats are mammals |
| T4 | Zoologists know that cats are mammals |
| T5 | Babies don't know that cats are insects |
| T6 | Babies don't know that cats are mammals |
| T7 | Zoologists don't say that cats are insects |
| T8 | Babies say that cats are insects |
| T9 | Babies say that cats are mammals |

**Table 6.6:** SENTENCE TYPES from Chemla & Bott replication.

Proportion 'Agree' responses in this replication are comparable to the original study. Participants rejected critical test items a large proportion of the time (over 75% 'Disagree', more than in the original study). Participants rejected false controls T2, T3, and T8 almost at floor. However, responses on true controls varied: at ceiling for T4 (Zoologists know that cats are mammals), but a little under 75% for T6, T7, and T9. One

possible reason for the variation in these latter responses could be that a case could be made for the appropriateness of a 'Disagree' response.

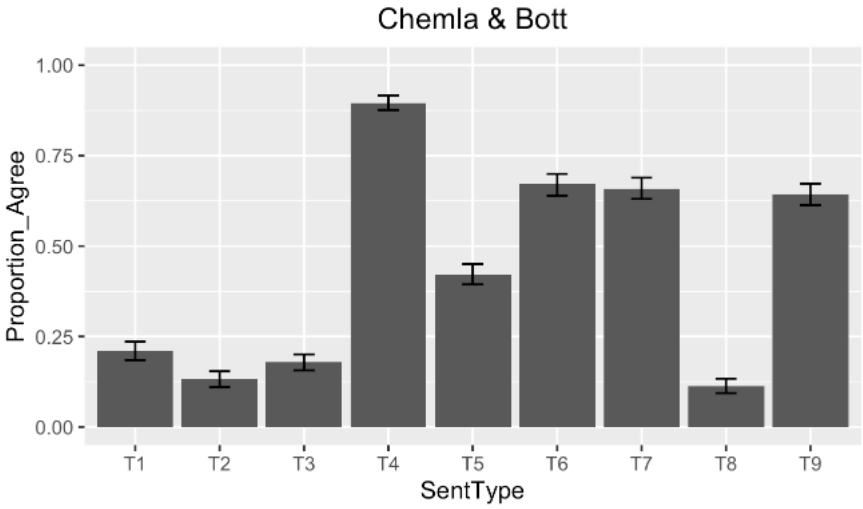


**Figure 6.4:** Proportion 'Agree' for Chemla & Bott (2013) Replication.

Sentences T5 (Babies don't know that cats are insects) and T7 (Zoologists don't say that cats are insects) are useful comparisons for T1 although they were not included in the original experiment. T5 includes the factive know but predicates knowledge of a subject who would not plausibly have the requisite world-knowledge of zoological and taxonomic categories. Participants actually accepted these significantly more than the critical T1 trials, accessing the local interpretation around 45% as compared to 20%. We expected a ceiling 'Agree' response to sentence T7, but participants only agreed to these 70% of the time. It is possible that some participants thus responded to these on the basis of the false complement.

Figure 6.5 plots reaction time as a function of SENTENCE TYPE. The data are not normally distributed, and violate homogeneity of variances (Levene's test $F(8,8351) = 36.546$, $p < 0.0001$) so we use non-parametric Kruskal-Wallis tests. We found an overall significant effect of SENTENCE TYPE on REACTION TIME ($\chi^2(8) = 445.92$, $p < 0.0001$), PROPORTION 'AGREE' ($\chi^2(1) = 6.3474$, $p = 0.01$), and an interaction between the two ($\chi^2(17) = 526.19$, $p < 0.0001$).

Figure 6.6 presents the critical trial T1, alongside foils T5 and T7. In these trials, we find a significant effect of PROPORTION 'AGREE' ($\chi^2(1) = 19.071$, $p < 0.0001$), but not

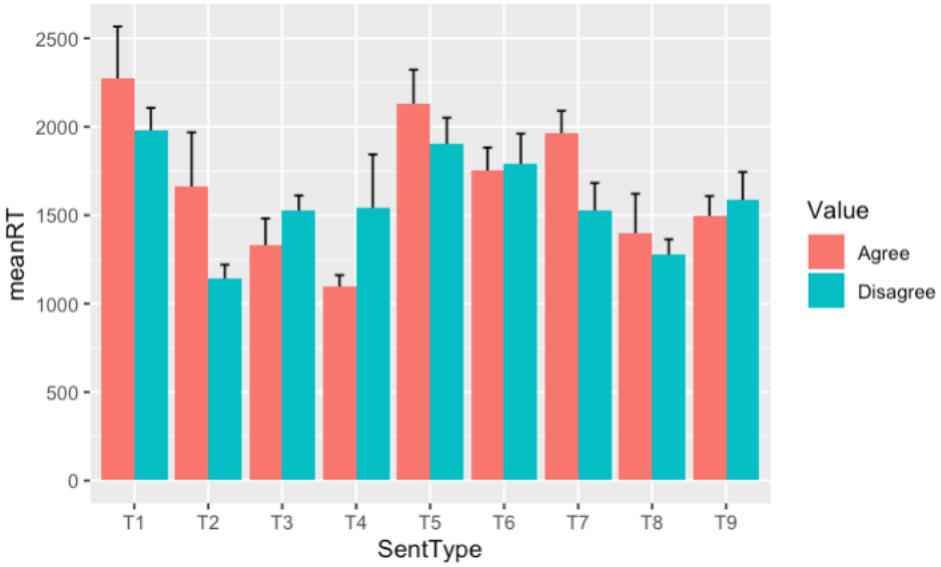


**Figure 6.5:** Reaction Time results for Chemla & Bott Replication.

SENTENCE TYPE. Overall, participants who agreed with these sentences responded significantly slower than participants who disagreed.

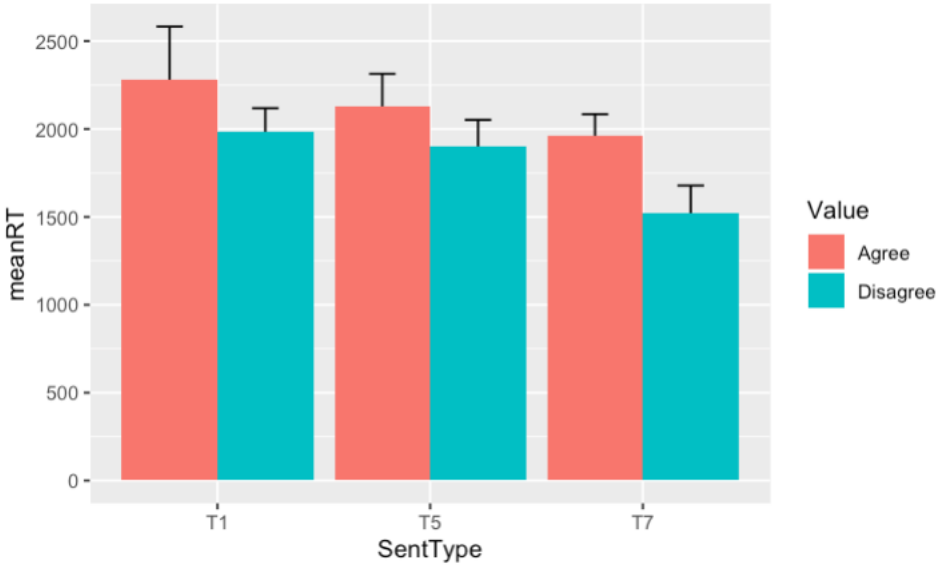


**Figure 6.6:** Reaction Time results for critical SENTENCE TYPES for Chemla & Bott Replication.

### 6.3.6 Discussion

We replicated the finding from C&B that local accommodation readings are processed slower than global accommodation readings. Note that this result held for not only

the critical T1 test targets, but also T5(Babies don't know that cats are insects) and T7 (Zoologists don't say that cats are insects). It is unsurprising that T5 should exhibit the effect, given that the sentence involves a factive verb that presupposes the truth of its complement. It is interesting that we find the same effect with T7, which does not have a presupposition trigger. Perhaps, given that zoologists are supposed to be experts about zoology, participants are treating say here as carrying a factive presupposition, and thus as ambiguous between a local and global reading. However, participants judge T7 true significantly more than T1 (70% vs. 20%), suggesting that they are not treating T7 exactly the same as T1.

## 6.4 Experiment 5c: Questions Task

### 6.4.1 Design and Materials

The third and final block of the experiment was the QUESTIONS TASK. This study manipulated a 2x4x2x3 factorial design, with MODAL (MODAL,NOMODAL), ANSWER (Mention-One (MO), Mention-Some (MS), Mention-All (MA), Mention-False (MF)), and VERB (know,say) as within-subjects factors, and GOAL (NONE,MA,MS) as between-subjects factor.

The premise of the experiment was that friends were getting together to play a game of cards, and two of those friends were relatively unfamiliar with cards and/or the game. Thus, the participant and the experiments were working together to help these two individuals learn. Figure 6.7 presents a sample trial. Each test trial displays a picture with six people labeled with letters of the alphabet (A-F). Under each person is a card, which represents the card that the character has/can play. Under the picture is a question-answer dialogue: a root who-question is asked, one character answers, and the other character gives a know-who or say-who report. Question predicates were always has a(n) ___ to play in the NOMODAL condition and can play a(n) ___ in the MODAL condition. The blanks were filled with 16 different nouns that tokened different properties of the cards (i.e., the suit, number, face, or color). Given the within-subjects factors, there were 16 unique sentence-types. Participants were shown


**Figure 6.7:** Experimental trial from questions section of Experiment 5.

each sentence type twice, with different token card properties. Presentation of sentence type and card property was pseudorandomized in a Latin-square design with 16 lists.

There were four phases of the experiment: familiarization with a deck of cards, experimental instructions and introduction to the game, practice/training, and then the test phase. During familiarization, we introduce the names of card properties, like faces, suits, etc. During the instructions, the study was introduced and the goal of the game identified (/manipulated). The practice phase included four trials where participants were introduced to the trial structure above, and trained to attend to Dana and Melissa's responses.

In the NO GOAL Condition, no explicit goal or game was specified. Rather instructions stated simply:

(175) **INSTRUCTIONS FOR NO GOAL CONDITION**
Some friends have gotten together to play a game of cards. Dana and Melissa have never played cards before. We will ask Dana a question, and she will give an answer. Then Melissa will say something about what Dana said. Your task to to say whether you agree or disagree with what Melissa says.

As is evident, no other information about the kind of game to be played was provided in this condition. The reason for this was two-fold: to test the assumptions about the card game that participants would impute into the experimental scenario, and to test the hypothesis that the lack of an explicit goal would lead participants to rely on semantic information encoded in the question form (i.e., the presence/absence of the existential priority modal can) to guide their evaluations of answers.

In contrast, the other two conditions included explicit information about the particular game that the friends would be playing. This information was included in two places: in between the very first sentence and the introduction of Dana and Melissa; and in a lead-up to the root question in the training/practice trials. The leading sentence was dropped in the test trials.

Let us first consider the MA GOAL condition below.

(176) **INSTRUCTIONS FOR MA GOAL CONDITION**
Some friends have gotten together to play a game of cards. The goal of this game is to collect points. If a player plays a certain card, they will receive a point. The players with the most points at the end can move on the the second round. Thus, it is important to keep an accurate account of the points.

(177) **MA GOAL QUESTION LEAD-UP**
Players with a __ get a point. Who has a __ to play?

These instructions encode a mention-all goal exactly for the reason discussed above: they explicitly say that the goal is to collect points, and that an accurate account of each players points will be crucial for determining whether that player passes to the next round. Thus, the mention-all answer is the most appropriate and salient answer for this goal because we must determine for each person whether they receive points on any given trial.

It was a challenge to find an acceptable mention-some goal for these stimuli without creating an overly complicated set of instructions that might make this condition more taxing on participants than the other two conditions. Another challenge was to keep the goal compatible with giving a mention-all answer as well—we did not want the goal to just restrict the domain of answers in some way, such that one could argue that the answer was exhaustive of that subset.

In the end, we modified trials in two ways:

(178) **INSTRUCTIONS FOR MS GOAL CONDITION**
Some friends have gotten together to play a game of cards. The goal of the game is to play a card that matches either the number, the color, or the suit of the current card. Players go one at a time.

(179) **MS GOAL QUESTION LEAD-UP**
Let's say you play a __. Who has a __ to play?

### 6.4.2 Participants

240 participants were recruited through two platforms: the Rutgers Linguistic and Cognitive Science undergraduate subject pool, and online through Mechanical Turk. 3 participants (undergraduate pool) participated twice, so the data from their second participation was removed. 11 participants were removed for reporting a native language other than English. 3 of those reported two native languages, one of which was English. We removed them from analysis to be conservative.

### 6.4.3 Predictions

#### Goal Manipulation

In this study, we predict a significant effect of MODAL in the NO GOAL conditions, but not in either the MA or MS GOAL conditions. While the NO GOAL condition is meant to be neutral with respect to mention-some and mention-all goals, we suspect that participants in this condition will pattern with participants in the MA GOAL condition, given that they may expect the card game to be inherently mention-all. We pursued this card game context despite this, rather than choosing a more neutral context, for several reasons. Practically, it provided a fairly straightforward way to depict the possible set of answers, and permitted a range of possible question predicates based on card properties while maintaining a fairly consistent visual display. Additionally, the possibility that hearers may recruit these strong prior expectations about the context into the evaluation of (non-) exhaustivity is theoretically interesting because it can reveal the influence of prior expectations on (non-) exhaustivity resolution. Future research

will be able to look at a wider range of contexts that may track different prior expectations about (non-) exhaustivity.

#### Semantic Theories

Let us turn to the predictions that our semantic theories make. Theories on which questions are semantically weak/strong exhaustive (Karttunen 1977; Groenenijk & Stokhof 1982, 1984; Heim 1994) predict that MA ANSWERS will receive a higher rate of agreement than MS/MO ANSWERS. We might also reasonable predict on these kind of theories that MS/MO ANSWERS may be agreed to higher in the MS GOAL condition than in the other two conditions.

It might also be reasonable to assume that modal theories, which hold that modal questions are ambiguous between a (strong) exhaustive and a non-exhaustive meaning, predict significant differences between MODAL and NO MODAL questions: MODAL questions should give rise to significant higher rate of agreement than NO MODAL questions, in the MS GOAL condition.

Along with modal theories, other ambiguity/underspecification theories predict that (non-) exhaustivity is resolved relative to the context. Reasonably, then we may expect on these theories that MS/MO answers will be accepted higher in the MS GOAL condition, MA ANSWERS in the MA GOAL CONDITION and perhaps equal acceptance in the NO GOAL condition—or, taking into account the caveat about card games, higher agreement with MA ANSWERS in the NO GOAL condition.

### 6.4.4 Data Analysis

Statistical models ranged from non-parametric Kruskal-Wallis tests where ANOVA assumptions were not met, as well as logistic regression models. Regression models were computed using RStudio (RStudio Team, 2019) and the `Ordinal` package (Christiansen, 2019), using cumulative link models specified with a logit link for binomial data with the function `clmm`(). Models were also fitted with random effects for Subject and Items. Model comparisons were conducted using Likelihood ratio tests

with the R function `anova()`.

6 participants were removed for incorrectly answering false controls over 50%. These were Mention-False reports for know-wh targets.

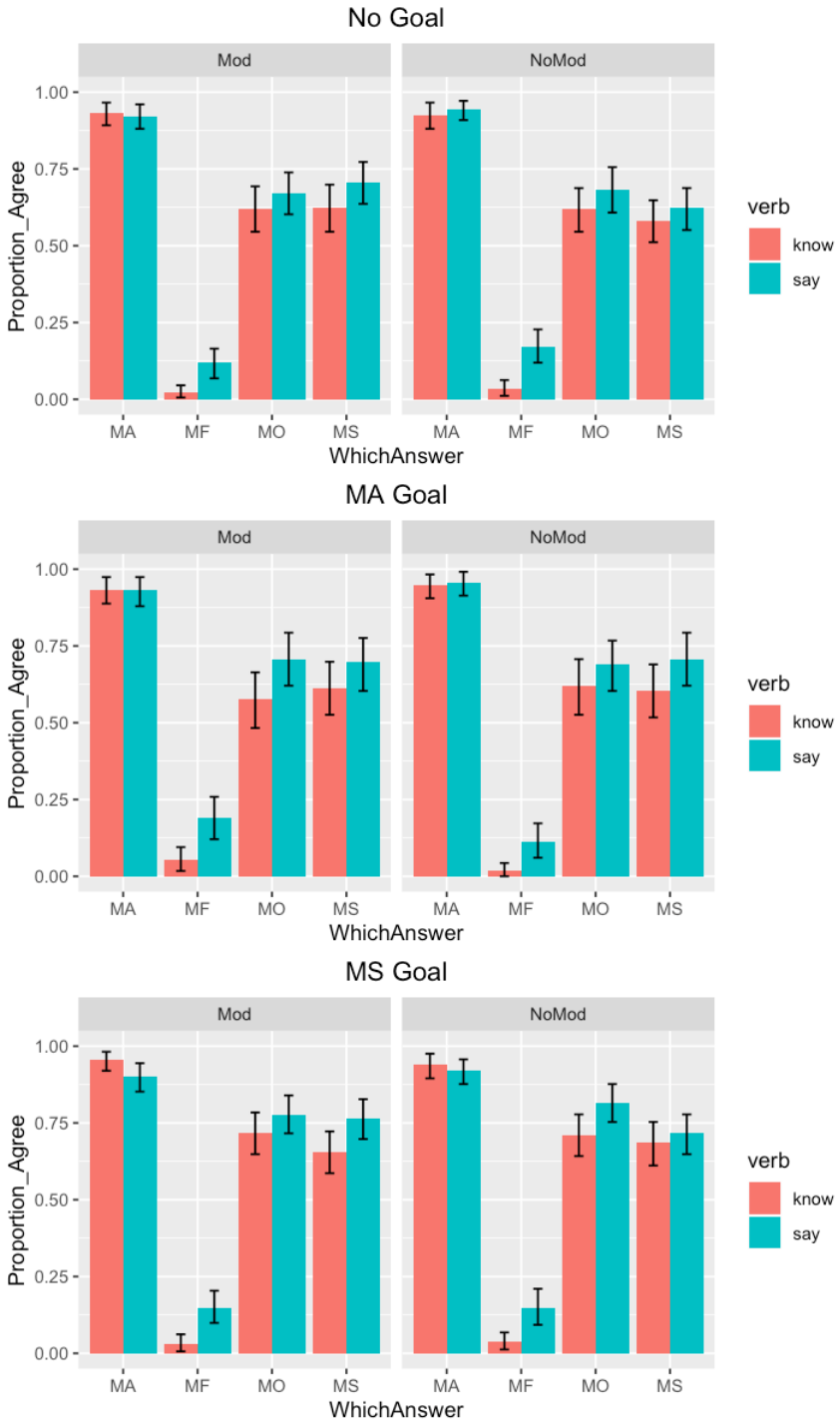


**Figure 6.8:** Results from Experiment 6, Questions Task.

### 6.4.5 Results

Results are presented in Figure 6.8. Overall, participants consistently rated MA ANSWERS near ceiling, and MF answers near bottom. There was no significant difference between MS/MO ANSWERS, but a main effect of ANSWER ($\chi^2(3) = 2878.7$, $p < 0.0001$) revealing significant differences between MA/MF and MS/MO ANSWERS.

This effect was modulated by the GOAL manipulation, which significantly affected participant responses alone ($\chi^2(2) = 12.324$, $p < 0.005$) and interaction with ANSWER ($\chi^2(11) = 2905.9$, $p < 0.0001$) as well as MODAL ($\chi^2(5) = 12.447$, $p = 0.03$). The effect of the GOAL manipulation was driven by the MS GOAL condition, which differed significantly from the other two conditions (MA $p = 0.03$, MS $p = 0.002$). Further, the two interactions reveal that participants agreed to targets in MS/MO ANSWER conditions significantly higher in the MS GOAL condition than in the other two (MS vs. MA: $p = 0.0003$; MS vs. NONE: $p < 0.0001$). The interaction with MODAL was significant only in aggregation.

Additionally, there was a significant effect VERB ($\chi^2(1) = 30.705$, $p < 0.0001$) that yielded 2-way interactions with ANSWER ($\chi^2(7) = 2925.5$, $p < 0.0001$) and GOAL ($\chi^2(5) = 43.597$, $p < 0.0001$), as well as an interaction between the three ($\chi^2(23) = 2954.3$, $p < 0.0001$). These effects and interactions reveal that while participants on average responded to know-who statements with a higher proportion of 'Disagree' responses, the magnitude of the difference varied based on what kind of answers the character in the story gave, and whether the goal of the game was MS or not: they were indeed rated highest in the MS GOAL condition.

### 6.4.6 Discussion

We predicted that the effect of MODAL would come out in the NO GOAL condition, because we hypothesized that participants would use the question form here (if anywhere) when no goal was explicitly presented. This prediction was not borne out. We found no significant Main effect of MODAL for any subset of factors, however there

was an overall interaction with GOAL, but no pairwise comparisons between conditions were significant.

At the same time, we acknowledged the possibility that participants' prior experience with card games could potentially lead them to impute an exhaustive-compatible goal in the NO GOAL condition. Thus, their responses would pattern with the MA goal condition. Indeed, there were no significant differences between those two conditions.

Note that the MS GOAL condition significantly raised the acceptance of MS/MO answers. These responses, while degraded from ceiling/MA, are closer to ceiling than they are to bottom. Thus, if participants are merely tolerant of partial answers, it is a different kind of tolerance than, for example, the tolerance of false answers we saw in Experiment 1 or in Philips & George (2016). There, when a character gives a false answer in addition to a weak exhaustive answer, participants were more likely to *reject* know-wh reports than to accept them. In contrast, here we see the reverse: participants are more likely to *accept* than to reject MS/MO ANSWERS.

Further, that this high acceptability holds across both know-wh and say-wh (despite statistically significant differences) speaks to the general availability of mention-some across different embeddings, and the importance of context in licensing interpretation.

## 6.5 Correlation Analysis

In this section, we present a correlation analysis between the three studies discussed above. Before discussing those results, we introduce potential connections between the three studies.

### 6.5.1 Links and Predictions

#### Scalar Implicature and Presupposition

The link between scalar implicature and presupposition is debated. Some have suggested that (some) presuppositions are generated by the same mechanism as scalar implicatures (Simons 2001; Abusch 2002, 2010; Chemla 2009; Romoli 2009, 2014). However, experimental results like Chemla & Bott (2013) seem to challenge this view (see also Romoli & Schwarz 2015; Bill, et al. 2016). These studies have found that the two phenomena behave differently, suggesting different processing mechanisms. While sentences with SI-triggering words or phrases are slower to process than their literal meanings, globally accommodating a presupposition (what we might think of as a pragmatic mechanism) can be faster than not accommodating it (what we might think of as a semantic mechanism).

#### Questions and Presupposition

To date, none have explicitly suggested a link between (non-) exhaustivity in questions and presupposition. Researchers have discussed questions as having a presupposition that one member of the Hamblin set is true (cf. Karttunen & Peters 1976; Karttunen 1977; Comorovski 1989, 1996; Dayal 1991a, 1991b; Cross 1991; Krifka 2011). Further, a speaker who asks a question might presuppose that the hearer can answer the question. But (non-) exhaustivity has not been suggested to be a presupposition of any kind.

#### Exhaustivity in Questions *as* Scalar Implicature

Theories of grammatical scalar implicature (cf. Chierchia, Fox & Spector 2012) employ exhaustivity operators with similar meanings as the ones used in many semantic theories of questions (cf. Fox 2014, 2018; George 2011; Nicolae 2015; Xiang (2016)). Some of these theories treat the exhaustivity operator as mandatory in all questions.

Other semantic theories of questions treat exhaustive readings as scalar implicatures, where questions are semantically weak (or non-)exhaustive and pragmatically strengthened to strong exhaustivity (cf. Spector 2007; Spector 2006; Schulz & van Rooij 2006; Zimmermann 2010). The advantage to this latter kind of approach is that, as Asher & Lascarides (1998) point out, this view of (non-) exhaustivity in questions is consistent with standard assumptions about the relationship between semantics and pragmatics: Pragmatic inferences are strong, but defeasible; semantic inferences are weak but non-defeasible.

**Using Logical Responders to Diagnose Literal Meaning**

It might be reasonable to assume that to be a logical responder in the B&N task means that one's response more directly reflects the output of a truth-conditional semantic mechanism. If that's right, then it seems reasonable that all theories would predict that logical responders will pattern with whichever truth-conditional representation that theory takes to reflect the literal meaning.

Theories where questions are semantically weak or strong exhaustive (e.g., Groenendijk & Stokhof 1982, 1984; Karttunen 1977; heim 1994) would then predict that logical responders would have a high rate of agreement in mention-all answer conditions, and low rate of agreement in mention-some conditions. We might further expect that this result would hold regardless of GOAL condition for logical responders. This kind of a theory would predict that only pragmatic responders would rate mention-some condition high (in the MS GOAL condition), because mention-some is pragmatically derived.

Modal theories would predict that logical responders will be sensitive to the MODAL form of the question, in the MS GOAL condition: participants should rate mention-some conditions higher in modal question conditions than in non-modal question conditions.

For ambiguity/underspecification theories, we might expect that both logical and

pragmatic responders would be sensitive to the GOAL manipulation, rating mention-some conditions high in the MS GOAL condition, and mention-all high in the MA GOAL condition.

Finally, theories which hold that strong exhaustivity is a pragmatic inference from a non-exhaustive baseline (cf. Asher & Lascarides (1998); Schulz & van Rooij 2006; Spector 2006, 2007; Zimmermann 2010) would predict that pragmatic responders would calculate exhaustivity inferences in mention-some conditions.

**Locating context-sensitivity**

Another motivation behind testing different kinds of speakers, is to draw parallels between computational pragmatic models, as discussed in Section 6.1.3, which posit literal and pragmatic hearers that differentially reason about contextual or speaker goal information. In the RSA framework, literal hearers respond strictly on the basis of the output of truth-conditional semantic mechanisms, while pragmatic hearers respond additionally on the basis of contextual information, and/or inferences about the speaker's goals/intentions.

By using the B&N task as an independent measure of these two kinds of speakers, we hope to pinpoint the location of the context-sensitivity of (non-) exhaustivity in questions. If logical responders (as proxy for literal hearers of the RSA model) are sensitive to the goal manipulation, then it would provide support for the idea that context is necessary to establish a literal meaning of questions.

If only pragmatic responders (as proxy for pragmatic hearers of the RSA model) are sensitive to the goal manipulation (rating answers differentially based on GOAL), then we might have support for the idea that the context-sensitivity is not necessary to establish the literal meaning of the question.

**Caveat**

Often goals of communicative exchange involve the maximization of information conveyance. Thus we might expect overall preferences for MA ANSWERS because these

independently convey the most information in the context/reduce uncertainty abou the true state of the world (cf. van Rooij 2003, 2004; Schulz & van Rooij 2006; Zimmermann 2010)). This result would not necessarily indicate an underlying exhaustive semantics, but a general pragmatic principle. Thus, it is possible that pragmatic responders would embody this by responding more exhaustively than logical responders.

The predictions for these theories may be further complicated when we remember the discussion about the typical goals associated with card games. The pragmatics of ambiguity resolution, the specification of context-sensitive variables, and most likely the fixing of the referential domain of a wh-expressions, will likely be sensitive to a variety of factors (linguistic, situational, psychological). These and many other factors, like frequency and co-occurrence statistics, subjective experience and world knowledge, all these factor together in the complete picture of the psychology of language understanding and production. The question of when an observed contextual aspect of meaning belongs in the explanatory realm of a semantic theory, or is the purview of a psycholinguistic theory is a complicated one, and depends on the explanatory goals of the semantic and psycholinguistic theories, and the extent to which they do or do not overlap. Thus, there are several open questions about the predictions for logical and pragmatic responders, as well for local and global responders.

### 6.5.2 Method

To determine a correlation between studies, first the proportion 'logical' and proportion 'global' were computed for each subject and compiled into one dataframe. For Bott & Noveck stimuli, proportion logical was the average rate of 'Accept' for sentence type T1 (aggregating over VERB, which was not a significant effect). For Chemla & Bott stimuli, proportion global was the average rate of rejection on critical T1 trials. Once those two numbers were calculated, they were combined together by participant ID so that each participant had a score from the two experiments. Then, that dataframe was merged with the QUESTIONS TASK dataset for statistical tests.

### 6.5.3 Results

#### Correlations between BN TASK and CB TASK

There was a weak but significant negative correlation between PROPORTION LOGICAL and PROPORTION GLOBAL ($r_s$ = -0.07, $p$ <0.0001). Figure 6.9 plots these two variables against each other. Participants who agreed with (responded based on a logical reading of) underinformative statements like Some elephants are mammals were slightly *less likely* to reject (respond based on the global reading of) statements like Zoologists don't know that elephants are insects. In other words, logical responders were more likely to access local presupposition readings.

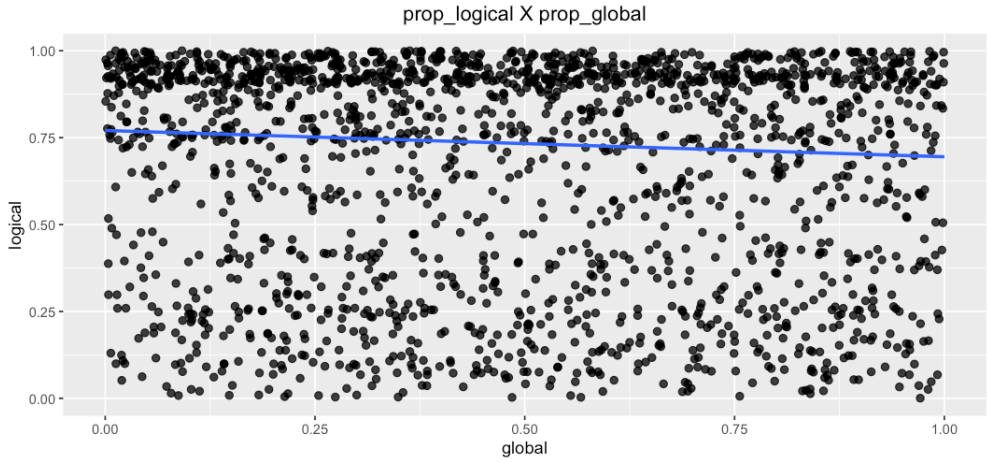


**Figure 6.9:** Bott & Noveck responses (proportion 'logical') plotted by Chemla & Bott responses (proportion 'global').

#### Correlations between CB TASK and QUESTIONS TASK

There was neither a significant correlation between responses on the CB TASK (proportion 'global' responses, $r_s$ = 0.02, $p$ = 0.1), a significant difference of that measure on responses in the QUESTIONS task ($\chi^2(2) = 4.48$, $p$ = 0.1), nor did participant's GLOBALITY significantly improve model fit ($\chi^2(1) = 1.185$, $p$ = 0.3). As we see in Figure 6.10, regression lines are generally flat, except in the NO GOAL condition.

Figure 6.11 plots results from the QUESTIONS TASK split by participants' GLOBALITY: participants who answered over 50% global on the CB TASK were labeled 'global

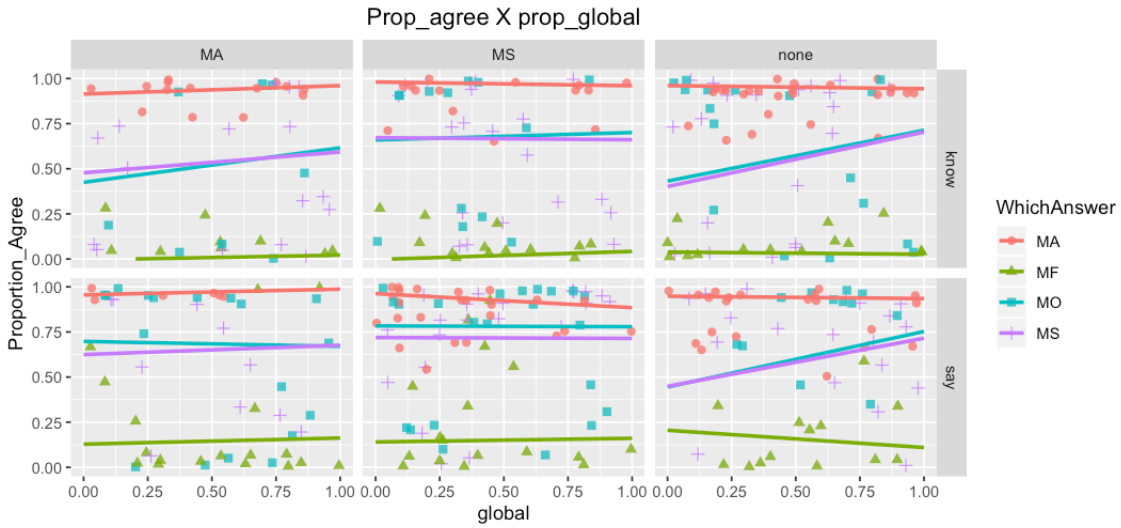


**Figure 6.10:** Responses from Questions Study split by response to scalar implicature in Chemla & Bott stimuli. The columns plot GOAL condition, the rows plot VERB, and colored shapes plot ANSWER condition.

responders', while participants who responded under 50% global were labeled 'local responders'. 66% of subjects were global responders, about 10% were local, and 24% are at 50%. While we do not find a correlation, global responders did agree in MS/MO conditions significantly more than local responders ($\chi^2(1)= 4.2778$, $p$=0.04). Neither kind of responder was significantly affected by GOAL.

### Correlations between BN TASK and QUESTIONS TASK

Using a Spearman non-parametric test, we found a weak correlation with BN TASK responses that was significantly different from zero ($r_s$ = 0.07, $p$ <0.0001). We confirmed this effect with both a non-parametric Kruskal-Wallis test ($\chi^2(6)$=107.36, $p$ <0.0001), and a model comparison showing that a regression model including the participant's score on the BN TASK (PROPORTION LOGICAL) as a predictor of response on the QUESTIONS TASK significantly improved model fit ($\chi^2(1)$ = 9.8278, $p$ <0.005).

Figure 6.12 plots the proportion 'Agree' in the QUESTIONS TASK by LOGICALITY. Here we see that regression lines for MS/MO answers are positively sloped in all three GOAL conditions, driving the positive correlation.

Figure 6.13 presents the results from the QUESTIONS TASK split by participants'

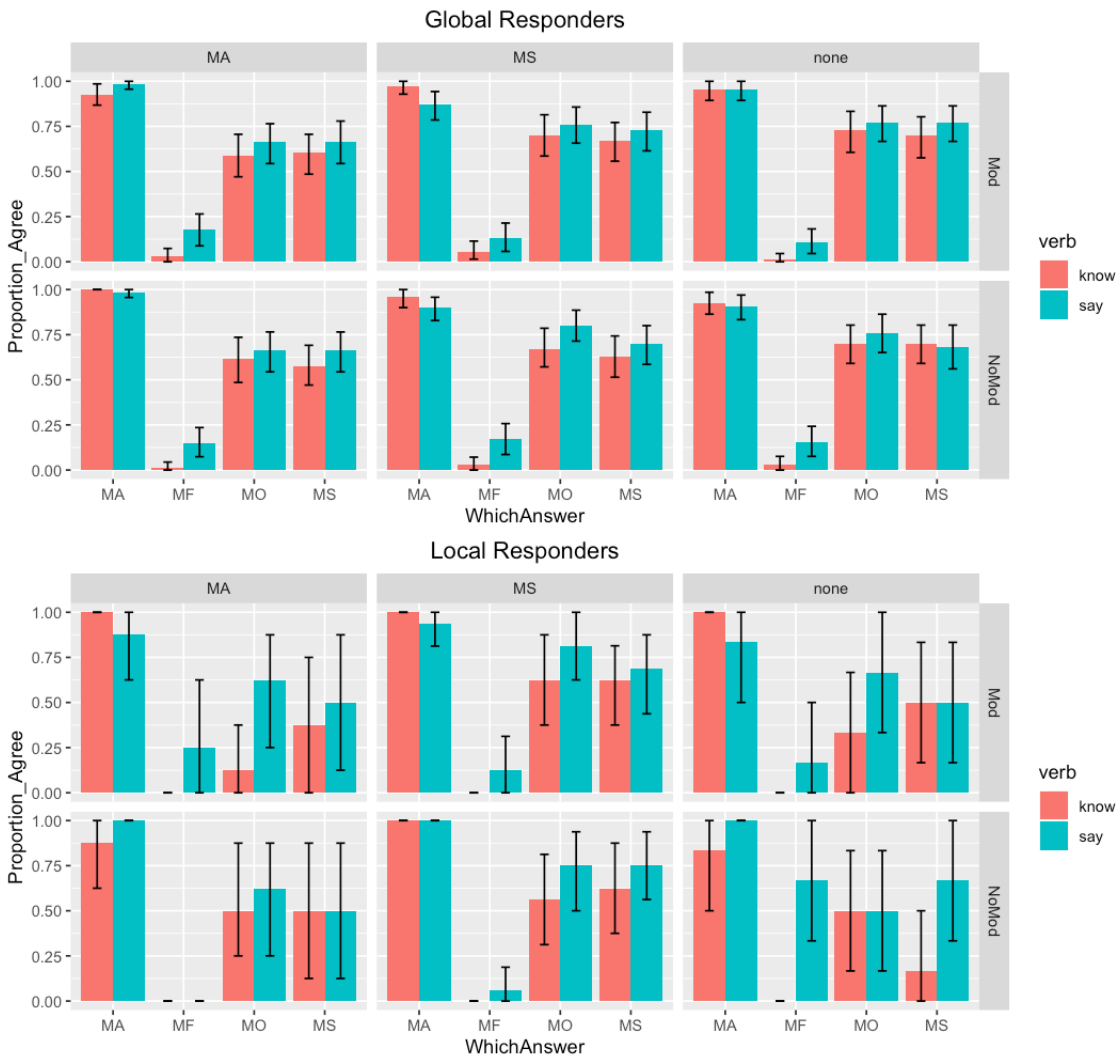


**Figure 6.11:** Responses from Questions Study split by response to presupposition projection under negation in Chemla & Bott stimuli.

LOGICALITY: participants who answered over 50% logical on the BN TASK were labeled 'logical responders' (top graph), while participants who answered under 50% 'logical' were labeled 'pragmatic responders' (bottom graph). About 70% of subjects were logical responders, 26% pragmatic, and the rest responded at chance. Note the significant increase in 'Agree' responses to MS/MO answer conditions amongst logical responders ($\chi^2(1) = 60.205$, $p$ <0.0001). Both logcal and pragmatic responders were significantly sensitive to GOAL(logical responders: $\chi^2(2) = 8.5605$, $p$=0.01; pragmatic responders: $\chi^2(2) = 30.592$, $p$ <0.0001), rating MS/MO answers higher in the MS GOAL condition. Neither logical nor pragmatic responders rated MS/MO answers differently

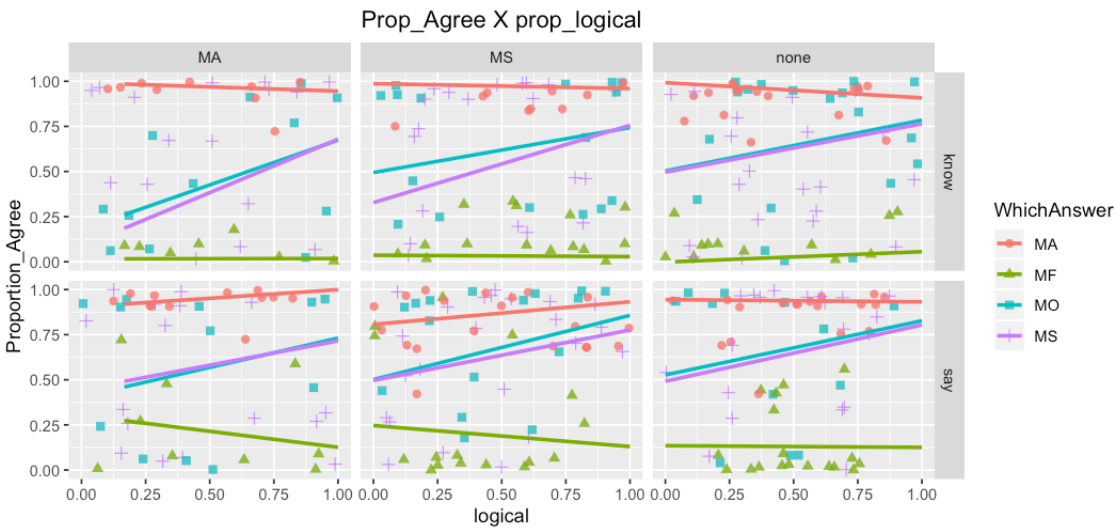


**Figure 6.12:** Responses from Questions Study split by response to scalar implicature in Bott & Noveck stimuli. The columns plot GOAL condition, the rows plot VERB, and colored shapes plot ANSWER condition.

between the MS GOAL and NO GOAL conditions, and both groups rated those answers significantly lower in the MA GOAL condition. MODAL was not significant in either group.

### 6.5.4 Discussion

We found a weak negative correlation between the B&N task and the C&B, suggesting that participants who accepted statements like Some cats are mammals on the logical reading (i.e., the true statement some and possibly all cats are mammals), were slightly less likely to accept statements like Zoologists don't know that cats are insects on the global presupposition readings (i.e., the false statement Cats are insects and Zoologists don't know that).

While we did not have concrete expectations about the connection between the C&B task and (non-) exhaustivity in questions, global responders rated mention-some conditions significantly higher than local responders. The difference was mostly driven by the NO GOAL condition.

We found that participant logicality predicted how they responded on the questions task: logical responders rated mention-some significantly higher than pragmatic

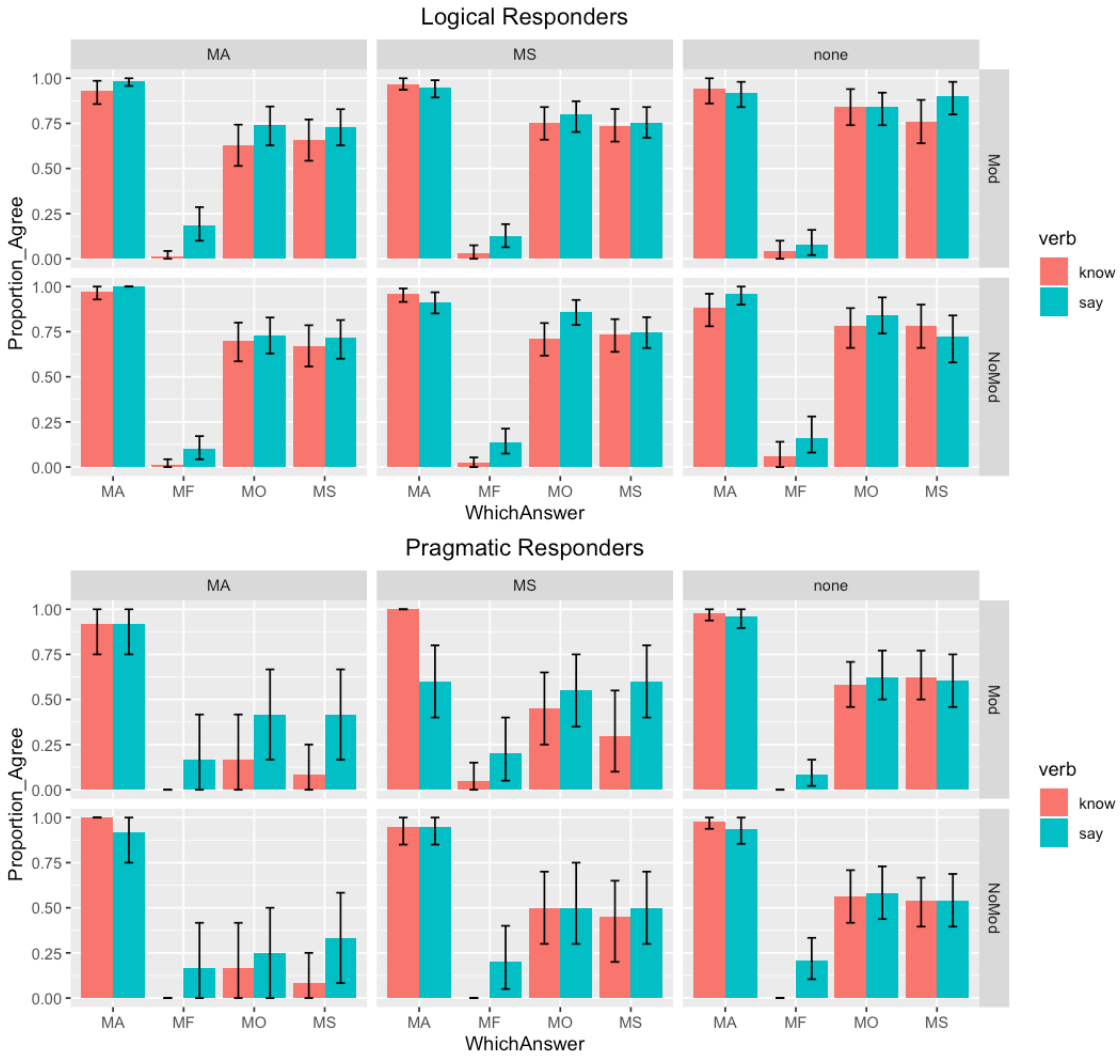


**Figure 6.13:** Responses from Questions Study split by response to scalar implicature in Bott & Noveck stimuli.

responders across all three GOAL conditions. Further, these responses are incredibly high, almost at ceiling. This result suggests that, if logical responders are responding based on the output of a semantic mechanism, this suggests the presence of a non-exhaustive semantic representation.

Mention-some answers with exhaustivity inferences (A and only A, or A and B, and only A and B) are false in the task. Given the near-floor rejections we see in the MS/MO conditions amongst pragmatic responders, it seems plausible that these participants calculated exhaustivity inferences. Further, these response rates are more on par with how participants responded in Experiment 1 to answers that provided false answers

in addition to the weak exhaustive answer. These pragmatic responders constituted a minority of overall participants.

Mention-all answer conditions are rated high across the board. This is consistent with our caveat about general pragmatic principles of maximizing informativeness, but it's also consistent with the availability of an exhaustive semantic representation. However, the fact that logical responders did not rate *only* these answers high, suggests that non-exhaustive answers also constitute logical ways to respond.

We found a significant effect of GOAL in both logical and pragmatic responders, although the effect was much greater in pragmatic responders. We hypothesized that context-sensitivity in logical responders could provide evidence for context-dependence at the literal/semantic level. The fact that mention-some responses are not accepted at ceiling, but degraded *from ceiling* (as opposed to from floor), is unsurprising given the overall bias of the card game context towards exhaustivity (cf. discussion in Section 6.1.1).

## 6.6 General Discussion

We found no effect of modal even in the NO GOAL condition. There could be several different explanations for this. First, one could argue as in response to Experiments 3a and 3b, that the experimental task allowed participants to complete the task without paying attention to the question form manipulation. Thus, so the reasoning goes, these results do not show that the modal doesn't make a difference to question meaning, but reflect a task-related artifact.

Recall the digression about the naturally exhaustive nature of card games. Given this, it is possible that with the addition of hearer expectations, the NO GOAL condition was sufficiently informative for a mention-all goal, thus question form was unnecessary as a cue to a discourse goal. The fact that we found no significant difference between the MA and NO GOAL conditions reveals that participants treated the two conditions on par.

Further, the contextual bias for exhaustivity could have been so strong as to cancel out the non-exhaustivity we built into the MS GOAL condition. Without this proper contextual licensing, then, the MS GOAL condition failed to be sufficiently non-exhaustive to actually license the mention-some representation. This interpretation of the results is consistent with our main hypothesis because it still posits a goal-driven (non-) exhaustivity.

However, one could argue that the lack of difference between those two conditions reveals that questions are semantically mention-all. One reason to reject this explanation is the fact that logical responders agreed in MS/MO ANSWER conditions almost at ceiling, while pragmatic responders agreed at most around 50% of the time. If responding logically tracks a penchant for responding closer to a truth-conditional semantic representation, then this result would suggest that there is a non-exhaustive semantic representation (in modal-less as well as modal questions). That we found this effect *in spite of* the context being biased towards exhaustivity, is perhaps stronger evidence.

Further, that pragmatic responders rejected MS/MO ANSWER conditions more than not, is consistent with them calculating exhaustivity inferences. This interpretation supports the idea that exhaustivity inferences are a general pragmatic phenomenon, rather than a hard-wired aspect of the question meaning (van Rooij & Schulz 2004, 2006; Schulz & van Rooij 2006; Spector 2006, 2007; Zimmermann 2010).

That logical responders were also sensitive to the GOAL manipulation is suggestive of context-sensitivity at the literal level. Hawkins et al.'s literal $A_1$ hearer will respond to the question asked with the answer that reduces the most uncertainty in the world. Given that exhaustive answers reduce the most uncertainty, the results are compatible with the RSA model predictions. Note, under this conception of a literal hearer, we might have expected to see mention-some answers accepted more than mention-one answers, given that the former conveys more information (reduces more uncertainty

about the world) than the latter. Of course, the exact model predictions would depend on how model parameters are actually cashed out, but this might be a reasonable expectation. We did not see any significant differences between mention-some and mention-one answers.

# Chapter 7

# Conclusions and General Discussion

Empirically, (non-)exhaustivity is like many other phenomena at the interface of semantics and pragmatics. It is resolved via an interplay between fine-grained compositional semantic information and general expectations about the context and world knowledge. Theoretically, the issue is vexed by an apparant asymmetry between readings, variations in readings due to linguistic form, and general context-sensitivity. These ingredients are compatible with quite distinct underlying semantics. How do we account for these variations in a principled way? Before attempting to answer that question, let us review what we have seen so far.

In Chapter 2, we reviewed several empirical generalizations about the distribution of mention-some answers to root questions, and readings of embedded questions. On the one hand, it appeared that mention-some was semantically constrained by the linguistic form of the question. In particular, we noted that the type of wh-question, the presence/absence of an existential priority modal, and in embedded questions, the matrix embedding verb, all appeared to impose restrictions on whether the the question was interpreted mention-some or mention-all. At the same time, these restrictions appeared to be no more than defeasible baseline interpretations that could shift appropriately with context.

Also in Chapter 2 we also reviewed theoretical semantic accounts of mention-some. We saw that semantic theories fall into two main categories: single representation theories, and ambiguity theories. Single representation theories, like the name suggests, posit that questions are not ambiguous but have a single underlying form. That single form could specify a (weak/strong) exhaustive/mention-all denotation (Groenendijk

& Stokhof (1982), (1984); Karttunen 1977), a non-exhaustive/mention-some denotation (Asher & Lascarides (1998); Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010), or be underspecified for (non-)exhaustivity (Ginzburg (1995); van Rooij (2003), 2004). No matter what the underlying semantic theory posits, all theories will require some pragmatic mechanism to explain however the range of interpretations are derived.

In Chapter 3, I conducted two experiments to test on a larger scale, and across a wide range of questions, the observed dual-licensing of mention-some. In Experiment 1, I determined baseline interpretations based on three linguistic form factors, the wh-word (who vs. where), the presence/absence of modality (finite non-modal clauses vs. infinitival clauses with a covert priority modal), and a matrix embedding verb (know vs. predict). We found significant differences for all three factors, as well as interactions, but the modal difference was the largest.

In Experiment 2, we manipulated context using high- and low-stakes as a first approximation of weak/strong exhaustive and non-exhaustive goals. We found that low-stakes contexts made singleton and intermediate non-exhaustivity significantly more acceptable than exhaustivity, while high-stakes contexts made weak/strong exhaustive and intermediate non-exhaustivity significantly more acceptable than singleton non-exhaustivity. Low-stakes contexts further made non-exhaustivity more acceptable than exhaustivity in non-modal finite clauses. This result suggests that constraints on the linguistic form of the question do not have a grammatical blocking effect, but instead facilitate baseline interpretations of (non-)exhaustivity. Not only is non-exhaustivity sensitive to discourse goals, but exhaustivity is as well (Schulz & van Rooij 2006; Spector 2007; Zimmermann 2010). The acceptability of a given level of (non-)exhaustivity is determined by its sufficiency for meeting contextual demands.

Across the two experiments, we also discussed the differences seen within individual experimental items. This post-hoc analysis suggested that variations in (non-)exhaustivity are further modulated by the expectations and world-knowledge that hearers imported into individual scenarios.

Together, these results support an analysis of (non-)exhaustivity that is resolved relative to the goal-relevant information encoded in the context, and additionally the world knowledge and expectations that are generated from each of those. The linguistic form of the question modulates those expectations, especially when the context does not make goal-relevant information clear enough.

In Chapter 4, we quantified the conditional probability of (non-)exhaustivity given the linguistic signal (Experiment 3a) and of (non-)exhaustivity given contextual discourse goals (Experiment 3b). In Experiment 3a where we did not explicitly manipulate contextual goals, we found that participants were sensitive to the linguistic form of the question but that evaluations of mention-some acceptability were all relatively high. Thus, while mention-some/mention-one conditions were rated lower in non-modal question conditions, median ratings were overall on the high end. These effects were driven by the dependent measure: when the measure was *likelihood* of (non-)exhaustivity, form differences were significant, but this was less often the case when the measure was *acceptability*. I argued that these task differences reflect another form of goal-sensitivity in the resolution of (non-)exhaustivity. In Experiment 3b we manipulated high and low stakes contexts as in Experiment 2, and found that hearers rated (non-)exhaustivity based on context, and not linguistic form. There was no main effect of modal in any condition. This result gave rise to the hypothesis that, when the context is informative with respect to a goal, defeasible goal-relevant information encoded in the linguistic signal is valued by hearers *less* than the contextual information.

In Chapter 5, we looked at the baseline distribution of linguistic cues in a corpus study, and then in an experimental task (Experiment 4) manipulating contextual goals as in Experiment 2 and 3b. The goal of these two studies was to understand the information available in the input to the language learner, and to quantify the probabilistic relationship between (non-)exhaustivity, linguistic form, and context.

Using the linguistic form factors established from the literature and experimental studies as cues for (non-)exhaustivity, the corpus study revealed that these cues are

infrequent, and often provide conflicting evidence for (non-)exhaustivity. While how-questions were the most frequent question-type (between who, where, and how), finite non-modal clauses are the most frequent clause type, and know the most frequent embedding verb. Thus, finite (non-modal) know-how-questions constitute the majority of questions found in naturalistic speech. On their own, these cues differentially link to (non-)exhaustivity; in combination, they render non-exhaustive interpretations much more frequent in the input (Asher & Lascarides (1998); Ginzburg (1995)). Yet, why is exhaustivity assumed to be more prevalent amongst semanticists? This suggested the need for a much deeper and fine-grained analysis of question predicates, as well as the surrounding contextual information. Both these elements could provide valuable insight into how goal-relevant information is encoded and updates hearer expectations about (non-)exhaustivity. In Experiment 4 we manipulated context again as high and low stakes, and found that speakers did not overwhelmningly produce questions that would reduce hearer undertainty about (non-)exhaustivity. This suggested that informative contexts reduce the need for informative questions.

Finally, Chapter 6 tested (1) the hypothesis that goal-informative contexts neutralize linguistic form factors, while under-informative contexts facilitate them, and (2) tapped in to hearer preferences to pinpoint the source of context-sensitivity.

Using a card-game experimental task (cf. Cremers & Chemla 2017; Phillips & George 2018), we manipulated context in three ways: to encode exhaustive goals, non-exhaustive goals, and to be explicitly underspecified for goals. On the one hand, we expect that form factors (presence/absence of a modal) would be significant in the unspecified context, per the hypothesis. At the same time, we expected that participants' prior expectations about typical card-game goals would be exhaustive, thus that responses in the unspecified goal condition would pattern with responses in the exhaustive goal condition. We found a significant effect of context which bore in the expected way (non-exhaustive answers were significantly more acceptable in the non-exhaustive goal condition), but found no effect of modal in any condition.

We also used two independent measures of semantic/pragmatic phenomena (scalar

implicature and presupposition projection) to guage hearer preferences and diagnose the resolution of (non-)exhaustivity against these more established phenomena. We found that (non-)exhaustivity resolution correlated with interpretation of sentences with scalar implicatures: literal/logical hearers who access the unstrengthend meaning accept non-exhaustive answer conditions near ceiling in the non-exhaustive goal condition, while pragmatic hearers reject them closer to floor. Further, both groups of hearers respond significantly differently with respect to our contextual manipulation. I argue this provides evidence that (non-)exhaustivity resolution is not a purely pragmatic matter, but is crucial to fixing literal meaning.

## 7.1 Fine brushstrokes

### 7.1.1 Interpretational variability and the underlying grammar

The data presented here, I argue, favor semantic approaches which in principle allow all wh-questions—regardless of their linguistic form—to give rise to any degree of (non-)exhaustivity. Essentially, that questions on the surface, are ambiguous. This descriptive observation is compatible with two distinct hypotheses about the underlying grammar. On the one hand, we could have underlying semantic ambiguity, where each distinct interpretation corresponds to a legitimate semantic representation. On the other hand, there could be one underlying representation.

We can take a look at some representative theories and identify what aspects of the theories are grammatical (or semantic), and what aspects are extra-grammatical (pragmatic). What I've shown in this work is that (non-)exhaustivity falls out from the close link between compositional semantic structure and information frequently considered to be extra-linguistic. This puts the current phenomenon squarely in line with much other research, for example like anaphora resolution (Asher & Lascarides 2003).

For Lahiri's (2002) semantics, this parallel is quite apt because his semantics appeals to context-sensitive variables which pick up their values from context. Lahiri (2002)'s goals are to capture (1) variability in question meaning due to the presences of

adverbs of quantification (quantificational variability effects), as well as variability due to different matrix interrogative-embedding verbs. Like many theories, Lahiri appeals to an ANS operator to capture the variability, but he adds a parameter C that relativizes answers to "context". This parameter serves two functions: it allows lexical semantic restrictions from the matrix verb to play a role in restricting the answer set, and it allows "further conditions derived from some contextually salient property (or some property easily inferable from the context)," (Lahiri 2002, pp. 93). By this latter condition, Lahiri refers to typical factors thought to constrain referential domains. He goes on to suggest that this variable may not be sufficient for capturing non-exhaustivity, and posits a covert quantificational adverb whose force varies contextually (a "pragmatically variable default adverbial," pp. 94). Thus, a sentence like Dana knows Q would have a meaning close to (180).

(180) enough $p$.[ANS$(p, \mathrm{Q}) \wedge \mathrm{C}(p)$][Dana knows in $w$ that $p$]

The covert quantifier is not always overtly pronounced, but syntactically present, and the contextual variable $C$ is in the restriction of the covert quantifier. Note that for Lahiri, there are two loci of context-sensitivity that conspire to constrain the set of answers, and they arise because of the lexical semantic properties of the two components (the ANS operator and the covert adverbial quantifier). Further, these pragmatic variables interact with the narrow sentential context as well as the broader situational context. As with other kinds of context-sensitive variables, these two variables here may introduce vagueness that is conpounded when embedded under know.

Asher & Lascarides (1998) model (non-)exhaustivity as arising from the interaction between compositional semantics, discourse, and the intentional content of interlocutor's mental states (for example, their plans). They do this using a very different formalism, Segmented Discourse Representation Theory (SDRT), whose goal is to provide a formal dynamic semantics of discourse, thus modeling the interaction of linguistic and extralinguistic information. Crucially, rhetorical (or coherence) relations determine how information in the discourse is composed, and which relation is used is determined both by semantic and syntactic information in the utterance, and by

mental state information.[1]

For Asher & Lascarides, a question denotes a set of propositions, the (direct) answers to the questions, and includes non-exhaustive ones. They capture the dependence of (non-)exhaustivity on the questioner's plan and mental state because these non-linguistic pieces of information helps guide the inference that the hearer makes about how utterances in the discourse attach to one another, which relation is the best to deploy. Root questions and answers can be attached in a discourse via two different relations. *Question-Answer Pair* takes a question and a response as arguments, and returns true if the response is in the set denoted by the question (a direct answer). *Indirect Question-Answer Pair* relates a question and a response, and returns true if a direct answer can be infered from the response (that a direct answer normally follows from the response). These relations also hold at the embedded level because they represent models of the discourse participants' knowledge states, and SDRT representations can be arguments to propositional attidude verbs (Asher 1986, 1993).

The theories of Lahiri and Asher & Lascarides have very different explanatory goals. Lahiri's theory represents a traditional approach to the study of linguistic meaning that separates out linguistic meaning from linguistic use. As such, it attempts to abstract away from the pragmatic aspects of language. To give a completely explanatory theory of language, and even more so to give a theory of how linguistic meaning changes in discourse, Lahiri's theory would have to be embedded in a theory of discourse, and in a theory of intentional content. However, this is not his explanatory goal. In contrast, it is Asher & Lascarides's aim to give a formal account of discourse that shows how hearers make the context-sensitive and/or pragmatic inferences that they make. Indeed, their theory is intended to be cognitively-plausible, while traditionally linguistic theories avoid such questions of cognitive plausibility.

Further, while neither theory gives an explicit account of how prior expectations are involved in interpretation, neither theory is inherently incompatible with this data. For Lahiri, as a representative of traditional semantic theories, linguistic expectations

[1]More correctly, by the hearer's mental model of their interlocutor's mental states.

are a product of language use and language experience. They are therefore not in the explanatory realm of a semantic theory per se. Like Lahiri, the theories on the table which posit underlying ambiguity (Beck & Rullmann 1999 and George 2011, Ch. 2) are in principle compatible with the results presented here. For Asher & Lascarides, the interaction between expectations and compositional semantic information is potentially salient to modeling discourse. SDRT allows for these two kinds of information to interact. While Asher & Lascarides (1998) haven't attempted to model all the other interesting phenomena occuring with questions that many semanticists like Lahiri have, SDRT provides a framework in which such accounts could be integrated and their interaction with extra-linguistic information formally modeled and tested. Thus, it provides a hope for examining the interaction between context and linguistic form in much more concrete and scientific way.

### 7.1.2 Baseline interpretations derive from hearer expectations

Baseline interpretations of (non-)exhaustivity in questions are derived, not from an underlying semantics, but from extra-grammatical factors concerning the prior likelihood of a given level of (non-)exhaustivity, given the linguistic form of the question. This is because, when we access an intution about meaning we implicitly construct background contexts against which we evaluate meaning.

The psycholinguistic evidence showing that prior expectations often drive interpretation is abundant. For example, world knowledge (Chambers et al. 2004; Kehler et al. 2008), lexical semantic, syntactic, prosodic properties of the utterance (Carroll, Tanenhaus & Bever 1978; Tanenhaus 1978; Tanenhaus, Leiman & Seidenberg 1979; Altmann & Steedman 1988; Crain 1980; Altmann 1985; Trueswell, Tanenhaus & Kello 1993; Spivey-Knowlton, Trueswell & Tanenhaus 1993; Trueswell, Tanenhaus & Garnsey 1994; Altmann & Kamide 1999; van Berkum, Brown & Hagoort 1999; Cummins & Rohde 2015; De Marneffe & Tonhauser 2019), context and QUD (Marslen-Wilson, Tyler & Seidenberg 1978; Swinney 1979; Sedivy and Spivey-Knowlton 1993; Sedivy, Tanenhaus, Chambers & Carlson 1999; Degen & Goodman 2014).

Consider an analogy: the difference between the two minimal pairs in (181), from Steedman (2000), citing Bever (1970). These are variations on a classic garden path sentence, The horse raced past the barn fell. They have the same syntactic structure, but the difference is in the subject of the sentence.

(181) a. The doctor sent for the patient arrived.
b. The flowers sent for the patient arrived.

(181a), unlike (181b), gives rise to a garden-path. The differences in transitional probability between the pairs doctor/flower and flowers/sent are said to be blamed: the first pair has a higher transitional probability than the second pair. This fact derives from our experience and world knowledge: "because flowers, unlike doctors, cannot send for things," (Steedman 2000, p. 241). Importantly, those expectations will change with context. If the sentence occured in a children's story where we expect animacy from typically inanimate things, (181b) would give rise to the predicted garden path effect because in that world, unlike the first, flowers *can* send for things. Without explicitly manipulating context in this way, we might be inclined to conclude that the difference between (181a) and (181b) was grammatical; however this conclusion would be wrong. Our expectations, which may differ with context, guide the interpretive processes of language understanding. Without this manipulation, we fail to capture an important generalization about interpretation, and thus about meaning.

The fact that an interpretation may accessed by "default", typically, more often, or first, could reflect several different factors, including a lack of proper contextual manipulation as shown above. We see this in studies of quantifier scope ambiguity as well (Bolinger 1965; Akmajian & Jackendoff 1970; Jackendoff 1972; Ladd 1980; Ward & Hirschberg 1985; Kadmon & Roberts 1986; Musolino 1998; Musolino, Crain & Thornton 2000; Baltzani 2002, 2003; Fodor 2002; Anderson 2004; Musolino & Lidz 2004; Syrett & Nisula 2014), and the lesson we learned was that interpretive preferences with regards to ambiguous utterances should not license the inference that the dispreferred interpretation is not grammatically available, without a serious attempt at providing a context to license the purported reading. Pragmatics determines both

the "*a priori*" and "*a posteriori*" interpretations, in virtue of filling in required contextual information implicitly ("*a priori*") or explicitly ("*a posteriori*") via expectations.

Even if a given question form is underspecified or underlyingly ambiguous, not all resolutions of non-exhaustivity are *a priori* likely given hearer expectations. Both the questions in (182) can be asked by a smoker looking to light their cigarette. While (182a) clearly indicates a goal, (182b) could be understood in the same context as indicating the same goal.

(182) a. Who has a light?
b. Who has fire?

Note that, if we had not been discussing the lighting of cigarettes, it would perhaps be more difficult to make sense out of (182b)—what *kind* of fire do you need?—unless you are familiar with smokers, and in particular, French smokers. You might have the impulse to answer (182b) in a certain way, depending on your prior expectations about the likely goals that a speaker asking that question would have. Those expectations may derive from any number of things, but they do not show that the question cannot have a goal that is different from your best-guess. The manner in which a question is asked may betray all sorts of extra-linguistic facts about the speaker asking it, including the kinds of goals that they are likely to have.

Similarly, if a speaker wanted to be completely transparant about why they are requesting information, they might ask not (183a) but (183b) or (183c).

(183) a. Who came to the party?
b. Who is everyone that came to the party?
c. Who is someone that came to the party?

But speakers are not that transparent, as we saw in Chapter 5 (the corpus study and Experiment 4). They do not go out of their way to unambiguously communicate their goals and intentions. Given that the hearer's job involves a high degree of uncertainty about the speakers goal and the speaker's intended meaning, it seems natural that prior expectations would fill in some of those holes.

In Chapter 3 we analyzed individual test scenarios and found significant differences in some cases. We said that the world knowledge associated with these items permitted participants to impute additional information into the test items, which further constrained the resolution of (non-)exhaustivity. In particular, for some high-stakes test scenarios in Experiment 2, participants did not rate (highly informative) non-exhaustive answers differently from exhaustive answers. It is plausible that participants imputed for example, an addition constraint of time-sensitivity into the scenarios, and this rendered non-exhaustive answers as acceptable as exhaustive ones.

When we quantified hearer preferences in Experiment 5, we found that there were significant differences between "logical/literal" hearers who access some and possible all reading, and "pragmatic" hearers who access stronger some but not all readings of ambiguous sentences with the existential quantifier. This was yet one more way that hearer-specific factors drive (non-)exhaustivity resolution.

### 7.1.3 The source of strong and weak exhaustivity on an underspecified semantics

I have argued that questions are underspecified for non- and weak/strong exhaustivity. Ginzburg (1995) and Asher & Lascarides (1998) argue weak and strong exhaustivity are too strong when we consider how-, why-, and even where- questions. Note, the claim is not that one cannot ever construct a context where it is possible to determine a strong/weak exhaustive set. On the contrary, the central claim of this thesis is that one could construct a contexts to make any answer felicitous, and that the preference for certain kinds of answers is determined by pragmatics. I have also argued that our a priori expectations about how a question should be answered (or an embedded question be interpreted) derives from our expectations about the typical goals of a speaker who makes the utterance. This happens because question underspecification is resolved relative to speaker goals, and thus hearers will impute goal-relevant information into the discourse if the requisite information is not already forth-coming.

Asher & Lascarides have suggested that one reason for this asymmetry in questions

has to do with whether the cognitive task at hand is compatible with the strongest interpretation, i.e. "the maximally specific proposition in the meaning of the interrogative," (169). They cite Dalrymple et al. (1998) for this, the strongest interpretation principle. For non-who-questions, it's not cognitively reasonable to demand the strongest interpretation. I think the bears further discussion because this idea relates very much to Grice's Maxim of Quantity.

The Maxim of Quantity has two parts. Quantity 1 states that speakers should make their contributions to the conversation as informative as is required, and Quantity 2 states that they should not make their contributions *more informative* than is required.

In a question/answers dialogue, when the hearer doesn't know the questioner's goal, the pressure to be informative could drive them to be as exhaustive as possible in order to be a good conversational participant, and discharge their responsability as answerer.

For a who-question, the referential domain will often be greatly constrained by the common ground between interlocutors. This lessens the cognitive task of constructing a weak or strong exhaustive set. Let us take a familiar question, Who will chair their committee?, asked in a familiar context: a graduate student preparing for their doctoral candidacy qualifications. In the worst case scenario, there might be forty faculty in the student's department who are possible chair candidates, in virtue of being faculty members. However, by the time the student approaches candidacy, already most of those candidates will be weeded out based on the match between the student's and potential chair's research interests, amongst other practical factors like the faculty's administrative duties. In this context, because the community is rather small, there will be shared knowledge between the members of the department. It will be known—or at the very least suspected based on the faculty's knowledge of their colleagues' research programs—who the short list of candidate chairs would be, and it would be rare to find a list greater than two. In terms of cognitive tasks, it is therefore quite easy to construct the weak or the strong exhaustive answer set.

Shared knowledge can in principle constrain the referential domain for any question. However, it seems that the "pre-packaged" restrictions that we saw for who-questions will not necessarily be available for non-who-questions. If we keep the context of an academic department we can examine how other answer sets may constrained by world knowledge. A question like, How do I get in to grad school?, will have an answer set with a somewhat different structure, and the questioner's world knowledge is potentially vastly different than the answerer's. One candidate answer, You get in to grad school by demonstrating your ability to think critically and thoroughly about a particular subject, may not give enough concrete information. Another candidate answer that reiterates the list of typical application requirements (e.g., writing sample, GRE scores, three recommendation letters), will be underinformative, despite the fact that it constitutes an exhaustive list at some level. In some sense the first answer subsumes the second, because following the guidelines in the application list demonstrates the candidate's competence. Finally, let us not dismiss the additional reality of the situation, the ingredients which must be just right, but are not under the candidate's control: the pool of applicants, issues of funding, the composition of the admissions committee. Even in this constrained domain, it isn't clear that one can provide a weak exhaustive answer. However, the answer may invole a set of causally-dependent steps much in the way a recipe does. At what point does the student know how?

There will also be interactions with the linguistic form of the question, which encodes some degree of information about the world. Chairing a committee will typically be a single-person job, especially in academic circles, though co-chairing is possible *as stipulated by the practice of a given context*. This a fact about language grounded in social ontology. The linguistic form of the question thus reflects the natural constrains on the referential domain of expressions.

## 7.2 Conclusion

I've argued that questions are underspecified for (non-)exhaustivity, and as such, must be resolved relative to a context which makes explcit discourse goals. When a question is evaluated without an explicit context, hearers impute one using their expectations about typical contexts given the linguistic utterance, and using default pragmatic assumptions consistent with Grice's Maxim of Quantity. While hearers are driven to give weak and strong exhaustive answers, and likewise so interpret embedded question utterances, this is strategy is tempered by cognitive constraints: often it is not possible, plausible, or even necessary to be weakly or strongly exhaustive. Thus, the asymmetry between who- questions, which on the one hand appear to typically require weak or strong exhaustive answers/interpretations of embedded counterparts, and other wh-questions, which on the other hand appear to disprefer weak/strong exhaustive answers, can be explained as a balance between being maximally informative, and being sufficiently informative.

# Bibliography


Asher, N., & Lascarides, A. (1998). Questions in dialogue. *Linguistics and Philosophy*, *21*(3), 237–309. https://doi.org/10.1023/A:1005364332007

Baayen, R. (2007). *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge University Press.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Beaver, D., & Clark, B. (2003). *Always* and *only*: Why not all focus-sensitive operators are alike. *Natural Language Semantics*, *11*, 323–362.

Beck, S. (1996). *Wh-constructions and transparent logical form* (Doctoral dissertation). University of Tubingen.

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*, 105–157.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python*. O'Reilly Media Inc.

Boër, S. E., & Lycan, W. (1975). Knowing who. *Philosophical Studies*, *28*(5), 299–344.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457.

Breheny, R., Ferguson, H., & Katsos, N. (2012). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443–467.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? an on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463.

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in cnversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(6), 1482–1493.

Chambers, C., Tanenhaus, M., & Magnuson, J. (2004). Action-based affordances and syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 687–696.

Chemla, E., & Bott, L. (2013). Processing presuppositions: Dynamic semantics vs pragmatic enrichment. *Language and Cognitive Processes, 28*(3), 241–260.

Chemla, E., & George, B. (2015). Can we agree about 'agree'? *Review of Philosophy and Psychology*, 1–22.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics (C. Maienborn, K. von Heusinger, & P. Portner, Eds.). In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning*. Berlin: Moulton de Gruyter, 2297–2332.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Capbridge, MA.
Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2013). Inquisitive semantics: A new notion of meaning. *Language and Linguistics Compass*, *7*, 459–476.
Clark, H. (1996). *Using language*. Cambridge University Press.
Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.
Cole, P., Hermon, G., & Huang, C. (2006). Long-distance anaphors: An asian perspective (M. Evaraert & H. van Riemsdijk, Eds.). In M. Evaraert & H. van Riemsdijk (Eds.), *The blackwell companion to syntax, vol. 3*. Malden, MA: Blackwell.
Comorovski, I. (1996). *The interpretation of interrogative phrases*. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-94-015-8688-7_2
Cremers, A., & Chemla, E. (2016a). Experiments on the acceptability and possible readings of questions embedded under emotive-factives. *Semantics Archive*.
Cremers, A., & Chemla, E. (2016b). A psycholinguistic study of the exhaustive readings of embedded questions. *Journal of Semantics*, *33*(1), 49–85.
Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science*, *60*, 721–736.
Dabrowska, E. (2010). Naïve v. expert intuitions: An empirical study of acceptability judgements. *Linguistic Review*, *27*, 1–23.
Dayal, V. (1996). *Locality in wh quantification: Questions and relative clauses in hindi*. Dordrecht: Kluwer.
Dayal, V. (2015). *Questions*. Oxford, Oxford Surveys in Semantics; Pragmatics.
Degen, J. (2013). *Alternatives in pragmatic reasoning* (Doctoral dissertation). University of Rochester.
Degen, J. (2015). Investigating the distribution of "some" (but not "all") implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(11), 1–55.
Degen, J., & Tanenhaus, M. (2015). Processing scala implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.
Elman, J., Hare, M., & McRae, K. (2004). Cues, constraints, and competition in sentence processing (M. Tomasello & D. Slobin, Eds.). In M. Tomasello & D. Slobin (Eds.), *Beyond nature-nurture: Essays in honor of elizabeth bates*. Mahwah, NJ: Erlbaum.
Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *Linguistic Review*, *22*, 365–380.
Ferreira, F., & Clifton, C. (1986). The indepdndence of syntactic processing. *Journal of Memory and Language*, *25*(3), 438–368.
Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production [PMID: 30231000]. *Annual Review of Psychology*, *70*(1), 29–51.
Ferreira, V., & Dell, G. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*, 296–340.
Ferreira, V., & Hudson, M. (2011). Saying "that" in dialogue: The influence of accessibility and social factors on syntactic production. *Language and Cognitive Processes*, *26*(10).
Ferreira, V., & Schotter, E. (2013). Do verb bias effects on sentence production reflect sensitivity to comprehension or production factors? *Quarterly Journal of Experimental Psychology*.
Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*. Bradford Books, MIT Press.

Forster, K. (1979). Levels of processing and the structure of the language processor (W. E. Cooper & E. Walker, Eds.). In W. E. Cooper & E. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to merrill garrett*. Hillsdale, NJ: Erlbaum.

Fox, D. (2014). *Mention-some readings* [Class notes]. Class notes.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frazier, L. (1987). Sentence processing: A tutorial review, In *Attention and performance xii: The psychology of reading*. London: Erlbaum.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–325.

Frege, G. (1892). Uber sinn und bedeutung (Geach & Black, Eds.). In Geach & Black (Eds.), *Translations from the philosophical writings of gottlob frege*. Blackwell, Oxford.

George, B. R. (2013a). Which judgments show weak exhaustivity? (and which don't?) *Natural language semantics*, *21*(4), 401–427.

George, B. R. (2011). *Question embedding and the semantics of answers* (Doctoral dissertation). University of California, Los Angeles.

George, B. R. (2013b). Knowing-'wh', mention-some readings, and non-reducibility. *Thought: A Journal of Philosophy*, *2*(2), 166–177.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Science*, *14*, 233–234.

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*, 88–124.

Gibson, E., Piantadosi, S., & Fedorenko, E. (2011). Using mechanical turk to obtain and analyze english acceptability judgements. *Language and Linguistics Compass*, *5*, 509–524.

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.

Goro, T. (2007). *Language specific constraints on scope interpretation in first language acquisition* (Doctoral dissertation). University of Maryland, College Park.

Grice, H. P. (1975). Logic and conversation (P. Cole & J. Morgan, Eds.). In P. Cole & J. Morgan (Eds.), *Speech acts*. New York: Academic Press, 41–58.

Grimshaw, J. (1979). Complement selection and the lexicon. *Linguistic Inquiry*, *10*(2), 279–326.

Grodner, D., Klein, N., Carbard, K., & K.Tanenhaus, M. (2010). ''some,'' and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42–55.

Grodner, D., & Sedivy, J. (2011). The effects of speaker-specific information on pragmatic inferences. *Processing and Acquisition of Reference*, 239–271.

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. (1982). Semantic analysis of wh-complements. *Linguistics and Philosophy*, *5*(2), 175–233.

Guerzoni, E. (2007). Weak exhaustivity and 'whether': A pragmatic account, In *Proceedings of salt*.

Guerzoni, E., & Sharvit, Y. (2007). A question of strength: On NPIs in interrogative clauses. *Linguistics and Philosophy*, *30*(3), 361–391.

Hamblin, C. L. (1963). Questions aren't statements. *Philosophy of Science*, *30*(1), 62–63.
Hamblin, C. (1958). Questions. *Australasian Journal of Philosophy*, *36*(3), 159–168.
Hawkins, R., & Goodman, N. (2019). Why do you ask? the informational dynamics of questions and answers.
Hawkins, R., Stuhlmüller, A., Degen, J., & Goodman, N. (2015). Why do you ask? good questions provoke informative answers, In *Proceedings of the 37th annual conference of the cognitive science society.*
Heim, I. (1994). Interrogative semantics and Karttunen's semantics for *know*, In *Proceedings of iatl*.
Heller, D., Grodner, D., & K.Tanenhaus, M. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.
Hirotani, M. (2005). *Prosody and lf interpretation: Processing japanese* wh*-questions* (Doctoral dissertation). University of Massachusetts, Amherst.
Hobbs, J. (1979). Coherence and coreference. *Cognitive Science*, *3*, 67–90.
Horton, W., & Gerrig, R. (2002). Speakersõ experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*, 589–606.
Ito, K., & Speer, S. (2008). Anticipatory effects of intonation: Eye movements during instructed visual speech. *Journal of Memory and Language*, *58*, 541–573.
Jon Sprouse, C. S., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, *134*, 219–248.
Julien Musolino, S. C., & Thornton, R. (2008). Navigating negative quantificational space. *Linguistics*, *38*(1).
Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, *1*(1), 3–44.
Kehler, A., Kertz, L., Rohde, H., & Elman, J. (2008). Coherence and coreference revisited. *Journal of Semantics, Special Issue on Processing Meaning*, *25*, 1–44.
Kehler, A., & Rohde, H. (2019). Prominence and coherrence in a bayesian theory of pronoun interpretation. *Journal of Pragmatics, Special Issue on Prominence in Pragmatics*, *154*, 63–78.
Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.
Kitagawa, Y., & Fodor, J. (2006). Prosodic influence on syntactic judgements (G. Fanselow, C. Fery, M. Schlesewsky, & R. Vogel, Eds.). In G. Fanselow, C. Fery, M. Schlesewsky, & R. Vogel (Eds.), *Gradience in grammar: Generative perspectives*. Oxford University Press.
Klinedinst, N., & Rothschild, D. (2011). Exhaustivity in questions with non-factives. *Semantics and Pragmatics*, *4*(2), 1–23.
Kratzer, A. (1981). The notional category of modality (H. J. Eikmeyer & H. Rieser, Eds.). In H. J. Eikmeyer & H. Rieser (Eds.), *Words, worlds, and contexts*.
Kratzer, A. (1991). Modality (A. von Stechow & D. Wunderlich, Eds.). In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research*.

Kroll, M., & Rysling, A. (2019). The search for truth: Appositives weigh in, In *Proceedings of salt 29*.

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 University Press on Demand.

Levinson, S. (2000). *Presumptive meanings - the theory of generalized conversational implicature*. MIT Press.

MacDonald, M., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Marslen-Wilson, W., Tyler, L. K., & Seidenberg, M. (1978). Sentence processing and the clause boundary.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Neely, J. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226–254.

Nicolae, A. C. (2014). *Any questions? polarity as a window into the structure of questions* (Doctoral dissertation). Harvard University.

Phillips, C. (2010). Should we impeach armchair linguists. *Japanese/Korean Linguistics*, *17*, 49–64. http://ling.umd.edu/~colin/wordpress/wp-content/uploads/2014/08/phillips2010_armchairlinguistics.pdf

Phillips, C., & Parker, D. (2014). The psycholinguistics of ellipsis [published online Nov 27, 2013]. *Lingua*, *151*, 78–95. http://ling.umd.edu/~colin/wordpress/wp-content/uploads/2014/08/phillips-parker2014.pdf

Phillips, C., & Wagers, M. (2007, January 1). Relating structure and time in linguistics and psycholinguistics, In *Oxford handbook of psycholinguistics*. Oxford University Press. http://ling.umd.edu/~colin/wordpress/wp-content/uploads/2014/08/phillipswagers2007.pdf

Phillips, J., & George, B. (2018). Knowledge wh and false beliefs, experimental investigations. *Journal of Semantics*.

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Pollard, C., & Xue, P. (2001). Syntactic and non-syntactic constraints on long-distance reflexives (P. Cole, G. Hermon, & C. Huang, Eds.). In P. Cole, G. Hermon, & C. Huang (Eds.), *Long-distance reflexives*. San Diego: Academic Press.

Portner, P. (2009). *Modality*. Oxford: Oxford University Press.

Posner, M., & Snyder, C. (1975). Attention and cognitive control (R. Solso, Ed.). In R. Solso (Ed.), *Information processing and cognition*. Hillsdale, NJ: Erlbaum.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*, 358–374.

Reis, M. (1992). The category of invariant *alles* in wh-clauses, In *Arveitspapiere des sonderforschungsbereichs 340*. University of Tubingen, 1–33.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, 91–136.

Roelofsen, F., Theiler, N., & Aloni, M. (2014). Embedded interrogatives: The role of false answers, In *7th questions in discourse workshop, göttingen*.

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presupposition and indirect scalar implicatures. In *Experimental perspectives on presuppositions* (pp. 215–240). Cham, Springer International Publishing. http://www.springer.com/us/book/9783319079790

Rubinstein, A. (2012). *Roots of modality* (Doctoral dissertation). University of Massachusetts.

Russell, B. (1905). On denoting. *Mind*, *14*(56), 479–493. http://www.jstor.org/stable/2248381

Schulz, K., & Van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, *29*(2), 205–250. https://doi.org/10.1007/s10988-005-3760-4

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgements and linguistic methodology*. U Chicago Press.

Schwarz, F. (2007). Processing presupposed content, *24*(4), 373–416. https://doi.org/10.1093/jos/ffm011

Schwarz, F. (2015a). *Experimental perspectives on presuppositions*. Cham, Springer International Publishing. http://www.springer.com/us/book/9783319079790

Schwarz, F. (2015b). Presuppositions vs. asserted content in online processing. In *Experimental perspectives on presuppositions* (pp. 89–108). Cham, Springer International Publishing. http://www.springer.com/us/book/9783319079790

Schwarz, F., Aguilar-Guevara, A., Le Bruyn, B., & Zwarts, J. (2014). How weak and how definite are weak definites? In *Weak referentiality* (pp. 213–235). Amsterdam/Philadelphia, John Benjamins. https://benjamins.com/#catalog/books/la.219.09sch/details

Schwarz, F., & Tiemann, S. (2017). Presupposition projection in online processing, *34*(1), 61–106. https://doi.org/10.1093/jos/ffw005

Sharvit, Y. (2002). Embedded questions and 'de dicto' readings. *Natural Language Semantics*, *10*(2), 97–123.

Shiffrin, R., & Schneider, W. (1977). Conrolled and automatic human information processing. *Psychological Review*, *84*, 127–190.

Shuhong Lin, N. E., Boaz Keysar. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556.

Simons, M. (2001). On the conversational basis of some presuppositions. (B. J. Rachel Hastings & Z. Zvolenszky, Eds.). In B. J. Rachel Hastings & Z. Zvolenszky (Eds.), *Proceedings of salt 11*, Ithaca, NY: CLC Publications.

Spector, B. (2007). Modalized questions and exhaustivity. *Semantics and Linguistic Theory*, *17*, 282–299.

Spector, B., & Egre, P. (2015). A uniform semantics for embedded interrogatives: An answer, not necessarily the answer. *Synthese*, *192*(6), 1729–1784.

Sprouse, J. (2011). A validation of amazon mechanical turk for the collection of acceptability judgements in linguistic theory. *Behavioral Research Methods*, *43*, 155–167.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. *Journal of Linguistics*, *48*, 609–652.

Steedman, M. (2000). *The syntactic process*. MIT Press.

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344.

Swinney, D. (1979). Lexical access during sentence comprehension (re)consideration of contextual effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.

Tanenhaus, M., Leiman, J., & Seidenberg, M. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 427–440.

Tanenhaus, M., & Trueswell, J. (1995). Sentence comprehension (J. L. Miller & P. D. Elmas, Eds.). In J. L. Miller & P. D. Elmas (Eds.), *Speech, language, and communication. handbook of perception and cognition*. San Diego, CA: Academic Press.

Theiler, N. (2014). *A multitude of answers: Embedded questions in typed inquisitive semantics* (Doctoral dissertation). Universiteit van Amsterdam.

Tomioka, S. (2007). Pragmatics of lf-interention effects: Japanese and korean *wh*-interrogatives. *Journal of Pragmatics*, *29*, 1570–1590.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *ournal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528–553.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285–318.

Uegaki, W. (2014). Predicting the variation in exhaustivity of embedded interrogatives. *Sinn und Bedeutung*, *19*.

v. Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the n400. *Journal of Cognitive Neuroscience*, *11*(6), 657–671.

van Rooij, R. (2004). The utility of mention-some questions. *Research on Language and Computation*, *2*, 401–416.

van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language, and Information*, *13*, 491–519.

von Fintel, K. (1994). *Restrictions on quantifier domains* (Doctoral dissertation). University of Massachusetts, Amherst.

von Fintel, K., Fox, D., & Iatridou, S. (2016). Definiteness as maximal informativeness (L. Crnič & U. Sauerland, Eds.). In L. Crnič & U. Sauerland (Eds.), *The art and craft of semantics: A festschrift for irene heim*. Cambride, MA, MIT Working Papers in Linguistics.

Wu, S., & Keysar, B. (2006). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*, 169–181.

Xiang, Y. (2016a). Complete and true: A uniform analysis for mention-some and mention-all (N. B. Polina Berezovskaya & A. Sch oller, Eds.). In N. B. Polina Berezovskaya & A. Sch oller (Eds.), *Proceedings of sinn und bedeutung 20*.

Xiang, Y. (2016b). *Interpreting questions with non-exhaustive answers* (Doctoral dissertation). Harvard University.

Xiang, Y. (2016c). Solving the dilemma between uniqueness and mention-some, In *Sinn und bedeutung 20*.

Xiang, Y., & Cremers, A. (2017). Mention-some readings of plural-marked questions: Experimental evidence, In *Proceedings of north east linguistics society 47*.