UNIVERSITY OF CALIFORNIA,
IRVINE

Disambiguating Information in Speech and Context

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Language Science

by

Noa Attali

Dissertation Committee:
Professor Lisa Pearl, Co-Chair
Associate Professor Gregory Scontras, Co-Chair
Assistant Professor Connor Mayer
Assistant Professor Xin Xie

2024

DEDICATION

To Ima, Aba, Amnon, Yonatan, and Didi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

To my advisors, Lisa and Greg, thank you from the bottom of my heart for your support these last years – starting in the pre-pandemic world when every one of those ridiculous horses wouldn't jump the fence, all the way to when *everything wasn't perfect*, according to NPR, in the post-pandemic world. If research is measuring a mile in inches (which sometimes makes me want to toss both the ruler and the compass) you showed me how to take steady steps in the right direction, and I'm much richer for this knowledge. I will really miss our meetings, which were always fun and exciting, but I'm grateful to think that over time, you became not just my mentors but also my friends.

Connor Mayer and Xin Xie, my committee members, thank you so much for your support and your feedback and advice, particularly with respect to the prosody analysis in this dissertation.

Stefanie Wulff, I'm indebted to you for your guidance regarding the corpus analyses. Working with you inspired me to think about the question 'What do we do in everyday conversation?', which became the focus of this dissertation.

Ambilab – Ky-Vinh Mai, Jordan Jin, Aadya Sharma, and Meadow Quibodeaux – your work on the ambiguity corpus was invaluable. The corpus would not be anywhere near as comprehensive without your hard work in creating it, and all the hours we spent checking it wouldn't have been anywhere as fun without your team spirit. Honestly it's one of the biggest projects I've ever embarked on. Ky, thank you especially for your creative and effective programming, which was the backbone of the creation of the NPR corpus. Jordan, I can't thank you enough for your attentive work on the search functions and final versions of the NPR corpus.

I am grateful to the National Science Foundation for providing the financial support necessary for my research through the Doctoral Dissertation Research Improvement Award in Linguistics, without which this work would not be possible.

I would also like to thank my cohort and lab mates in the Language Science and Cognitive Science departments for their camaraderie, friendship and feedback. To Minkyu, Hongju, and my dear roommate Lingyu, my mirrors – even five years ago, sitting around on the floor of an unfurnished dorm, I knew that we will all make it (relatively unscathed too, as long as we're not splitting grey hairs).

To my family, you are my anchors and who I am. It's hard to find other words. Thank you for always supporting me.

VITA

Noa Attali

EDUCATION

Doctor of Philosophy in Language Science **2024**
University of California, Irvine *Irvine, CA*

Bachelor of Arts in Cognitive Science **2019**
The Honors College of Rutgers University *New Brunswick, NJ*

Bachelor of Arts in English **2019**
The Honors College of Rutgers University *New Brunswick, NJ*

ABSTRACT OF THE DISSERTATION

Disambiguating Information in Speech and Context

By

Noa Attali

Doctor of Philosophy in Language Science

University of California, Irvine, 2024

Professor Lisa Pearl, Co-Chair
Associate Professor Gregory Scontras, Co-Chair

In this dissertation, I investigate how people navigate ambiguity in everyday speech, with a focus on quantifier-negation sentences. Combining corpus analysis, behavioral experiments, and computational modeling in the Rational Speech Act framework, I explore preferred interpretations of quantifier-negation and examine the contexts and prosodies that shape these interpretations. In particular, to address a knowledge gap on naturalistic ambiguity use, I analyze *every*-negation uses in two large-scale corpora. I find that certain expectations about the world, made salient in context, predict interpretations; in general, listeners try to align their interpretations with what they already know about the world. I also find that certain pausing and pitch patterns predict interpretations, including the use of pitch to emphasize the quantifier. Altogether, despite the inherent variability in context and prosody, by integrating different methodologies, it is possible to identify specific ways in which context and prosody shape meaning.

Chapter 1

Introduction

How do we navigate ambiguity in everyday conversation? A widespread potential for ambiguity characterizes natural language and leads to many questions about when and why a certain interpretation of an ambiguous expression is preferred. In the series of studies described here, I take English quantifier-negation scope ambiguity (e.g., *Every vote doesn't count*) as a test case of ambiguity. This kind of construction is a good case study of a relatively ambiguous, difficult linguistic form which nevertheless, when spoken in naturalistic contexts, yields confident interpretations. In other words, context and prosody can go a long way towards explaining interpretation preferences of an otherwise ambiguous construction. I focus on understanding how specific aspects of the preceding linguistic context – a source of preceding disambiguating information – and prosody – a source of simultaneous disambiguating information – facilitate interpretations, by regularly constraining plausible or possible meaning.

1.1 Ambiguity as a puzzle and the potentially disambiguating role of context and prosody

Natural languages are "massively ambiguous" (Wasow et al., 2005, pp. 1) across linguistic levels: they are full of expressions with a form that maps to multiple potential meanings. This literature review will focus on sentence-level ambiguity, in particular ambiguity of the kind exemplified in (1).

(1) Every vote doesn't count.

a. No vote counts. *Surface scope*: $\forall x[\text{vote}(x) \rightarrow \neg\text{count}(x)]$ (every > n't)

b. Not all votes count. *Inverse scope*: $\neg\forall x[\text{vote}(x) \rightarrow \text{count}(x)]$ (n't > every)

There are two expressions in (1), *every* and negation (*n't*), which are logical operators in the sense that they perform some action on the rest of the sentence (Szabolcsi, 2011). Specifically, understanding the sentence involves interpreting which operator takes scope over the other: the utterance could have surface scope such that the first operator takes scope over the second (1a) or the utterance could have inverse scope such that the second operator takes scope over the first (1b).

A major question for understanding human communication is, is this kind of ambiguity a bug or a feature of communication? Ambiguity seems like a bug in that it should greatly challenge successful communication, or at least slow it down (Chomsky et al., 2002; Piantadosi et al., 2012; Frazier and Fodor, 1978; MacDonald et al., 1994). For example, ambiguity might introduce greater potential for comprehension errors (Chomsky et al., 2002), greater comprehension effort and computational difficulty (Piantadosi et al., 2012), and increased online processing difficulty if a listener has to reanalyze a sentence because their initial parse was incorrect (as in garden path effects; e.g., Frazier and Fodor, 1978; MacDonald et al., 1994).

(That being said, structural ambiguity does not necessarily introduce errors or slowdowns to online processing (e.g., Grant et al., 2020).)

When it comes to the ambiguity of interest here, involving logical operators, the challenge to successful communication is evident as a potential combinatorial explosion of interpretations. Giving consideration to them all is cognitively and computationally implausible. For example, (2), which has eight logical operators, is argued to have hundreds of thousands of valid (grammatical) interpretations if every possible scope interpretation were to be considered (Poesio, 1996).[1]

(2) A politician can fool most voters on most issues most of the time, but no politician can fool all voters on every single issue all of the time.

Yet listeners effortlessly interpret sentences like (2) (Poesio, 1996). The general observation is that disambiguation tends to be effortless, which makes sense given human online processing constraints such as working memory limitations (Saba and Corriveau, 2001). Specifically, we rely on fast inference-based reasoning to disambiguate what we hear: Saba (1999) argues that rather than considering all possible interpretations, "readers initially select the most plausible reading given an appropriately defined context, with the caveat that such an inference could be retracted in the future" (p. 8).

Ambiguity then could be a desirable property rather than a bug of natural languages. An utterance will be unambiguous (enough) for successful communication given the context in

[1]More specifically characterizing the combinatorial explosion, computational approaches to scope ambiguity resolution often describe $n!$ possible interpretations for a surface structure with n logical operators (e.g., in Grosz et al., 1987) (where those operators are often in effect limited to quantified nouns, because these are often the focus of interest in this literature (Saba, 1999)), although they may then limit the final number of potential permutations according to different structural or logical constraints (e.g., in Hobbs and Shieber, 1987; Fox, 1995; Park, 1995). For example, Hobbs and Shieber (1987) argue that (i) has only 42 valid scope interpretations.

(i) Some salesman of every department in most companies saw a few samples of each product.

which it's used (Achimova et al., 2022). That is, since sentences tend to be used in contexts that are already informative as to the speaker's intended meaning, a totally unambiguous language would be inefficient because it would be redundant, to some extent, with the context (Piantadosi et al., 2012). Givón (2014) argues that ambiguity is the direct result of efficient data compression and storage in natural languages.

Although less often mentioned in the ambiguity literature, prosody works very similarly to context in providing information that obviates the need for the sentence to be unambiguous on its own. In spoken language, suprasegmental acoustics provide a wealth of information, simultaneous with the potentially-ambiguous sentence, about the speaker's intended meaning (for a recent review, see Wagner and Watson, 2010).

On a broader view, the lexical string of any language expression on its own has an under-determined meaning, and listeners have to perform a series of inferences given context and prosody in order to understand what the speaker meant to say (Grice, 1975; Sperber and Wilson, 1986). Taking this broader view which assumes that ambiguity is a feature of language, the question is, what information is signalled in context and prosody that listeners can use to arrive at the intended interpretation of sentence-level ambiguity?

For sentence-level ambiguity, as exemplified by (1), I use scope ambiguity – specifically, utterances with a quantified subject and verb negation, hereafter quantifier-negation scope ambiguity – as the key case study of potential ambiguity use. I also consider the literature on other forms of scope ambiguity, especially any form containing quantifiers or negation, wherever it seems that the broader predictions or approach would apply similarly to quantifier-negation scope.

The especially salient aspect of context that might disambiguate scope ambiguity is the linguistic discourse immediately preceding the utterance. Based on the literature, I will argue that one disambiguating factor, which may receive expression in the preceding linguistic

discourse, is interlocutor expectation about the state of the world (e.g., Wason, 1972; Gualmini, 2004; Scontras and Pearl, 2021). World expectations influence in turn the relative plausibility of competing interpretations. The preferred interpretation is then the interpretation that is more likely to be true. The prior literature does not often discuss context as a disambiguating factor in exactly these terms (with a few exceptions, such as Scontras and Pearl, 2021), but I will show that this characterization of context could account for some interpretation preferences.

The aspects of prosody that might matter are the relative boundaries (perceptible phrases or sub-groupings of linguistic constituents) and patterns of relative prosodic prominence, as marked primarily by fundamental frequency (F0) and other phonetic cues like duration and intensity (Wagner and Watson, 2010). Specifically, the scope literature suggests that both boundaries (e.g., pausing, among other acoustic cues) and prominence (e.g., F0 rises and falls) are potentially disambiguating factors. First, in certain cases, the sub-groupings created by prosodic phrasing map transparently onto the constituents of a scope interpretation (e.g., Hirschberg and Avesani, 1997; Baltazani, 2000; Koizumi, 2009), with the idea being that prosodic phrases can to some extent divide an utterance into meaningful 'chunks' of information (Bolinger, 1989). Second, non-default interpretations are predicted to associate with a relatively more complex, fall-rise pitch contour (e.g., Jackendoff, 1972; Ladd, 1980; Ward and Hirschberg, 1985). Contra context, there is a more extensive literature on the role of the fall-rise contour for scope ambiguity involving negation, but the empirical evidence is in fact more mixed about whether prosody disambiguates (Syrett et al., 2012, 2014).

Overall, scopally ambiguous utterances are a good case study because they allow context, prosody, and their interactions to be investigated in fruitful ways. Investigating the disambiguating effect of prosody and context for scopally ambiguous utterances like (1) "lead[s] us through quite a global conspiracy of syntax, intonation, (lexical) semantics and pragmatics" (Büring, 1997, pp. 193), including conflicting empirical characterization and a high degree of

Figure 1.1: Prosody overrides or counters the effect of context (left), or prosody is redundant with context (right).

variation in interpretations and prosodies.

Further, considering both context and prosody in the same study contributes a new perspective on prosody. Specifically, an empirical characterization of the roles of context and prosody for interpretations can help clarify if there is any separate impact of prosody as compared to linguistic context. Can prosody convey information not otherwise conveyed by the preceding context, as illustrated by the left image in Figure 1.1? Or is prosody the reflection of context (i.e., prosody is redundant with context, not adding any additional information), as illustrated by the right image in Figure 1.1? Further, how consistent and reliable is the relationship between prosody and interpretations? Answering these questions would shed light on the extent to which the function of prosody is conventionalized (and not specific to individual speakers or contexts), as part of the linguistic/communicative knowledge of speakers and listeners.

Finally, both prosody and context have been the subject of a research approach that highlights how difficult it is to clarify a disambiguating role for either factor. For example, in a large-scale corpus study of *all*-negation, Neukom-Hermann (2016) writes that context has a clear disambiguating role, but that context-dependent disambiguating rules or tendencies could only be articulated to a very limited degree.

> The relevance of the context for the disambiguation of *all...not* constructions thus manifests itself on all linguistic levels (lexical, syntactic, semantic, pragmatic), and comprises knowledge of the situation, the culture, the world or a specialist

> field. Usually the complex interplay of all these factors is responsible for the disambiguation of a sentence. It is therefore only to a very limited degree that rules (or rather tendencies) can be formulated that would predict particular readings for particular structures. (pp. 130)

Similarly, in a corpus-based study on the pragmatics of intonation (prosody), Ward and Hirschberg (1985) cite a claim by Cutler (1977) that it should be very difficult to extract a consistent disambiguating meaning of an intonational contour. Although Ward and Hirschberg (1985) disagree with this claim, they in fact argue that the role of intonation is weaker and more incidental than past studies had argued, and that context is the most important disambiguating factor for *all*-negation (without further clarifying how context disambiguates).

> ... even the non-specific, non-referential effects exercised by intonation contours can be shown to be context-dependent to such a degree that the attempt to extract from them an element of commonality valid in all contexts must be reckoned a futile endeavor (Cutler, 1977, p. 106)

In this dissertation, although it's true that both prosody and context are variable factors with multiple linguistic functions in addition to their disambiguating one, I emphasize that both have a disambiguating function and it can be specified to a greater degree than past studies have done.

1.2 Looking ahead

This dissertation is structured as follows. Chapter 2 introduces the case study of scope ambiguity in greater detail, discussing the relevant literature on interpretations, with a focus on the evident variation of interpretations. My goal is to highlight that many characteristics

of both the form and contexts of these ambiguities likely work together to account for interpretation preferences.

I then turn to how plausibility, a contextual factor, may account for preferences. Chapter 3 proposes a general model of scope disambiguation, beginning by introducing computational cognitive modeling as a method and how it can help to navigate the variability underlying scope interpretations. I describe in detail a computational model of scope resolution from the literature, discussing how the model implements the role of context as world expectations, as part of a maximally explicit hypothesis about how interlocutors reason towards preferred interpretations of quantifier-negation. I then extend this model so that it makes predictions about preferred interpretations for *every*, *some*, and *no*-negation. That is, the model makes quantitative predictions about preferred interpretations (e.g., that a certain quantifier-negation utterance should be interpreted in one way 75% of the time), which I then evaluate against crowd-sourced judgments from a behavioral experiment. The behavioral data validate the model with little parameter fitting, suggesting support for the ambiguity resolution hypothesis which is expressed by the model.

I then focus on understanding a key aspect of this disambiguation hypothesis: world expectations and how they mediate the role of plausibility in interpretation preferences. Chapter 4 and 5 focus on gathering further data to validate this specific aspect of the role of context in the model.

Chapter 4 describes creating a large-scale corpus of ambiguity from spontaneous speech. Naturalistic conversation can offer great insight into these questions about language use, and the existing literature highlights some knowledge gaps and open questions concerning the attested, naturalistic contexts and prosodies of scope ambiguity. Thus, I mined cases of *every*-negation ambiguity from corpora and archives of radio and TV interview speech from the past twenty years. To gather interpretation preferences, native English speakers were asked in online experiments to provide their best guess at the speaker's intended meaning of

these naturalistic cases.

Chapter 5 takes the resultant ambiguity corpus and tests the central hypothesis from the model in Chapter 3, that a certain kind of world expectation should account for inverse scope preference of *every*-negation. I first discuss the model hypothesis about world expectations in general terms and in light of past studies. I then explore different ways to measure the world expectations of interest in the contexts of the ambiguity corpus, including through a behavioral study to gather judgments about these contexts. Through all these measures, I find that the specific world expectations predict scope preferences in the expected direction.

Chapter 6 turns to the role of prosody in the ambiguity corpus. I first review the literature on the role of prosody for scope interpretations, particularly quantifier-negation, with the goal of identifying some specific predictions for prosodies of surface or inverse scope. I then test these predictions on the prosody of the ambiguity corpus, finding some evidence that pausing before the negation predicts surface scope, a final rise in pitch predicts inverse scope, and that the pitch contour over the quantifier and negation generally predicts scope.

Chapter 7 summarizes the findings with respect to the context and prosody of quantifier-negation, particularly *every*-negation, to create a picture of ambiguity resolution that is grounded in naturalistic behavior and crowd-sourced judgments from native English speakers.

Chapter 2

Background

This chapter reviews the literature on scope ambiguity, with an emphasis on defining the kind of ambiguity that this dissertation focuses on, understanding the interpretation variation that past studies have found for this kind of ambiguity, and the different factors that have been proposed to account for interpretations.

2.1 Case study of ambiguity

This section introduces the case study of ambiguity in greater detail, defining scope ambiguity and some of the main approaches from the semantics literature for how to characterize different scope interpretations, representation at the level of logical form, entailment relations between interpretations, broader examples of potential scope ambiguity, and limitations of definition (Section 2.1.1). Section 2.1.2 defines and motivates quantifier-negation as the case study more specifically.

2.1.1 Scope ambiguity

I take sentences like (1) as a case study of ambiguity:

(1) Every marble isn't red.

(1) is potentially ambiguous in the sense that, under one of its interpretations, it could be truthfully uttered both about a situation where none of the marbles are red (e.g., the right image in Figure 2.1) and a situation where it's merely true that not all of the marbles are red (e.g., the left image in Figure 2.1). However, under its other interpretation, it would be false rather than true of the left image in Figure 2.1.

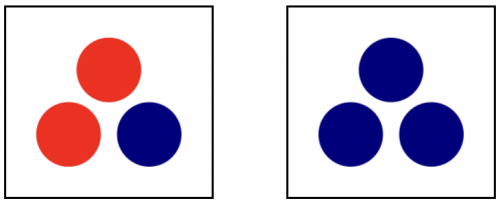


Figure 2.1: Illustration of scope ambiguity: two scenarios compatible with the meaning of *Every marble isn't red.*

This potential for ambiguity has to do with how the semantics of a linguistic expression (e.g., a phrase) is like a logical formula successively built up of smaller formulae that combine in a particular order. Each building-block formula is a constituent of the expression, and the order in which they combine reflects the overall constituent structure. If there are operators that apply to particular constituents (e.g., like *every* and *not* in (1)), the scope of each operator, in natural language, includes the constituents in this semantic structure to which it applies (Szabolcsi, 2011). For example in (1), *not* may apply to the constituent *is red*; another potential constituent for *not* is *every marble is red*. The problem for natural language interpretation, unlike in logic, is that there are no parentheses and brackets in the overt form of the expression to signal the relative scope of each linguistic operator. The relative semantic scope of operators does not necessarily coincide with their overt syntactic domain.

For (1) as repeated in (2), which has the two operators *every* and negation (*n't*), one potential order of operations is consistent with the overt order of the operators in the sentence, so that *every* takes scope over negation, resulting in the (2a) interpretation: it is the case for each of the marbles that each one isn't red. The other potential order of operations is inverse to the overt order, so that negation takes scope over *every*, resulting in the (2b) interpretation: it's not the case that each of the marbles is red; none or some are red.

(2) Every marble isn't red.

a. No marble is red. *Surface scope* (every > n't): $\forall x[\text{marble}(x) \rightarrow \neg\text{red}(x)]$

b. Not all marbles are red. *Inverse scope* (n't > every): $\neg\forall x[\text{marble}(x) \rightarrow \text{red}(x)]$

Scope terms. Surface scope is also referred to as isomorphic (e.g., Musolino, 1999), direct (e.g., Ruys and Winter, 2011), or high or wide scope of the first operator (e.g., Szabolcsi, 2011). The corresponding terms for inverse scope are nonisomorphic, indirect, or narrow scope of the first operator.

Ambiguity of logical form given the overt linguistic form. Where exactly does the ambiguity lie? Scope ambiguity is often characterized as an ambiguity in the overt linguistic form (i.e., in the surface string) which is disambiguated at the level of logical form (LF): a level of unambiguous representation distinct from surface form, preceding production and corresponding to the speaker's intended meaning (May, 1977; May and Keyser, 1985). This logical form directly represents those aspects of syntactic form which are relevant to interpretation. In other words, the difference between the surface and inverse scope interpretations of an utterance is a difference in their logical forms.

For example for (2), its two interpretations in LF would involve a difference in whether negation is lower or higher than the quantifier. The left image in Figure 2.2 shows the surface

scope interpretation of (2), where *every* takes higher scope over negation. The right image in Figure 2.2 shows the inverse scope interpretation of (2), where negation takes higher scope over *every*.

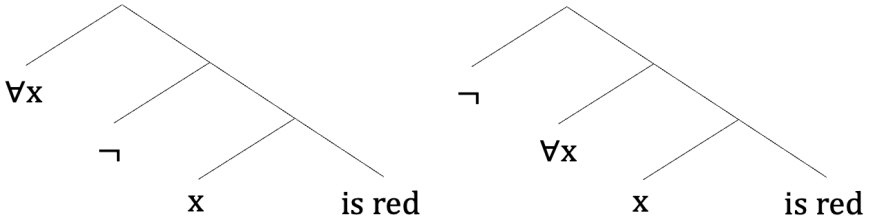


Figure 2.2: Logical form (LF) representations for the two interpretations of *Every marble isn't red*: surface scope (left), or inverse scope (right).

Entailment relations between scope interpretations. For quantifier-negation scope ambiguity, the two interpretations can be logically related to each other (such that there are certain situations that make *both* interpretations true). This becomes relevant for the design of experimental stimuli that test interpretation preferences. It's also relevant for questions about whether speakers always intend one interpretation or the other, or whether speakers may underspecify their intended scope.

For *every*-negation (as well as *all*-negation), as in (2), the surface scope interpretation entails the inverse scope one: if it's the case that no marble is red, then necessarily not all of the marbles are red. However, the inverse scope interpretation does not entail the surface scope one: if it's the case that not all the marbles are red, it may not be true that none are red, because it may be that some are red. Figure 2.3 illustrates how overall for *every*-negation, because of this asymmetrical entailment relationship, either both interpretations are true of a situation at once (when it's the case that *none*), only the inverse scope interpretation is true (when it's the case that *some but not all*), or neither are true (when it's the case that *all*). With other types of scopally ambiguous utterances, the entailment relations might be different, depending on what the two potential interpretations are.

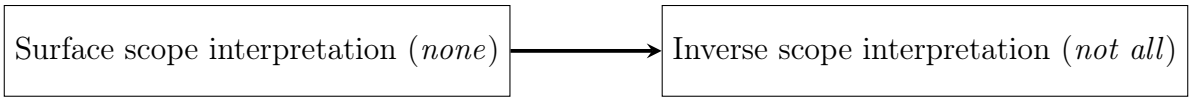


Figure 2.3: Entailment relation between the two scope interpretations of *every*-negation. If it is known that the surface scope interpretation is true, then the inverse scope interpretation must be true. However, if it is known that the inverse scope interpretation is true, it is not known whether the surface scope interpretation is true.

The fact that the surface scope interpretation of *every/all*-negation entails its inverse scope interpretation means that, perhaps for the purpose of efficient communication, speakers could underspecify their intended scope – they might be willing for the listener to arrive at either scope interpretation, when the world state that they intend to describe would be well-described by either interpretation (Neukom-Hermann, 2016). For example, imagine a context in which an interlocutor makes a bet that every marble they will draw from a bag will be red. In other words, in such a context, there's a salient question *Is every marble red?* If another interlocutor (who knows the contents of the bag) knows that the answer to this salient question is *no*, they could say *Every marble isn't red* with either scope interpretation (*none* or *not all*) in order to meet the communicative goal of conveying that the answer to the salient question is *no*.

Broader examples of potential scope ambiguity. For the purpose of further illustrating the phenomenon in question, and to show how scope ambiguity can arise in a broad variety of sentences, the following examples show scope ambiguity in different configurations of scope-taking operators (with the operators bolded). Notably, on first reading many of these sentences without context, they may not appear ambiguous.

(3) **Who** did **every**one meet?

a. Who is the person who met everyone?

b. For everyone, who did they meet? (Each person may have met different people.)

(4) **No** chimp ate **two** of the apples.

a. There are no chimps who ate two apples.

b. There are two apples which were not eaten by chimps. (There may be a chimp that ate two apples.)

(5) **Some** people are **always** right.

a. There exist some people who are always right.

b. It is always the case that some people are right.

(6) **Many** people **seem** to agree on that issue.

a. There are many people who seem to agree on that issue.

b. It seems to be the case that many people agree on that issue.

Limitations in defining where scope ambiguity could arise. It isn't always clear when scope ambiguity should be theoretically possible. Defining scope ambiguity depends on 1) whether and how syntactic structure determines semantic scope, and 2) what a scope-taking operator is. Based on key ideas argued for by Reinhart (1983) for example, 1) the scope of a linguistic operator corresponds to its domain in some syntactic representation that the operator is part of, even if not the overt syntactic form or before movement (Szabolcsi, 2011). But a challenge for an account of how structure determines scope is that not all expressions containing two scope-taking operators are actually ambiguous (in any context). For example, according to linguist intuitions in May and Keyser (1985), (7) is ambiguous between (7a) and (7b) but (8) can only be interpreted as (8a). So a structural account of scope ambiguity should describe semantic/syntactic representations and operations that allow *every* to take both wide and narrow scope in (7) but allow *every* to only take narrow scope in (8) (Kiss and Pafel, 2017). This dissertation doesn't focus on answering these questions of definition, but instead takes as a case study of ambiguity a construction which seems to clearly create

the potential for scope ambiguity.

(7) What did everyone buy for Max?

a. What is the single thing such that everyone bought it for Max?
Surface scope (what > every)

b. For each person, what are the things that (s)he bought for Max?
Inverse scope (every > what)

(8) Who bought everything for Max?

a. Who is the single person who bought everything for Max? *Surface scope* (who > every)

b. For each thing, who who bought them all for Max?
Inverse scope (every > who)

The second problem for defining scope ambiguity is settling on the expressions that take scope in the first place (Szabolcsi, 2011). It seems clear that explicit quantifiers (more specifically, quantified noun phrases (e.g., in Ruys and Winter, 2011) or quantificational determiner phrases (e.g., in Szabolcsi, 2011) can take scope in the exact sense of formal semantic/syntactic accounts of scope ambiguity: for example *each*, *every*, *all*, *most*, *several*, *some*, *few*, *no* (Kiss and Pafel, 2017). Negation and *wh*-phrases, as used in the examples above, also take scope (Kiss and Pafel, 2017); also some adverbials like *always VP* (Ruys and Winter, 2011).

But, some expressions that appear to participate in "scope(-like) phenomena" can't or shouldn't be coherently treated as scope-taking in formal accounts (Szabolcsi, 2011): for example, bare plurals (e.g., *dogs* in *Dogs barked everywhere* seems on a first analysis like it might contribute existential quantification, but a closer analysis would suggest that this existential import comes from the predicate or somewhere else). Additionally, some utterances

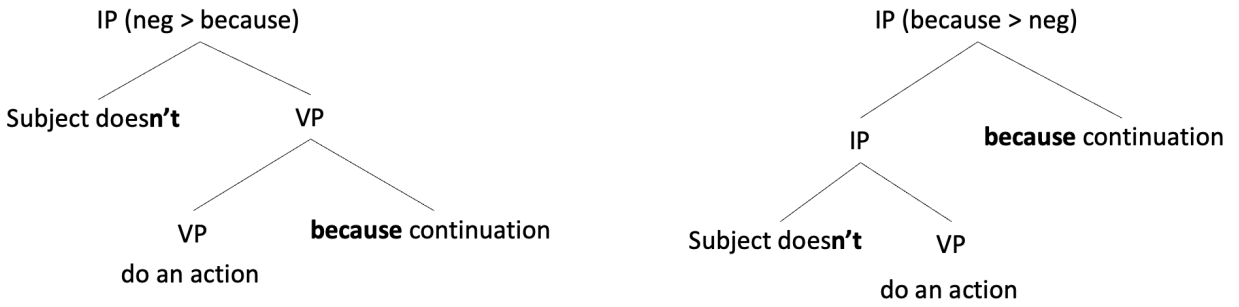


Figure 2.4: Possible analysis of the ambiguity of not-because utterances like *He's not watching TV because he's bored* (10). Local attachment of the *because*-clause corresponds to the surface scope interpretation (left image). High attachment of the *because*-clause corresponds to the inverse scope interpretation (right image).

exhibit an ambiguity that could be described in terms of order of operations, but which should potentially be treated as different from the ambiguity in (2). In the case of example (9) – using *the* and *every* – the ambiguity is arguably lexical, not structural (as in, an ambiguity of whether *the* should really be considered to involve quantification here). In the case of (10) – using negation and *because* – while the ambiguity is structural, it is specifically one of different adjunction sites of the adverbial clause (e.g., Johnston, 1994). Figure 2.4 shows a way that the two interpretations of a *not-because* utterance can be analyzed as corresponding to two adjunction sites.

(9) ... the father of every family ...

a. There's a single father of each family. *Surface scope* (the > every)

b. For each family, there's a different father. *Inverse scope* (every > the)

(10) He's not watching TV because he's bored.

a. There is TV watching, but its cause isn't boredom. *Surface scope* (not > because)

b. Due to his boredom, there is no TV watching. *Inverse scope* (because > not)

Constructions like negation-*because* are especially interesting in comparison to many other

scopally ambiguous constructions, because the truth conditions of its two interpretations are entirely distinct – there is no situation that verifies both interpretations. That is, neither interpretation entails the other, and an utterance would be incoherent if it were interpreted with both interpretations. For example, if both (10a) and (10b) were taken to be true, then there both is and isn't TV watching, and the subject both is and isn't bored. Therefore, for example, it is harder to imagine how speakers might produce an underspecified meaning for negation-*because* as compared with *all*-negation; it is hard to imagine a scenario in which a speaker would have a communicative goal that would be met by the listener arriving at either scope interpretation of negation-*because*.

2.1.2 Quantifier-negation scope ambiguity

This literature review focuses on interpretations of scopally-ambiguous configurations in English such as (2), where the overt string contains a quantified subject preceding negation in the same clause – quantifier-negation utterances. Some common examples in the literature on quantifier-negation are (11) and (12) below.

(11) Every horse didn't jump over the fence. (Musolino, 1999)

a. No horses jumped. *Surface scope* (every > n't): $\forall x[\text{horse}(x) \rightarrow \neg\text{jump}(x)]$

b. Not all horses jumped. *Inverse scope* (n't > every): $\neg\forall x[\text{horse}(x) \rightarrow \text{jump}(x)]$

(12) All the men didn't go. (Jackendoff, 1972)

a. No man went. *Surface scope* (all > n't): $\forall x[\text{man}(x) \rightarrow \neg\text{go}(x)]$

b. Not all the men went. *Inverse scope* (n't > all): $\neg\forall x[\text{man}(x) \rightarrow \text{go}(x)]$

All-negation like (12) is also compatible with a third, collective scope interpretation (which is sometimes but not always taken into account when *all*-negation is studied), in contrast to

its two distributive interpretations, as illustrated in (13).

(13) All the bills don't amount to $50. (Neukom-Hermann, 2016)

a. Not a single bill amounts to $50. *Surface scope* (distributive all > n't)

b. The sum doesn't amount to $50. *Surface scope* (collective all > n't)

c. Not all the bills amount to $50. *Inverse scope* (n't > distributive all)

The distinction between (13a) and (13c), compared to (13b), relates to an ambiguity in the meaning of *all* as distributive or collective. When *all* is taken to be distributive, something is said about every individual bill; when it is taken to be collective, something is said about the totality of bills (Neukom-Hermann, 2016).

In these examples (11), (12), and (13), the quantifier is universal, which reflects the general past focus on universal quantifiers. However, the same ambiguity could be possible with other quantifiers (e.g., *some*, *no*), modulo claims about how the semantics of certain quantifiers interact with the quantifier-negation construction to yield scope.

(14) Something will not happen.

a. There's something that will not happen.

Surface scope (some > not): $\exists x[\text{thing}(x) \,\&\, \neg\text{happen}(x)]$

b. There is nothing that will happen.

Inverse scope (not > some): $\neg\exists x[\text{thing}(x) \,\&\, \text{happen}(x)]$

(15) No citizen can't vote.

a. There isn't a citizen that can't vote.

Surface scope (no > n't): $\neg\exists x[\text{citizen}(x) \,\&\, \neg\text{vote}(x)]$

b. There aren't zero citizens that can vote.

Inverse scope (n't > no): $\neg\nexists$x[citizen(x) & vote(x)]

2.2 Variation and Context

Let's turn to how context can account for interpretations, taking as a starting point the fact that there's wide variation in interpretations given a variety of structural and pragmatic factors, so it's hard to shed light on a single factor in the absence of the others. I first review some evidence of interpretation variation (Section 2.2.1). Different uses of the same quantifier-negation construction receive different interpretations in corpora (Section 2.2.1.1) and experiments (Section 2.2.1.2). Furthermore, there is variation in interpretations of sentences with different quantifiers (Section 2.2.1.3).

Corpus and behavioral studies of quantifier-negation focus on *all*-negation and *every*-negation, showing both a preference for inverse scope yet general variation in preferred interpretations, even of similar or identical constructions (Carden, 1970; Heringer, 1970; Carden, 1973; Ioup, 1975; Beghelli and Stowell, 1997; Musolino, 1999; Musolino et al., 2000; Neukom-Hermann, 2016). This interpretation variation is likely due to differences in the specific form (e.g., the lexical content) and context (e.g., the preceding linguistic context) of the quantifier-negation use. These varied forms and contexts affect a range of factors that influence interpretation preference; past literature highlights semantic-syntactic factors (e.g., how the formal properties of quantifiers constrain calculations of grammatical scope) and pragmatic factors (i.e., factors that influence the felicity of an utterance's use or listeners' reasoning about which interpretation is more likely to have been intended).

2.2.1 Variation

Corpus studies of quantifier-negation in English focus on *all*-negation or *every*-negation, and all find that these constructions are 1) rare but attested, 2) used most often with inverse scope, but 3) with wide variation in the form, context, and preferred interpretation.

2.2.1.1 Inverse scope preference for *all/every*-negation in corpora, with variation across similar sentences

The most comprehensive corpus study, described in a dissertation by Neukom-Hermann (2016), finds that 56% of 490 *all*-negation uses from the British National Corpus (which is comprised of 100 million words, primarily written) are intended with inverse scope and only 17% with surface scope, as judged by the author. The majority of the remainder are judged to have a collective interpretation (the third type of interpretation which is available to *all*-negation but not to *every*-negation, where the quantifier still takes surface scope over the negation). Further, a few cases (about 2%) are "truly ambiguous ... the available context does not help to disambiguate the meaning" (pp. 78). Relating to the possibility of efficiently underspecifying meaning, Neukom-Hermann (2016) also points out a few cases of what she calls underspecification, where "the potential ambiguity between possible readings does not necessarily have to be resolved because the readings do not differ in a way that is important in the particular context" (p. 82).

The following examples selected from Neukom-Hermann (2016) illustrate these four different types of *all*-negation as attested in the corpus. I note further some subdivisions that she points out in how these items are used.

(16) *Surface scope all > neg*: The facts are the facts, and I am compelled to record them with a plainness of detail which in the end offers the only means of extending that

small degree of compassion, or perhaps even understanding, **which all men in whatever circumstance or however degraded should not be denied**. (No men in whatever circumstance, or however degraded, should be denied that small degree of compassion.)

(17) *Collective interpretation all > neg*: If **all that money we gave to Band Aid didn't do the trick**, it must be because there are just too many of them. (It is the case for the total sum of the money given to Band Aid, that that total sum didn't do the trick.)

26% of the 134 cases of collective *all*-negation were also judged to be *formulaic* by the first author, that is, having the form *all NP in the world V not* or *(as if) all this wasn't enough.* The limitation of this analysis, though, is that it is not clear how to *a priori* decide the criteria for formulaic expressions. (One possibility could be to determine that an expression is formulaic if its exact lexical form is significantly more frequent than expected in a corpus of quantifier-negation.)

(18) *Formulaic collective interpretation all > neg*: As if all this were not enough, schools have started managing their own financial affairs.

(19) *Inverse scope neg > all*: The value of doubt is that it can be used to detect error. We live in a fallen world. **All is not true**, so not everything should be believed; some things ought to be doubted. (Not all is true.)

56% of the 255 cases judged to have inverse scope were also judged to be *formulaic*, that is, having the form *all is not lost/well/perfect/good/gloom (and doom).*

(20) *Formulaic inverse scope neg > all*: Sock Shop admitted earlier this year that all was

not well with its American outlets.

An interesting issue is that if formulaic constructions were to be discounted from analysis, then the three interpretations in her analysis would be roughly of the same prevalence, though inverse still slightly more common (38%) and surface less common (27%).

(21) *Truly ambiguous*: So, I phoned up Joe, and Joe says **all the results weren't in** because that's the kind of ordeal next week er er next Friday, tomorrow.

a. None of the results were in.

b. Not all of the results were in.

The example of true ambiguity is one of the few cases found in the BNC corpus which comes from spoken rather than written language. Also, the reason it's difficult to resolve this ambiguity could lie in a lack of helpful context.

In the following underspecification case, I quote the exact phrasing of the interpretation paraphrases as written by Neukom-Hermann (2016):

(22) *Underspecification*: The CSA 1985 is a complex piece of statutory craftsmanship, and **all of its provisions do not directly concern us**.

a. Surface scope: None of the provisions concern us directly, but all or at least some of them concern us indirectly.

b. Inverse scope: Not all provisions concern us directly, but some of them do concern us directly.

Neukom-Hermann (2016) argues that in the underspecification case, the use of the adverb after the negation minimizes the difference in meaning (though doesn't do away with it entirely) between the surface and inverse scope interpretation. The adverb *directly* prevents

the proposition in the surface scope interpretation from being completely negative – instead of asserting that *none of the provisions concern us*, the speaker merely asserts that *none of the provisions concern us directly*.

Tentatively, underspecification suggests that resolving the ambiguity is not the only strategy for navigating the ambiguity of the construction: another strategy is for a speaker to make resolving the ambiguity unnecessary. If the speaker's communicative goal is underspecified for whether they intend the surface or inverse scope interpretation, then the listener would understand the content that the speaker considers relevant regardless of the scope interpretation that the listener arrives at. In other words, the communication was good enough.

The main findings of an inverse scope preference, in combination with variation, are replicated by the other smaller corpus studies of English. Musolino et al. (2000) cite in a footnote that 28 out of 30 *every*-negation uses collected from English spontaneous speech were intended with inverse scope (the method of collecting scope judgments is unclear). Similarly, in an unpublished manuscript by Taglicht as cited by Neukom-Hermann (2016), 52% of an *all*-negation corpus of 23 items were inverse, 13% surface, and 35% collective (Taglicht, ND).

The limitations of these studies are that two have a small sample size and one is based primarily on written language and relied on the primary researcher to determine the intended scope interpretation. It's worthwhile to return to these corpus studies with larger sample sizes and interpretation preferences collected from a random sample of English speakers, in order to further bridge the understanding of scope ambiguity that has been gained in the lab with an understanding of how ambiguity is used everyday. I do this in Chapter 4, although I mined data from different corpora than the ones that were used in these studies.

2.2.1.2 Inverse scope preference for *all/every*-negation in experiments, with variation across similar sentences

Open-ended interviews for adults' interpretations have found a general preference for inverse scope (in trivial contrast to what Carden calls the "all-readings-are-equal hypothesis", where both interpretations would be equally preferred) (Carden, 1973). These are findings for interpretations of sentences with universally-quantified subjects, which were elicited through linguistic interviews in which the sentences were spoken to adults without context and the adults' interpretations were then probed by the interviewer. Musolino et al. (2000) (N=15) found that for the *every*-negation sentence (11) (repeated below as (23)), 80% of participants arrived at inverse scope, 13% arrived at surface scope, and 7% said it was ambiguous. Likewise, Carden (1970, 1973) found a general preference for inverse scope interpretations of *all-n't*. In interviews (N=40), for (24), 40% said that only the inverse scope interpretation was possible, 50% said that both were possible but that they favored inverse scope, and 10% said that only surface scope was possible.

(23) Every horse didn't jump over the fence. (Musolino, 1999)

a. No horses jumped. *Surface scope* (every > n't)

b. Not all the horses jumped. *Inverse scope* (n't > every)

(24) All the boys didn't arrive. (Carden, 1973)

a. No boy arrived. *Surface scope* (all > n't)

b. **Not all the boys arrived.** *Inverse scope* (n't > all)

Musolino (1999) found that adults have no trouble accessing the inverse scope interpretation of *every*-negation. Participants (N=20) always endorsed *every*-negation statements such as (11) (*Every horse didn't jump over the fence*) as a description of a scenario of which the

inverse scope interpretation was the only true one (while the surface scope interpretation was false). This suggests that they can access the inverse scope interpretation, since they would have needed to interpret the sentence with inverse scope in order to judge it to be true of the scenario.

Complementing this apparent preference for inverse scope interpretations, it is more difficult to induce adults to interpret *all*-negation with surface scope as opposed to inverse scope. Heringer (1970) asked adults to interpret ambiguous sentences in context, using a written questionnaire where informants judged "acceptability" on a four-point scale. Table 2.1 reports the sentences, the scope interpretation favored by the context, and the proportion of responses per sentence that fell on the right side of the scale. Crucially, under contexts that favor surface scope interpretations, acceptance rates are lower: the highest rate of acceptance is 0.60—of (25) under a surface scope-favoring context—as opposed to the near-complete acceptance of (26) under an inverse scope-favoring context.

(25) All the applicants didn't fail the test we so carefully rigged, did they? (Heringer, 1970)

a. No applicant failed, did they? *Surface scope* (all > n't)

b. Not all the applicants failed, did they? *Inverse scope* (n't > all)

(26) All the candidates didn't refuse the nomination, did they? (Heringer, 1970)

a. No candidate refused the nomination, did they? *Surface scope* (all > n't)

b. Not all candidates refused the nomination, did they? *Inverse scope* (n't > all)

Although there is a clear average inverse scope preference for universally quantified quantifier-negation, as summarized in Table 2.2, a striking characteristic of all past findings is that identical constructions can be interpreted differently in different contexts.

N	Stimuli	Favored scope	Proportion accepted	Bootstrapped 95% CI (calculated by me)
53	All the applicants didn't fail the test we so carefully rigged, did they?	Surface	0.60	[0.47, 0.74]
	All the guests didn't arrive at the house until 9 o'clock.	Surface	0.40	[0.26, 0.53]
	All the students didn't pass the test, did they?	Surface	0.38	[0.25, 0.51]
	All the treasure seekers didn't find the chest of gold.	Surface	0.36	[0.23, 0.49]
	All the candidates didn't refuse the nomination, did they?	Inverse	0.98	[0.94, 1.02]
	All the boys didn't leave.	Inverse	0.79	[0.68, 0.90]
	All the drivers didn't start the race until 2:00 PM.	Inverse	0.36	[0.23, 0.49]

Table 2.1: Heringer (1970). Written questionnaire. A scope interpretation was induced by a context written in brackets after each item. Acceptability was measured on a four-point scale between "unacceptable"—"uncertain, but probably unacceptable"—"uncertain, but probably acceptable"—"acceptable"; any response of "acceptable" or "uncertain, but probably acceptable" was coded as acceptance. Note that my presentation of the data reworks Heringer's presentation, which was a report of the number of participants that fell into different groups of syntactic idiolects, i.e., the number of participants who accepted and rejected particular stimuli.

Phrase or Construction	**Scope Preference**	**Source and Evidence Type**
all-negation	inverse (neg > all)	Neukom-Hermann (2016, 300-item corpus)
all-negation	inverse (neg > all)	Carden (1970, 1973, interviews)
all-negation	inverse (neg > all)	Heringer (1970, acceptability judgments)
every-negation	inverse (neg > every)	Musolino et al. (2000, 30-item corpus, interviews)
every-negation	inverse (neg > every)	Musolino (1999, TVJT with adults)

Table 2.2: Adults tend to produce and interpret universally quantified quantifier-negation utterances with inverse scope.

Changes to the local linguistic context (i.e., the sentence containing the quantifier-negation clause) can flip interpretation patterns entirely. Carden (1973) found that for (27), 82.5% of participants said that only the inverse scope interpretation was possible, 7.5% said that both were possible but that they favored inverse scope, and none accessed only surface scope. On the other hand, for (28), 100% of participants said that only the surface scope interpretation was possible.

(27) All the boys didn't arrive, did they? (Carden, 1973)

a. No boys arrived, did they? *Surface scope* (all > n't)

b. **Not all the boys arrived, did they?** *Inverse scope* (n't > all)

(28) All the boys didn't leave until midnight. (Carden, 1973)

a. **No boy left until midnight.** *Surface scope* (all > n't)

b. Not all the boys left until midnight. *Inverse scope* (n't > all)

Likewise, Heringer (1970) found unexplained variation in the acceptability ratings of different sentences in their different contexts (see Table 2.1; although there is no sense of a control sentence or context to compare these different interpretations to). This variation seems most likely to be due to the immediate linguistic context.

Stability in individual interpretations. There is little evidence of stability in each person's interpretations, in that a person would consistently arrive at the same interpretation of the same or similar stimuli. Carden argues that interpretations do resemble clear intuitions but his data are unclear. Carden (1972) administered a slightly changed questionnaire after a month to 30 of the 40 subjects of Carden (1973), finding a per-speaker percent agreement (what he calls test-retest reliability, R = (1 - changes/responses)) of 83% for the *all-n't* sentences. He found a comparable percent agreement (80-83%) for four other kinds

of linguistic judgments, which, he argues, suggests that 83% reflects a high enough level of linguistic judgment stability to indicate that he's measuring something real. However, if we take the findings about the overall preference for inverse scope to reflect that inverse scope is a very likely guess for any subject, then baseline rates of percent agreement based on biased guessing could easily be close to 80%.

2.2.1.3 Variation across interpretations of sentences with different quantifiers

Variation becomes more evident when other scopally ambiguous constructions, beyond just *all*-negation and *every*-negation, are considered. These studies (that compare scope preferences given different scope-bearing operators, e.g., Ioup, 1975; Beghelli and Stowell, 1997) generate many predictions about preferred interpretations. The limitations of these accounts are that they are sometimes based on little data and may produce predictions that conflict with each other.

Ioup's (1975) Quantifier Hierarchy. According to one characterization, quantifier words fall along a hierarchy based on inherent tendency to take highest scope: *each* > *every* > *all* > *most* > *many* > *several* > *some* > *a few* (Ioup, 1975). This quantifier hierarchy is based on judgment data from different languages and syntactic constructions.

Wide scope of the quantifier translates to surface scope in a quantifier-negation construction. Thus, we can reformulate the hierarchy specifically for quantifier-negation in terms of the construction's tendency to yield surface scope:

(29) **Surface scope preference for quantifier-negation**: *each*-negation > *every*-negation > *all*-negation > *most*-negation > *many*-negation > *several*-negation > *some*-negation > *a few*-negation

The predictions from this quantifier hierarchy, in arguing for the strongest *surface* scope preference for universally quantified quantifier-negation, possibly conflict with the evident average preferred inverse scope for *all/every*-negation. One subtlety of the hierarchy, though, is that it doesn't predict which scope interpretation is actually preferred. It only predicts that inverse scope preference should be less for *each/every/all*-negation than for other forms of quantifier-negation.

An interesting prediction of this hierarchy is that the subtle differences between the universal quantifiers *each*, *every* and *all* should also lead to differences in scope interpretations. Ioup (1975) gives the following examples of how this difference should play out, writing that wide scope of the quantifier is clearly preferred for (30), but narrow scope of the quantifier, as in the surface scope interpretation (31a), is slightly preferred for (31). The constructions in these examples are not quantifier-negation, the quantifier is the second logical operator rather than the first (the *a*), so predicted wide scope of the quantifier translates to predicted inverse scope.

(30) ... a picture of every child ...

a. There is a single picture of all the children — *Surface scope* (a > every)

b. **Each child has a different picture** — *Inverse scope* (every > a)

(31) ... a picture of all the children ...

a. **There is a single picture of all the children** — *Surface scope* (a > all)

b. Each child has a different picture — *Inverse scope* (all > a)

Beghelli and Stowell's (1997) Quantifier Phrase types and predictions. On an even larger scale, Beghelli and Stowell (1997) explain scope preferences based on a grouping of quantifier phrase types into interrogative, negative, distributive-universal, counting, and

group-denoting (the largest and most heterogeneous class). Table 2.3 shows the classifications with examples.

Quantifier Phrase Type	**Examples**
interrogative	wh-phrases like *what*, *which man*
negative	*nobody*, *no man*
distributive-universal	*every* and *each*
counting	*few, fewer than five, at most six, more than five, between six and nine, more students than teachers*
group-denoting	indefinites like *a, some, several*, bare numerals like *one student, three students*, and definites like *the students*

Table 2.3: Quantifier phrase types according to Beghelli and Stowell (1997). Note that they later revise the characterization of *every* to be underspecified for distributivity.

Using this classification, Beghelli and Stowell (1997) predict that *wh*-phrases should usually take wide scope with respect to any other quantifier phrase in their clause. For example, both *Who did everyone meet?* and *Did everyone see what I brought?* should tend to be interpreted with the *wh*-phrase taking scope over *every*. (Here and in further examples, the interpretation that is predicted to be preferred is bolded.)

(32) Who did everyone meet?

a. **Who was met by everyone?** *Surface scope* (who > everyone)

b. For every individual of *everyone*, who did that individual meet? *Inverse scope* (everyone > who)

(33) Did everyone see what I brought?

a. For every individual of *everyone*, did that individual see what I bought? *Surface scope* (everyone > what)

b. **Was what I brought seen by everyone?** *Inverse scope* (what > everyone)

Beghelli and Stowell (1997) also write that a counting quantifier phrase in object position

should never be able to take inverse scope over *each*, *every*, and group-denoting quantifier phrases occurring in subject position (Beghelli and Stowell, 1997). For example, *Some/every/one of the students visited more/fewer than two girls* should only have surface scope.

(34) Each of the students visited more than two girls.

a. **It is the case for each student that that student visited more than two girls.** *Surface scope* (each > more than two)

b. More than two girls were visited by all the students. *Inverse scope* (more than two > each)

A group-denoting quantifier phrase should be scopally ambiguous with respect to a clausemate *each* or *every* (Beghelli and Stowell, 1997). For example, both *Each/Every student read two books* and *Two students read each/every book* should be ambiguous between a surface scope and inverse scope interpretation. This is a widely accepted generalization and many papers on scope ambiguity will illustrate the phenomenon on their first page with an example that uses a group-denoting quantifier phrase and a distributive-universal quantifier, with one in the subject and the other in the object position. For example, *A climber scaled every cliff* in Anderson (2004), *Some man danced with every woman* in Kiss and Pafel (2017), and *Every kid climbed a tree* in Kurtzman and MacDonald (1993), to name a few.

(35) Every student read two books.

a. For every student, that student read two books. *Surface scope* (every > two)

b. Each of two books were read by all the students. *Inverse scope* (two > every)

(36) Two students read every book.[1]

[1]Some people report that a third interpretation is possible here: together, two students read every one of the books (but there may have been some books that each student didn't read on their own).

a. For two students, that student read every book. *Surface scope* (two > every)

b. For every book, that book was read by two students. *Inverse scope* (every > two)

Third, a group-denoting quantifier phrase in a surface object position should usually be scopally higher than clausal negation. (Although in the text of the paper, it is not always clear whether Beghelli and Stowell predict that these interpretations would be preferred or merely possible.) For example, both *The students didn't read two/some books* and *No student read two/some books* should both be interpreted to mean that *There exist two/some books not read by the students under discussion*, where both forms of negation take lower scope.

(37) The students didn't read two books.

a. It is not the case that the students read two books. *Surface scope* (n't > two)

b. **There exist two books not read by the students.** *Inverse scope* (two > n't)

The exception, Beghelli and Stowell (1997) write, is that a bare numeral like *two books* could be interpreted as a counting quantifier phrase rather than as a group-denoting one, in which case *The students didn't read two books* could (or possibly should) be interpreted with surface scope. The relevant deeper distinction between what it means for a quantifier to be counting vs. group-denoting is unclear.

On the other hand, and most relevant to quantifier-negation, a group-denoting quantifier phrase *subject* should always take wide scope with respect to clausal negation or a clause-mate negative quantifier phrase (Beghelli and Stowell, 1997). They give the examples that *Two/some/the students didn't read this book* and *Two/some/the students read no books* should be interpreted with surface scope; or see (38), where the surface scope interpretation is bolded. So here it is predicted that the inverse scope interpretation, where negation takes scope over

the group-denoting quantified subject, should be impossible.

(38) Something will not happen.

a. **There is something that will not happen** *Surface scope* (some > not)

b. Nothing will happen *Inverse scope* (not > some)

The only exception to this rule is when the subject is quantified by simple indefinites (*a/an* or a bare plural); for example, *A student/Students didn't write this book* could be interpreted with inverse scope (Beghelli and Stowell, 1997). The general idea here is that the simple indefinites are acting not as quantifier phrases but rather as something else, which means they are not subject to the same rules as the quantifier phrases in the broader typology in terms of scope preference.

Specifically regarding *some*, another argument why the inverse scope interpretation of (38) is predicted to be impossible has to do with the semantics of negation and of *some*. Negation is antiadditive; broadly, additivity and antiadditivity of an operator are ways of characterizing the alternatives that the operator creates in the sentences in which it's used. (A specific test for antiadditivity is whether the operator applied to a subject predicated with *walks or talks* mutually entails the same thing predicated with *walks and talks*. For example, *no one* is antiadditive because *No one walks or talks* entails that *No one walks and no one talks*, and *No one walks and no one talks* entails that *No one walks or talks* (Szabolcsi, 2004).) While negation is antiadditive, *some* is a positive polarity item (PPI), which means that it always scopes above antiadditive operators like negation, except in certain contexts (Szabolcsi, 2004). Indeed, *some*-negation utterances have been treated in some studies as unambiguous, only compatible with the surface scope reading (Musolino, 1999).

However, since there are special contexts in which PPIs seem able to scope under negation, it may be that in just these contexts a *some*-negation construction could be able to receive an

inverse scope, *not* > *some* interpretation. Szabolcsi (2004) writes that those contexts that license an overt antiadditive operator taking wide scope over a positive polarity item are just those contexts which license weak negative polarity items like *ever*.

For example, just as an English speaker can say (39), so can they say (40) with the intention to convey that *didn't* scopes over *someone*. By implication, in contrast to the prediction in Beghelli and Stowell (1997) and more generally, speakers might be able to intend the *not* > *some* scope interpretation of *some*-negation in the same kind of context.

(39) I don't think that John ever called me.

(40) I don't think that John didn't call someone. (Szabolcsi, 2004)
= [didn't > some] is possible, as in *I don't think that John called no one.*

(41) It does not mean that something will not happen.
= [not > some] might be preferred, as in *It does not mean that nothing will happen.*

Table 2.4 summarizes some of the constructions and their preferred interpretations from the literature above.

Overall, quantifier-negation constructions with universal quantifiers tend to receive inverse scope interpretations in corpora and behavioral experiments. The ambiguity of the construction is evident in the fact that both interpretations are attested across speakers and utterances. Moreover, the ambiguity of the construction is evident in that similar sentences can nevertheless receive opposite interpretation preferences. The next section turns to the question of how this interpretation variation can be accounted for – first, with factors relating to structure, then, with factors relating to pragmatics, highlighting the role of plausibility as I see it.

Phrase or Construction	**Scope Preference**	**Source**
All the boys didn't arrive, did they?	inverse (neg > all)	Carden (1973)
All the boys didn't leave until midnight	surface (all > neg)	Carden (1973)
... a picture of every child ...	inverse (every > a)	Ioup (1975)
... a picture of all the children ...	surface (a > all)	Ioup (1975)
Every student read two books.	ambiguous	Beghelli and Stowell (1997)
Two students read every book.	ambiguous	Beghelli and Stowell (1997)
The students didn't read some books.	inverse (two > neg)	Beghelli and Stowell (1997)
Some students didn't read this book.	surface (some > neg)	Beghelli and Stowell (1997)
No student read some books.	inverse (some > no)	Beghelli and Stowell (1997)
Some students read no books.	surface (some > neg)	Beghelli and Stowell (1997)
(It doesn't mean that) some students didn't read this book.	inverse (neg > some)	Szabolcsi (2004)

Table 2.4: Some similar constructions that are nevertheless predicted to have different preferred scope interpretations.

2.2.2 Accounting for variation

Many factors work together to account for variation. One approach (e.g., Viau et al., 2010) to characterizing these factors, that I use here, is to separate them into structural factors (relating to speakers' knowledge of the formal properties inherent to the linguistic expressions, as they're used in the specific configuration) and pragmatic factors (relating to conversationalists' knowledge about how the quantifier-negation construction would be used in context). I review a range of accounts, before focusing in particular on plausibility.

Section 2.2.2.1 discusses how several structural factors predict or allow different interpretations, including the overt order of the logical operators, semantic-syntactic movement rules, lexical biases, the individual grammar of both adults and children, and general default parsing

strategies. However, pragmatic factors can confound structural ones: to illustrate this, I focus on a series of studies on children's scope acquisition that show how different contexts lead to different pictures of children's linguistic knowledge (Section 2.2.2.2). Section 2.2.2.3 describes pragmatic factors of interest according to this literature, including the relative plausibility of one interpretation over the other, as mediated by the interlocutor's knowledge of likely world states.

2.2.2.1 Structural factors

For various reasons, most (though not all) structural accounts tend to privilege surface scope interpretations, suggesting surface scope as a generally default or unmarked interpretation. In this section, I first describe these approaches that account for variation due to differences between sentences; I then describe approaches accounting for scope preference variation due to differences between speakers.

The leftmost quantifier phrase takes highest scope. Early accounts of scope in generative semantics directly privilege surface scope interpretations because they define that the leftmost quantifier in the overt string takes highest scope (e.g., Lakoff, 1971). For example, for the relative scope of *many* and *few*, Lakoff (1971) writes that "only the left-right order within the clause matters" (p. 241). In other words, if *many* comes before *few* in the overt clause, then the only interpretation should be *many* > *few*; if *few* comes before *many*, then the only interpretation should be *few* > *many*.

Quantifier phrases can move to scope positions at LF. Subsequent research develops semantic/syntactic accounts of the possibility of scope ambiguity. These theories of quantifier scope in generative grammar explicitly assume first that scope is determined by c-command relations at LF, and second that quantifier phrases are assigned scope by undergoing movement

to their scope positions in the derivation of the LF representations.

These accounts allow inverse scope. In one sense, they privilege surface scope interpretations, because inverse scope representation is derived from surface scope representation, or arriving at the inverse scope interpretation involves an extra operation. For example, May (1977) proposes a covert syntactic operation called QR (quantifier raising) at LF, which determines which operator takes widest scope by c-commanding the other (for refinements of this framework, see May and Keyser, 1985; Aoun and Li, 1993; Reinhart, 1997). Scontras et al. (2017) illustrate a simplified LF analysis, which follows Anderson (2004)'s derivations, in the two LF structures for (42) corresponding to its two scope interpretations. The inverse LF in (42b), as shown in the right image of Figure 2.5, involves the additional step of QR (raising *every pirate* above *a shark*) in comparison to the surface LF in (42a), as shown in the top image of Figure 2.5.

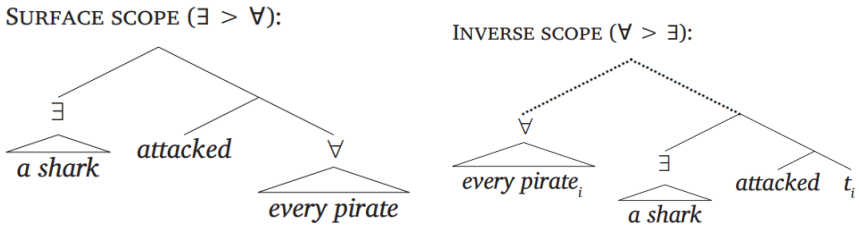


Figure 2.5: Logical form (LF) representations for the two interpretations of *A shark attacked every pirate*: surface scope (left), or inverse scope (right).

(42) A shark attacked every pirate. (Scontras et al., 2017)

a. There was a single shark that attacked multiple pirates. *Surface* (a > every)

b. For each pirate, there was a (different) shark that attacked him. *Inverse* (every > a)

Some studies tie the default representation of surface scope interpretations to default processing, predicting that surface scope interpretation should be preferred or easier to access in real-time comprehension: its less complex syntactic derivation involves less effort and

therefore a lower processing cost. Surface LF structure is claimed to be more economical (Tunstall, 1998), have less representational complexity (Pritchett and Whitman, 1995), or involve a lower processing cost (Anderson, 2004).

Lexical biases restrict quantifier phrase scope. One of the characteristics of QR accounts (e.g., May, 1977) is that all quantifier phrases technically have the same scope possibilities, such that QR would apply uniformly to different quantifier phrases (Beghelli and Stowell, 1997). But constructions with different quantifiers tend to receive different scope interpretations (see Section 2.2.1.3): to account for cross-quantifier interpretation variation, some structural accounts of interpretations 'hard-code' the scope preferences of specific phrases. These specific scope preferences are cast as due to the semantics of the quantifier (Kurtzman and MacDonald, 1993; Ioup, 1975) or to a larger semantic-syntactic system of how scope-bearing operators take relative scope (Beghelli and Stowell, 1997).

For example, Beghelli and Stowell (1997) explain lexical preferences in scope ambiguity by a system that integrates a grammatical theory with a typology of scope-bearing operators, stipulating their abstract features, and therefore their relative scope positions. The main point is that scope assignment is not uniform across quantifier phrases in this system: certain types of quantifier phrases can take scope by remaining in their first position at LF, while other types of quantifier phrases have to move to different LF scope positions. The framework proposed in Beghelli and Stowell (1997) accounts for some of the varied predictions in Section 2.2.1.3.

Adult speakers' grammar restricts their available scope representation. Individual syntactic dialect is used to account for variation across different adult speakers' interpretation behavior for the same sentences. (Although, some level of cross-rater disagreement makes sense for ambiguous sentences, since disagreement across different people's interpretations

is one way to characterize ambiguity in the first place.) Early studies frame differences in judgments as a difference in underlying syntactic access (i.e., a difference in whether an individual can or cannot access the syntactic structure, where this categorical access is a part of an individual's grammar). This experimental work with adults' interpretations of spoken and written *every-* and *all*-negation analyzes interpretation data in terms of how participants could be grouped into the syntactic dialect that their interpretations reflect (Carden, 1970; Heringer, 1970; Carden, 1973). For example, participants who only reported surface scope interpretations of the given stimuli were placed in the group of speakers whose syntax was analyzed as only allowing for surface scope.

A limitation of these analyses is the small number of stimuli sentences crossed with a low sample size of participants, which means that it's unclear whether the differences in apparent idiolects are robust. For example, if a group of ten people are found to only arrive at the surface scope interpretation of a set of ten sentences, it seems premature to conclude that these ten people never arrive at inverse scope interpretations. They should be provided with more sentences in a greater variety of constructions to confirm the hypothesis that they can't access the inverse scope interpretation. In the same vein, they should be provided with a greater variety of contexts than were provided in these studies, in order to best support either interpretation.

Children's grammar restricts their available or default scope representation. Similarly, children's developing syntax is used to account for inverse scope dispreference. Specifically, in experimental work with children's interpretations of spoken *every*-negation, five-year-olds appear to struggle to access inverse scope (Musolino, 1999; Musolino et al., 2000; Lidz, 2018). Musolino frames this result as the Observation of Isomorphism: children's developing processing abilities mean that they either can't yet generate, or can't access in the moment, inverse scope interpretations, but they don't encounter the same difficulty with

surface scope (leading to their low endorsement rates of inverse-verifying scenarios). The Observation of Isomorphism aligns with proposals that surface scope is a default and that there's extra processing difficulty for inverse (nonisomorphic) scope (Tunstall, 1998; Pritchett and Whitman, 1995; Anderson, 2004).

These acquisition studies use a truth value judgment task (TVJT; Crain and Thornton, 1998), which I review here. The TVJT method is intended to give children the best chance of indicating their linguistic knowledge, eliciting behavior that indirectly reflects their scope interpretation knowledge. Participants first see a contextualizing situation acted out visually with props. One contextualizing condition might be to see a small number of puppet characters fail at a task (e.g., three of three horses do not jump over a barn) – this would be an early-failure condition. An alternative condition might be to see all of the characters first succeed, rather than fail, in a task (e.g., three of three horses jump over a log) – this would be an early-success condition. Then, participants see a proper subset of the small number of characters succeed in another task (e.g., two of three horses succeed in jumping over a fence) – a some-but-not-all situation. This some-but-not-all situation is compatible only with the inverse scope, *not all*, interpretation and not with the surface scope, *none*, interpretation (e.g., it's false that *none*, but true that *not all*, horses jumped). A puppet says a quantifier-negation sentence (e.g., *Every horse didn't jump over the fence*) and participants are asked whether they endorse the puppet's statement.

If participants endorse, researchers infer that the participants can access the inverse scope interpretation of that construction, since they would have needed to access the inverse scope interpretation in order to conclude that the target sentence is okay. If they don't endorse, researchers often infer that they *cannot* access the inverse scope interpretation. The experimental design incorporates the conversational maxim that speakers tend to say true things (that there is a pressure to find a way to endorse). So, if a child nevertheless does not endorse, researchers reason that the child is not allowing the scope interpretation that would

make the utterance true.

Using only the early-failure condition, Musolino (1999) found that 5-year-olds almost always did not endorse *every*-negation, and always endorsed *some*-negation (Table 2.5). In contrast, adult controls always endorsed both, suggesting that children's behavior in these studies aligns with a surface scope preference.

N	Age	Tested scope	Stimuli	Proportion endorsed
20	4-7 y/o	Inverse	Every horse didn't jump over the fence	0.05
			Every boy didn't pet the polar bear	0.05
			Every caveman didn't ride on the giant turtle	0.10
			Every girl didn't ride on the merry-go-round	0.10
20	4-6 y/o	Surface	Some horses won't jump over the fence	1.00
			Some boys won't pet the polar bear	1.00
			Some cavemen won't ride on the giant turtle	1.00
			Some girls won't ride on the merry-go-round	1.00

Table 2.5: Early-failure, some-but-not-all TVJT experiments 1 and 4 in Musolino (1999), testing children's access to the inverse scope interpretation of *every*-negation and surface scope interpretation of *some*-negation, which in both cases is the *not all* rather than the *none* interpretation.

Overall, structural factors including the order of the logical operators in their surface form, covert operations in LF, lexical biases, and speakers' individual grammar account for some variation in interpretation preferences, often predicting surface scope preference. That being said, interpretation preferences result from multiple factors and some of the findings of the studies reviewed above are possibly confounded with pragmatic factors. An example of when pragmatic factors confound structural ones is described in the next section.

2.2.2.2 When pragmatic factors confound structural ones

A limitation in the TVJT studies on children's access to inverse scope is that children may struggle to accommodate and endorse the utterance if there is something unusual in the use of the utterance in context (Gualmini, 2004; Gualmini et al., 2008; Savinelli et al., 2017). In other words, someone might interpret the target sentence with the inverse scope interpretation, but nevertheless refuse to endorse, because the sentence is infelicitous in the story context (and not because it's false). One unusual aspect of the use of *every*-negation in the early-failure condition is the lack of affirmative information in the context (what this affirmative information might be is discussed below). That is, *every*-negation is a kind of negative utterance, and negative utterances are usually used in contexts that set up affirmative information to negate (Wason, 1961).

This limitation points to the importance of understanding the influence of context together with structure. In fact, differences in the contextual felicity of using a quantifier-negation utterance lead to improved endorsement of a quantifier-negation utterance under its inverse scope interpretation (Musolino, 1999; Gualmini et al., 2008; Viau et al., 2010).

In particular, Musolino and Lidz (2006) examined children's (N=20) ability to access the inverse scope interpretation of *every*-negation in two conditions. The first replicated Musolino's (1999) early-failure, some-but-not-all scenarios and target sentences. The second varied two factors: it replaced the early-failure with an early-success condition; further, the second condition added a clause with an explicit contrast to the target sentence, for example:

(43) Every horse jumped over the log, but every horse didn't jump over the fence.

Musolino and Lidz (2006) found an improved endorsement rate (50-60%) in the early-success, explicit contrast condition as compared with the replicating condition, which still yielded

a low endorsement rate (15%). Viau et al. (2010) found that even when only using an early-success context, with or without explicit contrast, endorsement was relatively high at about 60%.

The next section discusses more specifically how context influences interpretations. Pragmatic factors set up the conditions in which an utterance use is more felicitous; pragmatic factors also affect the competing plausibility of the two potential interpretations.

2.2.2.3 Pragmatic factors

Pragmatic accounts for inverse scope preference for *every/all*-negation have discussed how using quantifier-negation is more felicitous in certain contexts. First, the negation in quantifier-negation may be relatively more difficult to process when affirmative information is absent from the context; second, the conversation topic mediates the felicity of the use of any utterance (e.g., Roberts, 2012), including quantifier-negation. A complementary idea, however, is that the key variable has to do with the plausibility of the interpretations themselves, and context provides some information about this competing plausibility.

Affirmative contexts improve the felicity of using negation. Gualmini (2004) highlights that one pragmatic account of children's endorsement of *every*-negation for an inverse-verifying scenario is the felicity of using negation in context. The early-success contexts (e.g., when three of three horses successfully jump over a log) set up a more felicitous use for the negation in quantifier-negation (e.g., in *Every horse didn't jump over the fence*), as compared with the early-failure contexts (e.g., when three of three horses fail to jump over a barn), because the early-success contexts set up an expectation that goes unfulfilled (e.g., that the horses would all jump over the fence, just as they jumped over the log). Gualmini and colleagues (Gualmini, 2004; Gualmini et al., 2008) suggest that an unfulfilled expectation

can ease the processing burden of the negation in the target quantifier-negation sentences; this idea is explicit for example in a series of studies by Wason and colleagues.

As a broader source of the idea that context licenses negation when it contains information to be negated or contradicted, Wason (1961) suggested the translation hypothesis, that people tend to translate a negative into its affirmative form before evaluation, because most of their experience processing negation in natural language is as removal of misconceptions. Specifically, Wason writes that "... whereas assertions are generally used to give information about a state of affairs, denials are used to remove misconceptions about it. This subsidiary, or modifying role which denials play in language is assumed to be associated with negative statements in general (particularly when such statements occur in isolation), and tends to result in a compulsion to translate them into affirmative form before evaluating them" (Wason and Jones, 1963, pp. 306).

Wason suggests the translation hypothesis based on investigating the processing of negation in a number of tasks; for example, in a verification task (Wason, 1961), each participant (N=48) saw a total of 24 sentences of the following four constructions and was asked whether the sentence was true or false.

(44) True affirmatives: [Even number] is an even number (*or* [Odd number] is an odd number)

(45) False affirmatives: [Odd number] is an even number (*or* [Even number] is an odd number)

(46) True negatives: [Odd number] is not an even number (*or* [Even number] is not an odd number)

(47) False negatives: [Even number] is not an even number (*or* [Odd number] is not an

odd number)

Such a setup controlled for differences in meaning likely to characterize many affirmative and negative statements in natural conversation. Many negative statements that are not about whether a number is even or odd (e.g., *not blue*) are less concrete, specific, or could be satisfied in more ways than their corresponding affirmative statements (e.g., *blue*). One way to think about this difference is that an affirmative statement such as *blue* could be used to pick out certain states of the world, while the corresponding negative *not blue* could merely rule out certain states of the world.

A key finding was that people showed higher reaction times when responding to negative as opposed to affirmative sentences: reaction time decreased in the order of False negatives > True negatives > False affirmatives > True affirmatives. If a participant reported a strategy, that strategy always involved reducing negation to affirmation, either by first resolving any 'not even' to 'odd' (or vice versa) or by ignoring the negation, comparing the subject and predicate, and then taking negation into account by inverting the truth value.

Another key finding was that errors showed an interaction between polarity and truth value: True negatives > False negatives > False affirmatives > True affirmatives (Wason, 1961). In other words, although true affirmatives are easier to process than false affirmatives, false negatives are easier to process than true negatives. In a later paper, Wason (1972) writes that this processing difference between true and false negatives (which is also found in other studies) may be because false negatives are a better approximation of ordinary negation use than true negatives: "negatives, when they are used in language, are usually false rather than true. They are, of course, generally true in relation to states of affairs, but I shall argue that states of affairs are not the appropriate criteria against which to assess the specific semantic role which they play ... *The train wasn't late this morning* ... in one sense ... could count as a true negative, but that overlooks the reason for saying it. What the sentence does is

to falsify the preconception of my listeners (*His train is always late*). And in this sense the statement is a false negative" (Wason, 1972, pp. 32). In other words, I think Wason means that negative sentences are often used in a way which is false with respect to world knowledge, preconceptions, commonly-held beliefs, or some belief that is held by interlocutors. Wason (1972) further suggests that the strategy of reducing negatives to affirmatives in order to evaluate their truth could be an attempt to recover the preconception – which would be well-paraphrased by the affirmative – when the preconception is not explicit.

In accord with the translation hypothesis, the negation in *Every horse didn't jump over the fence*, when used in an early-success context, might function to remove the misconception that all the horses would succeed in jumping over the fence, since they all succeeded in jumping over the log.

Wason also suggests a non-linguistic basis for the translation hypothesis (though I think this idea is suggested mainly in speculation): that "[i]n general, the properties of objects are spontaneously described in affirmative terms, and hence it is reasonable to assume that the verification of a negative statement involves consideration of a positive one in order to make it accord with perception or judgment of the facts" (Wason, 1961, pp. 139). In other words, it is much easier for human cognition, given the way that our perception works, to think in affirmative rather than negative terms.

Conversation topics influence the felicity of the utterance. Another pragmatic factor in establishing the felicity of *every*-negation is the question under discussion (QUD; Roberts, 2012) or the Current Question, the most recent QUD in the stack that is being addressed by a discourse. A QUD is a kind of question, where a question is – odd as it may at first seem – equivalent to the set of its own answers. For example, the question *Did all the horses succeed in jumping over the fence?* could be thought about as the set of answers *All succeeded* and *Not all succeeded*. As a QUD, it would be one way of representing how interlocutors

think that a given utterance fits into a conversation: there needs to be compatibility between utterance and QUD (though the relationship is not deterministic; a given utterance is often compatible with many sets of QUDs). For example, suppose an interlocutor were to assert that *All the horses succeeded in jumping over the fence.* This utterance would fit into a conversation where the QUD was *Did all the horses succeed in jumping over the fence?* (as well as many other QUDs), because the asserted utterance rules out one of the answers to that QUD (it rules out that *Not all succeeded*, leaving only *All succeeded*), cutting down on the space of open propositions in the discourse. Importantly, narrowing down the space of open propositions is one way of advancing a conversation in a relevant way (e.g., Roberts, 2012).[2]

As another example, the same utterance would not fit into a conversation where the most recent QUD was *How many blue moles live in the ranch?* This blue mole QUD has answers *No blue moles live in the ranch, one blue mole lives in the ranch, ... QUANTITY blue moles live in the ranch. All the horses jumped over the fence* does nothing to cut down on the size of the set of answers, it does nothing to rule out any of these potential answers, and in that sense asserting *All the horses jumped* does not advance a conversation with the blue mole QUD.

Some quantifier-negation studies, (Hulsey et al., 2004; Gualmini et al., 2008) describe a more specific role for the QUD for establishing felicity: the Question-Answer Requirement. The idea is that an interpretation is felicitous if it resolves (i.e., entails a yes or no answer) to a question (QUD) made salient by the discourse.[3] By this reasoning, the early-failure contexts of *every*-negation sentences were infelicitous, because they may have set up the QUD *did*

[2]For example, Roberts (2012) puts it clearly by likening language use to a game, such that "we take the aims or goals of a language game ... to be to come to agree on the way things are in the world. Using Stalnaker's 1979 notion of the common ground (the set of propositions which the interlocutors in a discourse behave as if they all hold to be true, with a proposition realized technically as a set of possible worlds) and related context set (the intersection of the common ground, the set of worlds where all the propositions in the common ground are true), our goal is to reduce the context set to a singleton set, the actual world" (pp. 3).

[3]Note that this requirement, if it only takes into account a QUD that is able to be resolved in this way, is only taking polar questions into account, so it isn't clear what would happen with other kinds of QUDs.

none succeed? (i.e., the set of answers *none succeeded, some succeeded*). In this case, *not all succeeded* is a bad answer to the QUD, since it fails to resolve whether *none* or *some* is the answer, while *none succeeded* does resolve the QUD. Indeed, children tended to not endorse the target sentence in early-failure contexts. On the other hand, the early-success contexts as well as explicit contrast may have set up the QUD *did all succeed?* (i.e., the set of answers *all succeeded, not all succeeded*). Then, the fact that *not all succeeded* resolves the QUD (by answering *no*). Indeed, children were more likely to endorse the target sentence in early-success and explicit-contrast contexts.

To test the role of QUDs, Gualmini (2004) created new story contexts in certain TVJT tasks to manipulate QUD and raise endorsement rates. In one of their experiments (they also looked at other kinds of sentences), one *Did all?* context is for example that a friend delivers letters for another friend, who expects exactly four, but the deliveryman drops one on the way so that only three letters are delivered. The target sentence is then *Every letter wasn't delivered*. Children endorsed the utterance 80% of the time, which again contrasts with the less than 10% endorsement rate in the experiments with the early-failure contexts, and suggests that quantifier-negation is more felicitous when it answers a recent QUD.

QUDs might account for some interpretation variation in adult interpretations data as well. For example, in Carden's (1973) inverse scope-preferred sentence, (27) (*All the boys didn't arrive, did they?*), the QUD seems like the sentence itself, *Did all the boys arrive?*, to which a very informative answer is *not all*, that is, the inverse scope interpretation.

The case for the surface scope-preferred sentence, (28) (*All the boys didn't leave until midnight*), is more complicated because two QUDs come to mind when hearing the sentence out of context. The first is *When did all the boys leave?*—which would be answered well by the surface scope interpretation (i.e., they all left by midnight), and badly resolved by the inverse scope interpretation (i.e., not all left by midnight). In other words, it's answered well because the uncertainty is about when all the boys left, not whether some number of the boys left.

The other QUD is *Did all of the boys leave by midnight?*, which would be answered by both scope interpretations and therefore make both interpretations good. That is, the surface scope interpretation (i.e., they all left by midnight) answers *yes*, and the inverse scope interpretation (i.e., not all left by midnight), answers *no*, to *Did all of the boys leave by midnight?*.

World knowledge improves the plausibility of interpretations. An overview of the literature suggests another way that context disambiguates, though this is not often directly expressed in past studies, which is that context provides information as to how plausible an interpretation is, and all else equal, listeners may prefer the more plausible interpretation. This is plausibility in the cognitive sense (Saba, 1999), where the cognitive plausibility of an interpretation is its probability of being true according to interlocutors' world knowledge.

When it comes to evidence from corpora regarding the disambiguating value of plausibility, Neukom-Hermann (2016) offers examples like (24), for which she argues that the world knowledge that *Sainsbury's is a supermarket* causes the reader to prefer the *not all* reading.

(48) *Inverse scope due to world knowledge:* Many of you may have noticed that Good Housekeeping is now on sale at the checkout in Sainsbury's, which has gone down brilliantly with shoppers, as I discovered when I visited my local London branch. I can't think why **all supermarkets don't put GH at the checkout**.

Thus, Neukom-Hermann (2016) finds that world knowledge certainly can help to disambiguate, but she doesn't further characterize how it might do so. Her approach shows the strength of the general intuition that context matters, but on the other hand also shows the difficulty of simplifying the apparent complexity of factors that make up the context.

Similarly, Srinivasan and Yates (2009) mention that plausibility is a disambiguating factor, without a computational-level description of a disambiguating mechanism. Srinivasan and

Yates (2009) write that inverse scope is more preferred for (50) than for (49), because the surface scope interpretation in (50), that a single doctor lives in all the cities, is too implausible to be likely.

(49) A kid climbed every tree.

a. **There is a single kid who climbed all the trees.** *Surface scope* (a > every)

b. Each tree was climbed by potentially different kids. *Inverse scope* (every > a)

(50) A doctor lives in every city.

a. There is a single doctor who lives in all the cities. *Surface scope* (a > every)

b. **Each city has a different doctor living there.** *Inverse scope* (every > a)

Again, the broader role of plausibility discussed above applies to the example from Srinivasan and Yates (2009). The predicted preferred interpretation is the one that is more likely to be true, relative to the dispreferred interpretation.

Knowledge about the real-world probability of an interpretation being true may come from general world knowledge or any expectations about likely and unlikely states of the world as set up by the context. This knowledge is also closely related to the felicity of using the quantifier-negation construction in a given context, because the way that listeners reason about why or when a speaker would use a quantifier-negation utterance influences the way that listeners reason about the speaker's intended interpretation in that context.

Computational model for how plausibility disambiguates. In fact, though the term of plausibility isn't used in the model, a computational cognitive model from Scontras and Pearl (2021) instantiates the role of world knowledge in a way that is potentially consistent with this broader idea that plausibility disambiguates. The model shows how prior expectations facilitate a speaker's preference to produce quantifier-negation in certain contexts – for

example, certain expectations about likely world states are one way to facilitate a speaker's preference to produce *every*-negation for an inverse-scope-verifying context in a truth value judgment task. (Returning to Musolino (1999), these expectations would facilitate a speaker's preference to produce *Every horse didn't jump over the fence* for a not-all scenario where two of three horses jumped.)

In general, the model articulates the cognitive process that yields observed experimental behavior for scopally-ambiguous utterances, as a way of accounting for truth value judgment patterns for quantifier-negation. In the model, world expectations (among other factors, including questions under discussion) make different interpretations more or less informative and thus more or less likely. The key hypothesis, which is integrated as an assumption of the model, is that a pragmatic speaker chooses whether to say the potentially ambiguous utterance (or in the case of capturing TVJT behavior, endorse saying the utterance) by reasoning about whether a pragmatic listener, who hears the utterance, would arrive at the speaker's intended interpretation. This pragmatic listener arrives at an interpretation by reasoning about a rational and cooperative speaker, that is, a speaker who wants to maximize the probability that the listener will arrive at the intended understanding of the world state (while minimizing the cost of speaking). Utterances are more informative if using them means the speaker would be more successful at guiding the listener to the intended interpretation. For example, given certain world expectations (or questions under discussion), the model shows that it's informative to utter *every*-negation about an inverse-verifying scenario. Chapter 3, which focuses on modeling the role of context for scope judgments, first describes the model from Scontras and Pearl (2021) and its predictions in greater detail, in order to then build on this model to further account for scope interpretations.

Overall, context provides several kinds of information that may account for some of the attested variation in preferred interpretations of scope ambiguity. Plausibility is one such factor – the preferred interpretation of an utterance is the relatively more plausible one.

Plausibility may help cue the intended interpretation together with other cues, both structural ones as reviewed in the previous section, as well as other pragmatic factors like the QUD.

Chapter 3

General model of disambiguation

What factors combine to yield scope interpretations, and how do they combine? As mentioned in the previous chapter, a specific account of scope disambiguation which addresses this question comes from Scontras and Pearl (2021), who propose a computational model to account for child vs. adult behavior in past experimental work on *every*-negation. The model specifies a set of factors regarding how listeners integrate their grammatical knowledge of potential ambiguity (their knowledge of the two potential scope interpretations, and a truth-functional semantics) with their goals and beliefs as social agents using language to communicate (including both world knowledge and general principles of conversation). In particular, the model describe a general disambiguation mechanism that includes the role of context, where context is defined as prior beliefs about likely states of the world, such that certain prior beliefs favor an interpretation by making that interpretation more informative.

Since Scontras and Pearl focus on accounting for truth value judgments of *every*-negation (and quantifier-negation with the numeral *two*), one question is whether this model can account for interpretations of a broader range of quantifier-negation utterances, and whether it can do so for a listener who is deciding what a speaker meant (as opposed to making a

truth value judgment). Thus to test the disambiguation mechanism that this model proposes, I adapt and extend the original model to several different forms of quantifier-negation and model interpretation preferences directly. The modeled scenario is then: hearing an *every*, *some*, or *no*-negation utterance, what is the probability that a listener would arrive at an inverse rather than surface scope interpretation? And do the predicted probabilities match behavioral interpretation preferences?

Validating the adapted model, by finding that its predictions match behavioral preferences, would support its articulated role of context, which itself is implemented within a general mechanism of disambiguation. Reasoning given context is integrated with a set of other factors, including conversational cooperation, to yield interpretations.

Below, Section 3.1 introduces the computational modeling framework as well as the benefits of using the framework. Section 3.2 describes the original model; although this information is discussed in greater detail in Scontras and Pearl (2021), Section 3.2 highlights the main aspects of the model which are relevant to understanding why it makes the predictions that it does for *every*-negation truth value judgments. Section 3.3 then describes the extended model, again highlighting its predictions and the mechanism driving these predictions. Finally, Section 3.4 reports a behavioral experiment to gather interpretation preferences, tests extended model predictions against the results, and shows how the model can be fit to the results, with the idea that a better model fit reflects better explanatory power of the model. At the end, I briefly discuss open questions that the model raises about the role of context, namely, how does this role of context play out in naturalistic language data?

3.1 Background on computational models

In addition to behavioral and corpus studies, computational cognitive modeling is a useful tool for understanding the role of different factors for ambiguity resolution. A model is a concrete implementation of a hypothesis for how different factors interact to produce observable behavior (e.g., Goodman and Frank, 2016; Scontras and Pearl, 2021; Degen, 2022; Pearl, 2023). The modelled hypothesis is supported when model predictions are matched by human behavior (e.g., a model predicting inverse scope preference in a certain context would be supported by evidence that people indeed prefer inverse scope in that context). More specifically, the model is a "proof of concept" for its hypothesis, in that validating model predictions means finding proof that this is at least one way that this implementation of the hypothesis could account for human behavior (e.g., Pearl, 2014, 2023).

The models described below are articulated in the Bayesian Rational Speech Act (RSA) framework (Frank and Goodman, 2012; Goodman and Frank, 2016). In this framework, ambiguity resolution arises from rational and domain-general inferences that listeners regularly perform as they understand language. RSA models assume boundedly rational speakers who try to minimize the cost of speaking while maximizing the probability that listeners arrive at their intended interpretation, given limits on the linguistic knowledge and information available to listeners (e.g., limits on experience with certain syntactic forms). These models implement a series of recursive reasoning layers, with a speaker choosing utterances by reasoning about how a listener would interpret them, and a listener interpreting utterances by reasoning about the speaker who generated them. RSA models have been shown to capture various aspects of language use (for recent introductions to the framework with an overview of a range of models, see Scontras et al., 2018, 2021; Degen, 2022).

A benefit of modeling, as pointed out by Scontras and Pearl (2021), is that as a concrete implementation of a hypothesis (e.g., as a mathematical specification of how factors A and B

interact to yield an interpretation), a model can help separate the contributions from different processes and predict the relative contributions of each factor (e.g., it can predict whether varying factor A changes interpretations more than varying factor B).

Degen (2022) makes a similar claim specifically for RSA models, using terminology from Clark (1992), that this form of modeling connects between the language-as-product and language-as-action traditions. That is, the language-as-product tradition focuses on the cognitive processes by which speakers create and listeners recover linguistic representations (where these representations are the "product" of comprehension); the language-as-action tradition focuses on how people use language to perform acts in conversation (Trueswell and Tanenhaus, 2005). Probabilistic pragmatics like RSA, Degen (2022) writes, bridges between the two approaches because it both relies on the syntactic and semantic representations of words, phrases, and sentences (language-as-product) and provides a theory of how agents embedded in a social, communicative context with certain goals and beliefs should make decisions about the use of those linguistic units (language-as-action).

Another benefit of an RSA model of scope interpretations is that it yields probabilistic predictions for scope interpretation behavior (e.g., a modeled speaker may reason that an utterance has a 10% probability of being interpreted with inverse scope) which are able to capture the probabilistic nature of interpretation data such as average TVJT data (e.g., children may endorse an utterance 10% of the time in a particular context). This is a common practice in many RSA implementations, to model the population-level of the linguistic phenomenon (Scontras and Pearl, 2021). One subtlety, which the models don't differentiate between, is whether the modeled phenomenon is per speaker or across speakers (e.g., whether every predicted child has a 10% probability of endorsing vs. 10% of predicted children have a 100% probability of endorsing). But the main point is that these models are able to account for probabilistic phenomena (as opposed to predicting merely that speakers should arrive at one interpretation or the other).

Finally, another benefit of modeling is that many potentially hidden assumptions in the modeled hypothesis need to become explicit in the model articulation (Pearl, 2014). For example, one aspect of context, prior expectations about the state of the world, is coded as a prior in different levels of an RSA model, so there is no way for the model to keep hidden the role of that aspect of context. Instead, even if the model assumes that speakers have no prior informative expectations about the world (or suppose that the model is focusing on a phenomena across contexts, in a way which is intending to smooth out the role of context), this 'de-contextualization' would be explicitly coded as a uniform prior.

The next section turns to the specific model.

3.2 Original model of scope interpretations

Scontras and Pearl (2021) used modeling to describe child vs. adult truth-value judgment behavior in experiments like those reviewed in Section 2.2.2.2. This is the experimental data which is notable in part for how it shows a tangle of structural and pragmatic factors for disambiguation. The goal for Scontras and Pearl (2021) was to untangle the influence of these structural vs. pragmatic factors in different experimental contexts. Doing so helps articulate what exactly needs to change in underlying processes for children to become adult-like (e.g., how do certain changes in context systematically lead to changes in endorsement of *Every horse didn't jump over the fence*?).

More specifically, Scontras and Pearl (2021) describe how listeners integrate their grammatical knowledge of ambiguity (their knowledge of the two potential scope interpretations and a truth-functional semantics) with their goals and beliefs as social agents using language to communicate, including both world knowledge, questions under discussion, and general principles of conversation (e.g., interlocutors know speakers usually say things that are true

and informative). Hearing an ambiguous utterance (e.g., *Every horse didn't jump over the fence*), a pragmatic listener L1 reasons jointly about the true state of the world (e.g., the number of horses that jumped over the fence), the scope interpretation that the speaker had in mind (i.e., surface or inverse), and the QUD that the utterance addresses (e.g., *Did all the horses jump?*). L1's reasoning depends on a model of a cooperative speaker S1 who is trying to inform a hypothetical literal listener L0 about the true state of the world.

TVJT data is modeled at a higher level than L1; a truth-value judgment is in fact taken in RSA models as a form of language production rather than language interpretation (e.g., see Degen and Goodman, 2014). Both participant and speaker in these TVJT studies are already aware of the true world state (the scenario acted out with props), so arguably, when participants are asked if the speaker should describe the given scenario with the target sentence, they are not really asked to interpret the already-known truth conditions of the target utterance so much as they are asked to reason whether they themselves could or would produce that utterance in that scenario. Thus, Scontras and Pearl (2021) model endorsement as the choice of a pragmatic speaker S2 to produce the target utterance about the observed situation. S2 reasons about the probability that L1 (who, as stated above, reasons about S1's reasoning about L0) would arrive at the correct world state after hearing the utterance.

3.2.1 Model articulation

For the model, the context is that some quantity of successful outcomes is under discussion (e.g., the number of horses that successfully jumped over the fence). A world success base rate b_{suc} determines the probability that any individual will succeed (e.g., that an individual horse will succeed in jumping over the fence). The possible world states w are defined in terms of the number of successful outcomes: $w \in W = \{0, 1, 2\}$ (i.e., there are two individuals). The speaker can choose to say (endorse) the potentially-ambiguous utterance (e.g., *Every horse*

didn't jump over the fence) or say nothing (choose not to endorse the utterance): U ={amb, null}.

When interpreted with surface scope, modeled speakers and listeners understand that the ambiguous utterance means *none succeeded*; when interpreted with inverse scope, they understand it means *not all succeeded*. For the model, this shared knowledge is reflected in the truth-functional semantics for the utterances in (1), which determines which states are `true` for a given interpretation. The semantics offers a mapping parameterized by the scope interpretation $i \in I = \{surface, inverse\}$ from world states $w \in W$ to truth values $Bool$ = {`true`, `false`}. So, *every*-negation maps world 0 to `true` under surface scope and worlds 0 and 1 (i.e., w≠2) to `true` under inverse scope. The *null* utterance does not rule out any world states, mapping all of them to `true`.

(1) Utterance semantics $[\![u]\!]^i$:

a. $[\![\text{amb}]\!]^{surface} = \lambda\text{w. w} = 0$ (i.e., 'none')

b. $[\![\text{amb}]\!]^{inverse} = \lambda\text{w. w} \neq 2$ (i.e., 'not all')

c. $[\![null]\!] = \lambda\text{w. }$ `true`

The hypothetical literal listener L_0 hears an utterance u (e.g., *Every horse didn't jump over the fence*) and interprets it relative to its intended interpretation i (e.g., inverse). L_0 then reasons that there's an equal probability (i.e., a uniform distribution) over any state of the world w that is compatible with the literal semantics of u parameterized by i, using the semantics $[\![u]\!]^i$ from (1) (e.g., for the inverse scope interpretation of *every*-negation, $\text{p}(w = 0) = \text{p}(w = 1) = 0.5$). L_0 arrives at this uniform distribution by multiplying $\delta_{[\![u]\!]^i(w)}$ (i.e., 1 or 0) by the prior probability $P_0(w)$. $P_0(w)$ represents a uniform probability distribution – the hypothesized literal listener does not have informative prior beliefs, treating all world states as equally likely, and effectively learns to place zero probability on world states not compatible with the interpretation.

$$P_{L_0}(w|u,i) \propto \delta_{[\![u]\!]^i(w)} \cdot P_0(w) \tag{3.1}$$

The speaker's conversational goal is to address the topic of conversation by guiding the listener to a set x of intended world states. This set is determined by the question under discussion (QUD). That is, a QUD identifies a set of relevant world states for the listener. For instance, the QUD *all?* indicates that a speaker wants to resolve whether all the horses jumped (w∈{2}) or not (w∈{0,1}). The model implements a QUD as a mapping from worlds to partitioned sets of worlds x, as in (2). The full set of QUDs $q \in Q = \{$*all?, none?, how-many?*$\}$.

(2) QUD semantics $[\![q]\!]$:

a. $[\![$*all?*$]\!] = \lambda$w. w = 2

b. $[\![$*none?*$]\!] = \lambda$w. w = 0

c. $[\![$*how-many?*$]\!] = \lambda$w. w

To take conversational goals into account, the literal listener then infers the world state or set of world states x in the partition determined by the QUD, $[\![q]\!](w)$. The model implements this inference via the filter $\delta_{x=[\![q]\!](w)}$, which is 1 when $x = [\![q]\!](w)$ and 0 otherwise.

$$P_{L_0}(x|u,i,q) \propto \sum_w \delta_{x=[\![q]\!](w)} \cdot P_{L_0}(w|u,i) \tag{3.2}$$

The speaker S_1 selects u, knowing the particular intended world w, scope interpretation i, and QUD q as in (3.3). This calculation is based on the perceived utility of u, which depends

on the probability of u communicating the intended QUD answer x to L_0 and the cost of the utterance cost(u). That is, the speaker prefers an utterance the better it is at communicating x, but the speaker disprefers an utterance the greater its cost. Scontras and Pearl (2021) assume that both responses in the truth-value judgment task (*yes* or *no*) impose an equal cost, so utterance costs for this model are equal and the cost term cancels out. Additionally, this decision process is mediated by a softmax function *exp* and free decisiveness parameter α which controls how the speaker perceives the relative contrasts between potential options; $\alpha > 1$ means contrasts are sharpened; $\alpha < 1$ means contrasts are smoothed away; $\alpha = 1$ means contrasts are perceived as is.

$$P_{S_1}(u|w,i,q) \propto exp(\alpha \cdot (\log(P_{L_0}(x|u,i,q)) - cost(u))) \tag{3.3}$$

Hearing a quantifier-negation utterance, a pragmatic listener L_1 reasons jointly about the true world state w, scope interpretation i, and QUD q that would have been most likely to lead S_1 to produce the utterance that was observed. L_1 considers the prior probabilities of w, i, and q as well, as shown in (3.4). At this level, the listener's prior over world states $P(w)$ is informative, capturing expectations about which states are more or less likely in the context.

$$P_{L_1}(w,i,q|u) \propto P(w) \cdot P(i) \cdot P(q) \cdot P_{S_1}(u|w,i,q) \tag{3.4}$$

The level of the pragmatic listener would capture human listener behavior. As mentioned above, though, the goal of the model is to capture utterance endorsement in TVJT studies, which happens at the next layer of inference. A speaker S_2 as shown in (3.5) observes the state of the world (the scenario acted out with props, showing for example the number of

horses jumping over the fence) and chooses an utterance to convey that state of the world to L_1, marginalizing over other variables.

$$P_{S_2}(u|w) \propto exp(log \sum_{i,q} P_{L_1}(w, i, q|u)) \tag{3.5}$$

3.2.2 How the model shows the effect of pragmatic and structural factors

Model predictions depend on fixing the free parameters, which are 1) the decisiveness α for S_1, set to the default value of 1. Then there are the three priors: 2) the scope prior $P(i)$ (i.e., how easy it is to access surface vs. inverse scope), 3) the QUD prior $P(q)$ (i.e., listeners' beliefs about likely QUDs) and 4) the world prior $P(w)$ (i.e., listeners' beliefs about the general probability of the possible world states, based on the individual success base rate b_{suc}).

To test how these three factors (scope access, and the two expectations about the discourse context) influence utterance endorsement in *not-all* scenarios, Scontras and Pearl (2021) vary the three priors and see the resulting predicted interpretation preference, considering the model's predictions for the speaker S_2's marginal distribution over whether or not to endorse the *every*-negation utterance.

Figure 5.2 shows that the model indeed predicts that speakers should be more likely to endorse *every*-negation as their prior beliefs favor horses succeeding in jumping over the fence, in other words, a high positive expectation: the higher the prior probability that a horse will jump successfully, the higher the speaker's resulting preference to endorse. Endorsement also rises as prior beliefs favor inverse scope over surface scope, and the *all?* QUD. The

x-axis for the left panel is the base rate b_{suc}; as base rate increases, holding other priors at their uninformative default values, the prior expectation increases that a horse should succeed in jumping over the fence, and so does the probability of endorsement. The x-axis in the middle panel is the favored QUD $P(q = none?) = 0.9$ vs. $P(q = howmany?) = 0.9$ vs. $P(q = all?) = 0.9$, with non-favored QUDs given 0.05 probability, and all other priors held at uninformative default values. Highest endorsement is predicted when interlocutors expect that *all?* is the most likely QUD. The x-axis for the right panel is the scope prior $P(i)$. As the scope prior increases, the probability of inverse scope increases, and so does utterance endorsement.

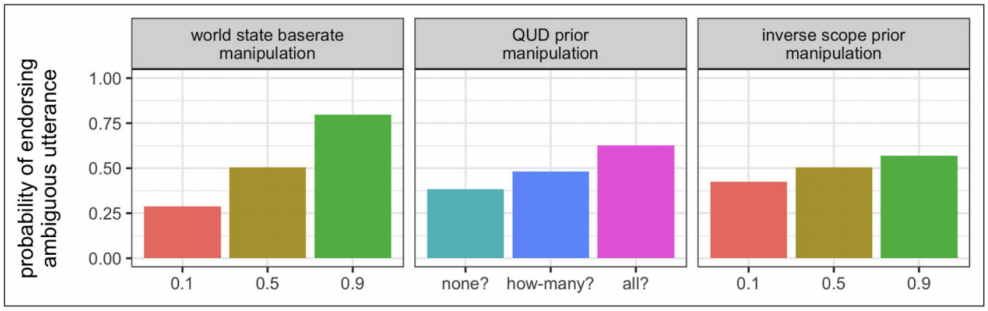


Figure 3.1: Predicted endorsement for *every*-negation (e.g., *Every horse didn't jump over the fence*) for an inverse-verifying scenario given different prior expectations (Scontras and Pearl, 2021).

Interestingly, the pragmatic priors (world state and QUD) have a greater impact on endorsement than the structural prior (scope interpretation). Scontras and Pearl (2021) suggest that the pragmatic factors have the strong impact that they do because of informativity. When expectations favor success, the utterance is a *maximally informative* way of conveying the *not all* scenario: in such a case, under either scope interpretation (*none* or *not all*) the listener who learns that prior expectations do not hold (that it is false that *all*) changes their posterior distribution the most from their prior distribution. Similarly, when *all?* is favored, both interpretations of the potentially ambiguous utterance resolve the QUD (answering *no* that *all*).

Thus, while the model implementation of expectations favoring success captures the broader idea of high positive expectations, the model mechanism for why high positive expectations matter is different from the reasoning described above for the broader role of plausibility. Specifically, the model is about speaker behavior, so its predictions are largely driven by considerations of informativity (an utterance is better when it is more useful at guiding the listener to the speaker's intended interpretation). In other words, speakers prefer to avoid saying things that are too unsurprising. However, the role of plausibility described above is about listener behavior and plausibility drives listener behavior to prefer interpretations that are true. So one pressure on listeners is to bring their interpretation of a potentially-ambiguous utterance in line with their existing understanding of the state of the world. Listeners use their prior knowledge of what is likely to be true to lend weight to certain interpretations over others. An open question is whether an extension of the model from Scontras and Pearl (2021) would predict the expected role of high positive expectations for listener rather than speaker behavior.

Overall, the RSA model of scope interpretation behavior demonstrates how context, in concert with other factors, helps disambiguate *every*-negation. In particular, world expectations have a strong influence on interpretation behavior: speakers are more likely to endorse an *every*-negation utterance for an inverse-verifying scenario in the context of an expectation like a high positive expectation. The model predicts greater endorsement given a high positive expectation because the speaker is driven to prefer an utterance that is more informative – that is, an utterance that creates relatively greater change between prior and posterior knowledge about the state of the world, and given a high positive expectation (that *all* is true), it is highly informative to learn either interpretation of *every*-negation (*none* or *not all*).

The model's mechanism of ambiguity resolution involving world priors is meant to be general, so we turn next to assessing if our model can account for quantifier-negation interpretation

preferences with other quantifiers.

3.3 Extended model of scope interpretations

To extend the model and generate testable predictions, I adapted the model space of utterances and semantics to include not only *every*-negation but also *some*-negation and *no*-negation. Making minimal assumptions, I then describe the predicted interpretation preferences.

3.3.1 Overview of changes to original model

The extended model adapts the original one in several ways, to better address the research questions about the role of context for listeners' preferred interpretations of quantifier-negation. First, the extended model removes the factor of the question under discussion, which Scontras and Pearl showed to play a similar role to world knowledge in accounting for child vs. adult interpretation behavior. Thus, extended model results more straightforwardly reflect the role of world knowledge, which is the aspect of context of interest here.

Second, the current modeling target is interpretation preference in context – belief in the relative probability of the two scope interpretations for a sentence that is judged in a linguistic context. In the language of the modeling framework, this modeling target is the pragmatic listener distribution over interpretations, rather than the pragmatic speaker's.

Third, the set of utterances is expanded from *every*-negation (and saying nothing) to include *some*-negation and *no*-negation. This choice of quantifiers allows three different classes of quantifiers to be investigated: universal *every*, existential *some*, and negative *no* (e.g., according to the classification system in Beghelli and Stowell, 1997). As discussed in greater detail in Chapter 2, these three kinds of utterances should have quite different preferred

interpretations. A key challenge for a computational model of disambiguation would be to account for the variation in preferences, and so having this range of quantifiers is a good test of how well this model captures disambiguation, based on its implemented hypothesis about the factors impacting disambiguation.

3.3.2 Model articulation

For the model, some quantity is under discussion, and the possible states of the world correspond to the different possible quantities. Specifically, the communication scenario features three marbles, each one blue or red; the possible world states w to be described are defined in terms of the number of marbles that are red: $w \in W = \{0, 1, 2, 3\}$ (see Figure 3.2). In this scenario, a speaker tries to communicate the number of red marbles to a listener. This scenario is equivalent to the one where a number of horses, out of three horses, did or did not jump over a fence; in general it would be equivalent to any scenario where a predicate applied to three entities is true or false.

(a) w = 0 (b) w = 1 (c) w = 2 (d) w = 3

Figure 3.2: Possible world states.

A speaker chooses to say one of the potentially-ambiguous quantifier-negation utterances $u \in U =$ {*every*-negation, *some*-negation, *no*-negation, *null*}. In other words, speakers can say *Every marble isn't red*, *Some marble isn't red*, *No marble isn't red*, or they can say nothing at all.

Speakers and listeners have the following interpretations, as also shown in the truth-functional semantics in (3):

- *Every marble isn't red* means *none are red* when interpreted with surface scope and *not all are red* when interpreted with inverse scope.
- *Some marble isn't red* means *not all are red* when interpreted with surface scope (i.e., there is some marble that is not red). It means *none are red* when interpreted with inverse scope (i.e., it is not the case that there is some red marble).
- *No marble isn't red* means *all are red* when interpreted with surface scope (i.e., for no marble is it the case that that marble is not red). It means *some are red* when interpreted with inverse scope (i.e., it is not the case that no marble is red, so at least one is red).

This shared knowledge about the meaning of either interpretation of a quantifier-negation utterance is reflected in the truth-functional semantics for the utterances in (1), which determines which states are `true` for a given interpretation. The semantics offers a mapping parameterized by the scope interpretation $i \in I = \{surface, inverse\}$ from world states $w \in W$ to truth values $Bool = \{\texttt{true}, \texttt{false}\}$. So for example, *every*-negation maps world 0 to `true` under surface scope and worlds 0, 1, and 2 (i.e., w$\neq$3) to `true` under inverse scope. The *null* utterance does not rule out any world states, mapping all of them to `true`.

(3) Utterance semantics $[\![u]\!]^i$:

a. $[\![every\text{-negation}]\!]^{surface} = \lambda \text{w. w} = 0$ (i.e., 'none')

b. $[\![every\text{-negation}]\!]^{inverse} = \lambda \text{w. w} \neq 3$ (i.e., 'not all')

c. $[\![some\text{-negation}]\!]^{surface} = \lambda \text{w. w} \neq 3$ (i.e., 'not all')

d. $[\![some\text{-negation}]\!]^{inverse} = \lambda \text{w. w} = 0$ (i.e., 'none')

e. $[\![no\text{-negation}]\!]^{surface} = \lambda \text{w. w} = 3$ (i.e., 'all')

f. $[\![no\text{-negation}]\!]^{inverse} = \lambda \text{w. w} > 0$ (i.e., 'some')

g. $[\![null]\!] = \lambda \text{w. } \texttt{true}$

All other aspects of the model articulation remain the same as in the original model, except that the question under discussion factor is removed. A listener interprets an utterance by reasoning about the speaker who generated it (and a speaker chooses an utterance by reasoning about how a listener would interpret it). Specifically, a pragmatic listener L_1 reasons about the speaker S_1 who generated the utterance, considering that S_1 was reasoning about an imagined literal listener L_0 when generating that utterance.

The hypothetical literal listener L_0 hears an utterance u and interprets it relative to its intended interpretation i; L_0 reasons that the state of the world w is any of the world states that are true, given the semantics $[\![u]\!]^i$ from (1). The model implements this reasoning as a filter on the possible world states $\delta_{[\![u]\!]^i(w)}$, which returns 1 when $[\![u]\!]^i(w)$ is `true` and 0 otherwise. L_0 then weights the true world states equally, returning a uniform probability distribution over those states w compatible with the semantics. L_0 arrives at this uniform distribution by multiplying $\delta_{[\![u]\!]^i(w)}$ (i.e., 1 or 0) by the prior probability $P_0(w)$; $P_0(w)$ represents a uniform probability distribution – the hypothesized literal listener does not have informative prior beliefs, treating all world states as equally likely.

$$P_{L_0}(w|u,i) \propto \delta_{[\![u]\!]^i(w)} \cdot P_0(w) \tag{3.6}$$

The speaker's conversational goal in this model is to guide L_0 to the intended world state. In this setting, the goal amounts to conveying exactly how many of the three marbles are red. The speaker S_1 selects u, knowing the particular intended world w and scope interpretation i as in (3.7). This calculation is based on the perceived utility of u, which depends in part on the probability of u and i communicating the intended world state w to L_0: $P_{L_0}(w|u,i)$. The other component of an utterance's utility is its negative cost, $c(u)$. Broadly, utterance cost can reflect different reasons for why utterance use is difficult or effortful: for example, an

utterance can be costlier than another if it is longer or less frequent. The speaker's decision process is mediated by a softmax function and free parameter α, which controls how the speaker perceives the relative contrasts between potential utilities; contrasts can be sharpened (α >1), smoothed away (α <1), or perceived as is (α=1).

$$P_{S_1}(u|w,i) \propto exp(\alpha \cdot \log(P_{L_0}(w|u,i)) - c(u)) \tag{3.7}$$

Hearing quantifier-negation, a pragmatic listener L_1 reasons jointly about the true world state w and scope interpretation i that would have been most likely to lead S_1 to produce the observed utterance. L_1 considers both the prior probabilities of w and i as well as the speaker's decision process $P_{S_1}(u|w,i)$, as shown in (3.8). At this level, the listener's prior over world states $P(w)$ is informative, capturing expectations about which states are more or less likely in the context.

$$P_{L_1}(w,i|u) \propto P(w) \cdot P(i) \cdot P_{S_1}(u|w,i) \tag{3.8}$$

The model predictions of interest are based on L_1 behavior, specifically the marginal posterior distribution on interpretations upon hearing the quantifier-negation utterance in context.

3.3.2.1 Model parameter setting

As before, generating predictions from the model depends on fixing the free parameters, which determine (i) the decisiveness α, (ii) the scope prior $P(i)$ (i.e., listeners' beliefs about the general probability of surface vs. inverse scope), (iii) utterance cost $c(u)$, and (iv) the

world prior $P(w)$ (i.e., listeners' beliefs about the general probability of the possible world states).

To implement minimal assumptions, $\alpha = 1$ (that is, no sharpening or smoothing of utilities) and the prior is uniform over scope interpretation, such that $P(\textit{surface}) = P(\textit{inverse}) = 0.5$ (that is, neither scope interpretation is preferred a priori).

For utterance costs, it seems more costly to say something than to say nothing, so cost(*null*) $= 0 <$ cost(*every/some/no*-negation). In addition, the relative costs of *every*-, *some*-, and *no*-negation were set to reflect their relative frequency in speech, such that less frequent utterances cost more. To estimate appropriate values, I mined *every*-negation, *some*-negation, and *no*-negation utterances from a naturalistic speech transcript corpus. See Section 4.1.2, which describes the methodology of identifying and extracting target utterances from the transcripts in the Corpus of Contemporary American English. The transcripts contained 390 occurrences of *every*-negation, 2,947 occurrences for *some*-negation, and 50 occurrences for *no*-negation. The relative costs of the utterances were then set to be inversely proportional to their relative frequency in the corpus: cost(*every*-negation) $= \frac{1}{\frac{390}{390+2947+50}} = 8.684615$, cost(*some*-negation) $= \frac{1}{\frac{2947}{390+2947+50}} = 1.149304$, cost(*no*-negation) $= \frac{1}{\frac{50}{390+2947+50}} = 67.741$.

This leaves world prior $P(w)$, which is specified such that individual marbles have a probability p_r of being red, and each world state contains three such marbles. So, the underlying distribution for $P(w)$ is a binomial distribution with three trials, each with success probability p_r, as in (3.9). This distribution corresponds to sampling three marbles with replacement, and models the resulting number of marbles that are red.

$$P(w = k) = \binom{3}{k} \cdot p_r^k (1 - p_r)^{3-k} \tag{3.9}$$

This base rate of marbles being red is set at $p_r = 0.5$, such that a marble is equally likely to be red or not.

3.3.2.2 Model predictions for scope interpretation preferences

As mentioned above, L_1 models a listener in the real world, and so the main predictions of interest are for pragmatic listener L_1's marginal distribution over scope interpretations of the three utterances. Figure 3.3 shows these predicted interpretation preferences.

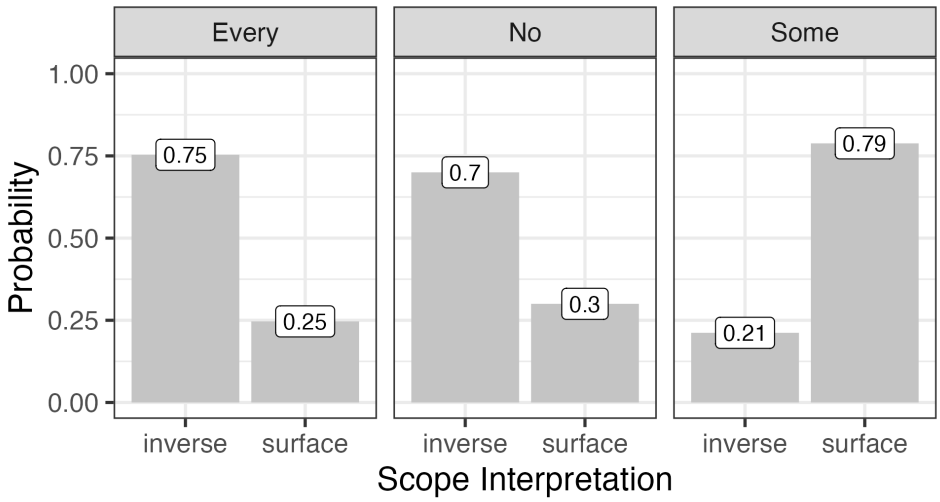


Figure 3.3: Pragmatic listener marginal probability distribution over scope interpretations, when it is only assumed that relative utterance costs reflect their relative frequencies of use in spontaneous speech (i.e., the rare *no*-negation is highly costly, *every*-negation moderately costly, and the relatively common *some*-negation is slightly costly; to say nothing costs nothing). Otherwise, $\alpha = 1$, the prior over scope interpretations is uniform, and each marble has a 50% chance of being red.

Under these parameter settings implementing minimal assumptions, the model predicts that the proportion of inverse scope interpretations depends on the quantifier. The probability that *every*-negation receives an inverse scope interpretation (0.75) is greater than the probability that *no*-negation receives an inverse scope interpretation (0.7), which is greater than the probability that *some*-negation receives an inverse scope interpretation (0.21).

How do these predictions compare qualitatively to predictions from the literature with

respect to different quantifiers? The clearest prediction from the literature is that *some* should generally scope above negation (Szabolcsi, 2004), so *some*-negation should usually or always receive a surface scope interpretation, because its inverse scope interpretation involves negation scoping over *some*. The predictions for *no*-negation utterances are less clear, in part owing to the difficulty introduced by double negation. *Every* should be able to scope under negation, and thus allow for inverse scope, but predictions for its preferred interpretation are also unclear, because while one line of studies suggest that adults prefer inverse scope interpretations, another line of experimental and theoretical studies suggest that surface scope is easier to access. Thus, model predictions are somewhat in line with the predictions from the literature, in that *some*-negation is also expected to receive the strongest surface scope preference. Even so, the model prediction of a 21% inverse scope preference may be too high, since *some*-negation should only receive inverse scope in restricted contexts.

Why does the model make the predictions that it does? In general, these predictions rest on the listener L_1 reasoning that the speaker S_1 maximizes the probability that the listener L_0 will arrive at the true world state. That is, the listener reasons that the utterance is true, and the probability that the utterance is true is higher under the certain scope interpretations for each quantifier.

First, consider why the *not all* scope interpretation is preferred for *every*-negation (inverse=0.75) and *some*-negation (surface=0.79). The reason is the same for both quantifiers. Since the marble redness base rate is $p_r = 0.5$ (i.e., chance), the most likely world states are those where exactly one or exactly two marbles are red (as shown in Figure 3.4). It is more likely for *not all* to be true (w could be 0, 1, or 2, and world states 1 and 2 are relatively most likely according to our prior) than for *none* to be true (w must be 0, and 0 is relatively unlikely according to our prior). The listener reasons that the utterance is true, and so reasons that the speaker most likely intended the meaning that is more likely to be true: the *not all* meaning (i.e., inverse for *every*-negation and surface for *some*-negation).

For the same reason, the *some* interpretation is preferred over the *all* scope interpretation for *no*-negation (inverse=0.7). In particular, it is more likely for *some* to be true (w could be 1, 2, or 3, and world states 1 and 2 are relatively most likely according to our prior) than for *all* to be true (w must be 3, and 3 is less likely according to our prior). The listener reasons that the speaker most likely intended the meaning that is more likely to be true: the *some* meaning.

Let's put these predictions again in intuitive terms, given these minimal-assumption model parameters where the most likely world states are the *some but not all* ones *a priori*. Upon hearing *every*-negation or *some*-negation, listeners will believe there is a high probability that *some but not all* is true; so, the speaker cannot have meant *none* and, therefore, meant *not all* instead (i.e., the inverse scope interpretation of *every*-negation and the surface scope interpretation of *some*-negation). Upon hearing *no*-negation, listeners will believe there is a high probability that *some but not all* is true, such that the speaker cannot have meant *all* and, therefore, meant *some* instead.

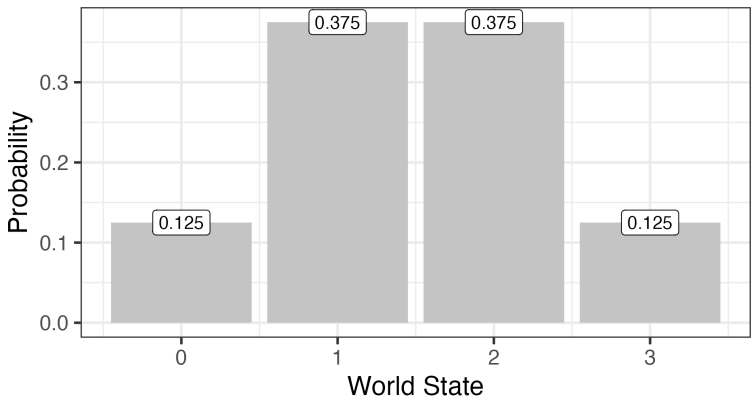


Figure 3.4: Prior probability distribution over world states when the probability of a marble being red is at chance (50%).

With these predictions from the unfit version of the model in hand, the next section considers whether the predictions are borne out in human interpretation patterns. If they are, that would demonstrate more general support for the extended model of scope disambiguation and its mechanism of ambiguous utterance interpretation.

3.4 Testing model predictions for *every*-, *some*-, and *no*-negation

To test model predictions, a short paraphrase-endorsement experiment elicited native English speakers' interpretation preferences for utterances with the quantifiers *every* vs. *some* vs. *no*. The stimuli were these three quantifier-negation utterances with no linguistic context, embedded in a communication scenario with two characters. A reference picture-selection experiment first validated that the relevant paraphrases of each potentially-ambiguous utterance, which were used in the paraphrase-endorsement task, were understood to have a meaning compatible with surface- vs. inverse-verifying scenarios. For example, that the inverse-scope paraphrase of *some*-negation, *None of the marbles are red*, was understood to have a meaning compatible with a scenario where none of the marbles were red.

3.4.1 Experiment 1: Paraphrase validation

Following the methodology of Scontras and Goodman (2017), Experiment 1 validates unambiguous paraphrases of the scope interpretations of the potentially ambiguous utterances. Given a paraphrase, participants were asked to select the picture that the paraphrase likely described.

3.4.1.1 Participants

102 participants were recruited through Amazon.com's Mechanical Turk (MTurk) crowdsourcing service, who had U.S. IP addresses and at least 95% approval ratings for at least 1,000 tasks on MTurk. Each received $0.50. 94 participants (42% female; mean age: 37) indicated that they understood the experiment and that English was their only native language; only

their data were included in the analyses reported below.

3.4.1.2 Design

The experiment began with a scenario intended to establish that the utterances to be interpreted were communication acts (Figure 3.5). A character, Mellow, is said to have a collection of marbles, three of which she places into a box. Participants were told that Mellow tells another character, Bluesy, about the box of marbles, and that their task is to help Bluesy interpret Mellow's utterance.

Participants then saw in random order three trials where they chose the scenario they thought an utterance described: one trial for the quantifier-negation utterance, one for its surface scope paraphrase, and one for its inverse scope paraphrase. The quantifiers *every*, *some*, and *no* were tested as a between-subject manipulation. On each trial, participants chose between an image consistent with the surface scope interpretation of the quantifier-negation utterance and an image consistent with the inverse scope interpretation (e.g., a participant in the *every*-negation condition chose between not-all-red-marbles and no-red-marbles, as in Figure 3.6). Image position (left vs. right) was randomized on each trial.

The surface/inverse scope paraphrases appear in (4) for *every*, (5) for *some*, and (6) for *no*.

(4) Every marble isn't red.

a. None of the marbles are red.

b. Not all of the marbles are red.

(5) Some of the marbles aren't red.

a. Not all of the marbles are red.

b. None of the marbles are red.

(6) None of the marbles aren't red.

a. All of the marbles are red.

b. Some of the marbles are red.

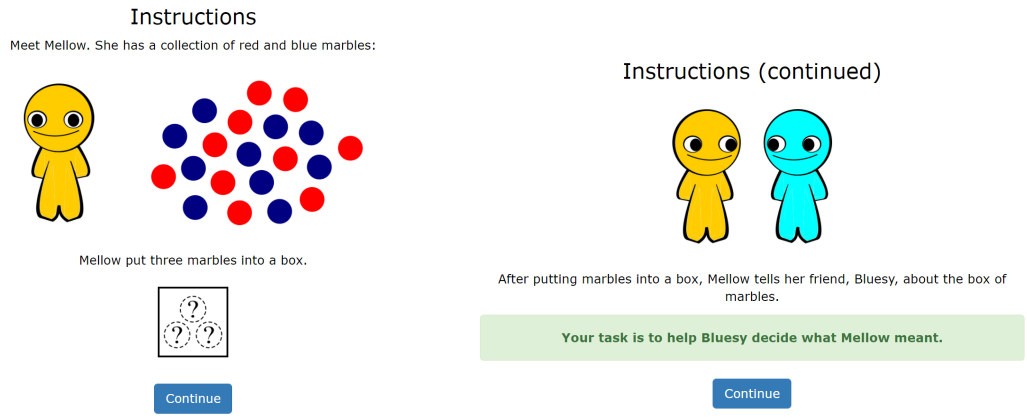
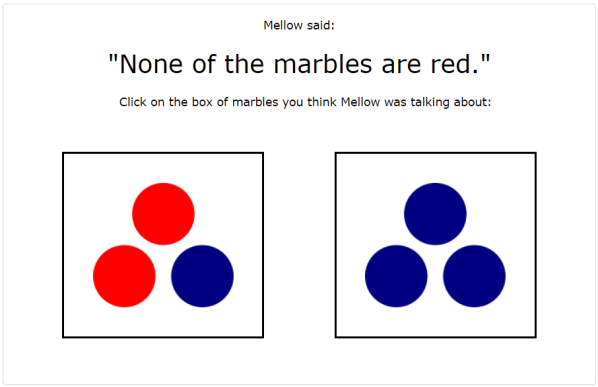
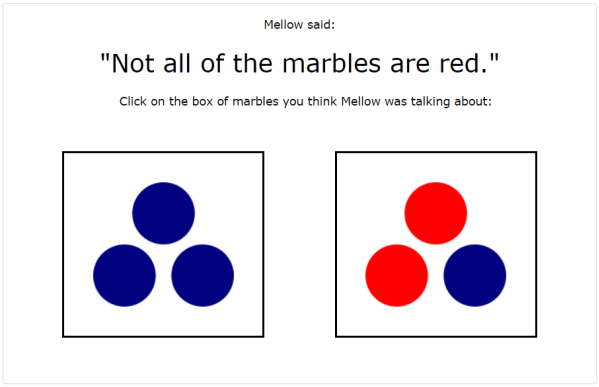


Figure 3.5: Instructions introducing the communication scenario.



(a) Validating a surface scope paraphrase: as intended, for this trial, participants chose at ceiling the image with three blue marbles.



(b) Validating an inverse scope paraphrase: as intended, for this trial, participants chose at ceiling the image with two red marbles.

Figure 3.6: Sample trials for the two scope interpretations of *every*-negation in Experiment 1.

3.4.1.3 Results

Figure 3.7 shows responses as the proportion of time that participants chose the inverse scope-verifying image, grouped by utterance type (ambiguous, inverse scope paraphrase, surface scope paraphrase) and quantifier condition (every, some, no). Participants chose at ceiling

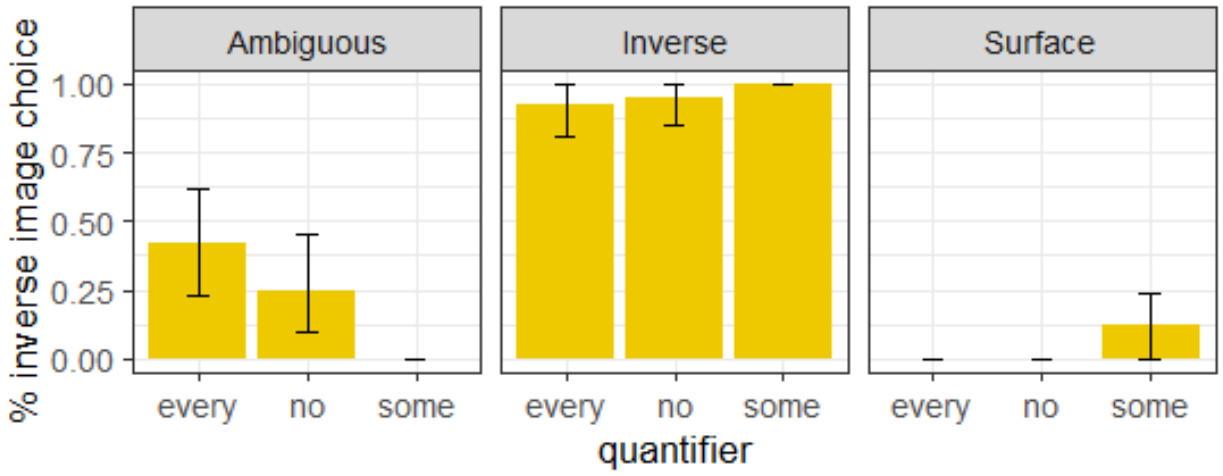


Figure 3.7: Experiment 1 results, showing that the Paraphrase validation results. Error bars are bootstrapped 95% CIs.

the image consistent with the intended scope interpretation for each of the unambiguous paraphrases: Figure 3.7, middle panel, shows that participants chose the inverse-verifying scenario nearly all the time (proportions near 1) for the inverse scope paraphrase of all the quantifiers. Likewise Figure 3.7, right panel, shows that participants chose the surface-verifying scenario nearly all the time (inverse proportions near 0) for the surface scope paraphrase of all the quantifiers.

It was not obvious that participants would respond in these clearly different ways to the competing scope paraphrases of each quantifier-negation, because for each set of competing paraphrases, one of the pictures was compatible with *both* paraphrases. That is, *not all* and *none* can both describe a state with zero red marbles – both can describe a surface-verifying scenario for *every*-negation and an inverse-verifying scenario for *some*-negation. Likewise, *some* and *all* can both describe a state with three red marbles – both can describe a surface-verifying scenario for *no*-negation. But despite this overlap in meaning, the picture-selection data suggest that *none* and *not all*, and *some* and *all*, are interpreted differently (as hoped) in the communication scenario.

One mechanism underlying their different interpretations in context may be reasoning about

alternatives in context: when participants saw the weaker alternative (*not all* instead of *none*, and *some* instead of *all*) as a potential description of the two images, they reasoned that Mellow would have used the stronger alternative had it been true, that the stronger alternative was therefore not true, and so that Mellow intended *not all* to mean *not all but not none* (i.e., *some*) and intended *some* to mean *some but not all*. On the other hand, it's the participants, not Mellow the speaker, who are faced with the two alternative images. The speaker can only be imagined to choose between alternatives of what utterance to say, not what set of marbles to describe.

Interpretations of the potentially-ambiguous utterances showed a non-significant trend (Figure 3.7, left panel) in line with the model predictions: *every* led to more inverse scope interpretations than *no*, which led to more inverse preference than *some*. Experiment 2 revisits this trend with a more sensitive measure of interpretation preferences.

3.4.2 Experiment 2: Paraphrase endorsement

Experiment 2 elicited interpretations of the utterances in the utterance space of the model (i.e., *every*-negation, *no*-negation, and *some*-negation) by asking participants to rate their validated paraphrases on a sliding scale.

3.4.2.1 Participants

60 participants with U.S. I.P. addresses were recruited through MTurk. Each received $0.50. Data were assessed for 47 participants (32% female; mean age: 36) who indicated they understood the experiment and English was their only native language.

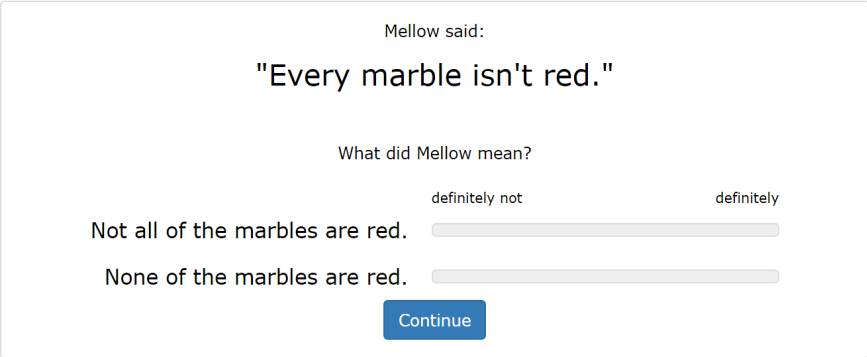


Figure 3.8: Sample paraphrase-endorsement trial.

3.4.2.2 Design

Participants saw the same communication scenario as in Experiment 1 (Figure 3.5). Participants then completed three trials (one for *every*, *some*, and *no*) in random order. Under the quantifier-negation, in order to highlight the ambiguity, they were presented with two sliders to rate each of the two paraphrases of the utterance (e.g., Figure 3.8). Paraphrases were the same as those given in (4), (5), and (6). Note that unlike the reference task experiment, no images of the referents were used.

3.4.2.3 Results

The responses on the inverse scope paraphrase sliders are reported below. To compare against model predictions, following the method used by Scontras and Goodman (2017), only model predictions for one slider response (this inverse scope slider response) are considered. In general, the slider responses per item were negatively correlated (correlation between surface vs. inverse slider decision for *every*: -0.51; *no*: -0.40; *some*: -0.67), suggesting that endorsing one interpretation led to reduced endorsement for the other interpretation.

How do the model predictions that implement minimal assumptions (i.e., unfit predictions from Figure 3.3) compare to behavioral responses? And can model predictions compare

better to behavioral responses if they are fit to the responses (i.e., fit predictions)? Figure 3.9 shows this comparison: endorsement rates (as yellow bars) grouped by quantifier for responses for the inverse scope paraphrases, with the fit model predictions (as dark grey bars) and unfit model predictions (as pale grey bars).

Regarding only the behavioral results, to assess significance, linear mixed effects models were used to predict the logit-transformed responses on each of the sliders by quantifier, with random intercepts for participant. All differences were significant. Considering the yellow bars from left to right in Figure 3.9: *every* allowed the most inverse interpretations (95% CI [0.65, 0.84]), *no* allowed an intermediate proportion (95% CI [0.27, 0.47]), and *some* allowed the fewest inverse scope interpretations (95% CI [0.07, 0.18]).

These behavioral results are qualitatively in line with the overall pattern of *every* vs. *no* vs. *some* interpretation preferences of the unfit model predictions, as shown by the pale grey bars in Figure 3.9. Inverse scope is most preferred for *every*-negation and least preferred for *some*-negation. More specifically, given utterance costs reflecting utterance frequencies and no other parameter fitting (maintaining minimal assumptions of $\alpha = 1$ and no expectations about the general probability of surface vs. inverse scope or the rate of marbles being red), the model is able to capture some of the pattern of average, cross-speaker interpretation preferences across quantifiers: model predictions fall just within the 95% CI for mean inverse scope probability for *every*, but overpredict the inverse scope preference for *no* and *some*.

The place where the unfit model qualitatively predicts the wrong preference is for *no*-negation: the unfit model predicts an inverse scope preference for *no*-negation, but the behavioral results show that *no*-negation is highly ambiguous, with a slight preference for its surface scope, *all* meaning.

To improve model fit and yield the model predictions shown in the dark grey bars in Figure 3.9, I increased the prior probability of a marble being red p_r from 0.5 to 0.67 and increased

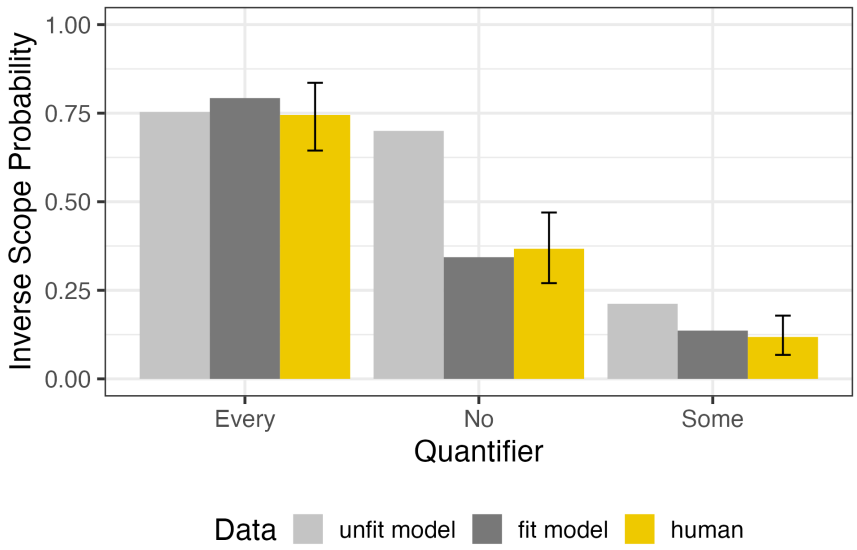


Figure 3.9: Results comparing model predictions and human data. Pale grey bars: Unfit model predictions for L_1 marginal distribution over interpretation i (the same as in Figure 3.3) with $pr = 0.5$, utterance costs based on utterance frequencies, $P(surface) = 0.5$, and $\alpha = 1$. Dark grey bars: Model predictions fit to human data for L_1 marginal distribution over interpretation i, with $pr = 0.67$, utterance costs based on utterance frequencies, $P(surface) = 0.5$, and $\alpha = 1.65$. Yellow bars: Degree of endorsement of the inverse scope paraphrase in the paraphrase-endorsement task. Error bars are bootstrapped 95% CIs.

the decisiveness parameter α from 1 to 1.65, keeping utterance costs realistic and scope priors uninformative. Correspondingly, the fit model is able to quantitatively match the preferred interpretations of each type of quantifier-negation utterance, correctly capturing *some*-negation and *no*-negation as well as *every*-negation.

By exploring the parameter space, I found that two changes were necessary in general to improve model fit: 1) increasing the base rate of success, since at decreased p_r, the model increasingly underpredicts inverse scope for *every*-negation and overpredicts inverse scope for *some*-negation. 2) Given the higher base rate values, α needs to increase, otherwise the model overpredicts inverse scope for *every*-negation and *no*-negation while underpredicting it for *some*-negation. With these two changes to parameter values, different settings of $c(u)$, $P(w)$ and $P(i)$ do not change the qualitative results reported here.

3.4.2.4 Discussion

The results of Experiment 2 show that average cross-speaker interpretation preferences vary across quantifier-negation utterances that have different quantifiers: participants prefer to interpret *every*-negation with inverse scope and *some*-negation with surface scope, while *no*-negation is ambiguous but shows a slight surface scope interpretation preference.

The extended ambiguity resolution model, without parameter fitting beyond incorporating minimal assumptions including utterance costs reflecting utterance frequencies, successfully predicts the relative pattern of inverse scope preference across quantifier. With parameter fitting – namely, fitting α and incorporating a higher expectation for the success rate – the model quantitatively captures the results as well.

The reason that the model, given an increased success rate expectation, successfully accounts for all three interpretation preferences, is that listeners prefer the most likely interpretation given their priors. When p_r increases from 0.5 to 0.67, it increases the probability on the *all* world state relative to the *not all* world states, and the *none* state becomes even more unlikely. Expecting this state of affairs, listeners of *every*-negation and *some*-negation still believe it unlikely that a speaker intended the *none* interpretation and, therefore, must have meant the *not all* interpretation. The greater change is with listeners of *no*-negation: now, since they believe the *all* world state more *a priori* likely than before, they put more probability on the *all* (surface scope) interpretation than they did before.

It is especially interesting that *some*-negation is almost entirely interpreted with its surface scope interpretation. *Some* has been called a positive polarity item, an expression that for the most part does not scope under negation (Szabolcsi, 2004). These modeling results offer an explanation for why *some* might behave as a positive polarity item in the first place: interpreting *some* under negation can result in an utterance that has an unlikely meaning and is therefore inefficient.

3.5 General Discussion

This section finds evidence for a computational model of how ambiguity resolution can proceed when sentences that have often been thought of as difficult or ambiguous are used as communication in context. The model successfully matches behavioral interpretation preferences and shows that one mechanism driving interpretation preferences is that listeners will try to align their interpretation with what they already believe to be true of the world.

Notably, with little parameter fitting beyond linking utterance cost to utterance type frequencies in a corpus, the model predicts observed variation in several quantifier-negation combinations. The unfit model accurately predicts the qualitative pattern of observed interpretations of *Every marble isn't red* vs. *Some marble isn't red* vs. *No marble isn't red* in a controlled experiment. *Every*-negation receives the highest proportion of inverse scope interpretations and *some*-negation receives the lowest. This finding demonstrates that the model of disambiguation can generalize beyond *every* in quantifier-negation utterances, highlighting the power of an RSA model for capturing interpretation preferences.

The fit model quantitatively matches behavioral preferences. A key assumption of this final version of the model is that it incorporates skewed priors about the world – namely, a belief in a relatively high success rate. The articulated mechanism underlying this pattern of interpretations is that listeners reason that speakers say things that are true, and the high success rate in the world prior lends relatively greater weight to world states that are compatible with these three interpretations for each utterance type: listeners expect that the *none* world state is unlikely, that the *all* world state is somewhat likely, and the *some but not all* states are most likely. With this assumption and expectations in mind, listeners then reason that speakers must have intended the *not all* rather than the *none* scope interpretations of *every*-negation and *some*-negation, and that speakers were slightly more likely to intend the *all* rather than the *some* interpretation of *no*-negation.

These results further demonstrate how the pressures driving listener behavior differ in some ways from the pressures driving speaker behavior. Specifically, in the RSA literature and more broadly, it seems understood that one pressure from the speaker's perspective is to be informative – to effect a change between the listener's prior and posterior distribution over world states, as a way of combating the cost of speaking. In other words, speakers are happy to surprise listeners. Or, in less simplistic terms, speakers prefer to avoid saying things that are too unsurprising. On the other hand, one pressure on listeners is to bring their interpretation of a potentially-ambiguous utterance in line with their existing understanding of the state of the world. Listeners use their prior knowledge of what is likely to be true to lend weight to certain interpretations over others.

3.6 Looking ahead

What does this ambiguity look like in the world, though? How does the broad mechanism, whereby listeners bring their interpretation into alignment with their prior beliefs, play out in concrete examples of ambiguity use? In fact, would we even see the model's predicted role of context attested in naturalistic language use?

Specifically, there is an open question highlighted in this chapter because the key factor of success rate belief is only specified in the model, not in language use as considered in an experiment or corpus. How would this belief be expressed or come into play for naturalistic language use? Moreover, this chapter has explored the model at the level of average interpretations of the same sentence (*QUANTIFIER marble isn't red*), but not at the level of interpretations of a range of different *every*-negation (or *some*-negation or *no*-negation) utterances. That is, it's not clear whether context would predict interpretations of different individual cases of the same kind of ambiguity.

Thus, the next two chapters focus on a naturalistic corpus of *every*-negation to explore within-quantifier variation and concrete measures of the role of context for interpretations. First (describing and extending the corpus work mentioned above for identifying the rate of quantifier-negation occurrences) Chapter 3 focuses on building corpora of *every*-negation from TV and radio archives and explores the within-quantifier variation found in these corpora. Chapter 4 then turns to assessing the role of success rate belief in the corpus data.

Chapter 4

Text and audio corpus of *every*-negation

In order to better understand how interlocutors navigate quantifier-negation ambiguity in everyday speech, one open question is how common quantifier-negation constructions and their preferred interpretations are. There have been a few corpus studies of quantifier-negation (e.g., Neukom-Hermann, 2016), but a limitation of these studies is that the data come from primarily written speech (Neukom-Hermann, 2016), have a small sample size of less than 30 items (Musolino et al., 2000; Taglicht, ND), or were annotated for interpretations based on the authors' judgment (Neukom-Hermann, 2016; Musolino et al., 2000; Taglicht, ND) and it's unclear how well these judgments would compare with other people's judgments. Much other research on quantifier-negation ambiguity is based on introspection and experimental studies of a restricted set of quantifier-negation utterances, mainly those with the universal quantifiers *all* and *every* (e.g., Jackendoff's 1972 *All the men didn't go*, Musolino's 1999 *Every horse didn't jump over the fence*).

Here, I build on these past studies to investigate *every*-negation sentences that are attested in production and interpreted by native English speakers. I created two corpora of *every*-negation ambiguity: the first by looking for all uses of *every*-negation in the transcript section

of the Corpus of Contemporary American English, and the second by looking for all uses of *every*-negation in All Things Considered and Fresh Air recording and transcript archives from National Public Radio. All these are relatively large-scale sources of data on spontaneous conversation, with the first being only in transcript form, and the second in recording and transcript form.

The goal of finding the *every*-negation uses in the transcript section of the Corpus of Contemporary American English was to answer baseline questions about the frequency and preferred interpretations of *every*-negation utterances in naturalistic speech. As described in Section 4.1, I investigated these questions for items in text form in their immediate linguistic context. Specifically, each case of potential ambiguity in context was judged by native English speakers for its most likely interpretation. I then discuss the frequency of items in the corpus and the patterns of preferred interpretations. I also tracked down the original audio of the *every*-negation items. This audio was available only for about 16% of the full set of ambiguity uses.

Overall in the first corpus, each case of ambiguity was recorded with its text, immediate linguistic context, preferred interpretation as gathered in a behavioral experiment, and the metadata that was available through the original Corpus of Contemporary American English; further, 16% of the cases are also recorded with the audio of the context and item itself, as well as key acoustic features analyzed for the item.

Since there were relatively few data points on naturalistic prosody from this first corpus, one motivation for creating a second corpus was to find a greater number of ambiguity uses with available original audio. I used National Public Radio archives, which are a high-quality source of spontaneous speech that combine audio and transcript data.

The second motivation for this second corpus was to answer questions about the way that interpretations would depend on the information in context and prosody. Thus, for the cases

of *every*-negation that I gathered from National Public Radio, I gathered crowd-sourced interpretations of the item as text alone, as text in context, the item alone in audio form, and the item in context in audio form. In Section 4.2.3.2, I discuss the frequency of items in the corpus and compare the patterns of preferred interpretations in these context and modality conditions.

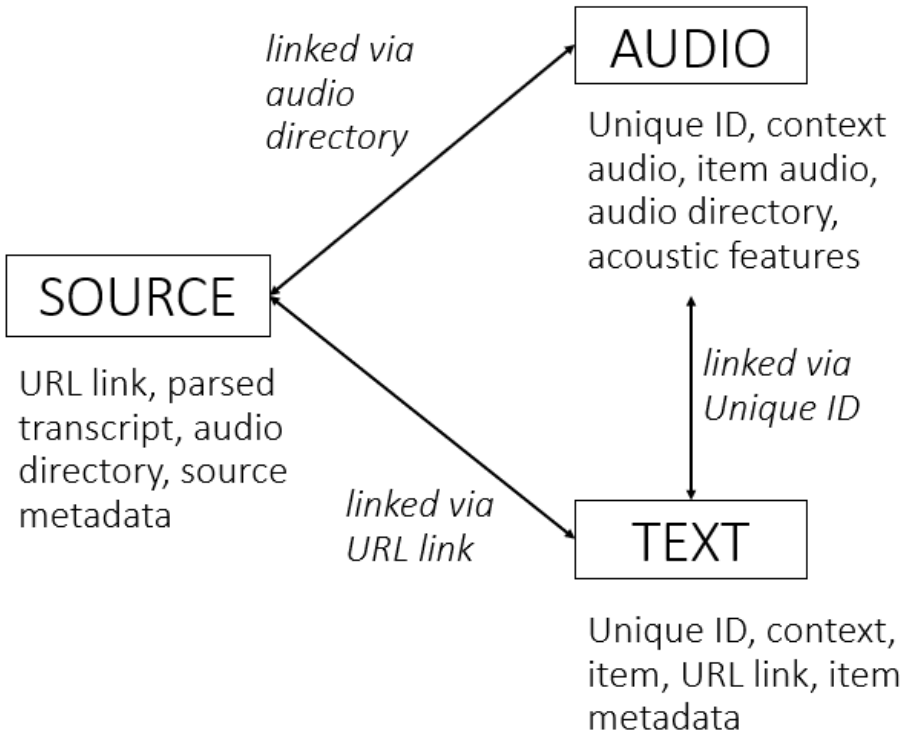


Figure 4.1: Corpus structure, including source, text, and audio information.

Overall in the second corpus, each case of ambiguity was recorded with its text, audio, context in text and audio form, preferred crowd-sourced interpretation depending on context and modality, key acoustic features, and whatever metadata was available from National Public Radio (see Figure 4.1). The result is a rich dataset that is grounded in naturalistic productions, allowing investigations into this ambiguity to take into account their attested frequency and variation in speech and context.

4.1 COCA *every*-negation

How often do people produce *every*-negation, and with what interpretations? Sections 4.1.1 and 4.1.2 describe how a corpus of naturalistic *every*-negation uses was created by mining all *every*-negation occurrences from the radio and TV transcripts in the Corpus of Contemporary American English. Section 4.1.3 then describes an interpretation-annotation experiment where naive participants were asked to indicate their scope interpretations for these *every*-negation uses in their immediate linguistic contexts. Section 4.1.3.3 explores the results and the variation attested in the corpus through examples of different items and interpretation patterns. Section 4.1.3.4 discusses some broader takeaways about naturalistic *every*-negation and the value of considering naturalistic data.

4.1.1 Data source

I used the speech section of the Corpus of Contemporary American English (COCA; Davies, 2015), a commonly-used corpus for English. The spoken section of COCA is made up of transcripts of spoken conversations from American radio and TV programs. The license I used gave me access to ≈9 million clauses, or ≈95 million words, from 1990 to 2012. (Different licenses for COCA provide very slightly different versions of the full corpus. Also, as of the time of writing, the most recent version of the corpus includes data up till the year 2023.)

Radio and TV provide data on a greater variety of speakers than psycholinguistics studies are typically able to access. COCA in particular provides mostly unscripted speech. Its web page reports that at least 95% of the spoken section is unscripted (Davies, 2024). On the other hand, a difference between the COCA data and completely natural conversation, is that speakers knew that they were on a recorded program and may have altered their speech accordingly (e.g., by reducing profanity) (Davies, 2024).

Also, while the corpus is said to be of American English, the data from this corpus isn't limited exclusively to American English dialects but rather to English which was used in American settings. Anecdotally I didn't see evidence of non-American English dialects in the ambiguity cases I extracted from COCA, but there aren't any controls for the speaker dialects in the data. For example, radio such as NPR was one of the sources of data for COCA, and there are some guest speakers on NPR who speak non-American English. So the data is best characterized as majority American English.

4.1.2 Corpus search for *every*-negation

To extract the *every*-negation occurrences, I defined target occurrences as those where quantified subjects precede and c-command sentence negation (with *not* or contracted *n't*). Quantified subjects included any phrase consisting of *every NOUN* (e.g., *every person is not here*) or the quantifier as a prefix (e.g., *everyone/everybody/everything can't be the same*). I also allowed modifications on the quantified subject (e.g., *everything we wish for hasn't yet come about*). Sentence negation included negation of the predicate of the quantified subject (e.g., *every kid doesn't go to an elite college, everybody isn't a famous clown*) or of a silent predicate (e.g., *everyone didn't*, such as in the context of *We thought everyone would become a famous clown, but everyone didn't*). To clarify, according to these criteria, a sentence such as *Everyone tried to not stand in front of the snowball* would have been excluded, since in this case the negation applies to a clause embedded beneath the predicate which begins with *tried*.

To develop the automated search, I randomly selected a year of COCA transcripts and manually searched it for uses of *every*-negation. That is, I used regex to extract every sentence containing the string *every* followed by the string *n't* or *not*, and manually filtered the hits to true hits of the target occurrence. I then wrote a regex search that returned each

of the occurrences in this development set. I applied this search to the rest of the COCA speech section, again hand-checking the results to ascertain true hits and filter out false positives.

In total, I identified 390 instances of *every*-negation. There are ≈9 million clauses in the COCA transcripts, which means that *every*-negation occurs at a rate of slightly less than 0.005%. *every*-negation uses are thus highly infrequent but do in fact occur in everyday English conversation.

4.1.3 Experiment 3: Preferred interpretations of *every*-negation

Is *every*-negation, as attested in everyday conversation, indeed ambiguous? If so, is any interpretation preferred? To answer these questions, I annotated the *every*-negation corpus with crowd-sourced scope interpretations.

4.1.3.1 Methods

Each of the 390 *every*-negation items was annotated with its preferred interpretation. Following Degen (2015), I gathered interpretations by asking participants to judge utterances in their immediate linguistic context. Interpretations were measured on a sliding scale using a version of the paraphrase-endorsement methodology used by Scontras and Goodman (2017).

Participants. 390 participants were recruited through Amazon.com's Mechanical Turk (MTurk) crowd-sourcing service, who had U.S. IP addresses and at least 95% approval ratings for at least 1,000 tasks on MTurk. Each participant received $2.00.

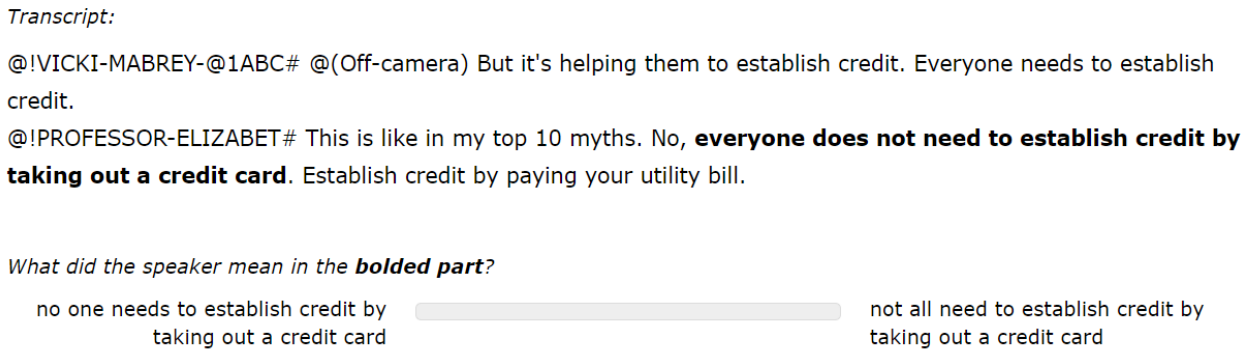


Figure 4.2: Sample paraphrase-endorsement trial from the corpus annotation of *every*-negation utterances. Participants saw the potentially-ambiguous phrase in bold (*Everyone does not need to establish credit by taking out a credit card.*), preceded by three sentences (*But it's helping them ...*) and followed by one sentence (*Establish credit by ...*). They were asked *What did the speaker mean in the* ***bolded part****?* They answered on a sliding scale between the paraphrases of the surface scope (*no one needs to establish credit by taking out a credit card*) and inverse scope (*not all need to establish credit by taking out a credit card*) interpretations, appearing in random order on either side of the scale.

Stimuli. An example trial is shown in Figure 4.2. Each of the 390 *every*-negation uses in the corpus was turned into an excerpt consisting of the three preceding sentences (or lines if punctuation was missing), the bolded potentially-ambiguous clause, and one following sentence (or line). For example, in Figure 4.2, the potentially-ambiguous clause is *Everyone does not need to establish credit by taking out a credit card*, the preceding context is *But it's helping them ...*, and the following context is *Establish credit by ....*

Table 4.1 shows examples of the paraphrases that were created for the surface and inverse scope interpretations of each item. The form of these paraphrases was validated in a separate experiment, as described previously in Section 3.4.1. In general, as Table 4.1 shows, for the ambiguous clause *quantified noun phrase–verb–negation–remainder* the surface scope paraphrase was *none/no one/nobody/nothing–verb–remainder* and the inverse scope paraphrase was *not all/not all things are–remainder*.

Design. The initial instructions asked participants to *choose the best paraphrase for the bolded part* for fifteen randomly-selected items; on each trial, participants were again asked

Subject	**Surface paraphrase subject**	**Inverse paraphrase subject**	**Example sentence**	**Example surface paraphrase**	**Example inverse paraphrase**
everybody	nobody	not all	everybody isn't happy	nobody is happy	not all are happy
everything	nothing	not all things	everything doesn't happen the way you want it to all the time	nothing happens the way you want it to all the time	not all things happen the way you want them to all the time
everyone	no one	not all	everyone can't be a military wife	no one can be a military wife	not all can be military wives
every NOUN	none	not all	every doctor can't do it	none can do it	not all can do it
SUBJECT's NOUN	nobody/no one's NOUN	not all people's NOUN	everyone's competing memoirs don't open up all the debates we've been talking about	no one's competing memoirs open up all the debates we've been talking about	not all people's competing memoirs open up all the debates we've been talking about
SUBJECT RELATIVE CLAUSE	none/nobody/no one/nothing (REL. CLAUSE)	not all (REL. CLAUSE)	everyone who uses LSD doesn't jump out of a window, obviously	no one (who uses LSD) jumps out of a window, obviously	not all (who use LSD) jump out of a window, obviously

Table 4.1: Each potentially-ambiguous clause had the form *quantified subject–verb–negation–remainder*. The paraphrases of the surface and inverse scope interpretations depended on the specific form of the subject (rows 2 through 6). When the noun phrase was modified by a relative clause (row 7), that information was kept in the paraphrases.

*What did the speaker mean in the **bolded part**?* Figure 4.2 shows an example trial. Beneath the conversation excerpt, participants rated the best paraphrase as a judgment on a sliding

scale between the surface and inverse scope interpretations. The two scope interpretations were randomly assigned for each item in left-right or right-left order.

Controls. To check that participants were reading and understanding the contexts of the items – and also as a way to demonstrate that context is useful for the task – two control trials were constructed to imitate the items from the corpus. The controls appeared in random order as the first two trials for each participant. These control trials contained clearly disambiguating information about the intended scope interpretation in the surrounding context. The disambiguating information always appeared as a restatement of the speaker's meaning.

The surface scope-disambiguating control item is in (1), and the inverse scope-disambiguating control item is in (2). For clarity, the disambiguating information is italicized, though it was not italicized in the experiment. Participants were considered to pass the surface control by placing the slider closer to the *none* paraphrase than to the *not all* paraphrase; they passed the inverse control by placing the slider closer to the *not all* paraphrase than to the *nobody* paraphrase.

(1) TONHAUSER: The ten board members voted last night. I was really surprised—I thought at least some of them would like Proposition 23. But *all ten of them voted against it.* Basically, **every board member didn't like Proposition 23**. *Not even a single one of them liked it.*

(2) SIDNER: Look, we completely fixed the issue. Indicators have improved across the board. Everybody's happy.
GROSZ: (VOICEOVER) No, **everybody isn't happy**. *Some are happy but others are deeply dissatisfied with what they call a 'band aid solution.'*

The rate of passing both controls was 53%. This relatively low pass rate may have been due to low English reading proficiency, low attention and motivation, or high task difficulty. Though we restricted MTurk participation to US IP addresses and to those MTurk workers who have completed at least 1,000 tasks in the past, and we also only analyzed data from self-reported native English speakers, some participants may not have fluently read English well enough, or they may have lacked motivation or engagement to read the items in detail. Participants in an online study, or on the MTurk platform in particular, may be disengaged with the experiment. A third factor is task difficulty: the paraphrase endorsement task is a kind of complex reading comprehension and logical inference task, because these sentences have multiple logical operators.

With the addition of the two controls, participants completed a total of 17 trials. Analysis was restricted to those participants who passed both controls and indicated English as their only native language. Thus out of the 390 participants, data was assessed for 208 (35% female; mean age: 41).

4.1.3.2 Results

Each item was judged by at least 2 and at most 14 different participants, with an average of between 8 and 9 ratings per item. Although the surface scope paraphrases randomly appeared on the left or right of the sliders, responses on sliders here are transformed and reported as though the surface scope paraphrases always appeared on the left. As a result, the final response measure for each trial varies from 0 (maximum endorsement of the surface scope interpretation) to 1 (maximum endorsement of the inverse scope interpretation).

As shown in Figure 4.3, people show a high degree of interpretation variation for these corpus *every*-negation utterances – suggesting that the construction is certainly potentially ambiguous – but also a general preference for inverse scope interpretations. The left panel of

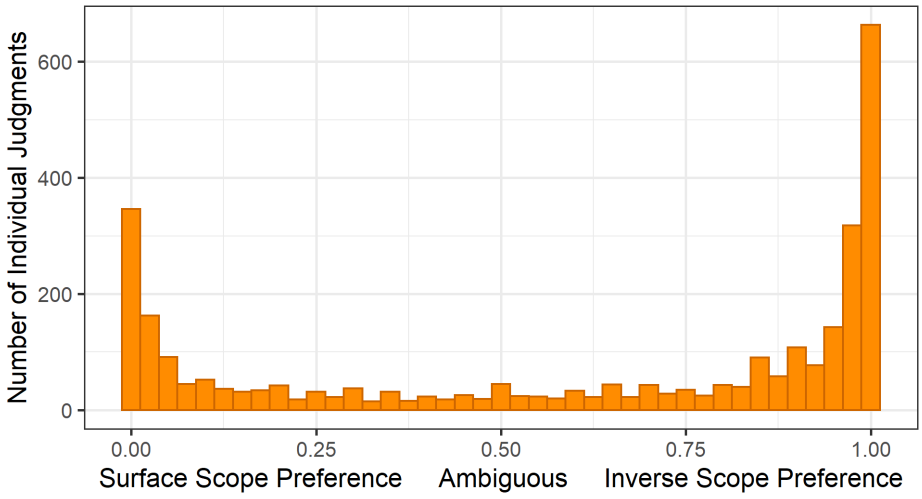
Figure 4.3 shows judgment-by-judgment interpretations. The two peaks at the endpoints suggest that many of these utterances in context elicit strong intuitions such that they are indeed unambiguous for a given listener in context: 29% of individual scores were below 0.25 (indicating a strongly surface scope interpretation) while 53% of individual scores were above 0.75 (indicating a strongly inverse scope interpretation). The right panel of Figure 4.3 shows the mean interpretations per item, and suggests that for some of our items, these strong intuitions are reliable across different participants' judgments: 12% of mean scores were below 0.25, and 38% of mean scores were above 0.75.[1]

4.1.3.3 Exploring naturalistic variation

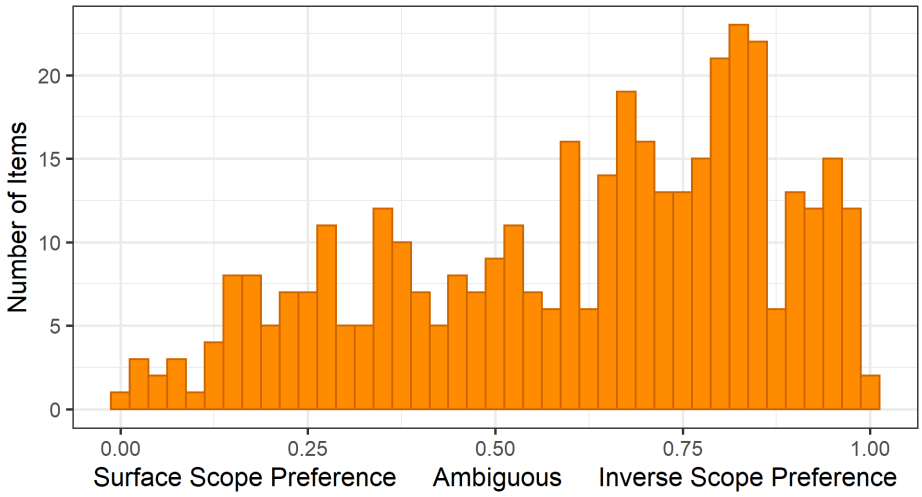
Figure 4.4 shows examples of four attested interpretation patterns: a strong surface scope preference for the item in (3) (top slider; mean response ≈ 0), a strong inverse scope preference for the item in (4) (second slider; mean response ≈ 1), and the two forms of true ambiguity (mean response ≈ 0.5). In (5) (third slider in Figure 4.4), we see high cross-rater disagreement, and in (6) (fourth slider in Figure 4.4) we see high cross-rater agreement. This last interpretation pattern is actually quite rare; in general, participants rarely placed the slider at the midway point between the two interpretation paraphrases, as is evident in the left panel of Figure 4.3.

What exactly does it mean to conclude that *every*-negation is potentially ambiguous? The corpus interpretation patterns suggest that this ambiguity mainly manifests in the fact that 1) there are uses that definitely have surface scope and there are uses that definitely have inverse scope, and 2) listeners can disagree about the intended scope interpretation of a

[1]One question that is raised by this experiment is whether participants inherently prefer the endpoints of the sliders regardless of the task. In another experiment that relied similarly on judgments collected via sliders, with results reported in a later chapter in Section 5.2.3.2, we did see judgments peaking at the midpoint of the scale, with few judgments at the endpoints. Those results demonstrate that participants indeed can use the full range of the scale when they want to; the reason they use the endpoints for the interpretation task here is likely to be clear interpretation preferences.



(a) Individual scope interpretations.



(b) Mean interpretations per item.

Figure 4.3: Distributions of individual and mean item responses from the *every*-negation corpus analysis.

given use in context. In other words it is not that common to observe a third way we might imagine ambiguity manifesting, where 3) listeners tend to agree that a given use in context is ambiguous on its own between two scope interpretations, as in example (6).

(3) (BEGIN VIDEO CLIP, SEPTEMBER 19, 2001) HOWARD LUTNICK, CEO, CANTOR FITZGERALD: Every person who came to work for me in New York,

everyone that was in the office isn't there anymore, every single one who was there isn't there anymore. You can't find them.

a. No one (that was in the office) is there anymore. *(every > n't)*

b. Not all (that were in the office) are there anymore. *(n't > every)*

(4) HOWARD KURTZ: At the risk of suggesting that this is not, perhaps, one of the great technological breakthroughs of the late 20th century, like, say, the microwave oven, the level of hype here has been incredible. I mean buying up 1.5 million copies of the London Sunday Times and giving them out for free? The press has- there's this fascination with high-tech computer subjects. We sometimes forget that **everybody in the world is not on-line**, is not going to go out and buy Windows. @ @ @ @ @ @ @ @ @ @, what does this tell us about the journalistic mind set, this hype?

a. Nobody (in the world) is on-line. *(every > n't)*

b. Not all (in the world) are on-line. *(n't > every)*

(5) Instead, he badmouths people, insults people, and has a crass attitude toward anyone who has got problems, or is weaker than he is as a governor and a wrestler. And I do @ @ @ @ @ @ @ @ @ @ money from it. I think it's unethical.

@!MAN: **Everything I've heard him say has not been ... good**, you know, hasn't been right.

@!MAN: I personally don't think he's taken much time to be governor.

a. Nothing (that I've heard him say) has been good. *(every > n't)*

b. Not all things (that I've heard him say) have been good. *(n't > every)*

(6) Just one week ago, Education Secretary Richard Reilly reported that 90 percent of America's schools like Jonesboro were free from violence. Now Jonesboro has become the sixth time students have fired on fellow students and teachers in the last two and

a half years. And Congress is already talking about new laws to prevent another one.

@(BEGIN-VIDEO-CLIP)

@!SEN-DICK-DURBIN-@: There is no reason why[2] **every child in America shouldn't be protected at least in some small way**, by assuming that every owner of a gun has to own it responsibly, keep it in a safe manner, keep it in a way where it can not be accessed by children.

@(END-VIDEO-CLIP)

@!PRESS: Is it that simple?

a. None (in America) should be protected at least in some small way. *(every > n't)*

b. Not all (in America) should be protected at least in some small way. *(n't >every)*

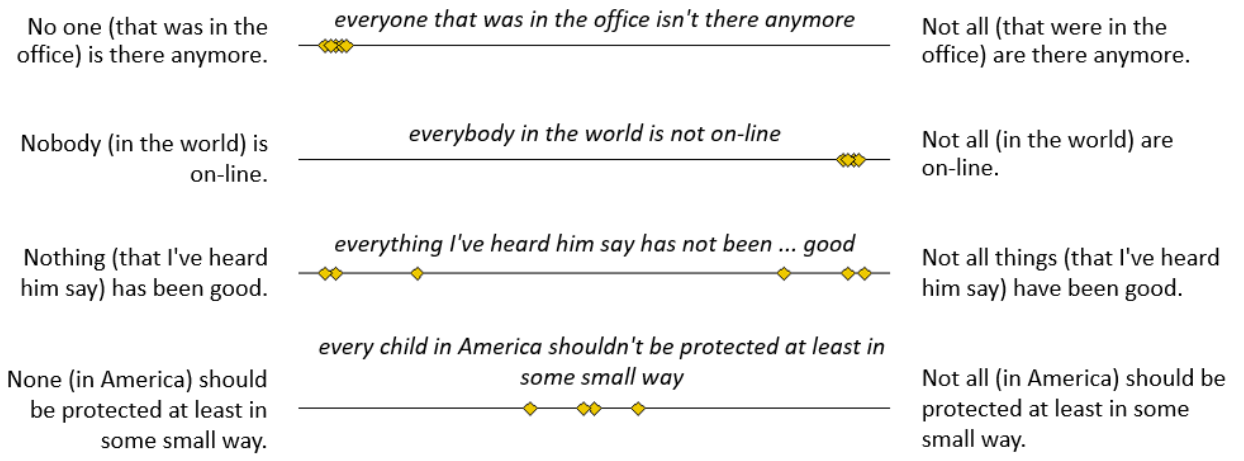


Figure 4.4: Individual interpretations of (3) (top slider), (4) (second slider), (5) (third slider), and (6) (fourth slider). In this figure, the horizontal line represents the length of a slider and each yellow diamond represents an individual judgment. The responses for these four items demonstrate four types of judgment patterns: unambiguous preference for surface scope (top slider) and inverse scope (second slider), ambiguity which reflects judgment disagreement (third slider), and true ambiguity on an individual judgment basis (fourth slider).

[2]It's worth noting that this preceding linguistic structure "There is no reason why" may have made interpreting this item more difficult or confusing to the participants.

4.1.3.4 Discussion

Altogether, naturalistic *every*-negation utterances are attested and potentially ambiguous, with the inverse scope interpretation generally preferred. Specifically, different uses receive a range of average interpretations, though inverse scope interpretations dominate.

Perhaps the main takeaway from the results of the corpus annotation is the variability in interpretations of a single type of utterance, *every*-negation. This picture of ambiguity only emerges when interpretations by many participants for many different items are taken into account, which demonstrates the value of a naturalistic corpus and crowd-sourced annotations. In general, to better understand linguistic ambiguity, this corpus study shows the value of data that include multiple instances of the same type of ambiguity and multiple judgments of each instance of the ambiguity. An individual judgment for a single utterance may not provide enough information about how people prefer to interpret that type of utterance, and the preferred interpretations of a single item or of a single listener might not form a reliable basis for predicting preferred interpretations of other items or other listeners. These utterances are ambiguous – that means interpretation variability across items and listeners.

4.1.4 Experiment 4: Preferred interpretations of *every*-negation as text or speech, with or without context

How do preferred interpretations of naturalistic *every*-negation depend on the amount of information in speech and context? To what extent would an inverse scope preference remain evident if people judged the speaker's meaning without knowing the immediate linguistic context, and how much would hearing the additional information in the speaker's speech affect interpretations? To answer these questions, I searched for the original audio recordings of the *every*-negation items by cross-referencing their metadata and transcripts with archives

of radio and TV that are available online, finding the original audio for 63 (out of the 390) items. I then annotated this subset of the *every*-negation corpus with crowd-sourced scope interpretations of the items as text or speech, with or without context.

4.1.4.1 Retrieving original audio of COCA *every*-negation

In order to retrieve the original audio of each use of ambiguity, I ran the metadata and transcripts of the ambiguities through YouTube, NPR, CNN, NBC, and a general Google search. In a minority of cases, I found an audio or video recording of the TV or radio segment that contained the ambiguity. I used Audacity to extract the sections of the audio file which corresponded to 1) the ambiguity itself and 2) the preceding context, ambiguity, and following context. In total, I recovered the original audio of 63 cases (out of the total 390 *every*-negation items).

To ensure a consistent auditory experience across all audio files used in the study – so that participants in a behavioral study using the audio as stimuli wouldn't need to adjust the volume as they proceeded through the experiment – I used Adobe Audition to normalize the loudness of the set of recordings. The normalization involved adjusting the integrated loudness of each file to a standardized target of -16 LUFS, where LUFS (Loudness Units relative to Full Scale) is a standard measurement that reflects perceived loudness and -16 is a standard commonly employed for music and podcast audio. This normalization process was designed to adjust the overall loudness level without altering the dynamic range within the files themselves, thus preserving the original prosody and any variations across the audio content.

4.1.4.2 Methods

Similarly to Experiment 3, each of the 63 *every*-negation items for which I found audio was annotated with its preferred interpretation, but in different modality (text or audio) and context (with or without context) conditions. As before, interpretations were measured on a sliding scale using a version of the paraphrase-endorsement methodology used by Scontras and Goodman (2017). Applying the corpus-annotation methods from Degen (2015) to audio as well as text data, I gathered interpretations by asking participants to judge utterances with or without their immediate linguistic context, as text or audio.

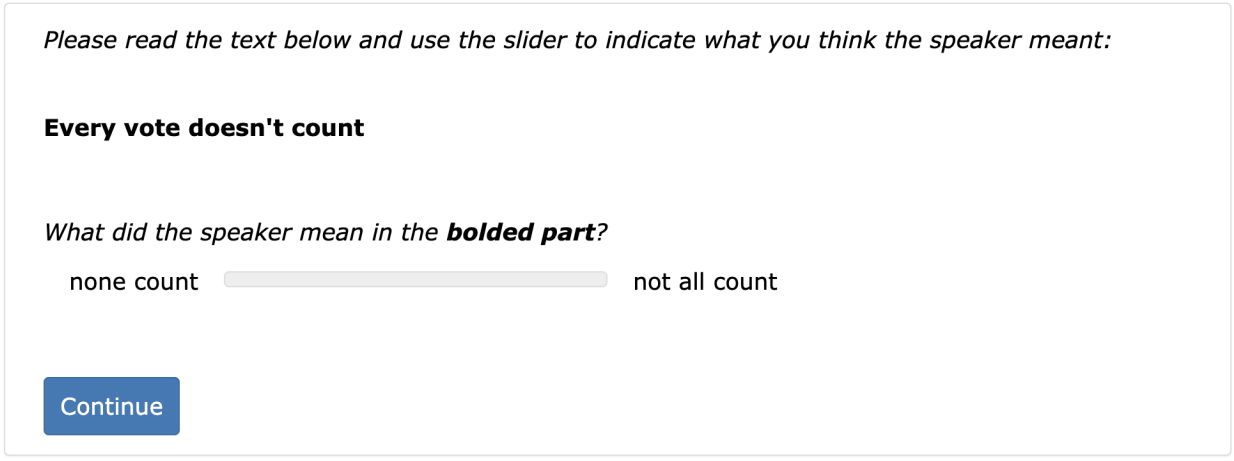
Participants. 108 participants were recruited through Prolific.com's (Prolific) crowd-sourcing service, who had U.S. IP addresses and indicated that they were monolingual English speakers. Each participant received $2.00. I switched to using Prolific, rather than MTurk, primarily because there may be more engaged users on this platform (so, for example, a higher pass rate on our controls).

Stimuli. The text-in-context stimuli were identical to Experiment 3 (an excerpt consisting of the three preceding sentences, the bolded potentially-ambiguous clause, and one following sentence). In contrast to Experiment 3, there was also an audio version of this excerpt (for the audio-in-context stimuli), and text and audio versions of the potentially-ambiguous clause on its own.

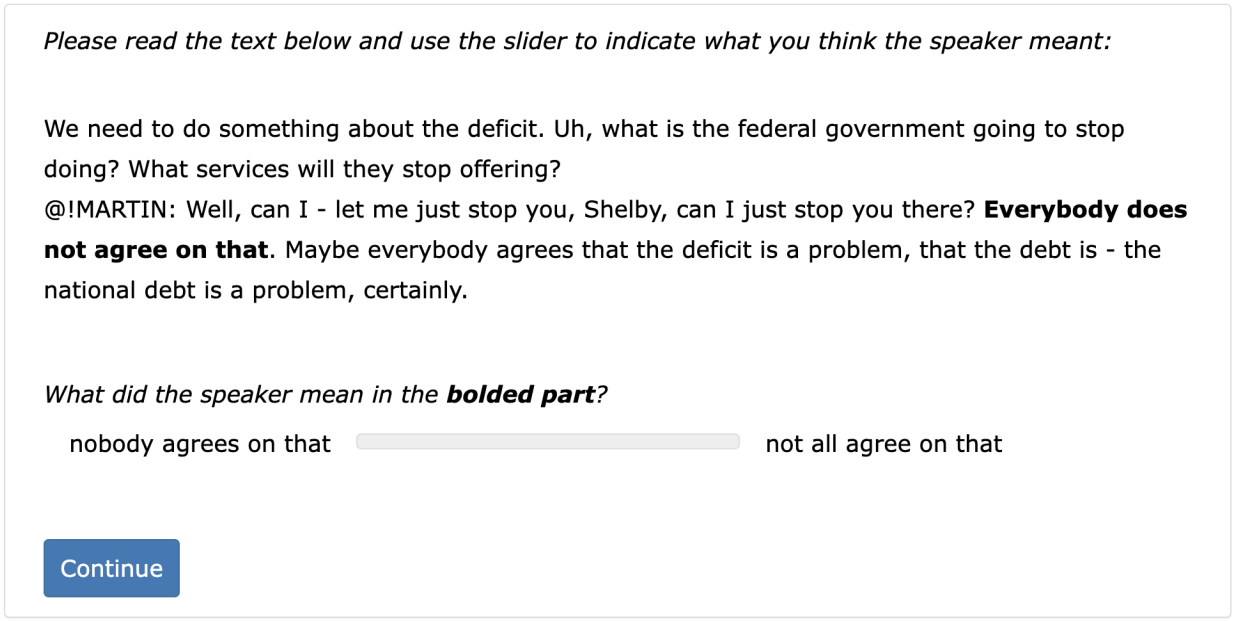
The paraphrases of the surface and inverse scope interpretations of each item were the same as in Experiment 3 (see Table 4.1).

Design. As in Experiment 3, the initial instructions asked participants to *choose the best paraphrase for the bolded part*, and on each trial, participants were again asked *What did the*

*speaker mean in the **bolded part**?* Participants judged the speaker's intended meaning on a sliding scale between paraphrases of the item's surface and inverse scope interpretations. The interpretation paraphrases appeared in random order on either side of the slider. In contrast to Experiment 3, participants judged twenty items: five randomly-selected items in each of four conditions (text-alone, audio-alone, text-in-context, audio-in-context), with the conditions scrambled. An example trial in each condition is shown in Figure 4.5.

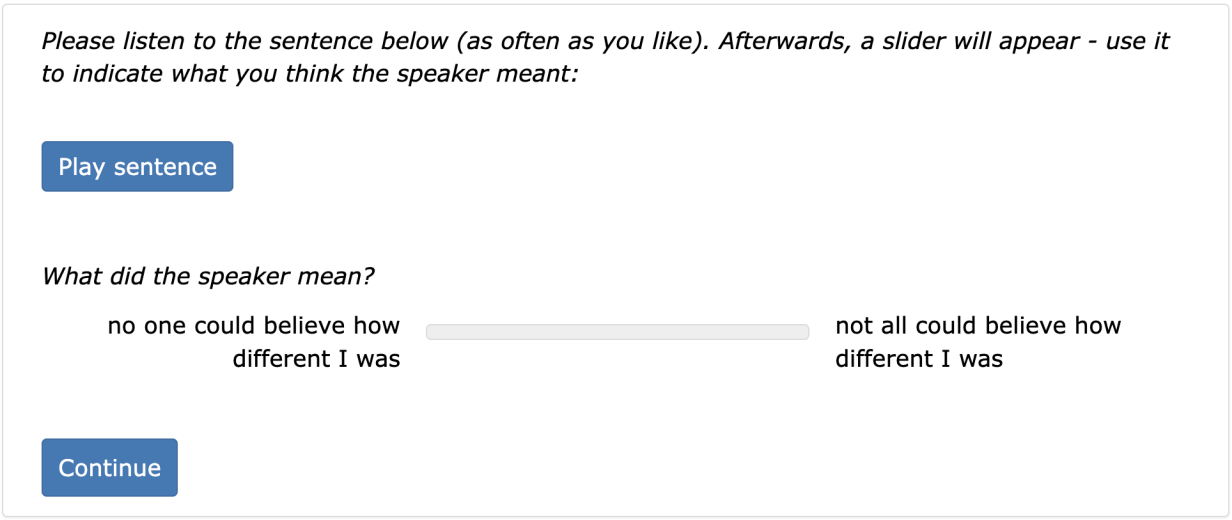


(a) Sample trial of the text without context condition.

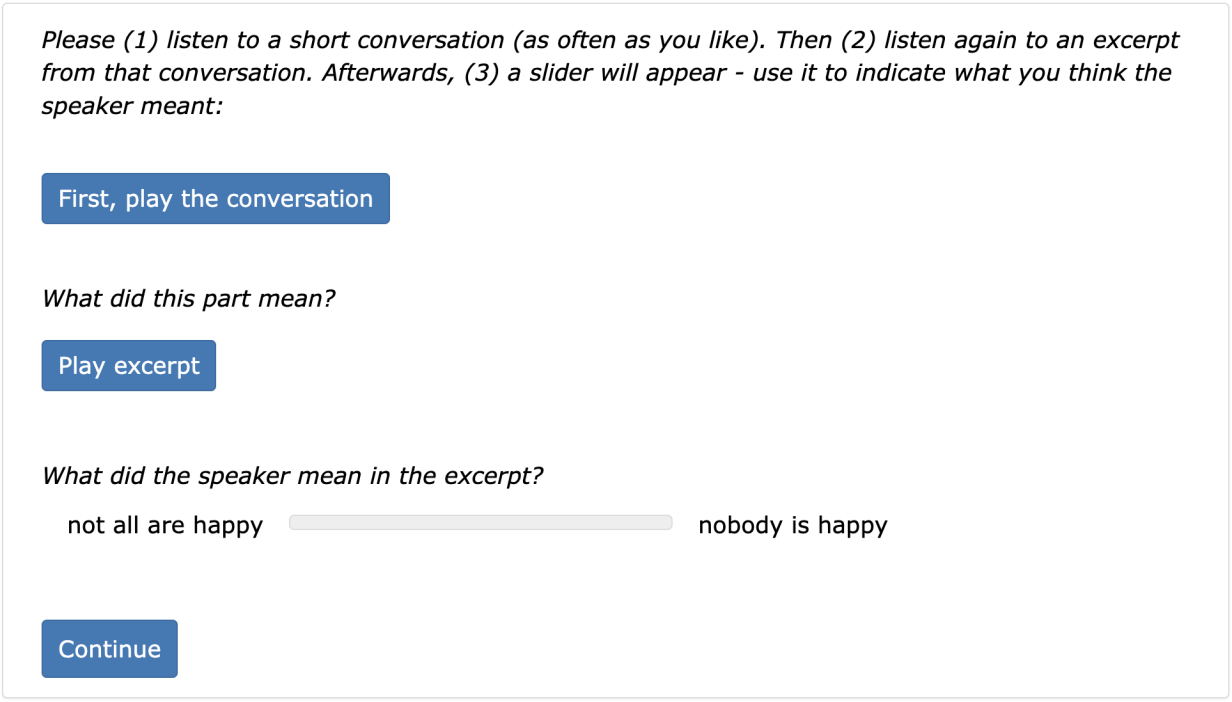


(b) Sample trial of the text with context condition.

Figure 4.5: Sample trials from the experimental task in each of the four conditions (text-only, audio-only, text-in-context, and audio-in-context) – continued on the following page.



(c) Sample trial of the audio without context condition.



(d) Sample trial of the audio with context condition.

Figure 4.5: Sample trials from the experimental task in each of the four conditions (text-only, audio-only, text-in-context, and audio-in-context) – continued from previous page.

Controls. To check that a participant's audio played correctly and at the right volume, participants first clicked a button that played a word (*computer*). Participants were then asked to type what they heard. They were considered to pass this control as long as their

response was *computer* or a plausible misspelling of the same word.

The attention and understanding controls, appearing in random order as the first two trials after the audio control, were identical to those in Experiment 3: the surface scope-disambiguating one (1) and the inverse scope-disambiguating control (2). Participants were considered to pass the surface control by placing the slider closer to the *none* paraphrase than to the *not all* paraphrase and vice versa on the inverse control. However, in contrast to Experiment 3, a randomly-chosen control appeared in audio rather than text form.

The rate of passing the audio control was 100%. The rate of passing both attention controls was 87%, which compares favorably with the past rate of passing on MTurk (53%).

Including the three controls, participants completed a total of 23 trials. Analysis was restricted to those participants who passed both the audio and item controls. Thus out of the 108 participants, data was assessed for 94 (57% female; mean age: 39.9 years old).

4.1.4.3 Results

Each item was judged by at least 2 and at most 15 different participants in each condition, with an average of between 7 and 8 ratings per item. As with the Experiment 3 results, the interpretation response varies from 0 (maximum surface scope preference) to 1 (maximum inverse scope preference).

The interpretation patterns are qualitatively similar to those in Experiment 3: wide variation but a general preference for inverse scope interpretations. This pattern is especially evident in the distribution of individual responses, which Figure 4.6 shows for the four modality and context conditions: regardless of condition, participants preferred to place sliders near an endpoint, using the inverse scope paraphrase endpoint. These peaks near the endpoints suggest that many of these utterances – whether encountered with or without context, in

audio or text form – elicit strong intuitions for a given listener: 27% of individual scores were below 0.25 (indicating a strongly surface scope interpretation) while 66% of individual scores were above 0.75 (indicating a strongly inverse scope interpretation). In other words, the general confidence of interpretations observed in Experiment 3, for items that were only text-in-context, was not due solely to the fact that participants encountered these items in context, since the items in this experiment were encountered both with and without context.

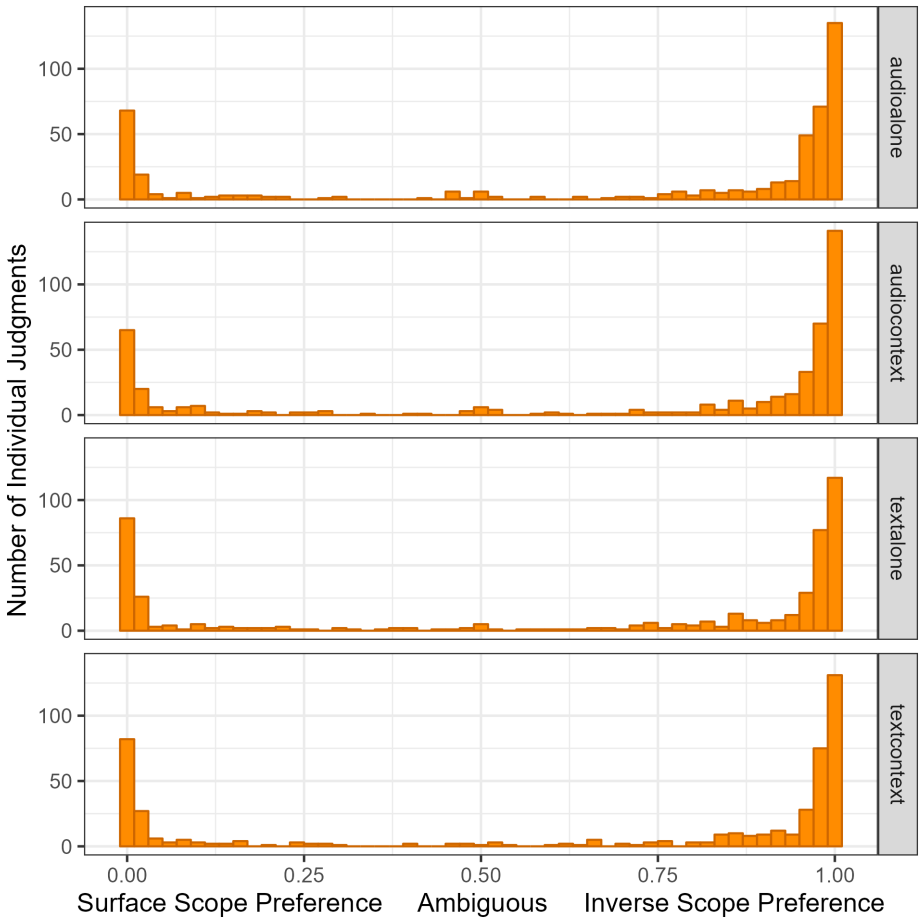


Figure 4.6: Individual scope interpretations, in each of four modality and context conditions, for the subset of the COCA *every*-negation corpus for which there was audio.

The average interpretations per item, in comparison to individual judgments, generally reflect a greater ambiguity per item, but still show the inverse scope preference. Mean interpretations were calculated using the non-parametric bootstrap method from the Hmisc package in R (Harrell Jr and Harrell Jr, 2019). This method was chosen as it provides a robust estimate

of mean interpretations that accounts for variability in the data, particularly when dealing with non-normal distributions or small sample sizes, and allows for more accurate confidence intervals around the means, for greater reliability in interpreting the results.

Figure 4.7 shows mean item interpretations in the four conditions, which in all conditions are distributed across the full range of potential interpretations, but peak at the range of responses that show inverse scope preference. For some items, the strong intuitions in the individual interpretations are still reliable across different participants' judgments: 15% of mean scores are below 0.25, and 56% of mean scores are above 0.75. These numbers indicate greater confidence of mean scores relative to Experiment 3 interpretations, for which 38% of mean scores were above 0.75 and 12% of mean scores were below 0.25.

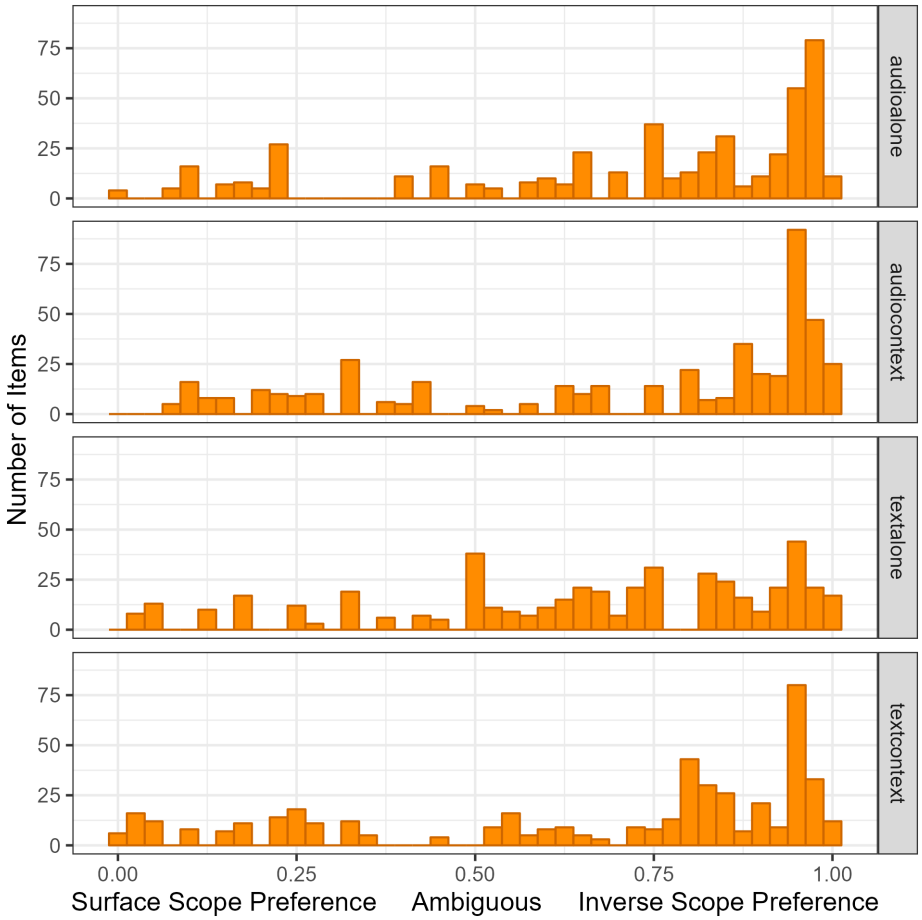


Figure 4.7: Mean interpretation per item, in each of four modality and context conditions, for the subset of the COCA *every*-negation corpus for which there was audio.

Figure 4.8 compares the distributions of mean item responses across conditions. The comparison suggests that items that are encountered without context are generally more ambiguous than when they are encountered with context: the audio-only and text-only condition distributions (the red and blue distributions in Figure 4.8, or rows 1 and 3 in Figure 4.7) are less peaked than their in-context counterparts (the green and purple distributions in Figure 4.8, or rows 2 and 4 in Figure 4.7). These differences suggest that listeners tend to agree more about the interpretation of an item when there is context; in other words, that context provides disambiguating information. Similarly, but to a lesser extent, the audio-only distribution seems more peaked than the text-only distribution, and the audio-in-context distribution seems more peaked than the text-in-context distribution, which would suggest that listeners tend to agree more about the interpretation of an item when there is audio; that audio and the prosodic information it provides also help to disambiguate.

To test the extent to which a condition facilitated disambiguation, one method is to calculate KL divergence from a uniform distribution as a measure of how each condition facilitated disambiguation. This measure is closely related to entropy – the uncertainty within each distribution – but has the advantage of providing a direct comparison of each distribution against a representation of complete ambiguity (the uniform distribution). Divergence from a uniform distribution increased between conditions in the following order: text-alone (0.202) $<$ text-context (0.324) $<$ audio-alone (0.487) $<$ audio-context (0.514).

That is, context added some confidence, audio added even more confidence, and the combination of context and audio information added the most amount of confidence. These two sources of information are partially redundant, as their contributions aren't additive. To return to Figure 1.1, which visualizes two ways of thinking about how prosody and context might relate to interpretations, this preliminary comparison of distributions suggests that prosody and context are certainly two sources of disambiguation in their own right, rather than the alternative, which is that context is the source of disambiguation and that prosody

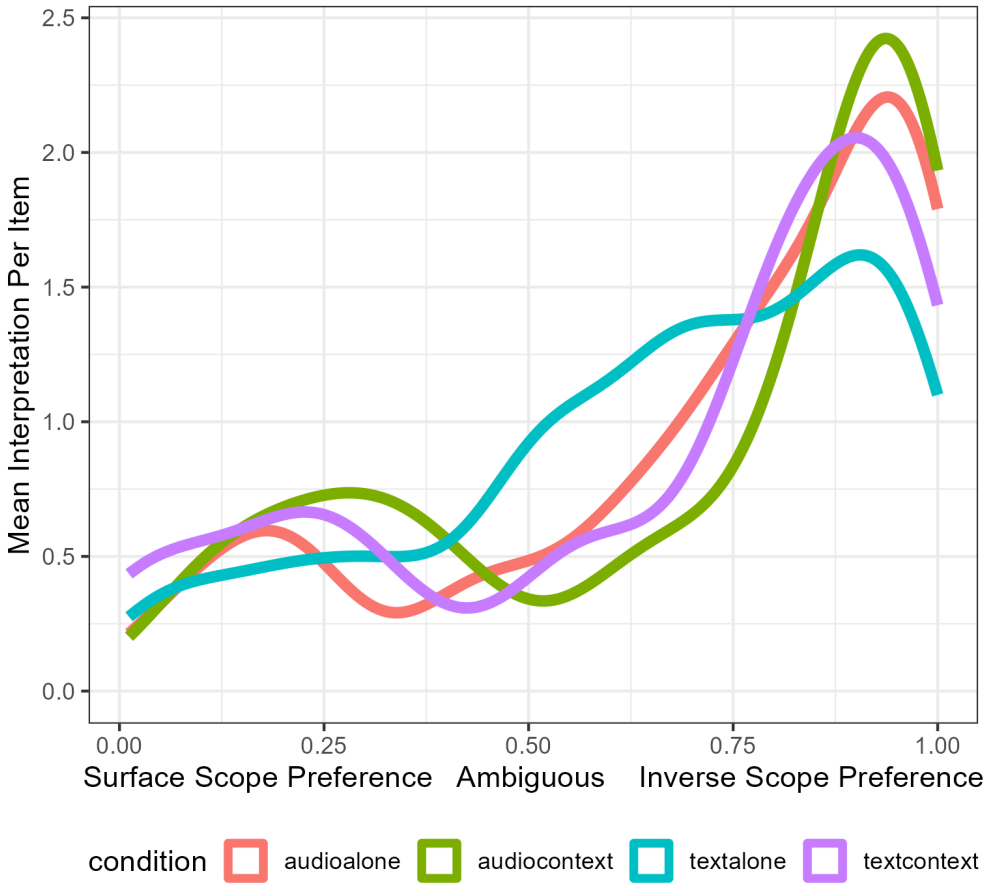


Figure 4.8: Mean interpretations per item for the COCA *every*-negation corpus across the four interpretation conditions.

is merely redundant with context.

Another method, for comparing how conditions facilitated disambiguation, is to compare the amount of additional information provided by context and prosody. For this comparison, I coded a variable INT-DIFF for each item, that encodes the absolute value difference in interpretations between the text-only condition and the three other conditions. For example, for a hypothetical item that received an average interpretation of 0.6 – 60% inverse – in text-only, 0.8 in text-in-context, 0.9 in audio-only, and 0.9 in audio-in-context, the corresponding int-diff values would be 0, 0.2, 0.3, and 0.3. I then used a mixed effects model predicting INT-DIFF by an interaction of CONTEXT and MODALITY, with random intercepts for item. Results are shown in Table 4.2.

The addition of context, the modality through which the utterance is delivered, and their interaction significantly influence differences in interpretation. For this model a main effect of context indicates how moving from a context-less to in-context interpretation affects the interpretation differences when the modality is held constant at text. There was a positive main effect of context: for text, the addition of context significantly changes interpretations.

For this model a main effect of modality shows the effect of switching from text to audio for interpretations made without context. There was a positive main effect of modality: the audio modality increases the interpretation differences compared to text. In other words, the presence of audio, like the presence of context, leads to greater interpretation differences.

Furthermore, there was a significant interaction between context and modality. The negative coefficient means that the effect of introducing context on the variability of interpretation differences is less pronounced when the modality is audio compared to when it is text. This is consistent with the idea that the audio information is partially redundant with the context information: when context is added to text-only ambiguity, interpretations change more than when listeners already had the audio information.

Fixed Effects	Estimate	Standard Error	t-value	p-value
Intercept	0.001424	0.01202	0.118	0.906
Modality	0.1455	0.005240	27.764	<2e-16
Context	0.1592	0.005254	30.304	<2e-16
Modality*Context	-0.1416	0.007424	-19.067	<2e-16

Table 4.2: Results of a mixed effects model with modality (text/audio) and context (no context/in context) predicting for each item the difference of its mean scope interpretation from its text-only, no-context interpretation (higher values indicating greater change), with random intercepts for item.

4.1.4.4 Discussion

Experiment 4 investigated how the presence of context and modality influences interpretations, using the subset of the 390 *every*-negation items for which the original audio was available. Individual interpretation distributions showed that these ambiguities still elicit strong intuitions about the speaker's intended meaning, regardless of whether context or audio information is available to the listener: listeners rarely indicated that a particular utterance was totally ambiguous to them. In other words, as in Experiment 3, individual interpretations don't tell the full picture for these ambiguities; we need to look at mean interpretations per item. In Experiment 4 as in Experiment 3, mean interpretation distributions show ambiguity as disagreement across different listeners' interpretations.

Here, as expected, context and audio play a clear disambiguating role: items interpreted in context or with audio, as opposed to those without context or without audio, received greater interpretation agreement. A mixed-effects model and entropy analysis provide additional insight into the degree of interpretation agreement in different conditions: context provides more confidence and greater interpretation agreement than audio, and the audio information is partially redundant with the context information.

All this said, a limitation of these results is that they are based on relatively few data (only 63 of the total 390 items from the corpus). The next section turns to creating a larger multimodal corpus, in order to replicate Experiment 4 on a larger dataset.

4.2 A larger multimodal corpus from NPR

4.2.1 Data source

I extracted the target occurrences from publicly available American radio, specifically, two podcasts from National Public Radio (NPR) archives, which have predominantly spontaneous speech (e.g., interviews) as opposed to speech that was read aloud. One was the All Things Considered podcast from the years 2008-2022. These archives consist of transcripts and associated recordings of spoken news, interviews and conversations (≈4.5 million clauses). Note that while transcript archives go back as far as the year 1990, I only mined for target occurrences from those years of archives for which associated audio was also available (during the year 2008 and later).

The second podcast was Fresh Air from the years 2008-2021. These archives consist of transcripts and associated recordings of spoken conversations between hosts and guests on subjects like popular culture, news, and issues (≈2 million clauses). Similarly, I only mined for target occurrences from those years of archives for which associated audio was also available (no earlier than 2008).

Like the COCA data, these NPR podcasts appear to provide mostly unscripted speech, although I could not find a concrete measure of the percentage of these NPR transcripts which are indeed unscripted. In my estimation, the majority of the speech in the interviews appear spontaneous, as opposed to scripted speech from the hosts or recordings played from other clips. An advantage of the NPR data is that non-spontaneous conversational speech is marked (i.e., by a "Singing" or "Reading" tag), so results could be filtered to only those cases which weren't marked this way. Like COCA, the NPR speech is also mainly but not exclusively North American English dialects.

Also, anecdotally, a bias in the NPR speech data is a skew towards apparently affluent and

educated speakers, though there is still a range of speakers of different ages and backgrounds. While it was difficult to find demographic data on the NPR speakers, perhaps one estimate of the speaker demographics is the NPR listener demographics. According to a 2012 Pew Research Center survey, the NPR audience tends to be highly educated (majority college graduates), about average in terms of gender (51% female), and have above-average incomes (NPR, 2012).

4.2.2 Corpus search for *every*-negation

As before, target occurrences for Corpus 2 were defined as any case where the *every*-quantified subject precedes and c-commands sentence negation (with *not* or contracted *n't*).

To develop a reliable automated search for the text of target occurrences in NPR, I created a new search that could match to 100% of the cases of *every*-negation in Corpus 1 (so Corpus 1 became a kind of development set for Corpus 2). In contrast to the regex search I used to create Corpus 1, the new automated search used spaCy, an open-source Python library that parses text for characteristics such as parts of speech and dependencies. Figure 4.9 shows an example of the most common dependency pattern that spaCy identified for *every*-negation.

I applied this search to the NPR All Things Considered archives to extract the text and context of target potential ambiguities, finding 270 potential *every*-negation cases. Trained annotators read through all the potential hits and verified the true matches, while also filtering out repeated cases of the same *every*-negation. (These repetitions were possible in the original data because sometimes the radio station ran the same segment multiple times.) The number of total, unique cases of *every*-negation was 213, making the search precision relatively high at 79%.

I then extracted the audio recordings associated with the web pages of the true hits and

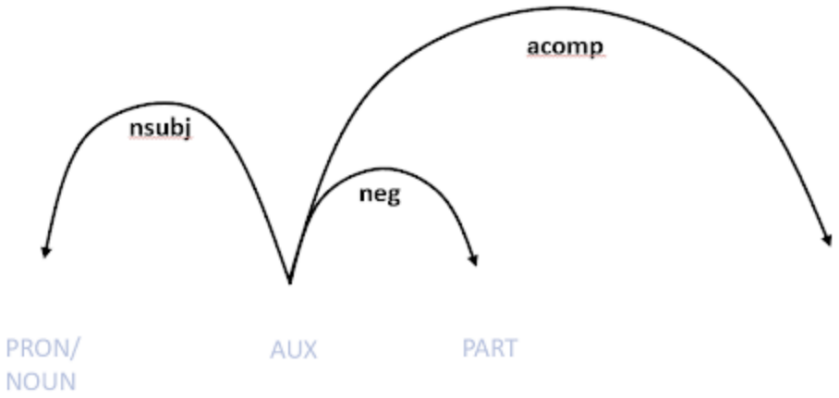


Figure 4.9: One of the spaCy dependency patterns which most often characterizes a true case of *every*-negation. The key aspect of the sentence, which is expressed by the dependency pattern, is that there's a single expression which is both a noun subject (*nsubj*) and negation (*neg*) in the same clause.

used the Montreal Forced Aligner (McAuliffe et al., 2017) to approximate the time stamps in each audio file which aligned with the context and text of the potential ambiguity. Trained linguists checked the alignment between each text and audio pair, using Praat or Audacity to recut the automated audio file where necessary. Overall, for each case of *every*-negation, two audio files were cut: one corresponding to the *every*-negation text on its own and one corresponding to the text in context. During this audio extraction and alignment process, I also filtered out any cases where the speaker was reading or singing the *every*-negation, or where there was no associated audio due to an error in the archive. In total, the number of unique cases of *every*-negation, which were neither reading nor singing cases, and for which we found the original audio, was 204.

I applied the same search to the NPR Fresh Air archives, finding 157 potential *every*-negation cases. Trained linguists read through all the potential hits and verified the true matches, while also filtering out repeated cases of the same *every*-negation, of which there were more in Fresh Air than in All Things Considered. The number of total, unique cases of *every*-negation was 91. Audio files were then extracted and automatically aligned, and trained linguists

checked and fixed the alignments and filtered out any cases where the speaker was reading or singing the *every*-negation, or where there was no associated audio due to an error in the archive. In total in the Fresh Air archives, the number of unique target cases with audio was 83. Overall, adding together the cases from All Things Considered and Fresh Air, Corpus 2 has 287 items.

As with the audio items in Corpus 1, to ensure a consistent auditory experience across all audio files without altering the dynamic range within the files themselves, I normalized the loudness of the set of audio files in Corpus 2 to an integrated loudness of -16 LUFS. Here, I used FFmpeg, a powerful multimedia framework, to analyze and adjust the loudness. Each file was processed with the ebur128filter, which adheres to the EBU R128 loudness standard, to measure its integrated loudness. Adjustments were then made to each file's amplitude to align its loudness to the target of -16 LUFS.

4.2.3 Experiment 5: Preferred interpretations of *every*-negation as text or speech, with or without context

Experiment 5 replicates Experiment 4 on the full Corpus 2: the experiment gathers crowd-sourced scope interpretations of *every*-negation as text or speech, with or without context. The goal, as before, was to understand how preferred interpretations of naturalistic *every*-negation depend on the amount of information in speech and context.

4.2.3.1 Methods

Similarly to Experiment 4, each of the 287 *every*-negation items from NPR was annotated with its preferred interpretation in different modality and context conditions. As before, participants indicated their preferred interpretation by choosing paraphrases of the two scope

interpretations on a sliding scale.

Participants. 660 participants were recruited through Prolific.com's (Prolific) crowd-sourcing service, who had U.S. IP addresses and indicated that they were monolingual English speakers. Each participant received $2.00.

Stimuli. The form of the paraphrases of the surface and inverse scope interpretations was the same as in Experiments 1 and 2 (see Table 4.1). The form of the stimuli in each of the four modality and context conditions was the same as in Experiment 4 (see Figure 4.5).

Design. As in Experiment 4, participants judged twenty items randomly distributed across the four context and modality conditions. The initial instructions asked participants to *choose the best paraphrase for the bolded part* and on each trial, participants were again asked *What did the speaker mean in the* ***bolded part****?* Beneath the text or audio button, participants rated the best paraphrase as a judgment on a sliding scale between the surface and inverse scope interpretations. The two scope interpretations were randomly assigned for each item in left-right or right-left order.

Controls. As in Experiment 4, participants first saw a task that tested whether they could play audio and indicate correctly what they heard. They then saw two attention and understanding controls in random order, one in audio and one in text form.

The rate of passing the audio and attention controls was 82%, which compares well with previous pass rates.

With the addition of the two controls, participants completed a total of 23 trials. Analysis was restricted to those participants who passed both the audio and attention controls. Thus

out of the 660 participants, data was assessed for 530 (52% female; mean age: 43.5 years old).

4.2.3.2 Results

Each item in each condition was judged by at least 1 and at most 20 different participants, with an average of between 9 and 10 ratings per item and condition. As with the Experiment 3 and 4 results, the interpretation response varies from 0 (maximum endorsement of the surface scope interpretation) to 1 (maximum endorsement of the inverse scope interpretation).

The results largely replicate those of Experiments 3 and 4. Figure 4.10 shows the distribution of individual responses in the four modality and context conditions: 30% of individual scores were below 0.25 (indicating a strongly surface scope interpretation) while 63% of individual scores were above 0.75 (indicating a strongly inverse scope interpretation). In other words, as before, regardless of whether an individual judgment was made with context or audio information, people tended to clearly resolve the ambiguity, usually with inverse scope.

.75.75

Figure 4.10: Individual scope interpretations, in each of four modality and context conditions, for the NPR *every*-negation corpus.

Figure 4.11 shows mean item interpretations in the four conditions: overall, 12% of mean scores are below 0.25 (similar to the results in Experiments 3 and 4), and 47% of mean scores are above 0.75 (similar to the results in Experiment 4).

In comparing the distributions of mean item responses across conditions, it seems that items that are encountered without context are generally more ambiguous than when they are encountered with context: the audio-only and text-only condition distributions (rows 1 and 3 in Figure 4.11) are the least peaked, while the audio-in-context and text-in-context condition distributions (rows 2 and 4 in Figure 4.11) are more peaked. That is, listeners tended to agree more about the interpretation of an item when there was context. Similarly, but to a lesser

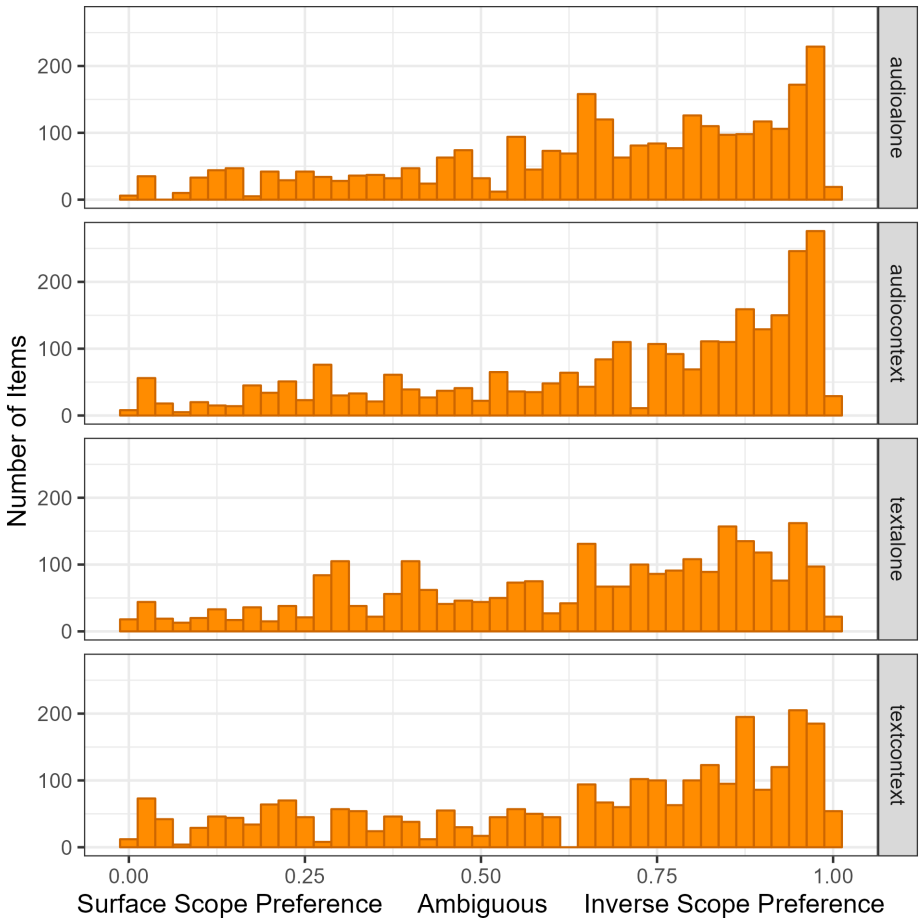


Figure 4.11: Mean interpretation per item, in each of four modality and context conditions, for the NPR *every*-negation corpus.

extent, the audio-only distribution seems more peaked than the text-only distribution, and the audio-in-context distribution seems more peaked than the text-in-context distribution.

I again used KL divergence from a uniform distribution as a measure of how each condition facilitated disambiguation. Divergence from a uniform distribution increased between conditions in the following order: text-only (0.16) > text-context (0.20) > audio-only (0.24) > audio-context (0.29). That is, as with the COCA data, context added some confidence, audio added even more confidence, and the combination of context and audio information added the most amount of confidence. Additionally, as before, these two sources of information are partially redundant, with their KL divergence scores not additive.

Using the interpretation difference measure (which encodes the absolute value difference in interpretations between the text-only condition and the three other conditions), I ran a mixed effects model predicting INT-DIFF by an interaction of CONTEXT and MODALITY, with random intercepts for item. Results are shown in Table 4.3.

As in Corpus 1, the addition of context, the modality through which the utterance is delivered, and their interaction significantly influence differences in interpretation. Again, for this model a main effect of context indicates how moving from a context-less to in-context interpretation affects the interpretation differences when the modality is held constant at text. There was a positive main effect of context: for text, the addition of context increases interpretation differences.

Again, for this model a main effect of modality shows the effect of switching from text to audio for interpretations made without context. There was a positive main effect of modality: the audio modality increases the interpretation differences compared to text. In other words, the presence of audio, like the presence of context, leads to greater interpretation differences.

Furthermore, there was a significant interaction between context and modality. The negative coefficient means that the effect of introducing context on the variability of interpretation differences is less pronounced when the modality is audio compared to when it is text. This is consistent with the idea that the audio information is partially redundant with the context information: when context is added to text-only ambiguity, interpretations change more than when listeners already had the audio information.

4.2.4 Discussion

In spite of variation, both context and prosody contribute significant information to the interpretations of naturalistic ambiguity, with context providing more confidence than audio,

Fixed Effects	Estimate	Standard Error	t-value	p-value
Intercept	0.002876	0.005203	0.553	0.581
Modality	0.1368	0.002204	62.083	<2e-16
Context	0.1604	0.002200	72.917	<2e-16
Modality*Context	-0.1408	0.003110	-45.283	<2e-16

Table 4.3: Results of a mixed effects model with modality (text/audio) and context (no context/in context) predicting for each item the difference of its mean scope interpretation from its text-only, no-context interpretation (higher values indicating greater change), with random intercepts for item.

and with the audio information partially redundant with the contextual information.

Combining together all the items from COCA and NPR, a corpus item was annotated with its information, as available, on its a) text transcript (e.g., *Every man didn't twirl*); b) its immediate linguistic context in text form (i.e., the three preceding and one following sentence); c) original audio (e.g., the .wav file for the production of *Every man didn't twirl*); d) its immediate linguistic context in audio form (i.e., the .wav file for the production of the three preceding and one following sentence); e) ambiguity type (i.e., *every*-negation); f) metadata (date of occurrence, source (e.g., NPR), title of radio or TV segment); g) a direction on how to find the occurrence in the larger data source (from COCA: the text ID of the transcript containing the occurrence; from NPR: a stable web link to the occurrence in the radio archives); h) crowd-sourced interpretation of text alone; i) crowd-sourced interpretation of the text in context; j) crowd-sourced interpretation of the audio alone; and k) crowd-sourced interpretation of the audio in context.

Constructing this corpus allows a variety of questions to be addressed. A broad set of these open questions relates to the naturalistic use of quantifier-negation as an instance of scope ambiguity. Scope ambiguity has been the focus of many linguistic studies, as a case study of the potential through natural language to express meaning that does not directly correspond to the overt order of a surface string of words. Yet there are many open questions about its naturalistic use, including how often scope ambiguity occurs in everyday speech, whether

it is actually ambiguous in context or with prosodic information, and if there is a preferred interpretation when both potential interpretations are attested. The corpus study shows that constructions with verb negation and a subject quantified by *every* are indeed attested in transcripts of conversational speech, although they are not common; that all interpretations preferences are attested across items, though inverse scope is generally preferred (in line with the behavioral results of average scope preference of *Every marble isn't red* from Chapter 3); and that the presence of context or prosody definitely provides disambiguating information, with that information partially redundant between the two information sources.

The next chapter turns to another, specific question that can be explored with this corpus regarding the disambiguating role of context, whether the theoretically-motivated model predictions from Chapter 3 are borne out in naturalistic data.

Chapter 5

Expectations in context

This chapter focuses on how expectations in context account for *every*-negation interpretations, taking a closer look at naturalistic and behavioral data. Context should help to account for some of the wide variation in interpretation preferences of quantifier-negation, especially the universally-quantified quantifier-negation focused on in the literature. But the question is, what specific aspect of context matters, and what role does this aspect of context play in people's computation for a preferred interpretation?

One role of context – expectations about the world – is described by the computational model discussed in Chapter 2 (which is itself based on a past model (Scontras and Pearl, 2021) that seeks to explain some findings from past experimental studies (e.g., Musolino, 1999)). According to the model, expectations about the true state of the world, which are salient to interlocutors and can even be expressed in the preceding linguistic context of the potentially-ambiguous utterance, can predict scope preference, because they make one scope interpretation relatively more plausible and therefore more likely. Indeed, the model fit human behavior when this sort of expectation was included.

The expectations are very specific to the case studies of *every*-, *some*-, and *no*-negation:

for example, a higher believed success rate is one kind of inverse-favoring context for *every*-negation. But the broader computation that interlocutors use is general: they assume that listeners are more likely to arrive at an interpretation that is more likely to be true of the world. This assumption itself rests on the broad understanding that speakers are cooperative.

In this chapter, I take a closer look at the aspect of context from the model in Chapter 2 that helped account for different quantifier-negation interpretations. I recast this modeled role of context in general terms, to show how it fits with naturalistic language data and past behavioral studies. 1) Section 5.1 discusses how the contextual factor from the model, the *world priors*, represents world expectation influencing interpretation plausibility. Focusing on *every*-negation, 2) I revisit the discussion from Chapter 2 about how *believed success rate*, which I term positive expectations here, matters specifically for scope preference (defining positive expectations in Section 5.1.1, discussing corpus-attested examples in Section 5.1.2 and ways of measuring positive expectations in Section 5.1.4). I return to the model from Chapter 2 to show how it systematically predicts that higher positive expectations lead to higher inverse scope preference (Section 5.1.5). I then show how these expectations about plausibility, especially positive expectations, can account for some past findings in the literature (Section 5.1.6).

With this definition in hand for one disambiguating role of context, I focus on how it could account for the interpretations in the *every*-negation corpus. The challenge becomes narrowing this factor of "context" down to a specific expression or measure of positive expectations. Using the COCA corpus, I test the correlation of inverse scope preference with different measures of positive expectations, relying for these measures either on the text itself of the preceding conversation (Sections 5.2.1 and 5.2.2) or on behavioral judgments of the beliefs set up in this text (Section 5.2.3). Section 5.2.4 compares these measures, identifying the behavioral one to be perhaps the better estimate of the idea of positive expectations. I apply this behavioral measure to the NPR corpus (Section 5.2.5), again testing its correlation with

inverse scope preference.

Section 5.3 summarizes the findings with respect to this one disambiguating role of context.

5.1 World expectations affect interpretation plausibility

Context can seem difficult to pin down as a factor, but the RSA models in Chapter 2 suggest that the relative plausibility of competing interpretations explains some variation in interpretation preferences. I use plausibility in the cognitive sense (Saba, 1999), where the cognitive plausibility of an interpretation is its probability of being true according to interlocutors' world knowledge.

As made concrete in the RSA models in Chapter 2, world knowledge influences relative plausibility of an interpretation in the following general way: interpretations are preferred when they are more likely to be true. This preference for true interpretations is based on the conversational goal, shared between speaker and listener, for the listener to arrive at the speaker's intended interpretation of an utterance, plus the shared conversational assumption that speakers tend to say things that are true (e.g., in accordance with the Gricean maxim of quality (Grice, 1975), to be truthful).

General hypothesis about why plausibility disambiguates: plausibility disambiguates because the listener uses world knowledge, plus the assumption that speakers say things that are true, to weight the relative probability of one interpretation over the other according to the relative probabilities that these interpretations are true. A simple execution of this cue is to rule out false states of the world, and therefore to rule out interpretations which are only consistent with false states of the world.

The next section turns to how world expectations affect interpretations specifically for *every-*

negation. The world expectation that facilitates inverse scope preference for these utterances is what I call a high positive expectation, which is the high believed success rate in the model.

5.1.1 High positive expectations

One kind of world expectations is a belief about the success rate of the non-negated predicate as it applies to each entity under discussion. To put this aspect of context in simpler terms, I call the success rate belief a positive expectation. A high positive expectation is the belief that the entities under discussion in fact have the property corresponding to the non-negated predicate. For example, for the *every*-negation utterance *Every horse didn't jump over the fence*, the corresponding high positive expectation is the belief that the horses under discussion are likely to succeed in jumping over the fence. As another example, for *Every vote doesn't count*, a high positive expectation would concern the prior probability that a vote counts: that a vote is likely to count.

The strength of a positive expectation would be the expected probability of success – in other words, the base rate p_r of success in the RSA models. The greater the believed probability of success of the non-negated predicate for the relevant entities, the greater the strength of the positive expectation. For example, suppose you know that the average horse can jump about three feet, while show horses can jump as high as seven feet, with the highest jump recorded at about eight. Imagine that the fence under discussion is a foot high; you would then hold a high (strong) positive expectation for the utterance *Every horse didn't jump over the fence*, in that you would expect that it is highly likely that a horse managed to jump over the fence. Conversely, imagine that the fence is about five feet high and the horses under discussion are ordinary horses; you would expect it to be relatively less likely that a horse managed the jump, so there would be a lower positive expectation.

Or consider the utterance *Every vote doesn't count* as it was used in one of the items from

the ambiguity corpus (with the *every*-negation bolded):

(1) The Democratic Party needs to be democratic far more so than it has been. Coming out of Florida, where every vote counts, where we now have learned a lesson, we can not be a part of an apparatus where **every vote doesn't count**. I believe that it's important that the system and the process be open.

said on CNN in 2000; corpus ID 104

In the context in (1), there is a strong positive expectation that votes count, which is set up by the speaker when they reference a situation in which every vote counts (the text of which is underlined above).

Positive expectations: For a quantifier-negation sentence, the greater the expected success rate of the non-negated predicate as it applies to each entity under discussion, the higher the positive expectation. The strongest version of a high positive expectation is that worlds consistent with the truth of the non-negated expression are `true` – a belief in the truth of the *all* world state – while a positive expectation in general is a belief in the *some* (*at least one*) world state.

This concept of positive expectations helps account for interpretations of *every*-negation (among other constructions). Figure 5.1 illustrates the key prediction, which is that the higher the positive expectation, the stronger the inverse scope preference. For example, if you knew that the fence was a foot high (right side of the figure) rather than five feet high (left side) – if you held a higher positive expectation – you would be more likely to think that *Every horse didn't jump over the fence* was intended with the inverse scope, *not all jumped* interpretation rather than the surface scope, *none jumped* interpretation.

The general intuition is that in the context with a high positive expectation, the speaker's use of *every*-negation is an emphatic way to express the message that the prior expectation does

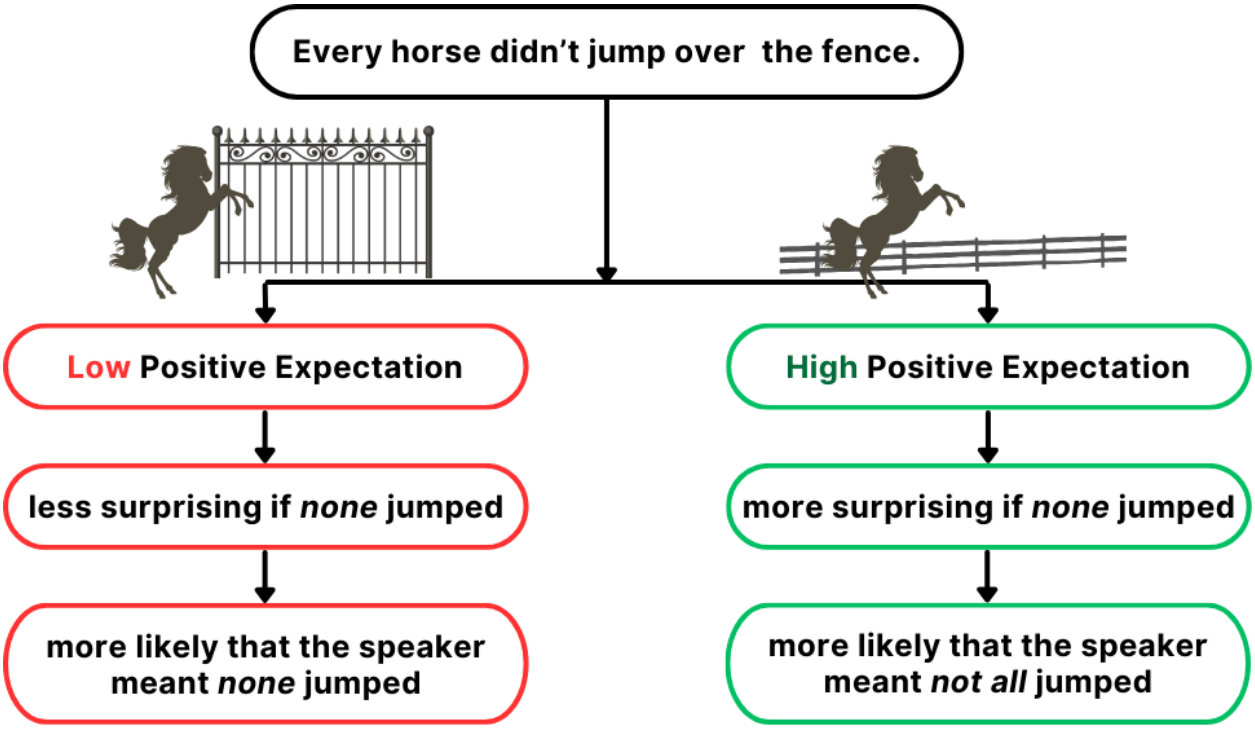


Figure 5.1: Positive expectations are a specific aspect of context that could predict scope preference for *every*-negation. For the utterance *Every horse didn't jump over the fence*, the positive expectation would be relatively low (as in the left panel) in a case where it would be unlikely for the horses under discussion to succeed in jumping over the fence. It would be relatively high (as in the right panel) if the horses were expected to succeed. The higher the positive expectation, the more surprising it would be if indeed *no* horse managed to jump, and the more likely it would be that the speaker of the utterance *Every horse didn't jump* intended to convey that not all, rather than none, succeeded in jumping (because at least one should have made it).

not hold. (All the horses should have jumped easily, the fence was low! Every vote counts, that's the democratic system!) More specifically, the high positive expectation influences interpretations because it increases the plausibility of the inverse scope interpretation relative to the surface scope one. First, holding a high positive expectation means that interlocutors place a greater prior probability on the *some* world states being true. As in the right rather than left side of Figure 5.1: holding a positive expectation for *Every horse didn't jump*, interlocutors expect that *some horses jumped*, so a lowered probability that *no horses jumped*. The two potential interpretations are that *no horses jumped* and *not all horses jumped*. So, believing that some horses jumped, listeners reason that the inverse scope interpretation is

more likely to be true than the surface scope one.

Similarly, holding a positive expectation for *Every vote didn't count*, interlocutors expect that *at least some votes counted*, so a lowered probability that *no votes counted*. The two potential interpretations are that *no votes counted* and *not all votes counted*. So, believing that some votes counted, listeners reason that the inverse scope interpretation is more likely to be true than the surface scope one.

From a listener's perspective, believing that the inverse scope interpretation describes world states that are more likely to be true, and assuming that speakers intend to say things that are true, the inverse scope interpretation then becomes more likely. This interpretation bias on the listener's part also leads speakers (who want to successfully guide listeners to their intended interpretation) to be more willing to use *every*-negation as a description of an inverse-verifying scenario given a high positive expectation.

All of this is a bit abstract, so in the next section we consider additional concrete examples of high positive expectations in the ambiguity corpus (Section 5.1.2). These corpus-attested examples demonstrate in particular that high positive expectations can be made salient at different levels of belief – they are not necessarily held to be true by both (or even one) of the interlocutors. The high positive expectation is just in the shared context.

5.1.2 High positive expectations in the shared context in conversations from the ambiguity corpus

A high positive expectation becomes a relevant prior for how interlocutors interpret ambiguity when it is in the shared context between them. There's a range of ways for people to enter propositions into the shared context, as evident in examples from the ambiguity corpus. (As with example (1), I underline the salient expectation and bold the quantifier-negation

utterance.)

For example, one of the interlocutors might hold a high positive expectation in the specific conversation. Excerpt (2) exemplifies a case where one interlocutor believes a high positive expectation (*everyone's cool*) while the speaker of the ambiguity disagrees (*everyone is not cool*). Excerpt (3) shows a case where the speaker of the ambiguity believed the high positive expectation in the past but uses their utterance to express disagreement with the expectation in the present.

(2) SAMBERG: (as Jesse) It's the perfect breakup.
JONES: (as Celeste) Yeah. Everyone's cool.
GRAYNOR: (as Beth) **Everyone is not cool**.
said on NPR All Things Considered in 2012; corpus ID 12304

(3) We were all so young and we were so naive that we kind of thought every experience would be like that ... And it isn't. **Every experience isn't like that**.
said on NPR Fresh Air in 2009; corpus ID 12537

High positive expectations can also be entered into the shared context when they are more global beliefs, or as a form of general knowledge. For example, in excerpt (4), the speaker of the ambiguity refers to a high positive expectation that was held by people who are not the interlocutors.

(4) This was a bipartisan consensus that if we just move to a more globalized market system, everybody will be better off. Well, in fact, **everybody was not better off**. The top 1 or 2 percent were significantly better off.
said on NPR Fresh Air in 2018; corpus ID 12445

People can also refer to propositions that are not necessarily considered to be true by anyone. In these cases the speaker's decision to enter the high positive expectation into the shared context is not an expression of direct belief that it is true. Excerpt (5) shows a speaker creating a high positive expectation for the current shared context by referring to an utterance that expressed it (*I said everything's going to be OK*).

(5) And I just, you know, I picked Nicole up and put her in my arms, and I said everything's going to be OK. And I knew in my mind, **everything's not OK**.

said on NPR Fresh Air in 2015; corpus ID 12471

In (6), the speaker brings up the high positive expectation in a context that already negates it (*you can't expect every doctor to know everything about every disease*). In (7), the speaker of the ambiguity brings up the high positive expectation as the antecedent of an *if*-statement. These are good examples of a high positive expectation that is rhetorical but nevertheless becomes salient in a conversation.

(6) When you have over 10,000 different diseases, you can't expect every doctor to know everything about every disease. But that's not actually the problem. The problem isn't that **every doctor doesn't know everything about every disease**; the problem is that for some diseases, no doctors know anything about those diseases.

said on NPR Fresh Air in 2020; corpus ID 12421

(7) You know, if everything was wonderful, you could ask the question, well, why would you talk about that difficult past? But **everything is not wonderful**.

said on NPR All Things Considered in 2018; corpus ID 12197

In addition, the high positive expectation might not receive expression in the preceding linguistic context at all, but may still be highly salient. Excerpt (8) shows a speaker playing

off the high positive expectation in The Lego Movie's song *Everything is Awesome*. This is essentially another example of a high positive expectation which is a form of world knowledge.

(8) BIANCULLI: Wilmore [...] cut to a clip of "The Lego Movie." [...] (SOUNDBITE OF TV SHOW, "THE NIGHTLY SHOW")

WILMORE: No, **everything is not awesome**.(LAUGHTER)

said on NPR Fresh Air in 2015; corpus ID 12478

These examples from the corpus highlight a similarity between the use of *every*-negation under inverse scope in a context with a high positive expectation, and metalinguistic negation (Horn, 1985). The next section discusses this similarity.

5.1.3 *Every*-negation as metalinguistic negation of high positive expectations

Maybe it's possible to describe inverse scope use in positive expectation contexts as metalinguistic negation. Horn (1985) suggests that natural language negation is pragmatically ambiguous between 'ordinary' and 'metalinguistic' uses. As an illustration of these two kinds of negation use (though, for an in-depth defense that there exists this pragmatic ambiguity of negation, please see Horn (1985)), consider the sentence (9) (as adapted from Horn).

(9) John didn't manage to solve the problem.

With a sentence like this one, ordinary negation shows two clear characteristics: it preserves a certain kind of information (10) but systematically changes another kind of information (11). Example (10) summarizes that one kind of information that ordinary negation preserves in this case is the conventional implicature, *it was difficult for John to solve the problem*,

of the non-negated proposition (*John managed to solve the problem*). That is, even if the speaker says that *John didn't manage to solve the problem*, if the negation in the sentence is ordinary negation, then the implicature remains that it was difficult for John to solve the problem. What has changed? (11) summarizes that the truth of the predicate, *John solved the problem*, has flipped under negation.

(10) *Manage to* contributes a conventional implicature which is preserved under ordinary negation

a. John managed to solve the problem – it was difficult for him to solve it.

b. John didn't manage to solve the problem – it was difficult for him to solve it.

(11) Ordinary negation affects the truth condition of the predicate

a. John managed to solve the problem – it is TRUE that *John solved the problem.*

b. John didn't manage to solve the problem – it is FALSE that *John solved the problem.*

In general, this conventional implicature – that it was difficult for John to solve the problem – represents one of the appropriateness conditions on the normal, felicitous utterance of both (10a) and (10b). In other words, if it were easy for John to solve the problem, there would be something inappropriate about the ordinary use of either (10a) or (10b). Thus stepping back, the idea of ordinary negation exemplified above is that it affects truth conditions, but it does not affect the appropriateness of use of an utterance.

There *is* another kind of negation which could be about the appropriateness of an utterance rather than its truth conditions: metalinguistic negation, according to Horn (1985). For example, the negation in the sentence *John didn't manage to solve the problem* could be metalinguistic rather than ordinary, in that it is used in a way that fails to show either of the effects of ordinary negation ((10) and (11)). This would be the case if the speaker said

(12): the conventional implicature of *manage to* changes while the truth condition of the predicate stays the same (i.e., it is TRUE that *John solved the problem.*). With this case of metalinguistic negation, it is the conventional implicature rather than the truth of the predicate which is objected to by the speaker.

(12) Metalinguistic negation of *John managed to solve the problem*: the conventional implicature is not preserved, while the truth condition of the predicate does not change.

a. John didn't *manage* to solve the problem – it was quite easy for him to solve! (And indeed he solved it.)

The kind of example in (12) has also been called 'contradiction negation' (Karttunen and Peters, 1979; Liberman and Sag, 1974).

So in general, metalinguistic negation is argued to affect the appropriateness of the normal, felicitous use of an utterance, where appropriateness encompasses a lot more than conventional implicatures. Horn gives examples like the following as similar cases of metalinguistic negation: objection to a conversational implicature (the insufficient informativeness in (13) and (14)), phonetics (15), morphology (16), "insufficient stylistic delicacy" (17), and to even more subtle connotations which are expressed with descriptions that are virtually identical in terms of their truth-conditional meaning ((18) and (19)).

(13) *Some* men aren't chauvinists – *all* men are chauvinists.

(14) John didn't manage to solve *some* of the problems – he managed to solve *all* of them.

(15) (So, you [miYonijd] to solve the problem.) No, I didn't [miYonijd] to solve the problem – I [maenijd] to solve the problem

(16) I didn't manage to trap two *mongeese* – I managed to trap two *mongooses*.

(17) Grandma isn't 'feeling lousy', Johnny, she's indisposed.

(18) Ben Ward is not a black Police Commissioner but a Police Commissioner who is black. (N.Y. Times editorial, 11/8/83)

(19) I'm not his daughter – he's my father.

To summarize, there is "a use distinction [for negation]: it can be a descriptive truth-functional operator, taking a proposition p into a proposition *not-p*, or a metalinguistic operator which can be glossed 'I object to u', where u is crucially a linguistic utterance rather than an abstract proposition" (Horn, 1985, pp. 136). Consider what u actually may be: an obvious pattern in these examples of metalinguistic negation is that a plausible u is an overtly expressed positive expectation. This positive expectation is not the only plausible context, but it seems a likely one, in that it clearly provides the material to be objected to.

(20) Plausible context of metalinguistic negation: overtly expressed positive expectation

a. (Some men are awful chauvinists.) *Some* men aren't chauvinists – *all* men are chauvinists.

b. (Luckily John manage to solve some of the problems.) John didn't manage to solve *some* of the problems – he managed to solve *all* of them.

c. (So, you [miYonijd] to solve the problem.) No, I didn't [miYonijd] to solve the problem – I [maenijd] to solve the problem

d. (You did it, you managed to trap two mongeese! How did you do it?) I didn't manage to trap two *mongeese* – I managed to trap two *mongooses*.

e. (Granny's feeling lousy.) Grandma isn't 'feeling lousy', Johnny, she's indisposed.

f. (That district has a black Police Commissioner.) Ben Ward is not a black Police Commissioner but a Police Commissioner who is black. (N.Y. Times editorial, 11/8/83)

g. (You're his daughter?) I'm not his daughter – he's my father.

Thus, it is helpful to consider a use of *every*-negation with inverse scope, in a context with a positive expectation, as a case of metalinguistic negation: the speaker uses negation to object to the assertability of the preceding positive expectation. This doesn't mean that negation in general should not be interpreted logically, only that negation can be about assertability rather than truth. That being said, the mechanism of world expectations affecting interpretation plausibility is broader and helps to account for why this use of *every*-negation would be a case of metalinguistic negation in the first place. It isn't just that the *every*-negation is an emphatic way to express the message that a proposition doesn't hold – it's a way to express that a prior *expectation* does not hold, whether or not such an expectation receives linguistic expression as a salient proposition, and whether or not this expectation is a high positive one or merely a positive one. Moreover, interpretations are then influenced due to the broader mechanism of increasing the plausibility of the inverse scope interpretation relative to the surface scope one.

5.1.4 Identifying positive expectations

How can we measure that positive expectations are salient in a conversation? First, as already previewed in the corpus examples above, for any universally-quantified quantifier-negation utterance, a strong version of the high positive expectation could be paraphrased by the non-negated quantifier-negation utterance itself (e.g., *Every vote counts* for *Every vote doesn't count*). The fact that this expectation is expressed suggests that the possibility of the belief is salient in the minds of the interlocutors. So, one way that a salient high positive expectation

could be measured is by the presence of the non-negated utterance in the preceding linguistic context.

More generally, the strongest version of a high positive expectation is a belief in the truth of the *all* world state. So another way to measure a salient high positive expectation would be to measure for any preceding expression that refers to a belief that the *all* world state is true. This is a less direct but more flexible way to measure. For example, it would identify a high positive expectation in excerpt (21), where the *everybody* in *everybody's not on at the same time* refers to the speaker's addressee, Colbert, Fallon, Kimmel, and O'Brien. While the speaker doesn't express a high positive expectation by actually expressing the non-negated *everybody's on at the same time*, she does express the belief that the *all* world state is true, by saying that the speaker's addressee, Colbert, Fallon, Kimmel, and O'Brien are on at the same time.

(21) Do you ever wish that you were not on at 11:30 opposite Colbert, and Fallon and Kimmel? And, you know, Conan O'Brien's on then, too. And, like, most of the late-night comics now are starting with political comedy, and I keep thinking like, shouldn't somebody be on at, like, 10 or 10:30? [...] can't we just, like, rearrange the schedule a little bit so, like, **everybody's not on at the same time**?

said on NPR Fresh Air in 2018; corpus ID 12447

These measures relate to strong versions of high positive expectations – the *all* world state – but we could also measure more generally for beliefs in the *some* world state, in order to capture that there's a salient positive expectation (a nonzero success rate), though not a particularly high one. For example, the expression of a belief that some votes do count would show that interlocutors have a salient positive expectation, which would still predict an inverse scope interpretation preference, though to a lesser extent than a high positive expectation would. Excerpt (22) is a naturalistic example of a case that would also be predicted to have

inverse scope (*everybody (who is dying) doesn't experience seeing their mother*), given the speaker's preceding assertion of existence for the non-negated predicate (*it's really common for people who are dying to see their mothers*).

(22) So it's – anybody who works in hospice will tell you - anybody – that it's really common for people who are dying to see their mothers. It's not a necessary step. It's not – **everybody doesn't experience it**, but it happens a lot.

said on NPR Fresh Air in 2016; corpus ID 12461

These methods of measuring for positive expectations in the text would fail to capture any relevant and salient expectation that isn't overtly expressed in the preceding linguistic context, as for example in excerpt (8), where the speaker plays off the title of the unmentioned song *Everything is Awesome*. (In that case even though the high positive expectation is in the preceding context, as knowledge of the song's title, it isn't in the immediate linguistic context.) So an additional, more powerful way to measure positive expectations could be to crowd-source judgments of the success rate of the non-negated predicate. For example, to measure the strength of the positive expectation for the utterance *Everything isn't awesome*, people could be shown the preceding context and asked *How likely is it that a thing is awesome?* on a scale between *very unlikely* and *very likely*. This measure could capture world knowledge regardless of what's specifically expressed in the preceding context (e.g., the knowledge that The Lego Movie has a song called *Everything is Awesome*), and it has the added advantage of measuring beliefs as gradient instead of as categorical, so it could be a more sensitive measure than the ones described above.

Finally, with respect to all these methods, it's worth noting that an expression or belief in the *none* world state is what would predict a surface scope preference. A good example is in excerpt (23), where the speaker of the *every*-negation utterance, *everyone else wasn't fazed*, first expresses a belief that the *none* world state is true: *no one around me was concerned*.

The potentially ambiguous utterance in this case would be predicted to have preferred surface scope.

(23) But when I got to New York and saw that there were insects or rodents or that there was trash on the street or that the park that I loved to nap in during the day was covered with rats at night, these are facts of the city. And I think the people – no one around me was concerned. And so even though I didn't have a context for it yet – I hadn't lived in the city for 10 years – I could see that **everyone else wasn't fazed**.

said on NPR Fresh Air in 2016; corpus ID 12463

In general, for a continuum between beliefs in the *none, some, all* world states, the closer a belief is to *all* the greater the predicted inverse scope preference.

With these examples in mind, to make the role of plausibility and high positive expectations more concrete, the next section returns to the computational model of scope ambiguity from Chapter 2, with a focus on how this model actually instantiates high positive expectations which affect competing interpretation plausibility.

5.1.5 How high positive expectations affect *every*-negation interpretations in the model

The model of scope ambiguity resolution from Chapter 2 demonstrated that a relatively high success rate – in other words, a high positive expectation – was a key component that allowed the model predictions to account for the different average interpretations of *every*, *some*, and *no*-negation.

Focusing on the role of positive expectations for *every*-negation, to review, Scontras and

Pearl (2021) propose a formalization of how positive expectations influence interpretation preferences. The model instantiates the role of world knowledge in a way that is consistent with this broader hypothesis about the role of plausibility. One of the factors in the model is prior world expectation, and the model shows how these priors facilitate a speaker's preference to endorse quantifier-negation in certain contexts. A key result was that prior world expectations which resemble high positive expectations are one way to successfully account for a speaker's preference to endorse *every*-negation for an inverse-scope-verifying context in a truth value judgment task. The extended model in Chapter 2 then demonstrates how these model parameters which describe positive expectations, with few other assumptions, also successfully account for listener preference to arrive at the inverse scope interpretation of *every*-negation.

Here, focusing on accounting for different interpretations of *every*-negations, I return to the extended model and vary the extent to which it assumes a high positive expectation, to show systematically that the greater the high positive expectation, the greater the predicted inverse scope preference for *every*-negation.

5.1.5.1 Initial parameter setting

Recall that model predictions depend on fixing the free parameters, one of which is the world prior $P(w)$, which represents listeners' beliefs about the general probability of the possible world states, based on the individual success base rate b_{suc}. This world prior $P(w)$ is the place to implement a high positive expectation. The greater the underlying expected success rate, the higher the positive expectation – the greater the belief in the *some* and *all* world states.

To test whether high positive expectations help the model predict inverse scope interpretations, I vary the world prior (about how many marbles are red) and see the resulting predicted

interpretation preference.

5.1.5.2 Results

Figure 5.2 shows that the model indeed predicts that listeners should be more likely to arrive at the inverse scope interpretation of *every*-negation (i.e., that not all the marbles are red) as their prior beliefs favor marbles being red: the higher the prior probability that a marble is red, the higher the pragmatic listener's resulting preference for the inverse scope interpretation.

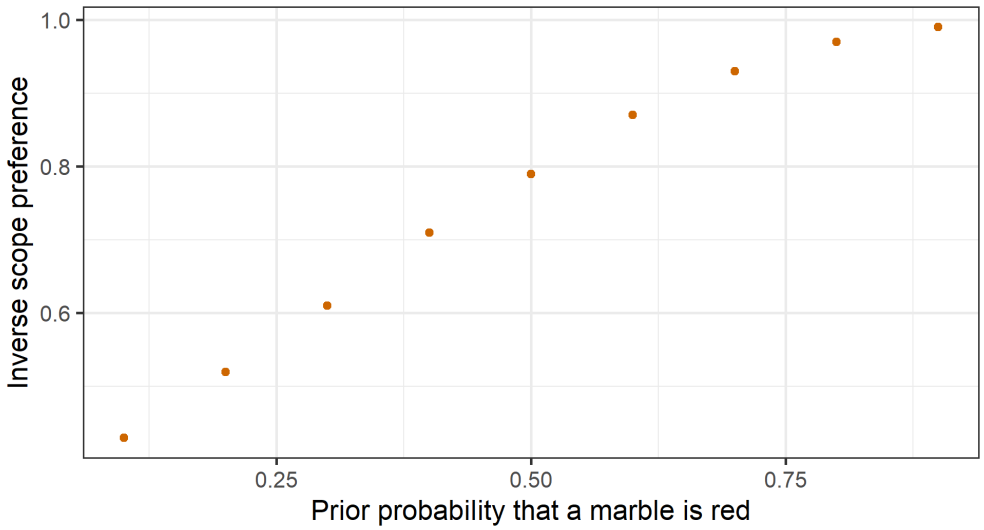


Figure 5.2: Predicted inverse scope preference for *every*-negation given the model's prior belief that a model is a red. As the probability that each marble is red rises, the extent to which there is a high positive expectation rises, and the predicted inverse scope preference also rises.

5.1.5.3 Discussion

The model indeed predicts that the inverse scope interpretation of *every*-negation becomes more likely as beliefs favor high positive expectations. The formal articulation of the model also allows us to better understand why this prediction is made: it rests on the listener's

reasoning that the utterance is true, and the probability that the utterance is true is higher under the inverse scope interpretation rather than the surface scope one. More specifically, there are more ways for inverse scope *not all* to be true (w could be 0, 1, or 2) than for surface scope *none* to be true (w must be 0). As the prior probability of a marble being red increases, the probability of world states 1 and 2 increases relative to the probability of world state 0, and so the probability placed by the pragmatic listener on the inverse scope scope interpretation correspondingly increases. In intuitive terms, the more that listeners hold a high positive expectation for *every*-negation and therefore believe there is a high probability that *some or all* is true, the more they reason that the speaker cannot have meant *none* and, therefore, meant *not all*.

Note that this reasoning underlying the model predictions for L_1 listener behavior (such as we would see in a paraphrase endorsement task) is different from the reasoning that Scontras and Pearl (2021) describe as underlying model predictions for S_2 speaker behavior (modeling a truth value judgment task). With truth value judgments, the modeled speaker's goal is to say something as useful as possible (modeling a participant's decision to endorse or not endorse an *every*-negation utterance as a description of a scenario in which its inverse scope interpretation is true). This usefulness for the S_2 speaker is defined by informativity and cost: in particular, without varying cost, an utterance is more informative the more that the pragmatic listener's (L_1's) posterior distribution over interpretations differs from the prior distribution, and in such a way that the pragmatic listener correctly arrives at the speaker's intended interpretation. In other words, learning that a strong prior belief is false is very informative. And, since prior beliefs shift at the level of the pragmatic listener L_1, they lead to differential utility for S_2 who reasons about L_1. Thus, differential utility for S_2, operationalized via informativity, determines utterance endorsement for truth value judgments.

In contrast, with interpretation preferences, the modeled listener L_1 has the goal of reasoning

about the intended interpretation of a speaker S_1, who reasons only about L_0. Prior beliefs do not shift at the level of L_0 in our model (L_0 has a flat prior on world states), so they cannot lead to differential utility for S_1. Thus speaker informativity is not affected by shifting prior beliefs when we only consider L_1 behavior; rather, it is the pressure on L_1 to reason about the ways that an interpretation can be true that is affected by shifting prior beliefs about the world.

With this definition of high positive expectations in mind, the next section takes a look back at the literature to describe specifically how the mechanism of positive expectations influencing plausibility can help to account for some attested variation in interpretations in past findings.

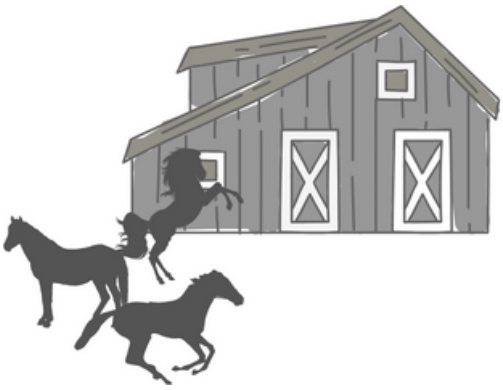
5.1.6 How expectations account for interpretations in past studies

World expectations influencing the relative plausibility of scope interpretations, in addition to the structural and pragmatic factors already mentioned, can help to explain some variation in past studies. This section describes the possible role of world expectations for children's TVJT behavior with *every*-negation (Musolino, 1999; Gualmini et al., 2008; Viau et al., 2010). I then review other pragmatic factors and discuss how they relate to plausibility.

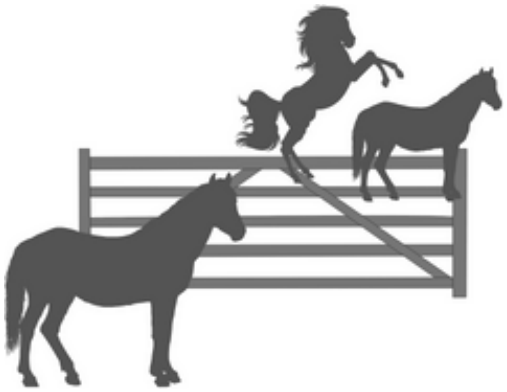
5.1.6.1 High positive expectations support children's apparent inverse scope preference for *every*-negation

Recall that context can facilitate children's inverse scope preference for *every*-negation, as in Figures 5.3 and 5.4. In a context like the one shown in Figure 5.3, children appear to have difficulty accessing the inverse scope interpretation of *every*-negation: in a truth-value judgment task, typically less than 10% of judgments by 4-6 year-old participants endorse

the utterance as a description of an inverse-verifying scenario (Musolino, 1999). However, in the contexts described in Figure 5.4, children increase their endorsement of the utterance to 50-60% of the time (Musolino and Lidz, 2006; Viau et al., 2010).



(a) Early failure: three of three horses fail to jump over a barn. The expected success rate of horses jumping is low.



(b) Inverse-verifying scenario: two of three horses succeed in jumping over a fence.

Figure 5.3: A context which potentially set up a low positive expectation (left image), and which led to low endorsement of the *every*-negation utterance "Every horse didn't jump over the fence", when this utterance is used to describe an inverse-verifying scenario (right image). In other words, this context leads to lower agreement that the *every*-negation utterance has inverse scope.

(a) Early success and/or contrast: three of three horses succeed in jumping over a log. The expected success rate of horses jumping is high.

(b) Inverse-verifying scenario: two of three horses succeed in jumping over a fence.

Figure 5.4: A context leading to high endorsement of *every*-negation. In this context, which potentially set up a high positive expectation, all horses succeed in jumping over a log (left image), with the additional optional description "Every horse jumped over the log, but..." These contexts lead to relatively high endorsement rates of the utterance "Every horses didn't jump over the fence", when this utterance is used to describe an inverse-verifying scenario (right image). In other words, this context perhaps leads to higher agreement that the *every*-negation utterance has inverse scope.

Arguably, the context in Figure 5.3 creates a low positive expectation, while the context in Figure 5.4 creates a high positive expectation. That is, a high positive expectation may have been made salient in just these contexts where children's behavior suggested an inverse scope preference (improved TVJT endorsement rates). For the utterance *Every horse didn't jump over the fence*, the high positive expectation would be that horses tend to succeed in jumping. An early-failure context, which did not facilitate endorsement rates of inverse-verifying scenarios, also did not obviously set up the expectation that horses succeed in jumping over the fence. In fact, it may have conveyed that horses are bad at jumping over things. On the other hand, the early-success and contrast contexts, which did facilitate endorsement rates of inverse-verifying scenarios, perhaps communicated a high positive expectation by conveying that horses are good at jumping over things, or that the experimenters or characters in the story (who participants believe know more about the state of the experimental world than the participants do) expected every horse to jump over the fence.

The higher the positive expectation, the more you would expect that *some jumped*, and the more you would be surprised to learn that *none jumped* rather than *some but not all jumped*. Listeners assume that the speaker is rational and cooperative, choosing to express the utterance that is more likely to guide the listener to their intended interpretation – for example, saying things that are true. Since it is likely that *some jumped*, the inverse scope *not all* interpretation is then reasoned to be likelier than the surface scope one. And this interpretation bias on the listener's part also leads speakers (who want to successfully guide listeners to their intended interpretation) to be more willing to use *every*-negation as a description of an inverse-verifying scenario.

5.1.6.2 Interpretation plausibility supports adults' inverse scope preference for *all-* and *every*-negation

To my knowledge only a few examples are used directly in the literature to show that world knowledge can help disambiguate universally-quantified quantifier-negation. One is the corpus-attested *all*-negation case from Neukom-Hermann (2016), repeated below in (24). Neukom-Hermann (2016) states that the world knowledge that *Sainsbury's is a supermarket* causes the reader to prefer the inverse scope reading.

(24) Many of you may have noticed that Good Housekeeping is now on sale at the checkout in Sainsbury's, which has gone down brilliantly with shoppers, as I discovered when I visited my local London branch. I can't think why **all supermarkets don't put GH at the checkout**.

Although Neukom-Hermann doesn't further characterize how it might do so, this case also exemplifies the broader hypothesis argued here for the role of world knowledge. The knowledge that *Sainsbury's is a supermarket* might disambiguate towards the inverse scope interpretation in the following way. Listeners reason from *Sainsbury's is a supermarket* and *Good Housekeeping is now on sale at the checkout in Sainsbury's*, to *there exists at least one supermarket (Sainsbury's) which puts GH (Good Housekeeping) at the checkout*. In other words, a positive expectation is set up as belief in the *some* world state. This belief means that *it is false that no supermarket puts GH at the checkout* – leading in other words to the belief that the surface scope interpretation is false for *all supermarkets don't put GH at the checkout*. In combination with the pragmatic reasoning that cooperative speakers usually say things that are true, the listener reasons that the inverse scope interpretation, which is the only remaining interpretation which *could* be true, *is* true: believing in the inverse scope interpretation is both consistent with (not ruled out by) the listener's knowledge, and

it would allow the listener to interpret the speaker's meaning as true of the world. More simply, the speaker *can't* have intended the scope interpretation with the proposition that both speaker and listener know is false of the world, so, the speaker must have intended the other scope interpretation, which, as far as the listener knows, could be true.

Similarly, Srinivasan and Yates (2009) mention that plausibility is a disambiguating factor, without a computational-level description of a disambiguating mechanism, but with an example consistent with the broader hypothesis here. Srinivasan and Yates (2009) write that inverse scope is more preferred for (26) than for (25), because the surface scope interpretation in (26), that a single doctor lives in all the cities, is too implausible to be likely.

(25) A kid climbed every tree.

a. **There is a single kid who climbed all the trees.** *Surface scope* (a > every)

b. Each tree was climbed by potentially different kids. *Inverse scope* (every > a)

(26) A doctor lives in every city.

a. There is a single doctor who lives in all the cities. *Surface scope* (a > every)

b. **Each city has a different doctor living there.** *Inverse scope* (every > a)

Again, the predicted preferred interpretation is the one that is more likely to be true, relative to the dispreferred interpretation.

5.1.6.3 How plausibility relates to other pragmatic accounts of *every/all*-negation interpretations

As also discussed in Chapter 1, an additional pragmatic factor in establishing the felicity of *every*-negation is the QUD (Hulsey et al., 2004; Gualmini et al., 2008). That is, a *did all?* QUD facilitates inverse scope preference on its own (Gualmini et al., 2008; Scontras and

Pearl, 2021). But the role of QUDs is consistent (for this data) with the role of plausibility suggested here too; these two factors may co-occur and be hard to disentangle. In general, manipulating the QUD and manipulating world expectations may appear to be the same context manipulation. Further, the effect of creating a *did all?* QUD – causing the inverse scope interpretation to be the better answer to the QUD than the surface scope one – and the effect of creating a high positive expectation – causing the inverse scope interpretation to be more plausible than the surface scope one – may be hard to disentangle experimentally.

For example, Gualmini's methods of manipulating the most recent QUD are accomplished through manipulating character expectations or participant expectations about the characters, which seems as useful a method of creating a salient high positive expectation as it is a method of creating a salient *did all?* QUD. As an example, in one experiment, a *Did all?* context which raised endorsement rates was that a friend delivers letters for another friend, who expects exactly four, but the deliveryman drops one on the way so that only three letters are delivered (Gualmini, 2004). (The target sentence is then *Every letter wasn't delivered.*) In this case, the context may have established both the QUD *Were all the letters delivered?* as well as the positive expectation *Every letter was delivered* (in the sense that most delivery situations may be expected to be successful). Additionally, the good answer to the QUD, *not all the letters were delivered*, is also the more plausible interpretation, in the sense that *no letter was delivered* is highly implausible. It would be interesting to find experimental ways of manipulating a QUD without manipulating character expectations, or vice versa. That is, finding a way to manipulate a QUD without manipulating world expectations would better clarify how QUDs surface in everyday conversations and what in effect (beyond the theoretical distinction) differentiates QUDs from merely world knowledge.

Taking stock, many factors matter for interpretation preferences, but one factor that could facilitate a preference for inverse scope interpretations of universally-quantified quantifier-negation is a high positive expectation in the preceding context. And a hypothesis for how

high positive expectations affect behavior is articulated in Scontras and Pearl's (2021) RSA model of disambiguation and in the model in Chapter 3.

An open question is whether positive expectations come into play in naturalistic speech. How often do speakers in fact say utterances such as *Every vote doesn't count*, intending the inverse scope interpretation *Not all the votes counted*, in contexts that set up the expectation that it is likely that votes count? Thus, the next section turns to an analysis of the contexts and interpretations in the ambiguity corpus – the set of *every*-negation utterances that people produced in everyday conversation. The analyses focus on whether the hypothesis concerning high positive expectations in context can make sense of some of the variability in the annotated corpus.

5.2 Positive expectations predict interpretations of *every*-negation

I combined corpus methods – returning to the naturalistic productions described in the previous chapter – with behavioral methods – gathering crowd-sourced judgments about context – to better understand the preceding world expectations of *every*-negation in the wild.

It's important to bridge the understanding of scope ambiguity that has been gained in the lab with an understanding of how ambiguity is used everyday. The experimental studies reviewed above, for which it's possible to see how positive expectations may have played a role in interpretations, investigated a limited range of ambiguous utterances, which may differ in many ways from quantifier-negation utterances in everyday speech. I know of only a few corpus studies of quantifier-negation, but one has a small sample size (Musolino et al., 2000) and the other is based primarily on written language (Neukom-Hermann, 2016) and relied on

the primary researcher to determine the intended scope interpretation. Here we can use the crowd-sourced interpretations of corpus-mined *every*-negation utterances to investigate the role of context for disambiguation.

Using the COCA section of the *every*-negation corpus, I first considered various methods of measuring high positive expectations in context, depending either on the text of the preceding context itself or behavioral judgments of the preceding context. Measures of the literal text included coding by hand whether the preceding context string contained a high positive expectation string (Section 5.2.1), as well as automatically coding the similarity between the preceding context string and a high positive expectation string (Section 5.2.2). To gather behavioral judgments, a context-annotation experiment asked participants to judge the extent to which the preceding contexts of those same items expressed a positive expectation (Section 5.2.3). I considered how all these measures of positive expectations relate to the crowd-sourced scope interpretation preferences of these naturalistic uses in their immediate contexts: for an individual item, does a measure of a high positive expectation in the preceding context predict an inverse scope interpretation preference?

Section 5.2.4 compares these different measures of positive expectation, suggesting that the behavioral one perhaps best captures the idea of positive expectations. I then gathered behavioral judgments in a second context-annotation experiment for the NPR section of the corpus (Section 5.2.5).

5.2.1 Categorical measure of high positive expectations

One way to measure for the salience of a high positive expectation is by its overt linguistic expression in context. For the high positive expectations of *every*-negation, this overt linguistic expression can come in the form of the non-negated utterance itself, which in fact would express a strong version of the high positive expectation. For example, for *Every vote doesn't*

count, a high positive expectation is the prior belief that votes *do* count. One unambiguous, strong version of this belief would be expressed by the expectation that it is highly probable that *every* vote counts. So, we would know that this expectation is salient for interlocutors if it were expressed as the non-negated counterpart *Every vote does count* in the preceding context of *Every vote doesn't count*.

Thus, I hand-coded categorically for the presence/absence of this kind of overt high positive expectation expression in the preceding context of each of the COCA items.

5.2.1.1 Results

59/390 (15%) of the items contained a categorical high positive expectation expression (that is, the non-negated counterparts of the potentially-ambiguous utterances). Of these 59 utterances that were identified via hand-coding to have high positive expectation expressions, 50/59 (85%) were on average better paraphrased by the inverse scope paraphrase than the surface scope paraphrase.

I also looked at p(high positive expectation|inverse) vs. p(high positive expectation|surface): how often items where the inverse interpretation was strongly preferred had a high positive expectation expression compared with items where the surface interpretation was strongly preferred. I found that 22% of highly inverse-preferred items (those with responses greater than 0.75) had high positive expectation expressions, as opposed to 6% of highly surface scope-preferred items (those with responses less than 0.25). These results suggest that the hand-coded high positive expectations do tend to co-occur with an inverse scope interpretation in the COCA sample.

This effect of categorically-measured high positive expectation is significant and in the expected direction. To assess significance, I used a mixed effects model predicting mean scope interpretation by the categorical measure, with random intercepts for participants.

Model results are shown in Table 5.1. In order to make sense of the fixed effect coefficients, which are on the log-odds scale, we transform them to the probability scale with the inverse logit function (e.g., see Ford, 2021). The predicted scope preference of an item with no categorical preceding expression of a high positive expectation is 0.62 (the inverse logit of the intercept). The predicted scope preference of an item that does have a categorical preceding expression of a high positive expectation is 0.77 (the inverse logit of the sum of intercept and categorical expectation: 0.4877+0.7089 = 1.1966). In other words, an average item with a categorical preceding expression of a high positive expectation has a clearly stronger inverse scope preference compared to an average item without such a preceding expression.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	0.4877	0.03152	15.470	<2e-16
Categorical High Positive Expectation	0.7089	0.07289	9.726	<2e-16

Table 5.1: Results of a mixed effects model with categorical high positive expectation per item (no/yes) predicting mean scope preference per item (log-odds; higher values indicating greater inverse scope preference), with random intercepts for the participants in the scope preference interpretation experiment.

The next section turns to the automatic measure of high positive expectations. While the goal of this automatic measure is the same as that of the categorical one – to capture the extent to which a high positive expectation was expressed overtly in the context – the automatic measure is more objective, can scale to larger datasets, and captures the continuous rather than categorical relationship between the use of an expression and the strength of inverse scope preference.

5.2.2 Automatic measure of high positive expectations

For an automatic, more objective, and scalable measure of the high positive expectation expression, I calculated the degree of lexical overlap between the preceding linguistic context and a string representing the high positive expectation (*pos_exp*). For each item (e.g., *Every*

vote doesn't count), I first coded *pos_exp* as the potentially-ambiguous clause without negation (e.g., *Every vote does count*). I then coded for the extent to which the *pos_exp* appeared in the preceding context as the longest common substring similarity (*LCS*; Needleman and Wunsch, 1970) between each preceding context string c and *pos_exp* pair, calculated using the R `stringdist` package (van der Loo, 2014).

Each LCS was equal to the longest sequence formed by pairing words from the preceding context string c and *pos_exp*, while keeping their order intact; the dissimilarity $d_{lcs}(c, pos\_exp)$ was then the number of unpaired words left over in both strings. Thus, dissimilarity ranges from 0 (completely similar) to the total words W in both strings combined (completely dissimilar), where $W = length(c) + length(pos\_exp)$. $d_{lcs}(c, pos\_exp)$ can be defined recursively as in (5.8) for different relative lengths of the two strings to be matched against:

$$d_{lcs}(c, pos\_exp) = \begin{cases} 0, & \text{if } length(c) = \varepsilon, \\ d_{lcs}(c_{1:length(c)-1}, pos\_exp_{1:length(pos\_exp)-1}), & \text{if } length(c) = length(pos\_exp), \\ 1 + min\{d_{lcs}(c_{1:length(c)-1}, pos\_exp), & \\ \qquad d_{lcs}(c, pos\_exp_{1:length(pos\_exp)-1})\}, & \text{otherwise.} \end{cases} \tag{5.8}$$

There are three possible outcomes for $d_{lcs}(c, pos\_exp)$, as described in equation 5.8. First, the value is trivially 0 for empty strings (ϵ) (5.8, line 1). Alternatively, the value is based on pairing each word from both strings if the two strings have equal length (5.8, line 2: $length(c) = length(pos\_exp)$). For example, see the first two examples in Table 5.2; in these two examples, $d_{lcs}(c, pos\_exp)$ is equal to the negative of the number of unpaired words left over in both strings. Third, the value is based on the minimum LCS-distance that can be

obtained from pairing all the words from the shorter string to an equal number of words from the longer string (5.8, line 3: otherwise). For example, see the third line in Table 5.2; here, d_{lcs}= -2 because all four words in *pos_exp* would pair to *every vote does count* in the context, and leave unpaired the two words *I believe*.

Preceding context c	*pos_exp*	$d_{lcs}(c, pos\_exp)$
Every vote does count.	Every vote does count.	0
What is going on?	Every vote does count.	-8
I believe every vote does count.	Every vote does count.	-2

Table 5.2: Automatically measuring the extent to which the preceding context contains an expression of a high positive expectation. The measure, $d_{lcs}(c, pos\_exp)$, is shown for different sample contexts c of the quantifier-negation utterance *Every vote doesn't count*, for which the high positive expectation *pos_exp* is *Every vote does count*.

Finally, LCS *similarity* is calculated as negative dissimilarity: $-d_{lcs}(c, pos\_exp)$. LCS similarity ranges from 0 to -W (i.e., the total number of words in the context c and *pos_exp*), with values closer to zero indicating more lexical overlap. Values closer to zero indicate a greater similarity between the context and the high positive expectation linguistic string, and so represent a higher probability that the context contained a linguistic string transparently encoding a strong expression of a high positive expectation.

5.2.2.1 Results

The effect of LCS-measured positive expectation is significant and in the expected direction, though very modest, as shown in Figure 5.5 for both mean interpretations and individual judgments. To assess significance, I used a mixed effects model predicting mean scope interpretation by the LCS similarity $-d_{lcs}$, with random intercepts for participants. Model results are shown in Table 5.3. The intercept of 1.043 represents the log-odds of the mean scope preference when the LCS context annotation is at its reference level (log-odds = 0); using the inverse logit function to transform the log-odds back into probabilities, the reference

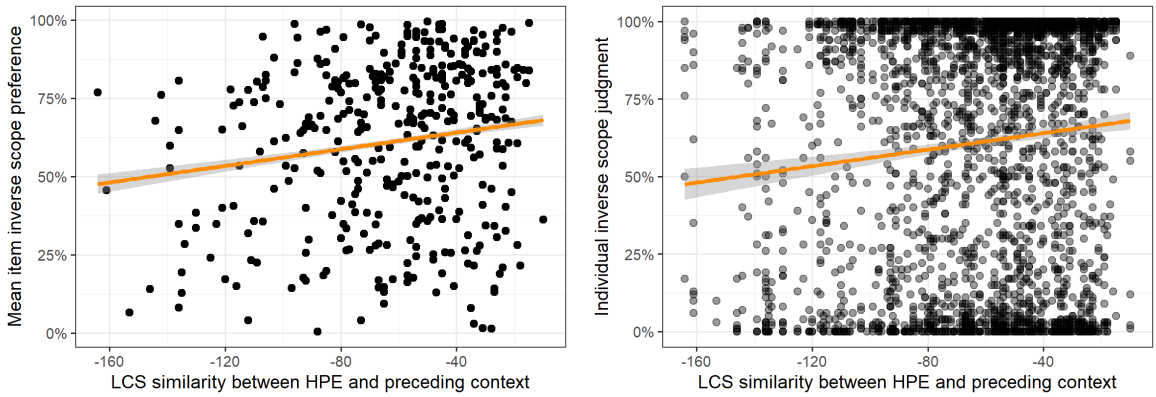


Figure 5.5: Preceding expression of a high positive expectation and inverse scope preference, for average item judgments and individual judgments.

predicted scope preference is 0.739 – in other words, the item is already quite likely to have inverse scope at the reference level of LCS similarity. For the coefficient for the LCS similarity predictor, applying the inverse logit (to 1.043+0.007303 = 1.050303), we get 0.741, meaning that a one-unit increase in the context annotation leads to a predicted probability of approximately 0.741 – in other words, with a higher positive expectation as estimated by the LCS metric, the item is very slightly more likely to have inverse scope.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	1.043	0.06011	17.345	<2e-16
LCS Positive Expectation	0.007303	0.0008384	8.711	<2e-16

Table 5.3: Results of a mixed effects model with LCS similarity per item (negative values to zero; values closer to zero indicating higher positive expectation) predicting mean scope preference per item (log-odds; higher values indicating greater inverse scope preference), with random intercepts for the participants in the scope preference interpretation experiment.

To determine whether a high positive expectation captures individual judgment variation above and beyond mean item-level variation, I used a separate model to predict logit-transformed individual item responses by LCS similarity, with random intercepts for participants and items. This model also found that LCS similarity was a significant predictor of an inverse scope preference ($p < .001$).

Interestingly, only expressions of high positive expectations that precede—but not follow—the

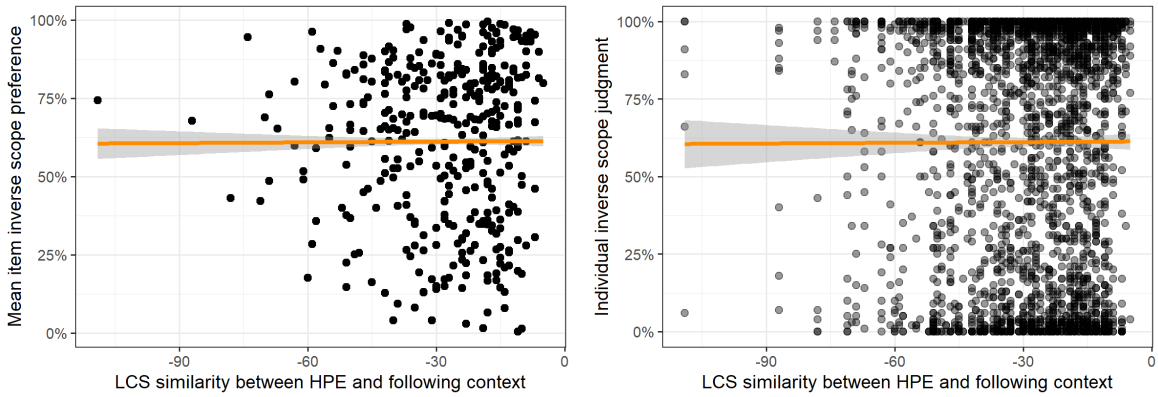


Figure 5.6: Following expression of a high positive expectation and inverse scope preference, for average item judgments and individual judgments.

ambiguous utterance reliably predict an inverse scope preference, as Figure 5.6 shows. More specifically, a version of both models that calculated LCS similarity using overlap with the following (rather than preceding) context found LCS similarity of the following context not to be a significant predictor of either item-level or judgment-level interpretations.

The next section turns to the behavioral measure of positive expectations. While the goal of this behavioral measure is similar to the previous two, it seeks to capture the extent to which there is a positive expectation in the context, whether or not it was a high positive expectation and whether or not it was overtly expressed in the conversation.

5.2.3 Behavioral measure of positive expectations

The preceding contexts of the 390 *every*-negation items (without the *every*-negation sentences themselves) were annotated with people's judgments of the extent to which the context contained a positive expectation. As before, judgments were measured on a sliding scale.

5.2.3.1 Experiment 6: Annotating positive expectations in COCA

Participants. 347 participants were recruited through Prolific, who had U.S. IP addresses and indicated that they were monolingual English speakers. Each participant received $2.00.

Stimuli. An example trial is shown in Figure 5.7. Participants saw an excerpt consisting only of the three preceding sentences (or lines if punctuation was missing) of items from the *every*-negation corpus.

Under each excerpt of the context, participants saw a question intended to measure how strongly they held a positive expectation given the context, in the form of *How likely is it that a random ...* combined with the non-quantified subject and non-negated predicate of the original *every*-negation item. Given that the ambiguous clauses took the form *quantified noun phrase–verb–negation–remainder*, the question took the form *How likely is it that a random–noun phrase–verb–remainder*. For simplicity in the remainder of this section, I call *a random–noun phrase–verb–remainder* a positive expectation. In the example in Figure 5.7, the original utterance was *everybody is not sitting home waiting for some pollster to call*, and so the positive expectation is *a random person is sitting home waiting for some pollster to call*.

One challenge in creating the questions is that sometimes content is missing from the *every*-negation constructions. That is, more implied aspects of meaning can be lacking for a question that is formulated solely on the basis of the *every*-negation item. For example, for an item of the form *Everybody's not*, the question as applied to the context becomes *How likely is it that a random person is?*, so that it is either misleading or no longer clear what the verb and remainder refer to. To solve this confusion, questions based on *every*-negation items which contained anaphoric or elided elements were replaced wherever possible with their antecedent in the preceding context. An example is shown in Figure 5.8. The original utterance was

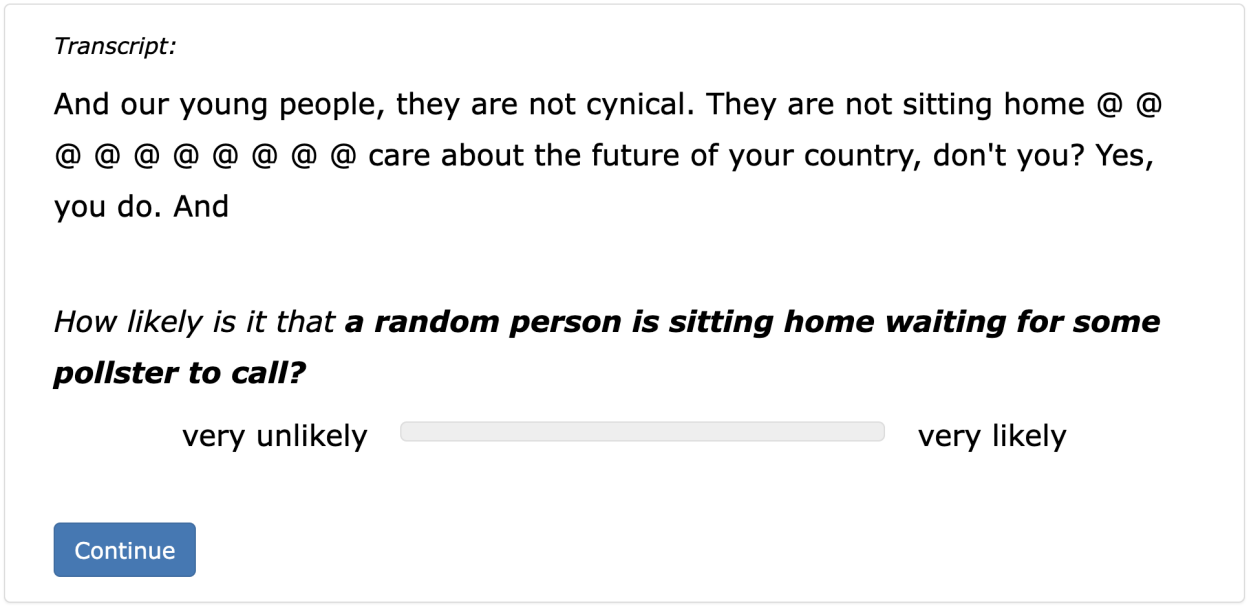


Figure 5.7: Sample trial from the behavioral context annotation of *every*-negation utterances. Participants saw this context for the original item "everybody is not sitting home waiting for some pollster to call". They were asked "How likely is that a random person is sitting home waiting for some pollster to call?" They answered on a sliding scale between "very unlikely" (always on the left) and "very likely" (always on the right).

Everybody's not, said in response to the other speaker's claim that *everybody is worried about the amount of this expense*. For this case, I formulated the positive expectation as *a random person is worried about the amount of this expense*.

Design. The initial instructions said to participants, before they accepted the task, that "You will be asked to judge statements about the world, given short written excerpts from real conversations." After participants accepted the task, more in-depth instructions then said, "You will see short excerpts of conversations from American radio and TV programs that took place between 1990 and 2012. With each conversation, you will also see a statement about the world. **Based on the conversation and your own knowledge of the world, your task is to judge how likely that statement is to be true.** Because we're peeking into the middle of real conversations, sometimes it may be hard to know exactly what the speakers were talking about, or it may be hard to see the connection between the conversation

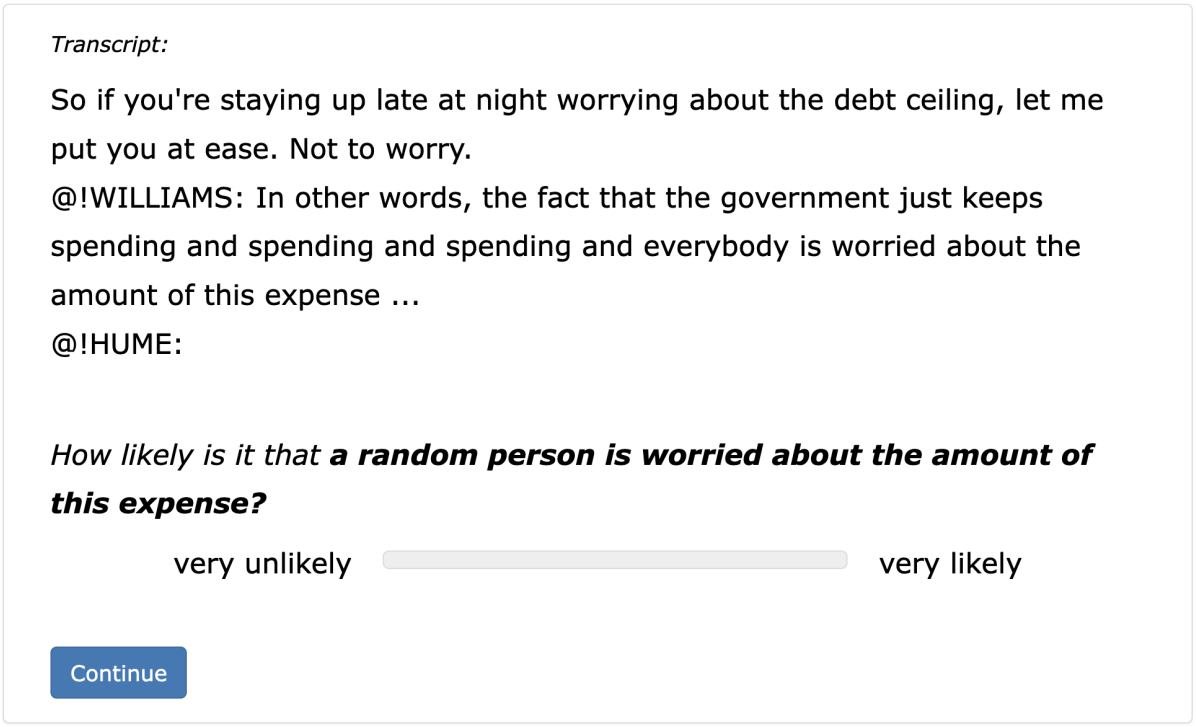


Figure 5.8: Sample trial from the behavioral context annotation of *every*-negation utterances. Participants saw this context for the original item "everybody's not". They were asked "How likely is it that a random person is worried about the amount of this expense?"

and the statement about the world. That's okay! Please indicate your best guess." There also followed several brief statements about the form of the excerpts and the fact that some conversations were about sensitive topics.

Each participant saw a total of fifteen randomly-selected items; on each trial, participants were again asked "How likely is it that *positive expectation*?" (see Figure 5.7). Beneath the conversation excerpt, participants gave a judgment on a sliding scale between "very unlikely" and "very likely".

Controls. To check that participants were paying attention, reading the contexts, and understanding the task, three control trials were constructed to imitate the rest of the items. Two of the controls appeared in random order as the first two trials for each participant, and the last control item appeared as the last trial for each participant, mainly to check continued

attention at the end of the task. These control trials contained clear information answering the question about the probability of a positive expectation.

The first two control items are described below in (27) – a low-probability control – and (28) – a high-probability control. For clarity, the text containing the key information (answering the question about the probability of a positive expectation) is italicized here, though it was not italicized in the experiment. Participants were considered to pass the low-probability control by placing the slider closer to the *very unlikely* side than to *very likely*; they passed the high-probability control by placing the slider closer to the *very likely* side than to the *very unlikely* side.

(27) @!TONHAUSER: The ten board members voted last night. No surprise - *every single one of them hated Proposition 23. All ten of the board members voted against it.* Basically,

How likely is it that ***a random board member liked Proposition 23****?*

(28) @!SIDNER: I'm glad to report that they completely fixed the issue at Greenwell. *Indicators have improved across the board and everybody's smiling. Everybody's happy.*

@!GROSZ: (VOICEOVER) Yep,
How likely is it that ***a random person at Greenwell is happy****?*

The third control item is described below in (29) and was a high-probability control.

(29) @!ROBERTS: You know, they said that that they completely fixed the issue at Silver Lake. *They reported improved indicators across the board and apparently everybody's smiling. I heard everybody's happy.*

@!ARIEL: (VOICEOVER): I heard the same thing,

How likely is it that ***a random person at Silver Lake is happy****?*

The rate of passing all three controls was 72% (while the rate of passing only the first two controls was 80%). Although this pass rate continues to be much higher on Prolific than on MTurk, speaking to the greater reliability of Prolific as a crowd-sourcing platform, still nearly 30% of participants didn't pass the controls. The pass rate may speak to high task difficulty, as the experiment asked participants to engage in complex reading comprehension.

With the addition of the three controls, participants completed a total of 18 trials. Analysis was restricted to those participants who passed all three controls. Thus out of the 347 participants, data was assessed for 250 (54% female; mean age: 39.6 years).

5.2.3.2 Results

As with the categorical and automatic measures of positive expectations, to test model predictions, I asked whether an item was more likely to receive an inverse scope interpretation in a context containing a higher positive expectation according to the behavioral measure.

Each item's preceding context was judged by at least 3 and at most 16 different participants, with an average of between 9 and 10 ratings per item. The final response measure for each trial varies from 0 (maximum endorsement that a positive expectation is very unlikely) to 1 (maximum endorsement that a positive expectation is very likely).

The left panel of Figure 5.9 shows individual judgments of positive expectations, which peak at the endpoints and midpoint of the scale (i.e., 0, 0.5, and 1), with a greater preference to place the slider below 0.5 than above it. The right panel of Figure 5.9 shows the mean context judgments per item, peaking slightly below the midpoint.

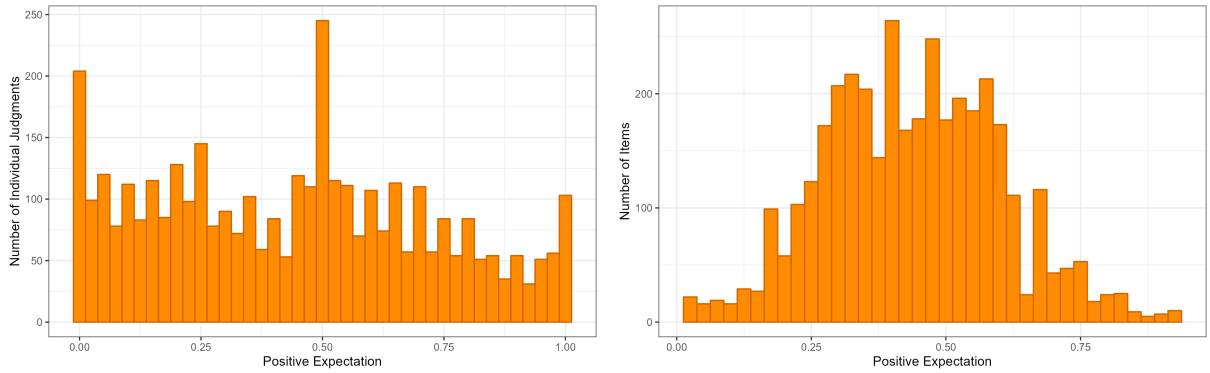


Figure 5.9: Individual judgments of positive expectations in context (left panel) and mean judgments per item (right panel) for the *every*-negation items from COCA.

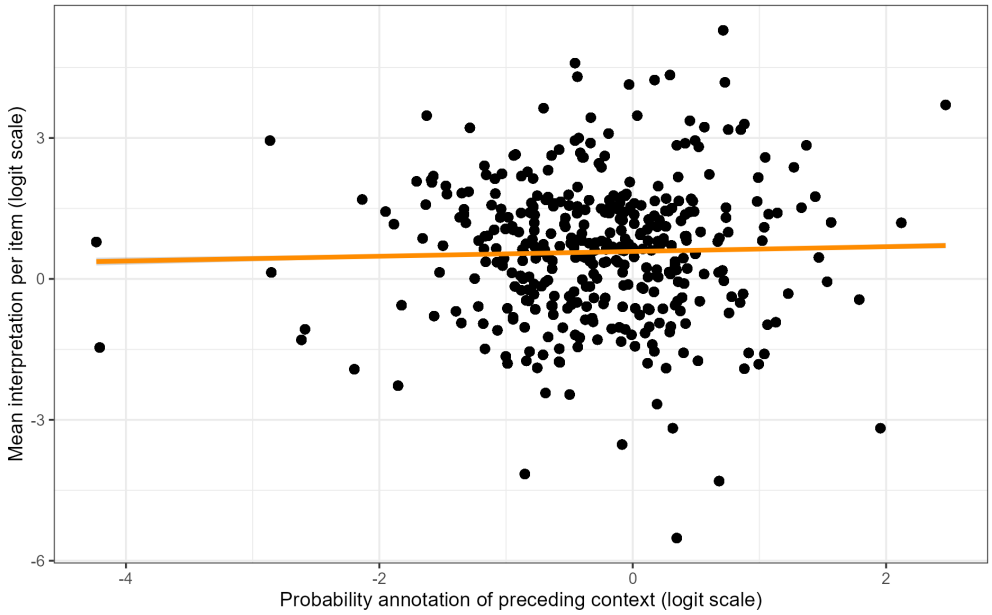


Figure 5.10: Positive expectation annotation of the preceding contexts for the *every*-negation items from COCA, predicting mean item inverse scope preference.

The effect of behaviorally-measured positive expectation is significant and in the expected direction, though modest. Figure 5.10 shows mean item inverse scope preference by positive expectation annotation. To assess significance, I used a mixed effects model predicting scope interpretation by context annotation, with random intercepts for participants. Model results are shown in Table 5.4. The intercept of 0.5887 represents the log-odds of the mean scope preference when the mean context annotation is at its reference level (log-odds = 0); using the inverse logit function to transform the log-odds back into probabilities, the reference

predicted scope preference is 0.64 – in other words, the item is already somewhat likely to have inverse scope. The coefficient for the context annotation predictor means that for each one-unit increase in the log-odds of the context annotation, the log-odds of the mean scope preference increases by 0.04312; applying the inverse logit (to 0.5887+0.04312=0.63182), we get 0.65, meaning that a one-unit increase in the log-odds of context annotation leads to a predicted probability of approximately 0.65 – in other words, with a higher positive expectation, the item is slightly more likely to have inverse scope.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	0.5887	0.02707	21.748	<2e-16
Behavioral Positive Expectation	0.04312	0.01036	4.163	3.15e-05

Table 5.4: Results of a mixed effects model with mean behavioral context annotation per item (log-odds; higher values indicating higher positive expectation) predicting mean scope preference per item (log-odds; higher values indicating greater inverse scope preference), with random intercepts for the participants in the context annotation experiment.

5.2.4 Measure comparison

All three measures – categorical high positive expectation, automatic high positive expectation, and behavioral positive expectation – demonstrate the predicted positive relationship with inverse scope preference in both individual judgments and mean item interpretations.

The advantage of the categorical annotation is that it clearly measures for those cases of salient high positive expectation, and indeed shows the strongest correlation with inverse scope. However, it only accounts for those cases when there is a very high positive expectation which was transparently linguistically encoded, but of course world expectations do not have to be encoded linguistically, encoded transparently, or encoded nearby even if they are linguistically encoded.

The advantage of LCS similarity is that it provides an automatic continuous measure to

improve the analysis of larger-scale data. However, like the categorical annotation, it only captures very high positive expectations and likely underestimates the presence of positive expectations. Further, the correlation observed between LCS similarity and inverse scope preference is the most modest one. Even given the restriction to a particular form of overtly expressed world knowledge in the preceding three sentences, LCS similarity potentially underestimates the presence of a high positive expectation for several reasons. First, it is affected by context length, such that LCS similarity is lower for longer contexts even if those contexts contain a clear high positive expectation. Second, LCS similarity looks for a high positive expectation based on the exact lexical items in the *every*-negation utterance. For instance, it would identify the high positive expectation in the context *Every vote does count* for the *every*-negation utterance *Every vote doesn't count*; yet it would miss the same expectation in the context *All votes should matter* because the individual lexical items differ (*every* vs. *all*, *count* vs. *matter*).

Thus, the behavioral measure may best capture the idea of positive expectations. It is not restricted to overt expressions of high positive expectation, nor is it restricted by the form of the expression. That being said, a potential disadvantage of the measure is that the task to elicit it may have been difficult or confusing to participants. One of the peaks in the distribution of responses estimating this positive expectation is at approximately the midpoint between very unlikely and very likely, which may have reflected that in many cases participants were unsure of the probability.

Table 5.5 shows examples of how these three measures play out in the COCA corpus. In this comparison, the LCS measure seems to be the most prone to reflect an aspect of the context which has nothing to do with positive expectations. For example, compare rows 1 and 3 in the table: while the categorical and behavioral measures indicate a high positive expectation in the context of the item in row 1 and not in row 3, LCS similarity suggests the opposite – a high positive expectation in row 3 than in row 1 – partly because the context happens to be

longer in row 1 than in row 3.

With the behavioral annotation thus the one which is potentially the best measure of positive expectations, I applied this annotation to the remainder of the ambiguity corpus, which was the NPR section.

5.2.5 Behavioral measure of positive expectations in NPR

The preceding contexts of the 287 *every*-negation items from NPR (without the *every*-negation items themselves) were annotated with people's judgments of the extent to which the context contained a positive expectation. The design replicated Experiment 6.

5.2.5.1 Experiment 7: Annotating positive expectations in NPR

Participants. 292 participants were recruited through Prolific who had U.S. IP addresses and indicated that they were monolingual English speakers. Each participant received $2.00.

Stimuli. The form of the stimuli (the excerpt of the context and the question about the positive expectation) was the same as in Experiment 6. As in Experiment 6, items which contained anaphoric or elided elements were replaced wherever possible with their antecedent in the preceding context.

Design. Design was the same as in Experiment 6. Participants were instructed to judge how likely statements were to be true. Each participant saw a total of 14 randomly-selected items; on each trial, participants were again asked "How likely is it that *positive expectation*?"

Context	Item	Mean scope (95% CI)	Categorical	Automatic	Behavioral
What's going on here? JACKSON: The done deal is the problem. The Democratic Party needs to be democratic far more so than it has been. Coming out of Florida, where every vote counts, where we now have learned a lesson, we can not be a part of an apparatus where	every vote doesn't count	0.82 (0.64–0.97)	yes	-50	0.79
Their grades have improved. The parents are involved. Everybody's happy. MR-HACKETT: But	everybody is not happy	0.84 (0.69–0.97)	yes	-15	0.73
I don't know. You know, if you live and let live, who's living and who's dying? You know, someone has to die eventually.	Everybody can't live forever	0.19 (0.05–0.36)	no	-25	0.01

Table 5.5: Sample COCA items shown with their preceding linguistic context, their mean scope interpretation (higher values indicating inverse preference), and a comparison of three measures of positive expectations in the context: a categorical one (hand-coded expression of high positive expectation: yes/no), an automatic one (LCS: values closer to 0 indicate greater positive expectations), and a behavioral one (sliding scale probability judgment: values closer to 1 indicate greater positive expectations).

Controls. To check that participants were paying attention, reading the contexts, and understanding the task, three control trials were constructed to imitate the rest of the items. These controls were the same as in Experiment 6 (as described in (27), (28), and (29)).

The rate of passing all three controls was 71% (while the rate of passing only the first two controls was 78%). This passing rate is comparable to that in Experiment 6.

With the addition of the three controls, participants completed a total of 17 trials. Analysis was restricted to those participants who passed all three controls. Thus out of the 292 participants, data was assessed for 207 (48.8% female; mean age: 42.7 years old).

5.2.5.2 Results

Each item's preceding context was judged by at least 1 and at most 22 different participants, with an average of between 10 and 11 ratings per item. The final response measure for each trial varies from 0 (maximum endorsement that a positive expectation is very unlikely) to 1 (maximum endorsement that a positive expectation is very likely).

As with the items from COCA, for the items from NPR, the effect of behaviorally-measured positive expectation is significant and in the expected direction, though modest. Since interpretations for the NPR items were elicited in different context and modality conditions (with or without context, as text or audio), I assessed the role of the behavioral context annotation for interpretations which were made for items in context, as text or audio (with the impact of modality also assessed). Figure 5.11 is of the raw data and shows how the context annotation and mean inverse scope preference per item have a weak but positive correlation, whether or not the item was encountered as text or audio.

To assess significance, I used a mixed effects model predicting mean item scope interpretation by context annotation and modality, with random intercepts for participants. Model results

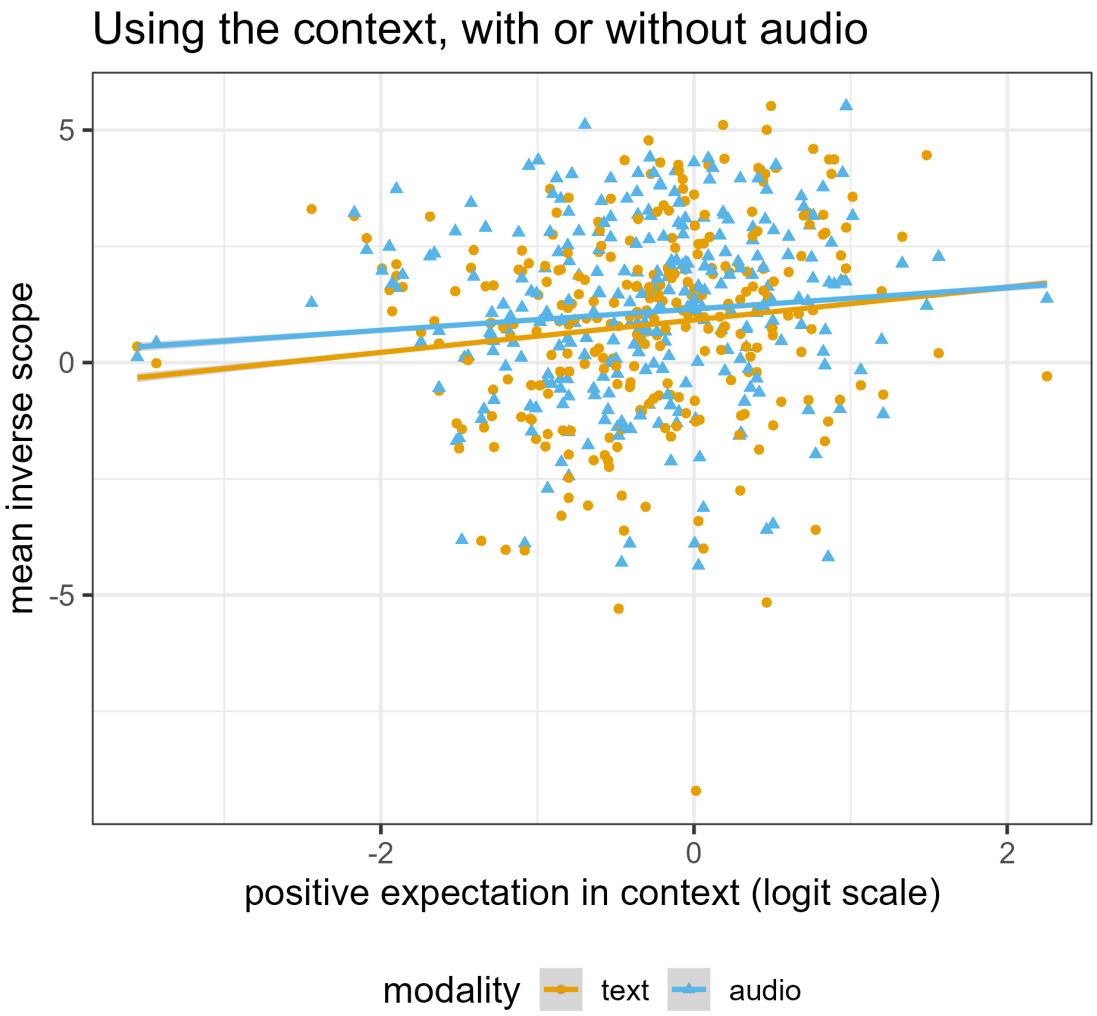


Figure 5.11: Positive expectation annotation of the preceding contexts for the *every*-negation items from NPR, predicting mean item inverse scope preference, depending on whether the interpretation was elicited for text-in-context or audio-in-context items.

are shown in Table 5.6. The predictions of this model are visualized in Figure 5.12. There is a main effect of modality (a positive significant coefficient): interpretations of audio as opposed to text were more likely to have inverse scope. More to the point, there is a main effect of the context annotation predictor: with a higher positive expectation, the item is also more likely to have inverse scope. There is also an unexpected interaction between modality and the context annotation: the inverse-scope-facilitating effect of positive expectations is weaker for items encountered as audio rather than as text. This may be due to the presence of information in the audio modality, which was absent in the text modality, which influenced

interpretations. Participants had more information to draw on in the audio than in the text modality.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	0.9076	0.03644	24.908	<2e-16
Modality	0.2366	0.01708	13.855	<2e-16
Behavioral Positive Expectation	0.3133	0.01441	21.736	<2e-16
Modality*Behavioral Positive Expectation	-0.1117	0.02031	-5.498	3.86e-08

Table 5.6: Results of a mixed effects model with mean behavioral context annotation per item (log-odds; higher values indicating higher positive expectation) and modality (text/audio) predicting mean scope preference per item when items were judged with context information (log-odds; higher values indicating greater inverse scope preference), with random intercepts for the participants in the context annotation experiment.

5.3 General Discussion

The ideas and analyses described in this chapter attempt to quantify some links between context and disambiguation, by providing a computational and empirical characterization of how disambiguation in context could proceed, and by testing these ideas on naturalistic language data. Broadly, this chapter explores how ambiguity resolution can proceed when sentences that have often been thought of as difficult or ambiguous are used as communication in context.

I started with the idea that context should help to account for some of the wide variation in interpretation preferences of the *every*-negation and *all*-negation constructions that the previous literature has focused on. In particular, if surface scope is indeed easier to access than inverse scope, context might explain the converging evidence for those cases where inverse scope is preferred for *every*- and *all*-negation. But the question is, what specific aspect of context matters – what is an inverse-scope-favoring context for universally-quantified quantifier-negation? – and what role does this aspect of context play in people's computation

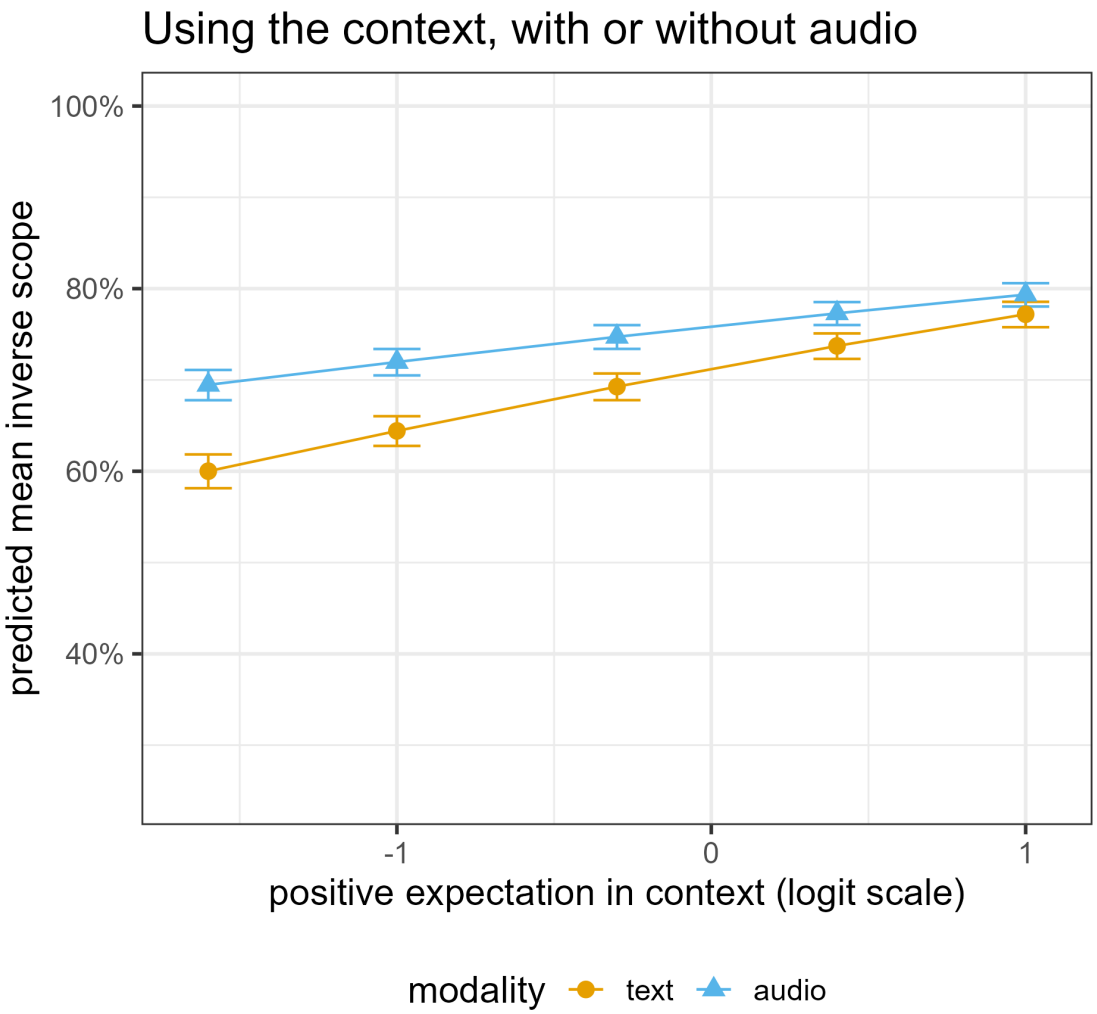


Figure 5.12: Positive expectation annotation of the preceding contexts and the predicted mean item inverse scope preference for an average item in the corpus, depending on whether the interpretation was elicited for text-in-context or audio-in-context items, according to the mixed effects model in Table 5.6.

for a preferred interpretation?

I suggested here that one role of context is to provide information on likely world states, which affects the relative plausibility of an utterance's interpretations, such that the preferred interpretation of an utterance is the relatively more plausible one – the one that is more compatible with likely world states. Plausibility helps cue the intended interpretation in combination with other structural and pragmatic factors (e.g., the QUD) – it is one way of

accounting for some of the attested variation in preferred interpretations of quantifier-negation. And a mechanism driving these interpretation preferences is that listeners will try to align their interpretation with what they already believe to be true of the world.

As a case study of this disambiguating mechanism, I focused on the skewed prior for *every*-negation which was identified in Chapter 3, and which I called *positive expectations*: the belief that the relevant entities have the property corresponding to the non-negated predicate. The RSA model from Chapter 3 describes how high positive expectations make more likely the inverse scope interpretation of *every*-negation: it is because with higher positive expectations, listeners expect that the *none* world state is unlikely, that the *all* world state is somewhat likely, and the *some but not all* states are most likely. Since listeners reason that speakers say things that are true, and this skewed prior lends relatively greater weight to world states that are compatible with the *not all*, inverse scope interpretation of *every*-negation, listeners then reason that speakers must have intended the *not all* rather than the *none* scope interpretations of *every*-negation.

I found evidence for model predictions for the disambiguating role of world expectations in a series of analyses of the ambiguity corpus. That is, inverse scope preference in the corpus is correlated with several different measures of positive expectations in context. Inverse scope is predicted by the presence of an expression of a high positive expectation (e.g., *Every vote counts* for *Every vote doesn't count*) in the preceding, though not the following context; by the automatic textual overlap of the high positive expectation expression and the preceding context; and by a behavioral annotation of the extent of a positive expectation for the preceding context. Thus, as predicted overall, positive expectations help account for the variation in interpretation preferences for *every*-negation utterances in the speech corpus.

These results align with previous modeling results in Chapter 3, extending the model validation to variation across different uses of *every*-negation. In general, the findings are consistent with the broader view that a sentence such as *Every vote doesn't count*, on its own, has an

under-determined meaning, so that listeners fill in meaning by reasoning with information such as context and communicative intent (Grice, 1975; Sperber and Wilson, 1986). Moreover, these findings accord with the prediction, based on this broader view, that spoken language used in a linguistic and social context should often be intended and interpreted with a single interpretation; that is, language in naturalistic context should show less ambiguity than the decontextualized text that we often study.

What about other factors that influence interpretations? The next chapter turns to prosody, a source of disambiguating information which is simultaneous with rather than preceding the ambiguity uses. The audio data in the naturalistic corpus allows us to explore some specific aspects of prosody that might reflect scope in everyday conversation.

Chapter 6

Prosody

The preceding chapters have focused on how context disambiguates and can help account for some variation in interpretation preferences, highlighting in particular how an interpretation is preferred when it is relatively more plausible than the other interpretation. Building on this discussion about disambiguating factors, this chapter turns to a distinct but related factor, prosody.

Prosody is the melody of an utterance: functionally, it can be thought of as those acoustic properties of speech that are not due solely to the choice of lexical items, but rather the semantic-syntactic relation and groupings between these items, as well as speaker emphasis, speech act, attitude, emotional state, and conversational state (Wagner and Watson, 2010). Alternatively, prosody can be thought of as its form rather than its function, as any suprasegmental acoustic properties such as syllable structure, intonation, and reflexes of prosodic structure, which are acoustically reflected in part in fundamental frequency, duration, and intensity (although some of these properties, such as syllable structure, can't be recovered only through these phonetic factors; Wagner and Watson, 2010). Using either ideas of prosody, those aspects of prosody most relevant for scope disambiguation are then prosodic phrasing

and prominence. There are some conflicts of definition between the function and form-based definitions of prosody, but both phrasing and prominence would count as prosody under either definition.

Prosody may play a similar disambiguating role to context, in that a quantifier-negation construction is underspecified for scope, and interlocutors use prosody, like context, to constrain the preferred interpretation (Grice, 1975; Sperber and Wilson, 1986). It would make sense that prosody helps disambiguate scope, because it's already known that prosody matters for different forms of structural ambiguity in speech production and comprehension, such as ambiguities of VP/NP-attachment, complementizer/parenthetical, and modifier scope (e.g., Kraljic and Brennan, 2005; Schafer et al., 2000; Snedeker and Trueswell, 2003; Hirschberg, 2013). Also, the role of prosody is not necessarily constrained to speech; the Implicit Prosody Hypothesis suggests that, although it is likely to vary with context, silent readers assign a default prosodic contour to what they read, thereby influencing syntactic ambiguity resolution: readers will prefer the syntactic analysis associated with the default prosodic contour (Fodor, 2002).

Prosody shows up several times in the footnotes of the literature reviewed in Chapter 1 for scope ambiguity, reflecting a widespread intuition that prosody can help disambiguate quantifier-negation. For example Carden (1973), who elicited interpretations of spoken ambiguous utterances, mentions in footnotes that he provided subjects with "neutral stress and intonation" at first, and then probed particular interpretations of each sentence by providing "stress and intonation patterns known to enforce one reading or the other" (p. 178). Similarly acknowledging the potential disambiguating role of prosody, Heringer (1970), who elicited interpretations of written utterances, mentions in a footnote that he considers the stimuli as "ambiguous when written, though perhaps not ambiguous when spoken because of stress and intonation" (p. 290).

In this chapter, Section 6.1 reviews the predictions from the literature about the prosody

of interpretations, highlighting a few predictions: that a prosodic boundary between the two logical operators would correspond to surface scope, and that a fall-rise contour would correspond to inverse scope. Section 6.2 tests these predictions in the corpus, and in a more exploratory analysis considers a range of other acoustic features that might reflect scope interpretation.

6.1 Prosody of quantifier-negation in the literature

The studies that directly address quantifier-negation prosody (Jackendoff, 1972; Ladd, 1980; Ward and Hirschberg, 1985; Syrett et al., 2012, 2014) tend to focus on *all*-negation disambiguation, and on the disambiguating role of prominence as measured by intonation (pitch). Intonation is the aspect of prosody which corresponds to the rises and falls of speech and can be measured with fundamental frequency F0 (Wagner and Watson, 2010; Xu, 2019). (However, note that many acoustic cues tend to co-occur with F0, such as intensity, loudness, and durations (e.g., Wagner and Watson, 2010; Xu, 2019), so that F0 may not be the only or primary cue to interpretations.) These studies suggest that an F0 peak on the quantifier, together with an utterance-final pitch rise, more likely corresponds to the inverse scope rather than the surface scope interpretation. The specifics of the acoustic characterization of the contour, and how and why it disambiguates, differ in many ways across studies. The little empirical data available on *all*-negation productions suggest that, although there is no evidence of a consistent prosodic scope signal across all speakers, for some speakers prosody does differ between sentences interpreted to have surface or inverse scope (Syrett et al., 2012, 2014).

Of relevance here too, a series of studies on negation-*because* ambiguity suggest that prosodic boundaries disambiguate the scope of negation. The empirical evidence for negation-*because* prosody is stronger and suggests some predictions when extended to quantifier-negation

ambiguity.

Section 6.1.1 reviews the disambiguating function of prosodic phrasing, both generally and for negation-*because* ambiguities in particular, because this prosodic factor in some ways paints a clearer and simpler picture than prominence. Section 6.1.2 then reviews the disambiguating function of prominence, mainly how the fall-rise contour might disambiguate *all*-negation. I then discuss why variation should be expected in a prosody-interpretation mapping (Section 6.1.3).

6.1.1 Prosodic phrasing

One potentially disambiguating aspect of prosody is phrasing. Intuitively, prosodic phrases divide an utterance into meaningful 'chunks' of information (Bolinger, 1989). Different levels of phrasing are marked by changes in F0 and often by phrase-final lengthening (a lengthening of the syllable preceding the phrase boundary), glottalization ('creaky voice') over the last syllables in the phrase, and some duration of pause; the acoustic/prosodic cues to larger, intonational phrases are also typically more marked than for smaller, intermediate phrases (e.g., Hirschberg, 1995; Wagner and Watson, 2010; Xu, 2019).

6.1.1.1 Phrasing can disambiguate syntactic ambiguities

There is a strong relationship between prosody and syntactic disambiguation across languages (e.g., for overviews, see Kraljic and Brennan, 2005; Hirschberg, 2013; Wagner and Watson, 2010; Xu, 2019). For example, consider the range of syntactic ambiguities in English alone that can be disambiguated by the presence/absence, or location, of a prosodic boundary (in the examples below, the discontinuity is marked with |), as adapted from a range of examples

provided in Hirschberg (1995, 2013).[1]

(1) **VP/NP-attachment**

a. Anna frightened the woman | with the gun (Anna held the gun)

b. Anna frightened | the woman with the gun (the woman held the gun)

(2) **Complementizer/Parenthetical**

a. Mary knows many languages you know (you both know many)

b. Mary knows many languages | you know (as you are aware ...)

(3) **Without/With an Appositive**

a. The animal that usually fights the lion is missing (the lion's normal opponent is missing)

b. The animal that usually fights | the lion | is missing (the lion is missing)

(4) **Restrictive relative/Non-restrictive**

a. My brother who is a writer needs a new job (I have multiple brothers but the writer brother is the one who needs a job)

b. My brother | who is a writer | needs a new job (I may or may not have other brothers)

(5) **Preposition/Particle**

a. John laughed | at the party (John laughed while at the party)

[1]It's interesting to note that these examples come from the context of research intended to improve automated text-to-speech systems. The source of the data on these prosodic patterns is "corpus-based techniques to improve default intonation ... [for example,] intonational phrasing decisions are based on syntactic and other information also inferred from text" (p. 1); I think automated phrasing was compared with human phrasing in a corpus. So, it is unclear from Hirschberg's report how robust these prosodic patterns are (to what extent a sample of English speakers would produce prosodic boundaries in the same way for these sentences). On the other hand, these are prosodic patterns that were deemed (by some measure) to meet a high standard of improving the naturalistic prosody and usability of a text-to-speech system.

b. John laughed at | the party (John ridiculed the party)

(6) **Simple complement/Parenthetical**

a. We only suspected | they all knew that a burglary had been committed (we only suspected that ... they all knew that a burglary had been committed)

b. We only suspected | they all knew | that a burglary had been committed (they all knew that we only suspected that a burglary had been committed)

(7) **Modifier scope**

a. This collar is dangerous to younger | dogs and cats (the collar is dangerous to both younger dogs and younger cats)

b. This collar is dangerous to younger dogs | and cats (it's dangerous to all cats)

(8) **Mathematical scope**

a. ten minus eight | divided by two (is one)

b. ten | minus eight divided by two | (is six)

As the last two examples show, prosodic phrases can signal not only syntactic structure but specifically scope boundaries: for example, when the discontinuity interrupts the phrase *dogs and cats*, the preceding modifier *younger* is more likely to no longer apply to the full phrase *dogs and cats* and, rather, only apply to the part of the phrase (*dogs*) that is contained within the prosodic phrase already containing the modifier *younger*.

The question is whether this phenomenon, where prosodic phrasing structure maps to scope structure, generalizes to other cases of scope ambiguity. The evidence seems mixed: it depends on the construction. As the next section shows, prosodic phrasing guides interpretations of neg-*because* constructions.

6.1.1.2 Phrasing can disambiguate neg-*because*

Prosodic phrases can help to limit the scope of negation to its own clause, as shown for neg-*because* ambiguities, which are one of the most well-studied scopally-ambiguous constructions with respect to their prosody (Hirschberg and Avesani, 1997; Frazier and Clifton, 1997; Baltazani, 2000; Hemforth and Konieczny, 2004; Kitagawa and Fodor, 2005; Nakao et al., 2007; Koizumi, 2009; Smith, 2011; Baumann and Rathcke, 2013).

For example, when (9) is uttered as a single intonational phrase, with no boundaries, it tends to be interpreted with surface scope. However, when a phrase boundary interrupts between the negated verb and the because-clause ((10)), the same phrase tends to be interpreted with inverse scope.

(9) Bill doesn't drink because he's unhappy.

a. **Bill does drink, but the cause is not his unhappiness. Surface (neg > *because*)**

b. Bill's unhappiness has led him not to drink. Inverse (*because* > neg)

(10) Bill doesn't drink | because he's unhappy.

a. Bill does drink, but the cause is not his unhappiness. Surface (neg > *because*)

b. **Bill's unhappiness has led him not to drink. Inverse (*because* > neg)**

Various sources stress that context can disambiguate negation-*because* utterances regardless of prosody, especially when the prosody is neutral. Specifically, for example, an utterance of *Bill doesn't drink because he's unhappy* when produced as a single phrase may still be interpreted with inverse scope, not necessarily surface scope, if context disambiguates (Hirschberg and Avesani, 2000). In such cases the single-intonational phrase contour is 'neutral' (Hirschberg, 2013).

6.1.2 Prosodic prominence

Most proposals highlight that inverse scope of *all*-negation is marked by a falling-rising intonational contour (e.g., Jackendoff, 1972; Liberman and Sag, 1974; Ladd, 1980; Ward and Hirschberg, 1985; Horn, 1989; Büring, 1997; Constant, 2012). Some details of the acoustic-phonetic/phonological realization of the fall-rise contour differ across accounts, but for most, there would be an abrupt pitch (F0) fall on or after the last syllable of the quantifier, as well as a phrase-final pitch rise. Also there are other acoustic cues that change with changes in F0 (e.g., duration, intensity) (Xu, 2019), but the fall-rise studies focus on F0.

According to one of the most acoustically-specific accounts in past studies, in Ward and Hirschberg's reading of different studies and their own corpus study, fall-rise is 1) a falling-rising F0 contour, which can be used on a wide number of sentences (e.g., see Figure 6.1). 2) In fall-rise, a pitch peak is realized late on at least one syllable, called the accented syllable; there may be more than one such accented syllable. In Ward and Hirschberg's notation, \x/ is used to identify x as an accented syllable of this kind, and this x can be a part of any lexical item as long as it is the lexically-stressed syllable of that lexical item (see examples (11)). 3) The contour is "scooped" in that the pitch peak is realized late in the accented syllable; in other words, the pitch starts very low on the accented syllable. 4) Within the two syllables following the pitch peak, there is a relatively abrupt pitch fall. 5) Finally, fall-rise involve a sentence-final pitch rise.

(11) A: Did Victor get tickets for the Fellini triple feature? (adapted from Ward and Hirschberg (1985))

a. B: Ve\ron/ica did.

b. B: He's con\sid/ering it.

c. B: He bought tickets for \some/ triple feature.

d. B: He got tickets for the Paso\li/ni triple feature.

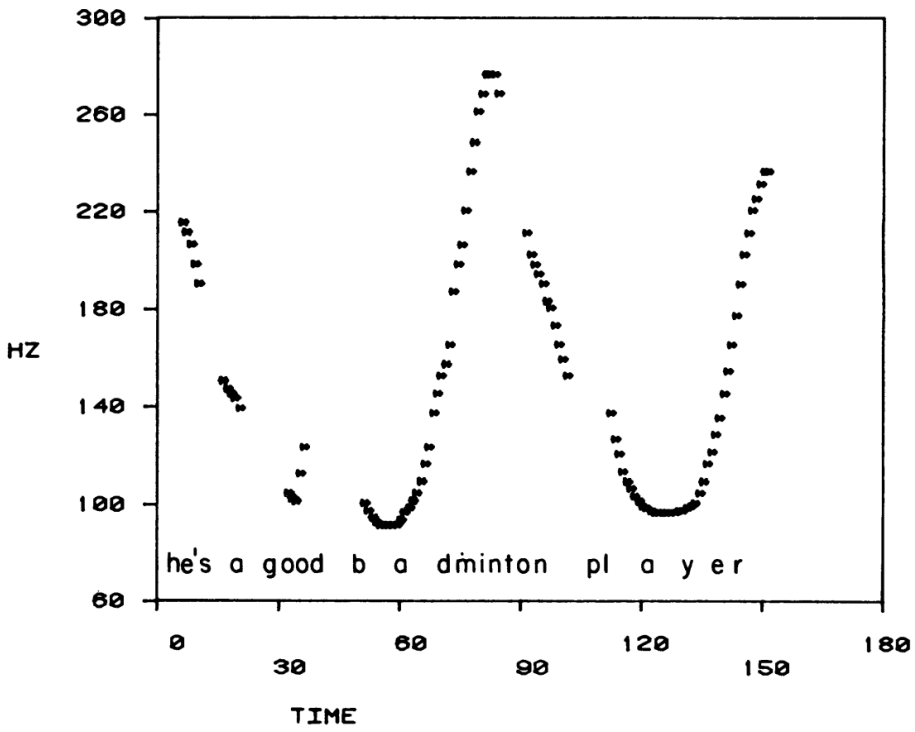


Figure 6.1: Example of fall-rise from Ward and Hirschberg (1985). The accented syllable is *bad* in *badminton*. The pitch is low by the time the speaker begins producing this syllable. The pitch then rises and falls sharply within the two following syllables. Finally, the pitch rises on the last syllable of the phrase (*er* in *player*).

e. B: He got tickets for the Fellini \dou/ble feature.

Using F0 prosody notation from Pierrehumbert (1980), the fall-rise as characterized by Ward and Hirschberg (1985) for a single-accented sentence would be L* + H L- H%[2]. In this notation, falls (low tones) are marked by L; rises (high tones) are marked by H; stress is marked by *, lack of stress by -, and boundaries by %. So, the nuclear accent is L* + H: the primary stressed syllable is an L tone with a low F0, and an F0 peak (an H tone) is in the following syllable. Then there's a low phrase accent (L-) and a sentence-final rise (a high boundary tone).

It isn't clear to what extent different accounts are committed to the acoustic account of

[2]The original text describes fall-rise as L* + H- L- H%, but I use the more modern ToBI standard which now reserves '-' for intermediate phrase accents.

fall-rise that is implicated by the L* + H- L- H% notation; Ward and Hirschberg (1985) use this notation, but not all accounts use it. To mark fall-rise in the examples below, I will use Ward and Hirschberg's notation of "\x/", where it is understood that there's the abrupt fall and rise on x, a fall just following x, and a phrase-final rise as well.

A limitation of past studies is that not enough data support the acoustic characterization of fall-rise. For example, when the fall-rise contour is provided with as many details as in Ward and Hirschberg (1985), the amount of supporting data could be improved: Ward and Hirschberg (1985) mention that they use a corpus of a year's worth of utterances from service encounter exchanges, radio and television programs, and informal conversations, totaling 'over a hundred' utterances, but it isn't clear how many of these utterances are fall-rise cases. Further, no statistical analysis is provided for these fall-rise contours. Additionally, even so, Ward and Hirschberg (1985) use relatively more data than most; for example, Jackendoff (1972) does not cite data beyond specific intuitions for several example sentences. This question of the acoustics of fall-rise is additionally important because if accounts use a different acoustic characterization for fall-rise, it's unclear whether they are taking the same falling-rising prosody as their object of study in the first place.

The next sections turn to the question, what does fall-rise do? Why does fall-rise mark the inverse scope interpretation of *all*-negation? I review the disambiguating mechanism suggested by several of the key accounts in past studies. Broadly, fall-rise is argued to mark some connection between the utterance and the preceding context, or it marks the speaker's attitude about this connection.

6.1.2.1 Fall-rise associates negation with the focus

Jackendoff (1972) predicted that a fall contour – for Jackendoff, a pitch peak on the initial quantifier and a sentence-final pitch fall – corresponds to intended surface scope. On the

other hand, a fall-rise contour – here, a pitch peak on the initial quantifier and sentence-final pitch rise – corresponds to inverse scope. (He called these two contours Bolinger's (1958) Accents A and B, respectively, though in fact both are forms of Bolinger's Accent A.)

Figure 6.2: Jackendoff (1972)'s falling 'A' accent (left) and fall-rise 'B' accent (right).

Simply, "a contrast in meaning ... is produced by a difference in the choice of pitch accent" (Jackendoff, 1972, pp.352). That is, the two intonation contours mark the information status of the negation: negation is either part of the information that is presupposed or asserted. A falling contour associates negation with the presupposition; a fall-rise contour associates negation with the focus (the information that's asserted).

Broadly, asserted information is new or not shared by the listener, while presupposed information is shared, or given, as part of the conversation scenario. Specifically for Jackendoff, the semantic representation of a sentence is divided into 'focus' and 'presupposition', where stress marks the focus constituents. The presupposition is formed when the focus in the semantic interpretation is replaced by a variable (see the examples below; the lexical item containing the focus syllable is capitalized). Furthermore, high pitch marks the focus. In an *all*-negation utterance, the high pitch is on the *all*. (Returning to Ward and Hirschberg's notation, this focus syllable is the locus of the sharp fall-rise.)

For an utterance like *All the men didn't go*, the speaker conveys surface scope against a negative presupposition, in other words, the presupposed information that there exists some quantity Q of men who didn't go. That quantity is *all*, so, none of the men went.

(12) ALL the men didn't go.

a. Intonation: Fall (A accent)

b. Presupposition: $\{Q : Q \text{ of the men didn't go}\} \neq \emptyset$

c. Assertion: all $\in$ {Q : Q of the men didn't go}

d. Interpretation: surface scope ("none went")

On the other hand, the speaker conveys inverse scope against a positive presupposition, in other words, the presupposed information that there exists some quantity Q of men who did go. It is asserted that *all* is not included in the quantity of men who went. So (while some men went), it is not true that all of the men went, and so not all of the men went.

(13) \ALL/ the men didn't go.

a. Intonation: Fall-rise (B accent)

b. Presupposition: {Q : Q of the men did go} $\neq \emptyset$

c. Assertion: all $\notin$ {Q : Q of the men did go}

d. Interpretation: inverse scope ("not all went")

Here, it happens to be the case that the most prominent pitch accent falls on the operator that takes widest scope. But the proposal is that an overall intonation contour corresponds to an overall interpretation; Jackendoff (1972) doesn't suggest a phonemic breakdown of the pitch contour in a way that directly associates a segment of the utterance with a certain scope.

To further clarify this proposal for how fall-rise disambiguates, consider the case for negation-*because* utterances: Jackendoff (1972) suggested further that fall-rise would disambiguate neg-*because* utterances in just the same way. (Note here, that the fall-rise contour would, similarly to *all*-negation, lead negation to take highest scope, but this means that fall-rise is predicted to lead to surface scope for neg-*because*, as the surface scope interpretation is neg>because.) As far as I understand, the reasoning would go as follows for his unfortunate example (14).

(14) Max doesn't beat his wife because he \LOVES/ her.

a. Intonation: Fall-rise (B accent)

b. Presupposition: $\{X : \text{Max beats his wife because } X\} \neq \emptyset$

c. Assertion: he loves her $\notin \{X : \text{Max beats his wife because } X\}$

d. Interpretation: surface scope ("Max loving his wife is not the reason that he beats her; he beats his wife because of another reason")

This proposal in Jackendoff (1972) was highly inspiring to subsequent studies, though there are many limitations to the analysis. One of the main limitations of this analysis is that it is in fact the least clear when it is applied to *all*-negation constructions (Jackendoff also explains how fall-rise disambiguates other kinds of simpler sentences in a clearer way). It's also not true that fall-rise removes negation from the presupposition. Ward and Hirschberg (1985) argue for example that in (15), there is a negative presupposition of *X doesn't like San Francisco*, even when the sentence is uttered with fall-rise.

(15) A: How can anyone with any sense not like San Francisco?

B: \Bill/ doesn't like it.

In general, Jackendoff proposes a semantics of intonation, where intonation marks focus and focus-related variables. Subsequent proposals for why fall-rise disambiguates involve a pragmatics of intonation: intonation disambiguates because it conveys speaker attitude or relates some utterance entity to discourse entities in a structured way.

6.1.2.2 Fall-rise signals that the focused expression is a member of a set

In contrast, Ladd (1980) suggested that fall-rise marks inverse scope through marking set-member relationship between context and focus. That is, fall-rise focus tends to signal that

the focused expression is a member of a set in the preceding discourse, or more broadly, it signals a 'narrowing' of the contextually-provided set. For example, in a sentence like *I fed the \cat/* the fall-rise intonation signals that *cat* is a member of the set *animals*: while the speaker fed the cat, they did not feed all the animals. In general, this semantics of fall-rise leads to a wide range of predictions about the effect of using fall-rise for different sentences and contexts, including not only single-focus declarative sentences but also interpretations of superlative uses and effects such as polite softening.

Turning to how fall-rise tends to shift the scope of negation, as an explanation of why fall-rise leads to the surface scope interpretation of neg-*because*, Ladd's reasoning is actually quite close to Jackendoff's: negation ends up being associated with the reason in the *because*-clause, rather than with the action that's negated in the first overt clause. If a fall-rise is produced over *John doesn't drink because he's un\happy/*, then the reasons (for why John drinks) becomes the contextually-defined set to which the speaker relates the focused item (the reason of being unhappy). The negative associates with this focused reason, the argument goes, and we infer that John drinks, not because of this reason of being unhappy, but due to some other reason.

Turning to *all*-negation such as *\All/ the men didn't go*, a fall-rise intonation on *all* suggests that the negative applies not to the entire group of men but rather to a subset of them, effectively shifting the interpretation to the inverse scope *Not all the men went.*

A challenge here is to understand how *not all* is a subset of *all* within this framework. One possibility is that *some but not all*, rather than *not all*, is the subset. Additionally, I believe that an aspect of the analysis which isn't mentioned overtly for *all*-negation, but which is mentioned for the neg-*because* case, is that Ladd considers surface scope to be default. Therefore surface scope preference is predicted in the absence of a reason to deviate from surface scope, in other words, in the absence of fall-rise. If this is indeed part of Ladd's framework, then reasoning over alternatives may also play a role with *all*-negation: the listener

knows that the speaker would have used the default falling contour to signal default surface scope. Since the speaker deviated from that intonation with fall-rise, and one interpretation of *all*-negation that is compatible with fall-rise is inverse scope, then the speaker probably intended inverse rather than surface scope.

6.1.2.3 Fall-rise conveys speaker uncertainty

A third influential argument comes from Ward and Hirschberg (1985), who suggest that fall-rise intonation expresses speaker uncertainty about how the utterance fits into the context – uncertainty about some salient relationship between discourse entities, including (but not limited to) Ladd's set-membership. For Ward and Hirschberg (1985), scalar *partial ordering* relations (which are defined formally) provide the basis for the felicitous use of fall-rise. For example, for the ordering 'is a part of' defined on the set 'parts of a dissertation', felicitous uses of fall-rise include:

(16) A: Did you read the first chapter?

a. B: I read the first \half/ of it.

b. B: I read the whole disser\ta/tion.

c. B: I read the \third/.

The speaker must then recognize that a particular value or entity in the discourse lies on a scale in the first place. The discourse function of fall-rise is then to convey speaker uncertainty about the use of the perceived scale (though a speaker may merely want to convey uncertainty for purposes of politeness, irony, or deference.) This uncertainty may be about whether it's appropriate to evoke a scale at all, about which scale to choose, or about the choice of some value on the scale.

Ward and Hirschberg (1985) argue that falling and other contours would be as appropriate as

fall-rise for all these kinds of uncertainty. They specifically argue that it's not the case that the function of fall-rise is to pick out a member of a set (Ladd, 1980), select a variable from the background (Jackendoff, 1972), or convey some sense of continuation or incompletion (Pike, 1945), since falling and other contours would be able to convey the same thing; the function of fall-rise is to convey uncertainty (although, further, there are other ways of conveying uncertainty, including other intonational ways).

An interesting aspect of all three arguments about the kind of uncertainty conveyed by fall-rise, is that this uncertainty could also be framed as uncertainty about what the QUD is or whether the speaker is resolving the QUD. This might be a more general kind of uncertainty than the uncertainty regarding scales that Ward and Hirschberg (1985) describe, but it's also simpler.

With this account of fall-rise, Ward and Hirschberg (1985) argue that there is no scope-disambiguating function of fall-rise: "that it is not fall-rise, but context or other co-occurring linguistic phenomena, that perform these [disambiguating] functions" (p. 770). Specifically, the scope interpretation of *All the men didn't go* would involve uncertainty about whether *all* is to be interpreted as scalar: they suggest the example of the context of a rehearsal for a men's choir, where the choirmaster asks his subordinate whether *all the singers had attended some rehearsal*, and in uttering *All the men didn't go* with fall-rise, the subordinate would convey uncertainty about whether the choirmaster was interested in a simple yes-no response, or whether he wanted to know what portion of singers had missed the practice.

Again, possibly, this argument could be rephrased as uncertainty about the QUD: that in using fall-rise, the subordinate would convey uncertainty about whether they are sufficiently addressing the QUD *Did all the men go?*, or whether the QUD is something else which is not sufficiently addressed by the utterance (e.g., *How many of the men went?*).

One of their main points, though, is that an utterance can have a certain scope due to

a disambiguating context, regardless of intonation. In example (17), they argue that the utterance has inverse scope even with a falling rather than a fall-rise intonation. In example (18), they argue that the utterance has surface scope even with a fall-rise intonation: in this case, the speaker conveys Type I uncertainty, wondering whether A simply wants to know which meeting at least some of the men missed, or whether he should evoke a quantifier scale (as he does).

(17) Inverse scope without fall-rise

a. George said that everyone had left for the game by five, but I know that all the men didn't go that early.

(18) Fall-rise without inverse scope

a. A: The foreman wants to know which union meeting some of the men missed.

b. B: \All/ of the men didn't go to the last one.

In sum, for Ward and Hirschberg (1985), fall-rise conveys a conventional implicature (Grice, 1975). Fall-rise meets all the criteria for a conventional implicature: it makes no contribution to the truth conditions of an utterance but it constrains an utterance's appropriateness. Its contribution is also detachable in the sense that it is always possible to substitute falling intonation for fall-rise, resulting only in the failure of falling intonation to convey uncertainty, such that a distinct, truth-conditionally equivalent utterance (the equivalent utterance with falling intonation) fails to convey the implicature of uncertainty associated with fall-rise. On the other hand, the fall-rise contribution is not cancelable, in the sense that a statement of certainty following the use of fall-rise would be weird and contradictory. For example, according to Ward and Hirschberg's intuitions, (19b) would be weird relative to (19a) and (19b), because (19b) simultaneously conveys speaker uncertainty and uncertainty about the same matter (whether the listener would like to eat pie).

(19) A: Do you have jello?

a. B: We have \pie/.

b. B: We have \pie, which we know you won't eat.

These sections have reviewed different accounts of why fall-rise intonation would map to inverse scope interpretations of *all*-negation. Although these accounts suggest different disambiguating mechanisms, they all agree on the association between fall-rise and inverse scope (though they disagree on the strength and consistency of this mapping). How much evidence is there for the association between fall-rise and quantifier-negation in the first place? One of the characteristics of all the studies mentioned till this point is that they have not cited a lot of examples of fall-rise used naturally with quantifier-negation, and none have cited behavioral data. The next section turns to the empirical evidence for a mapping between fall-rise and inverse scope of *all*-negation.

6.1.2.4 Empirical evidence for fall-rise use over *all*-negation

Syrett et al. (2012) found weak empirical evidence that different scope interpretations of *all*-negation utterances receive distinct prosodic cues. Syrett et al. (2012) measured 19 native English speakers' prosody in their productions of *all*-negation utterances (among other utterances with different forms of ambiguity) embedded in short contexts that favored particular scope interpretations. The three *all*-negation utterances and their contexts are reproduced below.

(20) *Surface scope-favoring contexts:*

a. The township decided to plant magnolia saplings a number of years ago to line a path through the park. They have experienced lovely blossoms every year. However, this year the area is experiencing less-than-standard rainfall, which

means that they expect the magnolias to straggle this year, with only a few surviving. In fact, I think the situation is much more dire than that. **All the magnolias won't bloom.** They'll just have to wait till next year.

b. With the weather turning colder, Mandy was going through her closet looking for her winter coat. She thought she had remembered that the lining on this particular jacket was in pretty bad condition, and it would all have to be removed. When she found it, she was pleasantly surprised. **All the wool lining wasn't worn.** Only the sleeves needed repair.

c. Some of the girls in the neighborhood decided to throw a party, where they would help each other apply makeup in preparation for the upcoming dance. The girls anticipated that some of their moms wouldn't let them wear eyeliner. It turns out that the moms were all on the same page. **All the moms didn't allow eyeliner.** This didn't come as a real surprise.

(21) *Inverse scope-favoring contexts:*

a. A few years ago, the township decided to plant magnolia saplings to line a path through the park. The saplings on the north side were planted mainly in sand and haven't been getting nearly enough nutrients. However, the soil near the south side is rich, and the magnolias are thriving there. **All the magnolias won't bloom.** But I bet the ones on the south side will.

b. Mandy was in need of a heavy winter jacket but had a limited budget. She was hoping to find one when she went to the thrift store, even though she knew there would be a chance that some of the lining would be in need of repair. Eventually, she found a nice, warm jacket. When she looked inside, she couldn't believe how lucky she was. **All the wool lining wasn't worn.** The mission was a huge success!

c. Several moms were helping their daughters get ready for the upcoming school

dance. This is a progressive school, and moms are usually lenient about certain things, so even the younger girls thought their moms would approve of eyeliner. But at the dance only the older girls were wearing it. **All the moms didn't allow eyeliner.** Only the moms of the older girls let their daughters wear it.

No reliable sentence-final fall vs. fall-rise F0 pattern was found across speakers for surface vs. inverse scope of the *all*-negation utterances. Fall contour was generally preferred overall. Surface scope received a fall contour more often (93.4%) than inverse scope received a fall contour (89.1%), but the difference was not significant. Further, there was no significant difference in both the pitch level and duration of the quantifier *all*, nor in the final word duration, between intended surface vs. inverse scope. However, Syrett et al. (2012) found that some participants did tend to produce a fall contour for surface scope and a fall-rise for inverse scope. Thus, the expected prosodic distinction between intended scopes appeared to be possible, but unpredictable and speaker-specific.

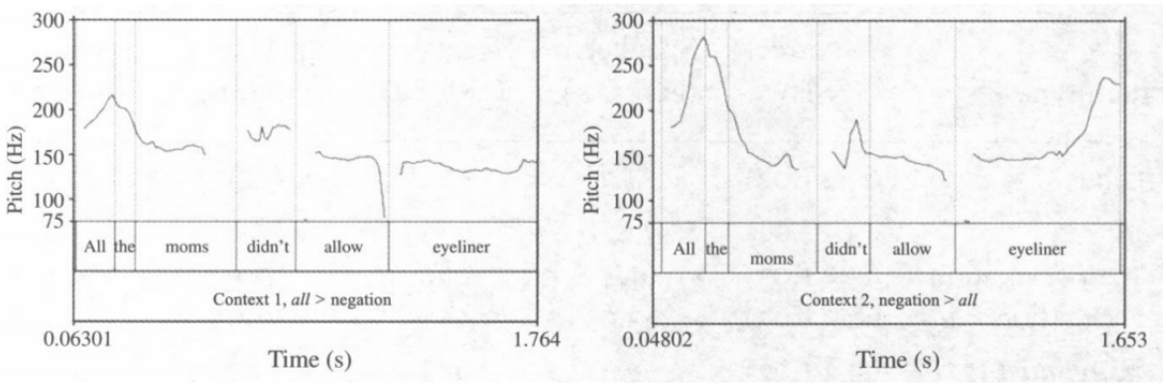


Figure 6.3: Syrett et al. (2014)'s examples of a speaker's fall (left) and fall-rise (right) intonational contour as gathered by Syrett et al. (2012).

The limitations of this study were that participants were asked to read aloud the short texts without an interlocutor for about an hour. This method might have suppressed or interfered with the role of prosody in naturalistic production, because participants might have felt that the purpose of their productions was not communication. A second limitation is that only a few prosodic measures were considered (the duration of *all* and the final word,

and utterance-final F0 contour), which leaves the possibility that the full utterance-level F0 contour, the durations of each word, and other factors like intensity, did differentiate intended scope.

There is also some evidence that when speakers use reliable surface-level cues to signal their intended interpretation, listeners use these cues to arrive at the interpretation that was intended by the speaker. Syrett et al. (2012) and Wu and Ionin (2020) conducted such perception studies and found that listeners were above-chance at inferring intended scope of *all*-negation utterances given the intonation contour, although they were sometimes not above chance at inferring the correct intonation contour given the context. For example, Syrett et al. (2012) found that listeners, hearing an *all*-negation utterance, correctly categorized it with the speaker's intended scope interpretation 63.9% of the time when surface scope was intended, and 66.2% of the time when inverse scope was intended. (Both average responses were significantly above chance level.) In this experiment, the stimuli were a range of utterances from four speakers from the previously-mentioned production experiment, including two minimal pairs of *all*-negation utterances in contexts favoring surface or inverse scope. Participants viewed the target sentence in the middle of the screen (e.g. *All the moms didn't allow eyeliner*) and heard a speaker's production of the sentence three times; participants then indicated their scope interpretation on a second screen, by choosing between two possible continuations (e.g. *A: They were all in agreement; B: Only the moms of the older girls let their daughters wear it*).

Listeners were less consistently successful at doing the opposite: using a discourse context (favoring one interpretation over the other) to choose the prosodic version of a sentence that was produced for that context. In this experiment, participants saw a context line by line in a self-paced cumulative window fashion (e.g., reproduced in (22)). Participants then heard two versions of the target sentence sequentially, each accompanied by the option 'A' or 'B' on the screen. The first version A was always a falling contour, and B a non-falling contour.

Participants then chose between A and B on a third screen.

(22) Context favoring the surface scope interpretation of *All the moms didn't allow eyeliner.*

a. Several young girls wanted to have a make-up party together.

b. Some of them thought their mothers wouldn't let them use eyeliner. In fact, the moms were all on the same page.

c. <<**All the moms didn't allow eyeliner.**>>

d. The girls were limited to mascara and blush.

When the context favored surface scope, listeners chose the correct falling contour 76.9% of the time. However, when the context favored inverse scope, listeners chose the correct non-falling contour 53.1% of the time (not above chance). This rate of responses for inverse-scope-favored all-negation was also the lowest of all the different forms of ambiguity tested in their study. Syrett et al. (2012) speculate that the non-falling contour was chosen for inverse-scope-favoring contexts only half the time because the inverse scope interpretation of *all*-negation is relatively difficult to access for a range of grammatical and pragmatic reasons. Another issue may be a preference for falling contours in general, as a default prosody.

6.1.3 Expecting and accounting for variation in prosody

Intonation might reliably disambiguate between surface and inverse scope interpretations of a quantifier-negation sentence, in a one-to-one mapping between intonation and interpretation (Jackendoff, 1972). But a variable prosody-scope mapping, as suggested by most later researchers, would be consistent with the general finding that different speakers and listeners implement prosody and prosodic functions with a high degree of variation (Cole, 2015; Xie et al., 2021). Even for the more well-documented phenomenon where prosody *can* disambiguate a range of syntactic ambiguities, speakers don't reliably and consistently

produce these prosodic cues (Kraljic and Brennan, 2005).

The deeper reason for this general finding may be that prosodic features like pitch, intensity, and duration are continuous variables in time, so they have many degrees of freedom (Chigusa Kurumada, personal communication). And prosody not only varies as a function of speaker-specific characteristics that can be normalized away (e.g., physiology, gender and sex) or can be learned as talker-specific distributional statistics (Xie et al., 2021) but also simultaneously marks for multiple features (e.g., syntactic boundaries, emotion) of a sentence or a speaker's meaning (Cole, 2015), which may override or interfere with the prosodic realization of scope. In other words, it might be hard to measure a strong or consistent prosody of scope not only because prosody is variably implemented but also because prosodic features are already geared towards other, also variable functions.

Furthermore, prosody may be optional and variable because producing prosodic contours which differ from more common or unmarked prosodies might be cognitively effortful from the speaker's perspective, such that an efficient speaker would prefer to use an unmarked prosody in general, especially when the speaker has reason to expect that the listener will successfully understand the intended meaning even without hearing the more effortful prosodic contour.

Additionally, these final-rise contours could be confounded with sentence continuation rises. A sentence continuation rise is the kind of rise in pitch you might expect over *log* in an utterance like *Every horse jumped over the log, but every horse didn't jump over the fence* (Syrett et al., 2014). It's possible that speakers who intend inverse scope may be more likely to immediately follow the inverse scope-intended phrase with a continuation of the utterance, thus producing a continuation rise on the end of the inverse scope-intended phrase, which looks identical or similar to scope prosody researchers' fall-rise contour.

Finally, turning to how we can understand the prosody of the corpus data, perhaps the most obvious source of prosodic variation is the variation in sentence structures: for example, the

corpus contains both the utterance *everybody wasn't fine* and the utterance *every particular moment there wasn't quite as fun or quite as happy as the movie college education*, which are likely to differ significantly in their prosodies only because of syntax and regardless of scope interpretations. There are also different statement types which are likely to differ in prosody regardless of scope (e.g., the *every*-negation in *When do you learn that everything was not as it appeared?* vs. the *every*-negation in *... we're still comfortable, and everything is not as bad as it seemed.*) On the other hand, the predictions about prosody from the literature tend to use examples of quantifier-negation which are in fact quite similar to each other, such as the following:

(23) All the men didn't go.

(24) All the moms didn't allow eyeliner.

These utterances are all declarative phrases. Furthermore, there is no additional modification on the subject, no adjuncts, no interruptions or restarts. The utterances are all single, complete, uninterrupted, and unmodified quantifier-negation. So, to the extent that predictions were formulated only for such uses, predictions may not hold for other kinds of utterances.

6.1.3.1 Accounting for some variation

The variation in the sentence types and structures in the corpus suggests two paths of analysis. One path would be to say that, if we are looking for a prosody of scope for *every*-negation, we should consider any and all uses of *every*-negation, regardless of structure. (Bad enough that prosody is likely to be variably implemented for other reasons.) However, a limitation of this approach is that it in some sense minimizes our ability to find a scope of prosody if we don't take any steps to identify wherever prosody may be obviously overridden by other

variables. Thus, another path is to return to the question of what exactly are the sentence structures that are the intended object of the predictions from the literature for a scope of prosody. That is, what are the key features of the sentence structures that we know are predicted to reflect a scope of prosody? We can then address some of the variation in the corpus by coding for whether an item does or does not contain the key features of the objects of study in the literature about scope prosody.

I identified the following key features:

1. Declarative (rather than interrogative)
2. Unmodified subject (rather than modified subject)
3. Uninterrupted (rather than interrupted)
4. Contains an overt, single expression of a quantified subject and negated predicate (rather than having any of these expressions be null or repeated)

A limitation of this list of key features is that it was inferred by me and necessarily incomplete; there are many other features of the syntax and use of the quantifier-negation that are considered in the literature. This list of features is intended as a tool to serve as a basis for better analysis of the corpus data, to be able to take into account broad differences in structure that may reflect in a prosody that has nothing to do with scope, rather than to explain why any of these differences should result in a certain prosodic difference.

6.2 Prosody of naturalistic *every*-negation

This section addresses several specific questions about the prosody of interpretations of naturalistic *every*-negation. Regarding the disambiguating role of prosodic phrasing: does

the existence or duration of a pause between the quantifier and negation signal surface scope? Second, regarding the disambiguating role of prosodic prominence: does a final pitch rise predict inverse scope preference? Finally, in general, can prosody over the quantifier and negation predict interpretations?

6.2.1 Coding the data

Before acoustic analysis, the *every*-negation items in the data were grouped according to several measures, mainly to address syntactic variation. Since there is a wide variation of item types in the data, these groupings were expected to potentially allow for better comparisons of prosody. Furthermore, within each item, I coded the timestamps of key syllables representing parts of the quantifier-negation that should be shared across all or most of the items, for example, the syllables of the quantifier and of the negation.

6.2.1.1 Statement type

The *statement type* coding estimates whether the *every*-negation use is a declarative or interrogative one.

To estimate statement type, each item in the corpus was coded for the type of text that immediately followed the *every*-negation: declarative main clause items were those that were followed by a period (example (25)), a comma or conjunction (example (26) and example (27)), and interrogatives were followed by a question mark (example (28)).

There are several limitations to this measure; it suffers from a problem of circular definition, since the text of the transcript – the decision to transcribe a period or a question mark – depends in part on the acoustics of the recording. Thus this measure is not intended to express any clear underlying distinctions between the different categories of statement types.

Rather from a practical standpoint, it was the best estimate I could imagine for capturing the variation in prosody that should be accounted for by statement type rather than intended scope, especially given that the predictions from the literature may only account for the kinds of quantifier-negation uses which are treated as if they clearly end in a period as part of a declarative sentence.

Examples of declaratives include:

(25) So I apologize if this is a painful question, but **everybody doesn't know this history**. Can you tell us, why are these languages disappearing?

(26) Joe says he hopes **everybody doesn't fall asleep during his show**, but if they do, he doesn't mind at all.

(27) President Obama ran as a person who's going to change politics but also was smart enough to know that when he got to Washington, **everything wasn't going to change overnight** and that he needed somebody who knew how to make deals on Capitol Hill, who had been through White House before and the crucible of living in the fishbowl ...

An example of an interrogative:

(28) In your remaining time in the Senate, this – the whole year, will you support, say, changing the cloture rule so that **every significant bill doesn't need 60 votes**?

6.2.1.2 Subject modification

The number and length of clauses in the quantifier-negation may also affect the prosody regardless of scope. One of the most clear distinctions between items was whether an additional phrase modified the subject (example (29)) or not (example (30)). The predictions in the literature regarding prosody are all based on sentences with no additional modifying clause on the subject, and so these predictions may not hold for the items with the additional modifying clause. For example, simply because these items are on average longer and more likely to contain prosodic boundaries within them, they may be more likely to exhibit pausing and less likely to exhibit the specific fall-rise contour hypothesized for inverse scope interpretations.

Subject restriction:

(29) Comey said this one started in July, and he said that's very young for an investigation of this kind, and **every day that Trump's ties to Russia are in the headlines is not a good day**.

No subject modification:

(30) See; **everybody can't root for the prisoners**.

6.2.1.3 Interruption

Corpus items were coded for whether they contained an interruption or restart; as mentioned above, predictions from the literature do not take interrupted sentences into account.

(31) Everything - it's very, very quiet - just doesn't feel the same.

6.2.1.4 Complete

Finally, predictions have been about uses of quantifier-negation that contain an overt, single expression of a quantified subject and negated predicate (rather than having any of these expressions be null or repeated). Thus, corpus items were coded for whether they deviated from a single, complete quantifier-negation use: deviations were considered to be multiple subjects (example (32)), trailing off (example (33)), or silent predicates such that there was no overt linguistic material following the negation (example (34)).

(32) Every school does not have – every school district does not have wraparound services for students in need.

(33) everyone I knew was not in ...

(34) I know **everybody can't**.

6.2.2 Within-item timestamps of key syllables

Although the different items are generally very different at the lexical level, almost all of the items have the quantifier *every*, a noun phrase subject, negation, and a predicate headed by a verb. I annotated each item for the timestamps of these key syllables, in order to better compare the prosodies of the different items. This segmentation system makes it possible to compare acoustic patterns at common points across all items.

First, I annotated each recording for its phone, syllable, and word timestamps. Phone and word timestamps were created automatically using the Montreal Forced Aligner (McAuliffe et al., 2017). Syllable timestamps were then created by hand, on the basis of the automated phone timestamps, to mark 1) the two quantifier syllables (*q1, q2*), 2) remaining syllables

before the negation, 3) the negation syllable (*neg*), 4) any remainder syllables in the predicate, and 5) any silences (*null*). One set of syllables, then, which is shared for many items, is the set of the quantifier syllables, the negation syllable, the syllable before the negation, and the several syllables after the negation. (That is, while all items contain a quantifier and negation, most but not all items contain at least several syllables after the negation). Another variable from this coding scheme is the existence and duration of pauses between the quantifier and the negation. A final set of syllables of interest are the last two syllables of an item's predicate, whatever they may be, as long as they are after the negation (most but not all items have at least two syllables following the negation).

6.2.3 Coding results

In the NPR corpus, 277 items are declarative and 8 are used as questions. 108 items have a subject modification (179 items have no such modification). 7 items are interrupted (280 are not). 11 items are incomplete (276 are not).

6.2.4 Pitch and scope in the NPR corpus

I investigated the first prediction, that the existence or duration of a pause between the two quantifiers signals surface scope. Then, I tested the second prediction of whether a final rise in mean F0 predicts inverse scope preference. Finally, I investigated whether an overall pitch contour reflects scope, where the pitch contour was measured as mean F0 at the syllables that were shared across many of the corpus items.

6.2.4.1 Does pausing between the operators predict surface scope

To estimate the degree to which a speaker produced a break between the quantifier and negation, I summed the duration of the total silences after the quantifier and before the negation in each recording. This measure, pausing duration, has the advantage that it is quantitative and it is also relatively conservative – pausing is only one of multiple ways to produce a break.

Subject modification and statement type were expected to capture some variance in interpretations and pausing, though there are no clear predictions from the literature about how exactly, so these two variables were included as predictors of interpretations in the model, so that the role of pausing could be captured as separate from modification and type. The few items with interruptions and restarts were excluded from analysis, since any pausing in these cases may be less plausibly a reflection of scope interpretation. Items with a silent predicate were included in analysis; however, whether or not the predicate was silent was not included as a predictor of interpretations, since the focus of this analysis is on the production *before* the negation.

To assess significance, I used a mixed effects model predicting mean item scope interpretation (this was the audio-in-context data from the corpus annotation experiment described in Section 4.2) by pausing duration, with additional predictors for subject modification, statement type, and the interaction of both these additional variables with pausing duration, as well as with random intercepts for participants in the interpretation-gathering experiment. Model results are shown in Table 6.1. The predictions of this model are visualized in Figure 6.4. All coefficients are significant. The positive intercept indicates (as is consistent with previous findings on this dataset) that interpretations are generally closer to inverse scope than surface scope, for an item that has no pausing, no subject modification, and is a declarative use.

In line with the primary prediction, there is a main negative effect of pausing duration: items

with greater pausing before the negation are more likely to have surface scope. There is also a main effect of subject modification and statement type: items with subject modification are more likely to have surface scope, and questions are more likely to have inverse scope than declarative cases. There is also an interaction between modification and pausing duration: the surface-scope-facilitating effect of pausing duration is weaker for subject-modified items than for unmodified cases. On the other hand, the interaction between statement type and pausing duration indicates that the surface-scope-facilitating effect of pausing is stronger for questions than for declarative uses.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	1.969	0.1385	14.217	<2e-16
Pausing Duration	-0.002305	0.0001968	-11.711	<2e-16
Modification	-1.725	0.05012	-34.417	<2e-16
Statement Type	0.5465	0.1398	3.910	0.0000934
Pausing Duration*Modification	1.831e-03	0.0002318	7.899	3.40e-15
Pausing Duration*Statement Type	-0.001430	0.0004285	-3.337	0.000852

Table 6.1: Results of a mixed effects model with total pausing duration before the negation, subject modification (unmodified/modified), and statement type (declarative/question) predicting mean scope preference per item (log-odds; higher values indicating greater inverse scope preference), with the maximal random structure of random intercepts for the participants.

Thus, the main prediction is borne out for the role of a phrasing break: using one measure of a break, pausing duration, before the negation, I found that surface scope is facilitated.

6.2.4.2 Does final pitch rise predict inverse scope

To estimate the degree to which a speaker produced a final pitch rise, I calculated the difference in the mean F0 between the last two syllables of each recording's predicate. This measure, final mean F0 change, similarly to pausing duration has the advantage that it is quantitative.

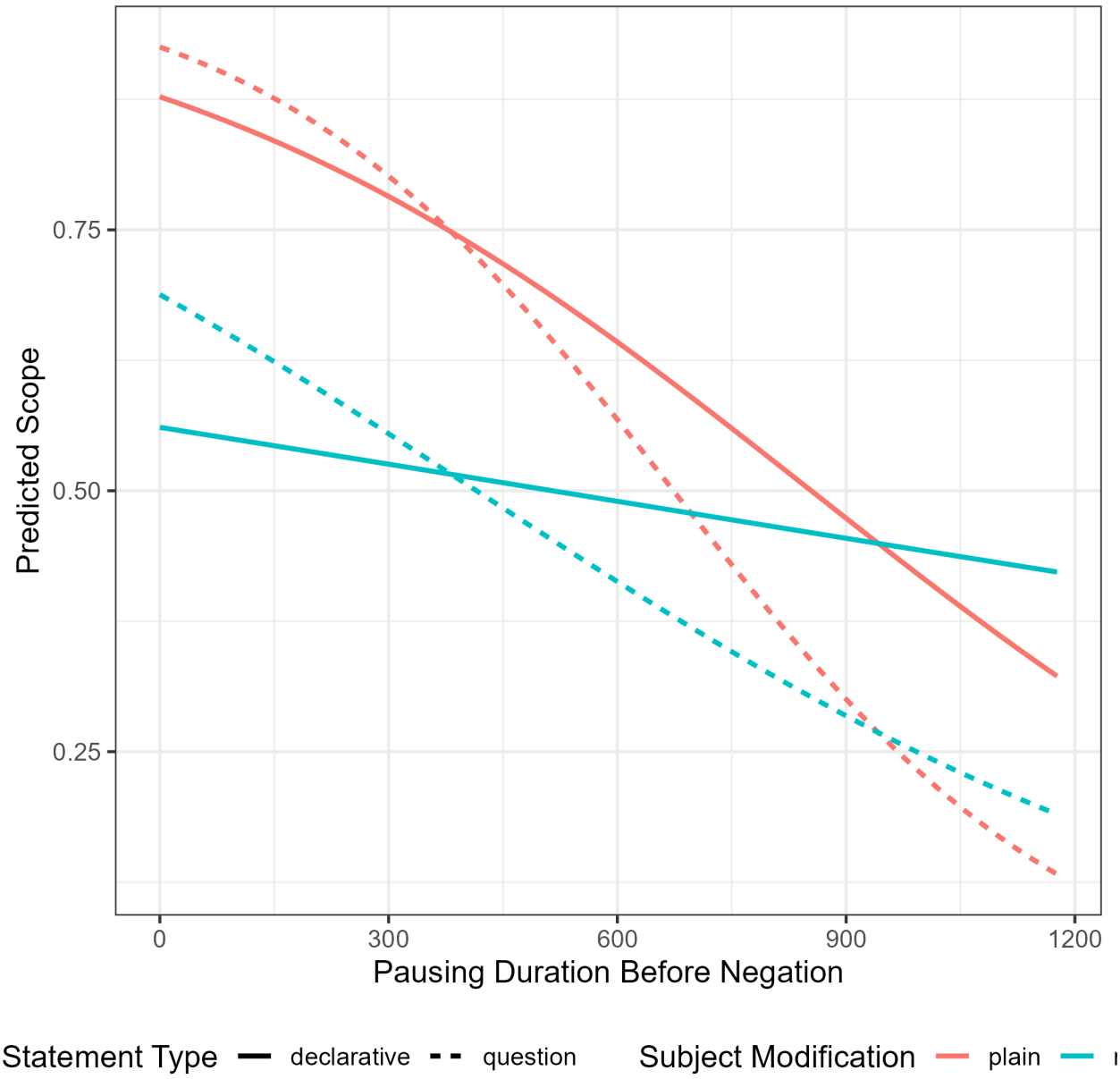


Figure 6.4: Pausing duration before the negation and the predicted mean item inverse scope preference for an average item in the corpus, depending on whether the item is declarative or interrogative and on whether the subject is modified or not, according to the mixed effects model in Table 6.1.

As before, subject modification and statement type were expected to capture some variance in interpretations and final rise, though there are no clear predictions from the literature about how exactly, so these two variables were included as predictors of interpretations in the model, so that the role of final rise could be captured as separate from modification and

type. The few items with interruptions and restarts were excluded from analysis, since any pausing in these cases may be less plausibly a reflection of scope interpretation. Items with a silent predicate were also necessarily excluded from analysis since they do not have an overt predicate.

To assess significance, I used a mixed effects model predicting mean item scope interpretation by final mean F0 change, with additional predictors for subject modification, statement type, and the interaction of both these additional variables with mean F0 change, as well as with random intercepts for participants in the interpretation-gathering experiment. Model results are shown in Table 6.2. The predictions of this model are visualized in Figure 6.5.

Again in line with the primary prediction, there is a main effect of final mean F0 change duration: items with greater rise are more likely to have inverse scope. There is also a main effect of subject modification and statement type: items with subject modification are less likely to have inverse scope, and questions are more likely to have inverse scope than declarative cases. No interactions are clearly significant.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	1.869	0.1449	12.903	<2e-16
Final Mean F0 Change	0.0009188	0.0002721	3.377	0.000738
Modification	-1.791	0.04858	-36.862	<2e-16
Statement Type	0.1679	0.1282	1.309	0.190546
Final Mean F0 Change*Modification	0.0002615	0.0004665	0.561	0.575093
Final Mean F0 Change*Statement Type	-0.006326	0.003616	-1.749	0.080331

Table 6.2: Results of a mixed effects model with the mean F0 difference between the last two syllables, subject modification (unmodified/modified), and statement type (declarative/question) predicting mean scope preference per item (log-odds; higher values indicating greater inverse scope preference), with the maximal random structure of random intercepts for the participants.

Thus, the main prediction is borne out for the role of prosodic prominence: using one measure of a final rise, change in mean F0 between the last two syllables of the predicate, I found

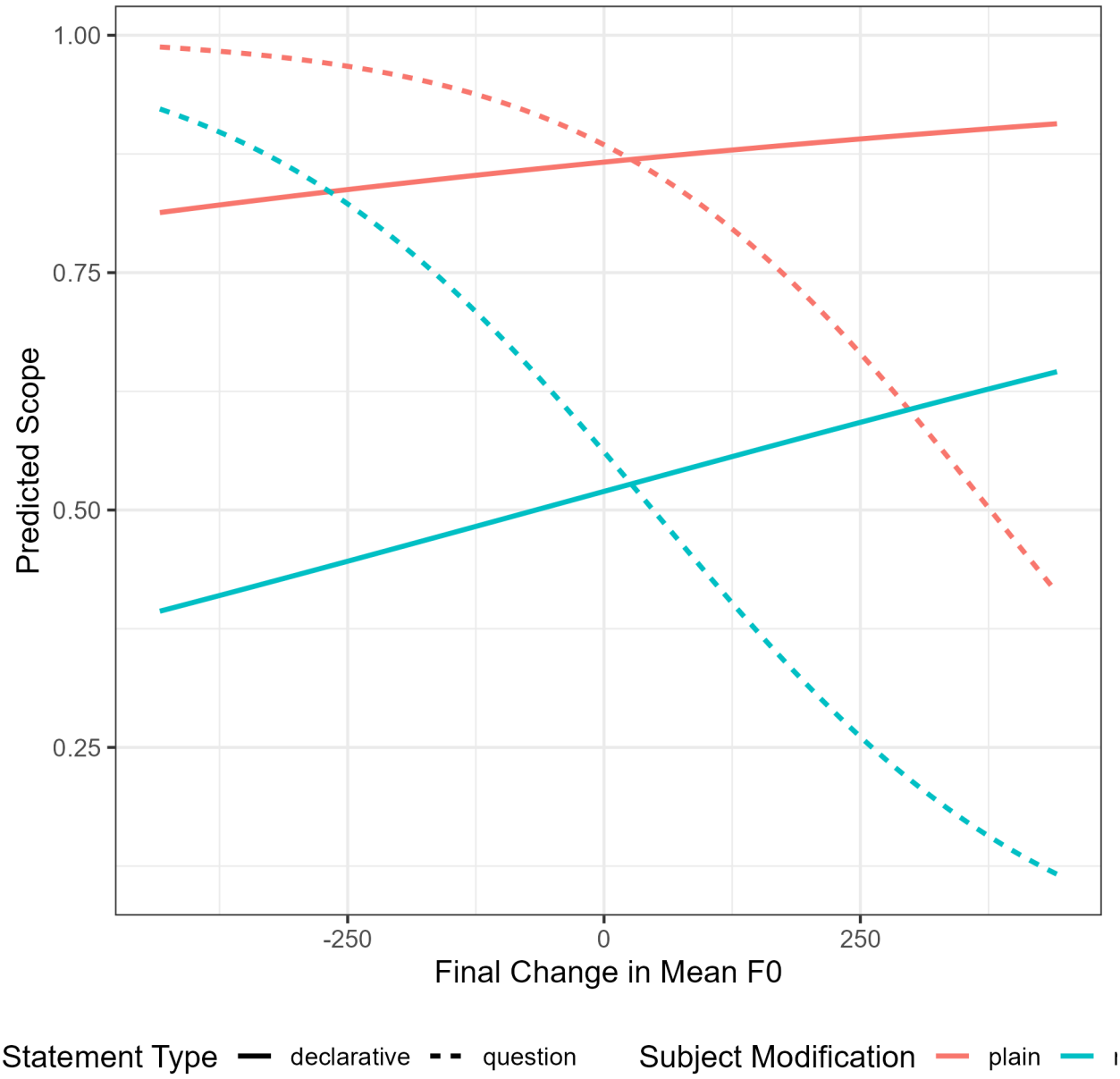


Figure 6.5: the mean F0 difference between the last two syllables and the predicted mean item inverse scope preference for an average item in the corpus, depending on whether the item is declarative or interrogative and on whether the subject is modified or not, according to the mixed effects model in Table 6.2.

that inverse scope is facilitated.

6.2.4.3 Does pitch reflect scope

In this analysis, I examined the influence of average item interpretation on the mean F0 trajectory across different syllables in a subset of recordings. Each recording was measured for the mean F0 trajectory at seven syllables: the first syllable of the quantifier (q1), the second quantifier syllable (q2), the syllable immediately preceding the negation syllable, the negation syllable (neg), and the three syllables immediately following the negation syllable. The primary goal was to understand how scope may be reflected in the shape of these F0 trajectories. I used generalized additive mixed models (GAMMs) with the bam function from the mgcv package in R, comparing a full model including scope interpretation with a nested model excluding it (for a tutorial please see Sóskuthy, 2017).

Because of the difficulty of interpreting the GAMMs models when many predictors are included in one analysis, instead of creating the same analysis as before, there were two separate analyses: the first was restricted to those items that shared the seven syllables points in common; the second analysis further restricted only to those items that were not questions, subject-modified, interrupted, or incomplete (in other words, removing those items that deviate in an obvious way from the items that are considered in the literature). Thus, the first analysis considered 213 recordings. The second analysis considered 120 recordings.

Analysis 1. Both models predict mean F0. The full model includes both the main effects of syllable and scope, the interaction between syllable and scope, and a random intercept for each recording. Specifically, it predicts mean F0 using smoothed terms for syllable to capture variations across syllables, a smoothed term for scope (mean scope interpretation per item, logit transformed) to account for its influence on average F0 values, and a tensor product interaction to flexibly represent the non-linear interaction between syllable and scope, using cyclic cubic regression spline bases with 6 degrees of freedom for syllable and 10 degrees of freedom for scope. Additionally, the model includes a random intercept for recordings to

accommodate variation between recordings. Model results are shown in Table 6.3, where the role of scope is not significant either as a main effect or interaction. However, for a more correct test of significance, this full model is compared to a nested model without scope.

Term	Estimate	Std. Error	t-value	p-value
Intercept	166.689	6.359	26.21	<2e-16 ***
Smooth Term	**edf**	**Ref.df**	**F**	**p-value**
Syllable	5.683	6.000	18.269	<2e-16
Scope	1.515	1.579	2.271	0.2005
Syllable*Scope	7.910	62.000	0.173	0.0942
Item	171.041	211.000	4.256	<2e-16

Table 6.3: Using all corpus items with at least three syllables after the negation (N=213), results of a generalised additive mixed model with syllable (quantifier-1, quantifier-2, before-negation, negation, after-negation, after-negation-2, after-negation-3) and scope (mean scope per item, logit-transformed) predicting for mean F0, with random intercepts for item.

The nested model serves as a baseline comparison by predicting mean F0 solely based on variations across syllables; it includes the main effect of syllable and the random intercept for recordings, but excludes the effect of scope. Specifically, it includes smoothed terms for syllable to capture their individual effects on mean f0 and a random intercept for recordings to account for recording-specific variability. I compared the two models using the compareML function from the itsadug package.

Model	Score	Edf	Difference	Df	p.value	Significance
Nested Model	8035.058	3				
Model	8030.893	7	4.165	4.0	0.080	

Table 6.4: Comparison of model fits for mean F0 trajectories; both models are for all corpus items with at least three syllables after the negation (N=213). The nested model does not include the influence of scope on F0; the full model does include scope.

The comparison indicates that the full model does not significantly improve the model fit compared to the nested model, with a p-value of 0.080. In other words, scope does not have a significant effect on the mean F0 trajectory across syllables, beyond the main effect of the

syllables alone. Specifically, the model predicts a mean F0 contour per syllable and scope as shown in Figure 6.6. Mean F0 begins relatively high at the quantifier and drops before the negation, regardless of scope. For example, consider the mean F0 on the second syllable of the quantifier: regardless of whether the item had an inverse scope (purple) or surface scope interpretation (red), the mean F0 was slightly lower at the second quantifier syllable relative to the first quantifier syllable, and relatively higher than the mean F0 before the negation.

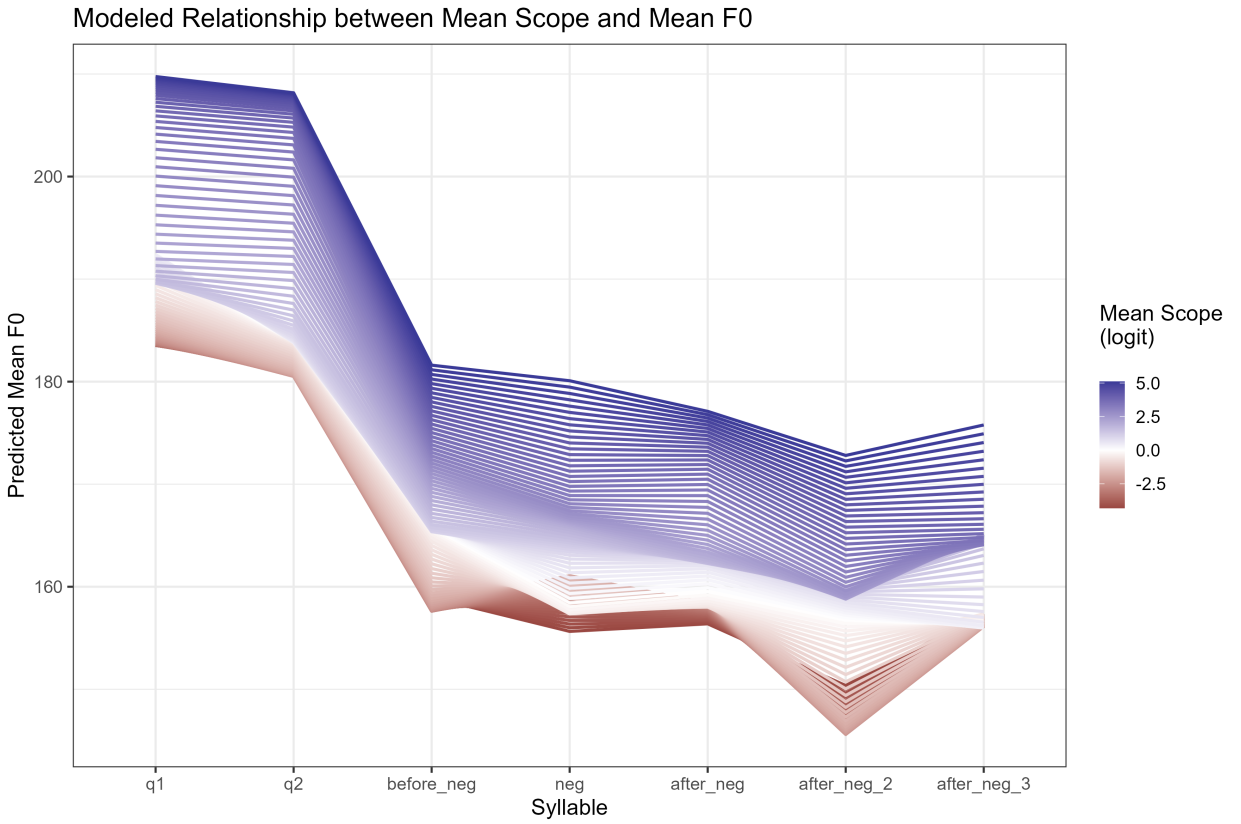


Figure 6.6: Modeled relationship between mean item scope (logit-transformed) and mean F0 at seven key syllables: the first two syllables of all the utterances (*every*), the negation syllable, and the syllables around it.

However, as mentioned above, this model considers all the corpus items which have at least three syllables after the negation, including those structures which have not been predicted in past studies to show a prosody of scope. The second analysis turns to items that were not questions, subject-modified, interrupted, or silent predicate.

Analysis 2. Again, both models predict mean F0, with the full model including both the main effects of syllable and scope, the interaction between syllable and scope, and a random intercept for each recording. Full model results are shown in Table 6.5, which shows in contrast to Analysis 1 that the role of scope is significant both as a main effect and interaction. To asses this significance more carefully, this full model is compared to a nested model without scope.

Term	Estimate	Std. Error	t-value	p-value
Intercept	169.343	6.663	25.41	< 2e-16 ***
Smooth Term	**edf**	**Ref.df**	**F**	**p-value**
Syllable	5.334	6.000	8.182	< 2e-16 ***
Scope	1.501	1.563	1.135	0.449750
Syllable*Scope	6.320	62.000	0.357	0.000246 ***
Item	96.089	118.000	4.368	< 2e-16 ***

Table 6.5: Using all corpus items with at least three syllables after the negation and that are uninterrupted, full, and declarative main-clause uses (N=120), results of a generalised additive mixed model with syllable (quantifier-1, quantifier-2, before-negation, negation, after-negation, after-negation-2, after-negation-3) and scope (mean scope per item, logit-transformed) predicting for mean F0, with random intercepts for item.

The nested model serves as a baseline comparison by predicting mean F0 solely based on variations across syllables; it includes the main effect of syllable and the random intercept for recordings, but excludes the effect of scope.

Model	Score	Edf	Difference	Df	p.value	Significance
Nested Model	4534.216	3				
Model	4525.771	7	8.445	4.0	0.002	**

Table 6.6: Comparison of model fits for mean F0 trajectories. The nested model does not include the influence of scope on F0; the full model does include scope.

The comparison indicates that the full model significantly improves the model fit compared to the nested model, with a difference in score of 8.445 and a p-value of 0.002. This suggests that scope has a significant effect on the mean F0 trajectory across syllables, beyond the

main effect of the syllables alone. The use of GAMMs allowed us to capture the non-linear interaction between scope and syllable regarding mean F0 – Figure 6.7 shows what this interaction looks like exactly. The more that an item had inverse scope (more purple lines), the more that speakers produced a pitch peak at the second syllable of the quantifier and a pitch drop to the negation. Conversely, the more that an item had surface scope (more red lines), the more that speakers started at relatively lower pitches, which rise to higher levels three syllables after the negation.

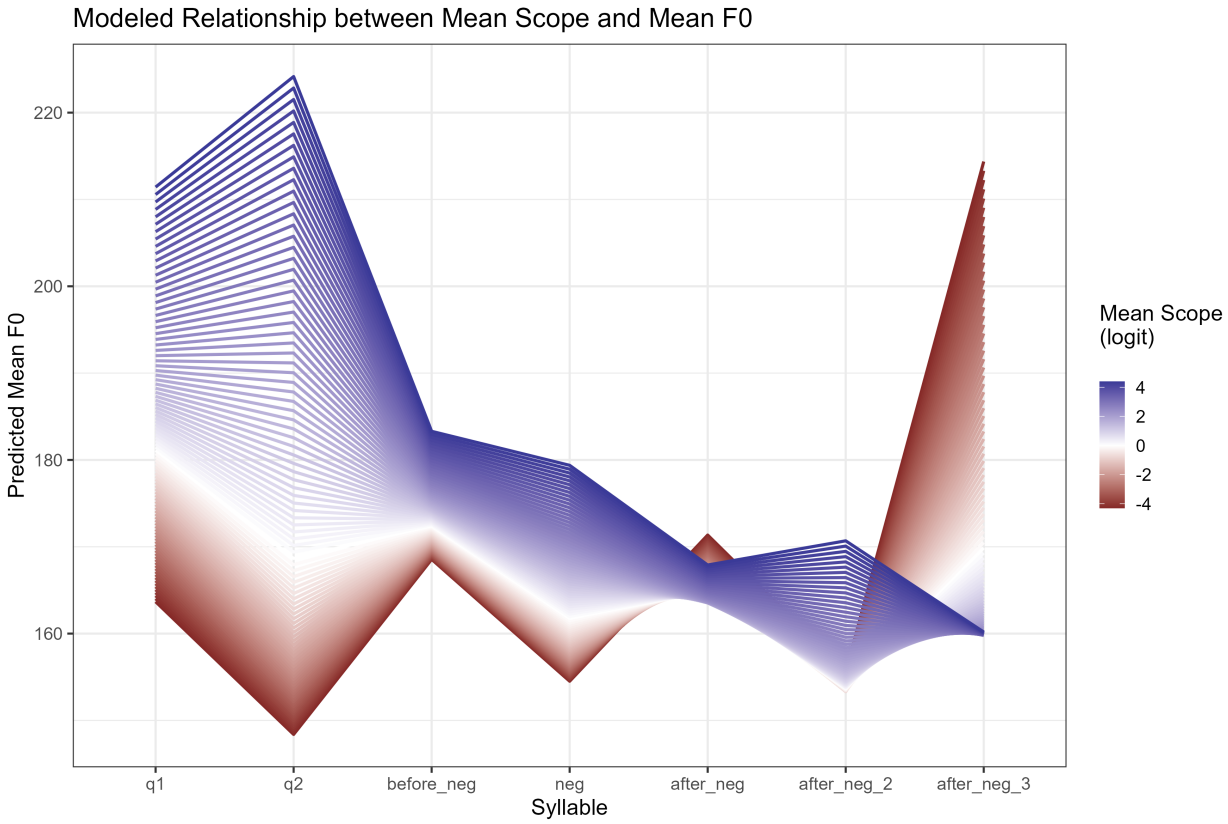


Figure 6.7: Modeled relationship between mean item scope (logit-transformed) and mean F0 at seven key syllables: the first two syllables of all the utterances (*every*), the negation syllable, and the syllables around it.

Thus, Figure 6.7 shows that speakers can use F0 to emphasize the quantifier more when they intend inverse scope. The pitch contour gradually flips to a relatively greater later emphasis the more that the speaker intended surface scope.

6.2.5 Discussion

I found evidence in favor of a few predictions from the literature regarding the prosody of naturalistic every-negation; in general, some aspects of pitch certainly reflect scope interpretation.

First, one measure of a prosodic phrasing break between the operators predicts scope: the total pausing duration between the quantifier and the negation is associated with greater surface scope. This finding is in line with the idea that the more that the scopally ambiguous utterance is produced in two 'chunks' which separate the two scope-bearing operators, the greater the surface rather than inverse scope interpretation.

Second, one measure of a final pitch rise also predicts scope: the change in mean F0 between the last two syllables of the predicate is associated with greater inverse scope. This finding is in line with the idea that the end of the utterance shows a key difference between the patterns of prosodic prominence for surface scope vs. inverse scope. The more that there is a pitch rise at the end of the utterance, the greater the inverse rather than surface scope interpretation.

Finally, a more exploratory analysis finds that scope predicts the mean F0 contour around the quantifier and negation. This analysis compares the mean F0 at seven syllables matched across the corpus items: the two syllables of the quantifier, the syllable before the negation, the negation syllable, and the three syllables afterwards. Thus, this analysis was restricted to the subset of corpus items that contained these syllables. For just those items that are similar to the sentences considered in past studies, the analysis finds that speakers use pitch to emphasize the quantifier more with inverse scope. Conversely, the more that surface scope is intended, the more that the pitch emphasizes the negation relative to the quantifier, although this later emphasis is not as marked as the utterance-initial emphasis with inverse scope.

With all these analyses, a challenge is that there are likely many additional aspects of the sentences and how they were used which should affect the prosody and may also affect the prosody of scope in an unforeseen way. The corpus items vary widely in terms of structure (e.g., the presence or absence of adjuncts), utterance length, the parts of speech and for that matter the lexical items themselves, and other factors. In order to assess how prosody reflects scope, I took into account whether items express certain key features of the quantifier-negation sentences that the prior literature considers. In other words, while there are many aspects of the sentences and their use which may additionally matter, I broadly took into account whether a corpus item was similar to the kinds of sentences previously considered in the literature. The main idea is that we should at least see a prosody of scope for those corpus items which are similar to the sentences in past studies, regardless of other aspects of these sentences which could affect prosody. The key features that I identified are that the utterance is declarative (rather than interrogative), has an unmodified subject (rather than a modified subject), is uninterrupted (rather than interrupted), and contains an overt, single expression of a quantified subject and negated predicate (rather than having any of these expressions be null or repeated).

When the naturalistic corpus is thought of in terms of these key features, the items with all these key features indeed demonstrate the patterns of a prosody of scope that I investigated – a surface-scope-facilitating pause between the operators, an inverse-scope-facilitating final pitch rise, and a mean pitch contour around the quantifier and negation which shows greater emphasis of the quantifier for greater inverse scope. For questions and items with a modified subject, pausing reflects surface scope and final rise reflects inverse scope, but there is no longer evidence for an overall pitch contour around the quantifier and negation which reflects scope.

Unexpectedly, corpus items that have a modified subject or are used as questions in fact show a different relationship with preferred scope than unmodified or declarative items, regardless

of prosody. Modified subject predicts surface scope, and questions predict inverse scope. Although it's not clear how to interpret these effects, one possibility is that the quantification of the modified subjects is slightly different than the quantification of the unmodified subjects, and as such may be less scopally ambiguous. Compare the unmodified (35) and the modified (36):

(35) Everybody wasn't blue.

(36) Everybody I talked to at the party wasn't blue.

(37) Everybody I talked to at the party wasn't blue – they weren't blue, Susan, I would have noticed.

In (36), the speaker of *everybody I talked to at the party* may not be quantifying the entities under discussion so much as they are specifying the referent, the *they* in (37). This difference that I speculate here between quantifying as in (35) and specifying as in (36) and (37) is similar to the ambiguity between the quantificational and referential use of an indefinite. Possibly, the more that the subject is specified rather than quantified, the less that interlocutors would understand the *every* in the subject to act as a scope-bearing operator, leaving only the surface scope interpretation, corresponding to 'it is the subject that is not ...' meaning.

With respect to the role of pausing, an additional finding that was not predicted was that pausing interacts with subject modification and statement type: the surface-scope-facilitating effect of pausing duration is weaker for subject-modified items than for unmodified cases, and it's stronger for questions than for declarative uses. However, subject-modified items are already more likely to show surface scope, and questions are already more likely to show inverse scope. Thus, both of these interactions with pausing duration may at best reflect a ceiling effect towards surface scope.

6.3 Discussion and Conclusion

I investigated the prosody of naturalistic *every*-negation, specifically focusing on whether features such as pauses and pitch can predict scope interpretations. Despite the many sources of variation in prosody and prosody of scope, in the naturalistic corpus, pausing duration before negation facilitates surface scope and final pitch rise facilitates inverse scope. In general, scope affects the pitch contour, with distinct prosodic patterns for surface and inverse scope, including a greater pitch rise on the quantifier for inverse scope.

First, I considered the hypothesis that the existence or duration of a pause between the quantifier and negation signals surface scope. I measured total pausing duration between the quantifier and negation and used a mixed effects model to predict scope interpretation. Indeed, I found that longer pauses before negation were associated with greater surface scope. Additionally, items with subject modification and declarative statements were more likely to have surface scope, and an interaction effect indicated that the influence of pausing on surface scope was weaker for modified subjects and stronger for questions.

I then investigated the hypothesis that final rise in pitch (mean F0) predicts inverse scope preference. I calculated the difference in mean F0 between the last two syllables of each recording's predicate and used a mixed effects model similar to the one above. I also found that greater final pitch rise was associated with inverse scope. Subject modification and statement type also influenced scope, with subject modification reducing inverse scope and questions increasing it.

Finally, I considered whether a broader pitch contour reflects scope interpretation. I measured mean F0 at seven syllables in each recording and used generalized additive mixed models to compare a full model (including scope interpretation) with a nested model (excluding scope). In the first analysis, which included all corpus items that had these seven syllables, scope did not significantly affect mean F0 trajectory. In the second analysis (120 recordings, excluding

questions, subject-modified items, interruptions, or silent predicates): scope significantly influenced mean F0 trajectory: inverse scope was associated with greater emphasis on the quantifier, and surface scope was associated with relatively greater emphasis on the negation.

Challenges to the analysis included variability in sentence structure, length, parts of speech, and other factors in the corpus items which could affect prosody in unforeseen ways. The analyses in this chapter primarily addressed variability by accounting for key features similar to those in previous studies to ensure the prosody of scope could be observed despite other variations.

Altogether, although this analysis of naturalistic speech was not set up to test many specific predictions from the fall-rise prosody-of-scope literature, the findings are largely in favor of a middle ground regarding some of these past expectations. There indeed exist mappings between prosody and scope, despite the lack of empirical evidence for such mappings in reading data of *all*-negation (Syrett et al., 2014), but the mappings are variable (e.g., as in Ward and Hirschberg, 1985); there is no one-to-one relation between intonation and interpretation (as in Jackendoff, 1972). On the other hand, the evidence of naturalistic associations between scope and various acoustic cues is clear enough that it also seems too much to suggest, for example as in Ward and Hirschberg (1985), that it is truly context and not prosody which disambiguates scope.

In other words, while it is not 'all about prosody', it is also not 'all about context'. Rather, in considering the relationships between context, prosody, and interpretations, I would predict that context and prosody are both (variable) sources of disambiguation in their own right, with additional relationships between context and prosody (and additional relationships between context and sentence form, and prosody and sentence form, etc.) which muddy the waters.

An exciting future direction would be to directly consider the relationship between context

and prosody. For example, one study could assess how often contextual cues to inverse scope co-occur with prosodic cues to inverse scope in the naturalistic data: would we observe that certain prosodic cues only occur after certain contexts and never without them? Can (a specific aspect of) preceding linguistic context predict (a specific aspect of) prosody? On the other hand, would we observe trade-offs, suggesting not only that these are two separate sources of information, but that speakers efficiently rely on them? Are speakers less likely to go to the effort of producing a non-default prosodic contour when the context clearly disambiguates? Another, behavioral study could put contextual cues to inverse scope in competition with prosodic cues to inverse scope in order to test whether listeners tend to privilege one source of information over the other.

Chapter 7

Conclusion

I asked how people navigate ambiguity, specifically the potential ambiguity in quantifier-negation sentences, combining evidence from corpus analysis, behavioral experiments, and computational modeling. What are the preferred interpretations of quantifier-negation, especially *every*-negation? What are the contexts and prosodies of these interpretations?

A core contribution of this dissertation has been to investigate these questions for naturalistic ambiguity. Previous studies primarily analyzed written data with small sample sizes or subjective annotations, lacking comparisons across judgments. To address this lack of data on naturalistic speech, I examined *every*-negation constructions in two large-scale corpora: the Corpus of Contemporary English transcripts and National Public Radio recordings. The first corpus aims to collect uses and interpretations of these utterances in text-in-context form, while the second corpus additionally collects audio data to allow study of prosodic cues. Both corpora provide text-only, context-dependent interpretations from native speakers, while the second corpus provides interpretations with or without context, with or without audio. The final corpus is a comprehensive dataset for analyzing frequency, preferred interpretations, contexts, and acoustic features of spontaneous ambiguity use. I found with this corpus

that examples of quantifier-negation construction are indeed ambiguous, eliciting a range of average interpretation preferences depending on the quantifier, the context, and the prosody.

To better understand how speakers resolve scope given context, I used a computational model in the Rational Speech Act (RSA) framework. In this framework, ambiguity resolution arises from rational and domain-general inferences that listeners regularly perform as they understand language. Just as RSA models have been shown to capture various aspects of language use, including truth value judgments of *every*-negation (Scontras and Pearl, 2021), I found that an RSA model accurately captures average interpretation preferences for *every*-negation, *some*-negation, and *no*-negation.

The central hypothesis of the model is that quantifier-negation interpretations reflect the behavior of cooperative interlocutors, who reason about preferred interpretations given salient, skewed beliefs about the state of the world. One influence of context, among these cooperative interlocutors, is to lend greater probability to interpretations that are more likely to be true, as listeners try to align their interpretations with what they already know about the world. This is only one pressure in conversation, and may even come into conflict with a pressure for utterances to be informative or surprising enough.

I found evidence for model predictions for the disambiguating role of world expectations in the naturalistic ambiguity corpus: inverse scope preference in the corpus is correlated with several different measures of positive expectations in context. Thus positive expectations help account for the variation in interpretation preferences for *every*-negation utterances in the speech corpus.

I also found that interpretation variation in the naturalistic speech corpus can be accounted for by specific aspects of utterance prosody. As with questions of utterance use and context, there is a knowledge gap regarding naturalistic prosody of scope; past studies also cast some doubt that any robust disambiguating patterns could be found. Excitingly, pausing before

negation and a final pitch rise are significant prosodic cues predicting scope interpretation in speech. These prosodic patterns align with parts of those patterns that are identified in previous studies. Moreover, specific prosodic patterns in the pitch contour are associated with different scope interpretations, for items that have the same sentence features as those that are considered in previous literature (declarative, uninterrupted, complete, and unmodified sentences): a more quantifier-emphatic prosody generally reflects inverse scope interpretations.

Future research could explore how prosodic elements might be incorporated into disambiguation mechanisms formally articulated by a computational model. One potential avenue for modeling the role of prosody is to model prosody as emphasis (e.g., emphasis on the quantifier), where the use of emphasis reflects greater speaker effort (a cost) as well as greater probability that the listener would hear the word correctly (a benefit) in a noisy channel framework. To account for other pragmatic phenomena in an RSA model, Bergen (2016) has modeled emphasis in this way. Another potential modeling paradigm would be to consider prosodic emphasis not in relation to a noisy channel but as a filter on the felicitous question under discussion, addressing how prosody might interact with discourse structure.

A pattern of note in the interpretation data is that while individual judgments are decisive – listeners usually appear to resolve the ambiguity strongly in favor of one interpretation or the other – the average item interpretations are variable and reflect cross-listener disagreements, even when interpretations are made using context and prosody information. In other words, even with the benefit of disambiguating information, many *every*-negation uses are still ambiguous to an extent. Some interpretation variation may be due to individual differences in scope preference; while this dissertation did not have sufficient data to fully explore this variable, investigating listener-specific preferences could shed much light on interpretations. More broadly, however, the communication may have been good enough regardless of the scope interpretation that the listener arrived at. Scope is only one kind of information that a speaker may wish to convey. For example, in a case with a high positive expectation in

context, communication may have been good enough if the speaker successfully conveyed that the prior expectation was unassertable. At the same time, decisiveness of individual judgments reflects efficiency too: with limited cognitive resources, it is good enough to keep just one interpretation in mind (and to revise it if necessary).

Thus, future research could also explore the possibility of underspecification in scope ambiguity. The behavioral interpretation tasks in this dissertation prompted an interpretation decision and did not provide any explicit and separate way for listeners to indicate that scope was underspecified. To test how scope interpretations unfold without prompting a decision, methods like eye-tracking or visual world paradigm tasks could be used to assess online responses.

Lastly, the relationship between prosody and context remains a central question for understanding disambiguation. In contrast to the possibility that disambiguation is essentially all about context, the findings in this dissertation suggest that prosody and context are partially redundant and both are sources of disambiguation in their own right. Future research could further investigate the influence of context on prosody, the co-occurrence of different cues in naturalistic corpora, the timing of certain cues, and their relative strengths in influencing scope interpretations. There may be instances where prosody serves as a cue for questions under discussion, particularly when contextual cues are weak. We may also observe efficient trade-offs in the use of different sources of cues, such that one type of cue is more likely to be present when the other type of cue is weaker.

Altogether, these findings suggest that disambiguation is facilitated by context and prosody. We started with the idea that ambiguity may be a kind of challenge to successful communication, a feature of natural language which needs explaining; with the ambiguity corpus, we arrive at a picture of ambiguity use which is integrated with multiple co-occurring cues to intended meaning. As text without context, uses of quantifier-negation may appear ambiguous and somewhat unusual. For instance for *every*-negation, why would the speaker

not use salient, unambiguous alternative phrases, such as *nothing* or *not everything*? Yet the original conversational context and recording is often rich in information that motivates and disambiguates it. In particular, one common use of *every*-negation is as an emphatic frame for contradicting a previous belief, with a quantifier-focused or final-rising prosody.

Bibliography

Section 4: Demographics and political views of news audiences, 2012. URL https://www.pewresearch.org/politics/2012/09/27/section-4-demographics-and-political-views-of-news-audiences/.

A. Achimova, G. Scontras, C. Stegemann-Philipps, J. Lohmann, and M. V. Butz. Learning about others: Modeling social inference through ambiguity resolution. *Cognition*, 218: 104862, 2022.

C. Anderson. *The structure and real-time comprehension of quantifier scope ambiguity*. PhD thesis, Northwestern University Evanston, IL, 2004.

J. Aoun and Y.-h. A. Li. *Syntax of scope*, volume 21. 1993.

M. Baltazani. The role of prosody in scope relations. *The Journal of the Acoustical Society of America*, 107(5):2856–2857, 2000.

S. Baumann and T. Rathcke. Disambiguating the scope of negation by prosodic cues in three varieties of german. *Lingua*, 131:29–48, 2013.

F. Beghelli and T. Stowell. Distributivity and negation: The syntax of each and every. In *Ways of scope taking*, pages 71–107. Springer, 1997.

L. Bergen. *Joint inference in pragmatic reasoning*. PhD thesis, Massachusetts Institute of Technology, 2016.

D. Bolinger. *Intonation and its uses: Melody in grammar and discourse*. Stanford university press, 1989.

D. L. Bolinger. A theory of pitch accent in english. *Word*, 14(2-3):109–149, 1958.

D. Büring. *The meaning of topic and focus: The 59th Street Bridge accent*, volume 3. Psychology Press, 1997.

G. Carden. A note on conflicting idiolects. *Linguistic Inquiry*, 1(3):281–290, 1970.

G. Carden. Multiple dialects in multiple negation. *Pap. 8th Regional Meet. Chicago Ling. Soc., ed. PM Peranteau, JN Levi, GC Phares*, pages 32–40, 1972.

G. Carden. Disambiguation, favored readings, and variable rules. *New ways of analyzing variation in English*, pages 171–82, 1973.

N. Chomsky, A. Belleti, and L. Rizzi. An interview on minimalism. *N. Chomsky, On Nature and Language*, pages 92–161, 2002.

H. H. Clark. *Arenas of language use*. University of Chicago Press, 1992.

J. Cole. Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31, 2015.

N. Constant. English rise-fall-rise: a study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, 35(5):407–442, 2012.

S. Crain and R. Thornton. Investigations in universal grammar, 1998.

A. Cutler. The context-dependence of" intonational meanings". In *Thirteenth Regional Meeting, Chicago Linguistic Society*, pages 104–115. CLS, 1977.

M. Davies. Corpus of Contemporary American English (COCA). 2015. URL `https://doi.org/10.7910/DVN/AMUDUW`.

M. Davies. Notes on the naturalness and authenticity of the language from these transcripts, 2024. URL `https://www.english-corpora.org/coca/help/spoken.asp`.

J. Degen. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1, 2015.

J. Degen. The rational speech act framework. *Annual Review of Linguistics*, 9, 2022.

J. Degen and N. Goodman. Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

J. D. Fodor. Prosodic disambiguation in silent reading. In *North east linguistics society*, volume 32, page 8, 2002.

C. Ford. Getting started with binomial generalized linear mixed models, Mar 2021. URL `https://library.virginia.edu/data/articles/getting-started-with-binomial-generalized-linear-mixed-models`.

D. Fox. Economy and scope. *Natural language semantics*, 3(3):283–341, 1995.

M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

L. Frazier and C. Clifton. Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research*, 26:277–295, 1997.

L. Frazier and J. D. Fodor. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325, 1978.

T. Givón. Deductive vs. pragmatic processing in natural language. In *Methods and tactics in cognitive science*, pages 137–190. Psychology Press, 2014.

N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

M. Grant, S. Sloggett, and B. Dillon. Processing ambiguities in attachment and pronominal reference. *Glossa: a journal of general linguistics*, 5(1), 2020.

H. P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

B. J. Grosz, D. E. Appelt, P. A. Martin, and F. C. Pereira. Team: An experiment in the design of transportable natural-language interfaces. *Artificial Intelligence*, 32(2):173–243, 1987.

A. Gualmini. Some knowledge children dont lack. *Linguistics*, 42(5):957–982, 2004.

A. Gualmini, S. Hulsey, V. Hacquard, and D. Fox. The question–answer requirement for scope assignment. *Natural language semantics*, 16(3):205, 2008.

F. E. Harrell Jr and M. F. E. Harrell Jr. Package 'hmisc'. *CRAN2018*, 2019:235–236, 2019.

B. Hemforth and L. Konieczny. Scopal ambiguity preferences in german negated clauses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

J. T. Heringer. Research on quantifier-negative idiolects. In *Chicago Linguistic Society*, volume 6, page 95, 1970.

J. Hirschberg. Controlling intonational variation using escape sequences in the bell laboratories text-to-speech system. 1995.

J. Hirschberg. Pragmatics and prosody. 2013.

J. Hirschberg and C. Avesani. The role of prosody in disambiguating potentially ambiguous utterances in english and italian. In *Intonation: Theory, models and applications*, 1997.

J. Hirschberg and C. Avesani. Prosodic disambiguation in english and italian. In *Intonation: Analysis, modelling and technology*, pages 87–95. Springer, 2000.

J. R. Hobbs and S. M. Shieber. An algorithm for generating quantifier scopings. *Computational Linguistics*, 13:47–63, 1987.

L. Horn. A natural history of negation. 1989.

L. R. Horn. Metalinguistic negation and pragmatic ambiguity. *Language*, pages 121–174, 1985.

S. Hulsey, V. Hacquard, D. Fox, and A. Gualmini. The question-answer requirement and scope assignment. *MIT working papers in Linguistics*, 48:71–90, 2004.

G. Ioup. Some universals for quantifier scope. In *Syntax and Semantics volume 4*, pages 37–58. Brill, 1975.

R. S. Jackendoff. Semantic interpretation in generative grammar. 1972.

M. J. R. Johnston. *The syntax and semantics of adverbial adjuncts.* PhD thesis, University of California, Santa Cruz, 1994.

L. Karttunen and S. Peters. Conventional lmplicature. In *Presupposition*, pages 1–56. Brill, 1979.

K. É. Kiss and J. Pafel. Quantifier scope ambiguities. *The Wiley Blackwell Companion to Syntax, Second Edition*, pages 1–36, 2017.

Y. Kitagawa and J. D. Fodor. Prosodic influence on syntactic judgments. *IULC Working Papers*, 5(2), 2005.

Y. Koizumi. *Processing the not-because ambiguity in English: The role of pragmatics and prosody.* City University of New York, 2009.

T. Kraljic and S. E. Brennan. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive psychology*, 50(2):194–231, 2005.

H. S. Kurtzman and M. C. MacDonald. Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279, 1993.

D. R. Ladd. The structure of intonational meaning (bloomington), 1980.

G. Lakoff. On generative semantics. *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, 232:296, 1971.

M. Liberman and I. Sag. Prosodic form and discourse function. In *Chicago Linguistics Society*, volume 10, pages 416–427, 1974.

J. Lidz. The scope of children's scope: Representation, parsing and learning. *Glossa: a journal of general linguistics*, 3(1), 2018.

M. C. MacDonald, N. J. Pearlmutter, and M. S. Seidenberg. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676, 1994.

R. May and S. J. Keyser. *Logical form: Its structure and derivation*, volume 12. 1985.

R. C. May. *The grammar of quantification.* PhD thesis, Massachusetts Institute of Technology, 1977.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.

J. Musolino. Universal grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifier-negation interaction in english. 1999.

J. Musolino and J. Lidz. Why children aren't universally successful with quantification. *Linguistics*, 44(4):817–852, 2006.

J. Musolino, S. Crain, and R. Thornton. Navigating negative quantificational space. *Linguistics*, 38(1):1–32, 2000.

C. Nakao, T. Goro, and J. Lidz. An experimental study on children's interpretation of negation and because-clauses. = *IEICE technical report:* , 107(138):105–110, 2007.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

A. Neukom-Hermann. *Negation, Quantification and Scope. A Corpus Study of English and German All... Not Constructions.* PhD thesis, University of Zurich, 2016.

J. C. Park. Quantifier scope and constituency. *arXiv preprint cmp-lg/9505027*, 1995.

L. Pearl. Evaluating learning-strategy components: Being fair (commentary on ambridge, pine, and lieven). *Language*, 90(3):e107–e114, 2014.

L. Pearl. Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, pages 1–21, 2023.

S. T. Piantadosi, H. Tily, and E. Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012.

J. B. Pierrehumbert. *The phonology and phonetics of English intonation.* PhD thesis, Massachusetts Institute of Technology, 1980.

K. L. Pike. The intonation of american english. 1945.

M. Poesio. Semantic ambiguity and perceived ambiguity. semantic ambiguity and underspecification, ed. by k. van deemter and s. peters, 159–201, 1996.

B. Pritchett and J. Whitman. Syntactic representation and interpretive preference. *Japanese sentence processing*, pages 65–76, 1995.

T. Reinhart. Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 6(1):47–88, 1983.

T. Reinhart. Quantifier scope: How labor is divided between qr and choice functions. *Linguistics and philosophy*, pages 335–397, 1997.

C. Roberts. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:6–1, 2012.

E. G. Ruys and Y. Winter. Quantifier scope in formal linguistics. In *Handbook of philosophical logic*, pages 159–225. Springer, 2011.

W. S. Saba. *An Inferencing strategy for resolving quantifier scope ambiguities.* PhD thesis, Carleton University, 1999.

W. S. Saba and J.-P. Corriveau. Plausible reasoning and the resolution of quantifier scope ambiguities. *Studia Logica*, 67(2):271–289, 2001.

K. Savinelli, G. Scontras, and L. Pearl. Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior. In *CogSci*, 2017.

A. J. Schafer, S. R. Speer, P. Warren, and S. D. White. Intonational disambiguation in sentence production and comprehension. *Journal of psycholinguistic research*, 29:169–182, 2000.

G. Scontras and N. D. Goodman. Resolving uncertainty in plural predication. *Cognition*, 168:294–311, 2017.

G. Scontras and L. S. Pearl. When pragmatics matters more for truth-value judgments: An investigation of quantifier scope ambiguity. *Glossa: a journal of general linguistics*, 6(1), 2021.

G. Scontras, M. Polinsky, C.-Y. E. Tsai, and K. Mai. Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A journal of general linguistics*, 2(1):1–28, 2017.

G. Scontras, M. H. Tessler, and M. Franke. Probabilistic language understanding: An introduction to the rational speech act framework. *Retrieved January*, 17:2021, 2018.

G. Scontras, M. H. Tessler, and M. Franke. A practical introduction to the rational speech act modeling framework. *arXiv preprint arXiv:2105.09867*, 2021.

R. J. Smith. Examining the role of prosody in the resolution of semantic ambiguity in l1 and l2 speakers of english. 2011.

J. Snedeker and J. Trueswell. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, 48(1):103–130, 2003.

M. Sóskuthy. Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*, 2017.

D. Sperber and D. Wilson. *Relevance: Communication and cognition*, volume 142. Citeseer, 1986.

P. Srinivasan and A. Yates. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1465–1474, 2009.

R. C. Stalnaker. Assertion. In *Pragmatics*, pages 315–332. Brill, 1979.

K. Syrett, G. Simon, and K. Nisula. Prosodic disambiguation of scopally ambiguous sentences. In *Proceedings of the Meeting of the North East Linguistic Society*, volume 43, pages 141—-152. GLSA (University of Massachusetts), 2012.

K. Syrett, G. Simon, and K. Nisula. Prosodic disambiguation of scopally ambiguous quantificational sentences in a discourse context. *Journal of Linguistics*, pages 453–493, 2014.

A. Szabolcsi. Positive polarity–negative polarity. *Natural Language & Linguistic Theory*, 22 (2):409–452, 2004.

A. Szabolcsi. Scope and binding. *Semantics: An International Handbook of Natural Language Meaning. Mouton de Gruyter*, 2011.

J. Taglicht. Some uses of all or: The king's horses and other curiosities or: All is not well in our grammar. pages 1–12, ND.

J. C. Trueswell and M. K. Tanenhaus. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. MIT Press, 2005.

S. L. Tunstall. *The interpretation of quantifiers: semantics & processing*. PhD thesis, University of Massachusetts at Amherst, 1998.

M. P. van der Loo. The stringdist Package for Approximate String Matching. *The R Journal*, 6(1):111–122, 2014. doi: 10.32614/RJ-2014-011. URL `https://doi.org/10.32614/RJ-2014-011`.

J. Viau, J. Lidz, and J. Musolino. Priming of abstract logical representations in 4-year-olds. *Language Acquisition*, 17(1-2):26–50, 2010.

M. Wagner and D. G. Watson. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945, 2010.

G. Ward and J. Hirschberg. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, pages 747–776, 1985.

P. C. Wason. Response to affirmative and negative binary statements. *British Journal of Psychology*, 52(2):133–142, 1961.

P. C. Wason. In real life negatives are false. *Logique et analyse*, pages 17–38, 1972.

P. C. Wason and S. Jones. Negatives: Denotation and connotation. *British Journal of Psychology*, 54(4):299–307, 1963.

T. Wasow, A. Perfors, and D. Beaver. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282, 2005.

M.-J. Wu and T. Ionin. Intonational effects on english scopallyambiguous sentences. *Ilha do Desterro*, 73(3):13–36, 2020.

X. Xie, A. Buxó-Lugo, and C. Kurumada. Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, 211:104619, 2021.

Y. Xu. Prosody, tone and intonation. *The Routledge handbook of phonetics*, pages 314–356, 2019.