Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods *

Judith Degen
*Stanford University*

**Abstract** A prevalent, but to date untested, assumption about lexicalized scalar implicatures such as those from *some* to *not all*, is that they fall into the class of GCIs and as such, constitute a homogeneous class of highly regularized and context-independent implicatures. This paper reports a test of this assumption in which linguistically untrained participants’ implicature strength judgments were collected for naturally occurring utterances containing the word *some* in a large-scale corpus-based web study. The results indicate that implicature strength is highly variable and systematically dependent on features of the linguistic context such as the partitive, determiner strength, and discourse accessibility. These results call into question the GCI status of scalar implicatures from *some* to *not all* and demonstrate the usefulness of corpora and web-based methods for challenging received wisdom, enriching the empirical landscape, and informing theory in pragmatics.

**Keywords:** scalar implicature, GCI, corpora, experimental pragmatics

1 Introduction

Ever since Logic and Conversation (Grice 1975), scalar implicature has been treated as an instance of Generalized Conversational Implicature (GCI). That is, scalar implicatures are assumed to arise as a matter of default, independently of context, though they may be canceled if not licensed by the context. This sets them apart from Particularized Conversational Implicatures (PCI), which rely heavily on the context of utterance. The GCI status of scalar implicatures is a fundamental assumption both for theories of the conditions under which scalar implicatures arise (e.g., Gazdar 1979, Horn 1984, Levinson 2000) as well as for theories of how scalar implicatures are processed (e.g., Levinson 2000).

Unfortunately, the data to support the categorization of scalar implicatures as GCIs —indeed the categorization of any sort of implicature as a GCI or PCI— has thus far consisted entirely of linguists’ intuitions, typically just the authors’, using a handful of examples. This was perfectly justified at the time that conversational implicatures were first investigated, when the tools to collect large quantities of regular

* Acknowledgments to be inserted.
language users’ judgments across different contexts were not available. However, the small number of experimental participants—the author(s)—and experimental items—the handful selected by the author(s)—introduce bias and call into question the generalizability of theories that are developed by this method (Gibson et al. 2011).

Fortunately, researchers now have access to large-scale corpora of spontaneous speech as well as the ability to collect judgments from a diverse population and a large number of experimental participants over the web. That is, we now have at our disposal the means to empirically test the validity of claims concerning the context-independence and defaultness of various types of conversational implicatures.

This paper takes a modest first step in this direction by testing the assumption—which I will refer to as the Homogeneity Assumption—that scalar implicatures from some to not all constitute a homogeneous, context-independent type of implicature that fall into the class of GCIs. This will be achieved by conducting a large-scale web-based study in which regular language users’ interpretations of utterances containing some, extracted from a corpus of spontaneous speech, are collected.

In the rest of this Section, I discuss what is at stake in testing the Homogeneity Assumption. In Section 2 I report the study which aims at testing two aspects of the Homogeneity Assumption: a) whether there is variation in the strength with which scalar implicatures from some to not all arise, and b) whether there is systematic context-dependence in this variation. In Section 3 I discuss the implications of the results a) for the status of scalar implicatures as GCIs and b) for theories that rely on scalar implicatures constituting a homogeneous class of implicatures.

1.1 What is at stake

Most linguistic and psychological processing theories of scalar implicatures make the Homogeneity Assumption to some extent (Gazdar 1979, Horn 1984, Huang & Snedeker 2009, Levinson 2000). It is thus crucial to be explicit about what exactly the Homogeneity Assumption entails and what the consequences would be, should it be overturned. This is what this Section is devoted to.

1 Most of the processing literature has been careful to remain non-committal about the status of scalar implicatures as GCIs, or has argued against the usefulness of GCI as a psychological term Breheny et al. (2006, 2013). However, this literature relies on scalar implicature processing being comparable across different experiments, and thus, across different linguistic and discourse contexts (e.g., Bott & Noveck 2004, Bott et al. 2012, Grodner et al. 2010, Huang & Snedeker 2009). It is in this sense that one of the aspects of the GCI claim is implicitly endorsed.
1.1.1 The Homogeneity Assumption and the GCI-PCI distinction

The Homogeneity Assumption can be stated as in (1) and includes the sub-assumptions in (1a) and (1b).

(1) The Homogeneity Assumption
Lexicalized scalar implicatures constitute a homogeneous class of implicatures.
   a. Strength invariance: Implicature strength is not variable.
   b. Context independence: Implicature strength is not systematically dependent on context.

I will demonstrate how the general assumption follows from the GCI-PCI distinction and elaborate on each of the sub-assumptions in turn.

The GCI-PCI distinction. Consider (3) as an answer to (2a). Dan can be taken to mean that not all of the students failed and in addition, that the exam was hard. In contrast, consider Dan’s utterance as a response to (2b): In this case, the scalar implicature that not all of the students failed still goes through, but Dan can no longer be taken to implicate (4a). However, now he can be taken to implicate that the teacher did not do a good job, which was not an available implicature in (2a). These kinds of observations of scalar implicatures seemingly arising independently of context have contributed to their status as GCIs in contrast to the more context-dependent PCIs in (4).

(2) a. Masha: Was the exam hard?
   b. Masha: Did the teacher do a good job?
(3) Dan: Some of the students failed.
    \(\sim\) Some, but not all, of the students failed. (GCI)
(4) a. \(\sim\) The exam was hard. (PCI)
   b. \(\sim\) The teacher did not do a good job. (PCI)

Grice characterizes the distinction between the two types of inferences as follows: PCIs are carried by “saying that \(p\) on a particular occasion in virtue of special features of the context, cases in which there is no room for the idea that an implicature of this sort is normally carried by saying that \(p\)” (Grice 1975: p. 56, emphasis in the original). In contrast, of GCIs he says “the use of a certain form of words in an utterance would normally (in the absence of special circumstances) carry such-and-such an implicature or type of implicature.” (Grice 1975: p. 56, emphasis in the original).
There is agreement in the literature that not all scales are created equal - that is, some scales are more readily involved in the generation of scalar implicatures than others. This is captured in the distinction between lexicalized and ad hoc scales (Hirschberg 1985, Horn 1989, Matsumoto 1995). Lexicalized scales are such that whenever the weaker element from the scale is observed, the stronger one functions as an alternative. Scales that have been proposed to be lexicalized are quantifiers like \langle all, some \rangle, sentential connectives like \langle and, or \rangle, modals like \langle must, can \rangle, or numerals like \langle three, two \rangle. In contrast, ad hoc scales require more context to become functional. Hirschberg (1985) has noted that any items that constitute a partially ordered set in which one item can be determined to be higher than another one can function as a scale. For example, the scale \langle send, write \rangle contains different stages an email may be in, where sending it follows writing it. However, for this scale to become functional, the context needs to be such that sending the email is a salient competitor to writing it. It is not the case that all utterances of I wrote the email compete with I sent the email.

The seeming context independence of lexicalized scales and the context dependence of ad hoc scales has been used to categorize lexicalized scalar implicatures as GCIs and ad hoc scalar implicatures as PCIs. Since the Homogeneity Assumption is formulated only for GCIs, not for PCIs, it should hold for lexicalized scales, but there is no expectation that it should hold for ad hoc scales. The most discussed and unambiguously agreed-upon case of a lexicalized scale is the \langle all, some \rangle scale. Thus, if the Homogeneity Assumption holds at all, it should hold for implicatures from some to not all.

In what way then do some-not-all implicatures constitute a homogeneous class of implicatures? Precisely in the way specified by the GCI-PCI distinction: GCIs set themselves apart from PCIs in that they usually arise (captured by sub-assumption (1a): STRENGTH INVARIANCE) and survive context shifts (captured by sub-assumption (1b): CONTEXT INDEPENDENCE).

A test of the Homogeneity Assumption is crucial not only to theories of the conditions under which scalar implicatures arise, but also to theories of how they are processed. Both of the main rival processing theories of scalar implicature - the Default model (Levinson 2000) and the Literal-First hypothesis (Huang & Snedeker 2009) - make the Homogeneity Assumption, though the respective status of the Assumption in the theories differs. For Levinson it is a core assumption of the theory; for the Literal-First hypothesis it is a background assumption that allows processing delays in computing scalar implicatures to be interpreted as evidence for a processing distinction between semantics and pragmatics, and in particular for a privileged position of computing literal content over computing.

---

2 These are not unrelated in principle, but often are in practice.
pragmatic implicatures. Thus, while the Homogeneity Assumption is necessary for the interpretation of delayed implicature processing results as support for the Literal-First hypothesis, overturning it would not call into question the theory itself, though it would call into question the testability of the theory. I defer a fuller discussion of the Literal-First hypothesis to the General Discussion and focus here on the consequences for the Default model.

Levinson (2000) postulates default, cognitively cost-free GCIs as a solution to what he calls the articulatory bottleneck problem: There is a significant articulatory bottleneck in the rate of information that can be transmitted via human speech (estimated as out-of-context phoneme information). In addition, he assumes that integrating contextual information to derive complex pragmatic inferences is hard and effortful. Nevertheless, linguistic communication proceeds at a miraculous speed. Thus, the communicative system must have evolved a solution that allows for rapid communication through a very limited channel. The solution, according to Levinson, is to make inference cheap for listeners on average - and the best way to do this is to allow for highly regularized inferences - GCIs - to be derived at no cost, thus balancing out the cost of deriving difficult contextual inferences - PCIs. This balance of costs would allow communication to proceed at the rapid rate at which it does.

Thus, both STRENGTH INVARIANCE and CONTEXT INDEPENDENCE are crucial to the Default model: if scalar implicatures from some to not all are in fact much weaker, less regularized, and more context-dependent than Levinson assumes, this would mean that scalar implicature processing involves much more cognitive cost due to implicature cancellations and integration of contextual information than previously assumed. If this result generalizes to other lexicalized scalar implicatures, this would mean that cognitive cost-freeness of GCIs does not clearly constitute a

---

3 This is a common, because intuitive, but unfortunately wrong, assumption in the linguistic literature. There is much evidence from the psycholinguistic literature that suggests that listeners can very rapidly integrate information from multiple contextual cues. For example, the visual context has immediate effects on whether a prepositional phrase is interpreted as a destination or as modifying a definite NP (Tanenhaus et al. 1995); an object’s affordances may immediately discard it as a potential referent (Chambers et al. 2004); whether a particular piece of information is in common or privileged ground can immediately affect the interpretation of definite NPs with prenominal scalar adjectives (Sedivy et al. 1999, Heller et al. 2008); and whether a speaker is deemed reliable with respect to the degree with which he overinforms can have rapid effects on contrastive inferences (Grodner & Sedivy 2011). Thus, processing contextual information may not in fact be costly, and so neither may processing PCIs which crucially depend on processing of contextual information. This assumption of Levinson’s is thus highly questionable, but I will not discuss it further here.

4 Note that there are no actual estimates of how rapidly communication should proceed under different assumptions about the cost of various inferences. That is, intuitions about the rate at which communication should proceed for different cost distributions, are in effect not more than that - intuitions.
solution to the articulatory bottleneck problem.

In the following I discuss the sub-assumptions in more detail and clarify the empirical predictions that the Homogeneity Assumption makes.

**STRENGTH INVARIANCE.** Scalar implicatures have traditionally been treated as a categorical phenomenon: either the implicature goes through or it does not. However, intuitively, implicatures are sometimes “felt” more strongly than other times (Russell 2012). Recent developments in probabilistic pragmatics have explicitly modeled scalar implicature as a matter of degree (Degen et al. 2013, Frank & Goodman 2012, Franke 2009, Goodman & Stuhlmüller 2013, Russell 2012). In these models, hearers are treated as having a certain degree of belief in the stronger alternative being true or false upon observing an utterance containing a weak scalar item. This is akin to assigning the stronger alternative a particular probability of being true - the lower the probability, the stronger the implicature. Underlying these models is the assumption that hearers have internalized a model of the speaker, that is, of the utterances a speaker is likely to produce, given that the speaker intends to communicate a particular meaning. Bayesian inference allows hearers to then reverse-engineer a distribution over likely intended meanings, resulting in a probability (or degree of belief in) the stronger alternative being true or false. This stands in contrast to the traditional view, where the outcome of the reasoning process is a belief in the stronger alternative with minimal probability 0 or maximal probability 1 - in other words, the implicature goes through or it doesn’t.

The sub-assumption of STRENGTH INVARIANCE captures not only that scalar implicatures are assumed to go through with probability 0 or 1, but also that they will always go through, with the exception of rare cases in which they are contextually canceled. This captures the consensus view both that scalar implicatures are a categorical phenomenon, and that they are rarely canceled, a view made explicit by various authors. E.g., Huang & Snedeker (2009) hypothesize that “the lower-bounded interpretation may be vanishingly rare in real-world communication”. Horn (1984) remarks that “as a generalized implicatum, the aforementioned [scalar] inference goes through in unmarked contexts, but it may be canceled”. Breheny et al. (2006) note that scalar implicatures “show a degree of regularity and have the intuitive feel of components of conventional meaning”.

**CONTEXT INDEPENDENCE.** It is easy to see that if STRENGTH INVARIANCE holds, so does CONTEXT INDEPENDENCE; if the implicature arises irrespective of context, then context must play no role in whether or not a scalar implicature
Distribution of scalar implicatures is derived.\textsuperscript{5} However, there are of course contexts in which the implicature is canceled, a fact often noted in the literature, but deemed to be a relatively rare occurrence (Levinson 2000). Under the strongest interpretation of the Homogeneity Assumption, then, CONTEXT INDEPENDENCE simply follows from STRENGTH INVARIANCE.

If there turns out to be more variability in implicature strength than expected, what is one to make of CONTEXT INDEPENDENCE? Given that there are cases in which the implicature is canceled, context cannot play no role whatsoever in the process of computing scalar implicatures. Importantly, however, the role of context is not predicted to be systematic\textsuperscript{6}—cases of cancellation are in some way marked (Horn 1984) or idiosyncratic. Thus, under the strong interpretation of STRENGTH INVARIANCE, context should not play any role at all in scalar implicature computation. Under a weaker interpretation that allows for some variability, context is allowed the role of idiosyncratic implicature cancellation, but implicature strength should nevertheless not be systematically predictable from features of the context.

\textbf{Empirical predictions.} The strongest form of STRENGTH INVARIANCE predicts that all utterances with \textit{some} should be interpreted as giving rise to a \textit{some-not-all} implicature. However, recent work (Degen et al. 2013, Frank & Goodman 2012, Goodman & Stuhlmüller 2013) has provided evidence that scalar implicature is a probabilistic phenomenon; treating the process of interpreting an utterance with \textit{some} as more or less strongly giving rise to the belief that the stronger alternative is false provides a better fit to participants’ judgments than assuming that an utterance either does or doesn’t give rise to an implicature. Therefore, participants’ task in the study reported below did not consist in simply giving categorical judgments. Instead, they were instructed to provide continuous implicature strength judgments. With this implicature measure, the strong version of STRENGTH INVARIANCE predicts all implicature judgments to be maximal. A weaker version, taking into account that participants’ judgments may be noisy and in some cases the implicature is canceled, predicts that implicature strength judgments should be generally high, and in some exceptional cases, strength judgments may be low. That is, STRENGTH INVARIANCE predicts either a unimodal distribution of ratings clustered at the strong implicature end of the scale, or a bimodal distribution of ratings that is heavily skewed towards the strong end. The empirical pattern incompatible with STRENGTH INVARIANCE is a lack of preference for strong implicature judgments.

\textsuperscript{5} Note that this entailment is asymmetric: if STRENGTH INVARIANCE holds, so does CONTEXT INDEPENDENCE. But if CONTEXT INDEPENDENCE holds, it is nevertheless possible for implicature strength to vary, e.g. because of noise processes in interpretation.

\textsuperscript{6} But see, e.g., Matsumoto (1995) for an attempt to capture systematically the conditions under which scalar implicatures are canceled.
CONTEXT INDEPENDENCE is compatible with the patterns predicted by STRENGTH INVARIANCE. However, CONTEXT INDEPENDENCE is also compatible with substantial variability in participants’ implicature strength judgments. Importantly, CONTEXT INDEPENDENCE predicts either no variability in strength, or if there is variability in strength, this variability should not be systematically predictable from features of context.

If either or both of the sub-assumptions of the Homogeneity Assumption are not borne out by the data, this would have serious consequences for the status of implicatures from some to not all as GCIs. If there is a large amount of variability in implicature strength, some could not be said to “normally (in the ABSENCE of special circumstances) carry such-and-such an implicature or type of implicature”, Grice (1975)’s characterization of GCIs. If, moreover, variability in implicature strength is found to be systematically dependent on and predictable from context, some-not-all implicatures would start to smack suspiciously of PCIs. I return to the consequences this has for the status of the GCI-PCI distinction more generally in the General Discussion.

1.2 An alternative view: probabilistic pragmatics

If the predictions of the Homogeneity Assumption are indeed not borne out, what would be an alternative framework within which to treat scalar implicatures and scalar implicature processing? Here I present a sketch of such a framework, which I will loosely refer to as probabilistic pragmatics. As with the traditional view of scalar implicatures as GCIs, I will present the probabilistic pragmatics framework in terms of the assumptions it makes.

(5) Probabilistic pragmatics

a. Scalar implicatures are probabilistic.
b. Scalar implicatures are context-dependent.
c. Listeners can efficiently integrate multiple, probabilistic, contextual cues to the speaker’s intended meaning.

In this framework, the problem of computing scalar implicatures is viewed from the listener’s perspective. Assumption (5a) reflects the view that scalar implicatures do not either categorically arise or not. Instead, as discussed above, scalar implicature strength reflects the listener’s resulting degree of belief in the stronger alternative being false (Degen & Tanenhaus in press, Frank & Goodman 2012, Russell 2012). Multiple factors contribute to this ultimate belief: at least a) the listener’s prior beliefs in the truth of the stronger alternative (world knowledge) and b) the contextual evidence that the speaker intends to convey the negation of the stronger alternative.
It is through the latter that assumption (5b) comes into play. Examples of contextual cues to the speaker’s intention that will be investigated in Section 2.3 are the partitive of and the discourse accessibility of the NP referent embedded under some. The following case is an example of a partitive, highly discourse-accessible (as indicated by pronominalization) some-NP, which received high implicature strength ratings from experimental participants.

(6) I sold some of them.

Under this probabilistic view of scalar implicatures, implicature strength can vary, but this variation should be predictable from features of context. Listeners are assumed to have rich, probabilistic knowledge of the contexts in which speakers intend to communicate the negation of the stronger alternative. By making use of the available contextual information, the speaker’s intention is reverse-engineered (or inferred) upon observing an utterance. This is assumption (5c) and another way in which the probabilistic pragmatics framework deviates from the traditional view that assumes that integration of contextual information is a difficult, cognitively costly process. Assumption (5c) is backed up by numerous findings from the psycholinguistic literature (Chambers et al. 2004, Grodner & Sedivy 2011, Heller et al. 2008, Sedivy et al. 1999, Tanenhaus et al. 1995).

Note that a consequence of this view of scalar implicatures is that the GCI-PCI distinction becomes obsolete. Each scale will likely be associated with different contextual features that listeners are sensitive to; and some may be more context-dependent than others. In consequence, conversational implicatures exhibit different degrees of context-dependence, instead of either being context-dependent or not. Under this view, the challenge lies in quantifying the cues that listeners are sensitive to in generating implicatures of varying strength and building explicit computational models of this process. This particular endeavor lies outside the scope of this paper, but see, e.g., Bergen & Goodman (2014), Goodman & Stuhlmüller (2013), Russell (2012) for examples of how contextual cues beyond utterance informativeness can be integrated into probabilistic models of scalar implicature.

1.3 Interim summary

In this Section I have demonstrated how the Homogeneity Assumption follows from the GCI-PCI distinction and worked out the two crucial empirical predictions it makes: a) there should be little to no variation in implicature strength in implicatures from some to not all (but if there is, there should be a preference for strong over weak implicatures); and b) to the extent that there is variability in implicature strength, it should not be predictable from or captured by features of context. I have also sketched an alternative view of scalar implicatures as a probabilistic,
context-dependent computation problem for listeners. Section 2 reports the study conducted to test the Homogeneity Assumption for scalar implicatures from some to not all.

2 Testing the Homogeneity Assumption

In the following I report a corpus- and web-based study that constitutes a first attempt at testing the Homogeneity Assumption. The study proceeds in two steps. The first step consists in determining whether there is variation in participants’ interpretation of utterances containing some. This constitutes a test of the STRENGTH INVARIANCE sub-assumption, which predicts little to no variation in implicature strength. The second step consists in determining whether there is systematicity to this variation; that is, whether certain features of the linguistic context reliably predict implicature strength. The CONTEXT INDEPENDENCE sub-assumption predicts that implicature strength should not be predictable from features of context. The contextual features, or cues to the interpretation, as I will sometimes refer to them from the hearer’s perspective, are a) syntactic partitivity of the some-NP, b) determiner strength, and c) discourse accessibility of the some-NP, which includes c.i) linguistic mention of the embedded NP referent, c.ii) topicality of the some-NP, and c.iii) modification of the head of the some-NP.

The study was conducted in three steps. First, a database of utterances containing some-NPs was generated by extracting all instances of utterances containing the word some from the Switchboard corpus. Second, implicature strength ratings were obtained for each case in the dataset via a web-based study. Finally, the obtained ratings were used to investigate the properties of interest: variation and systematicity in implicature strength.

2.1 The database

I used TGrep2 (Rohde 2005) to extract all 1748 occurrences of some-NPs that were not part of a disfluency from the Penn Treebank (release 3, Marcus et al. 1999) subset of the Switchboard corpus of telephone dialogues (Godfrey & McDaniel 1992). The corpus contains approximately 800 thousand spoken words in over 100 thousand utterances from about 650 telephone dialogues on various topics between two participants who did not know each other. The TGrep2 Database Tools (Jaeger 2006, Degen & Jaeger 2011) were used to organize the some-utterances into a database.

Because only those cases that do not syntactically prohibit a scalar implicature were interesting for the purpose of the study, 359 cases (20.5%) of some-NPs headed
by singular count nouns were excluded.\footnote{In the grand scheme of things one would not want to exclude these cases of \textit{some}, but rather include head noun number as a cue that listeners can use to restrict their interpretation of \textit{some} - that is, a singular count noun can be seen as a strong, but nevertheless probabilistic, cue against the implicature.} In a \textit{some}-NP, singular count nouns are compatible with two different meanings. The more common meaning is the specific indefinite reading, which cannot give rise to a scalar inference (see examples in (8a) and (7)). Singular count nouns in \textit{some}-NPs can, however, also receive a coerced interpretation as shown in (8). Under this reading, the implicature, made explicit in (8b) is possible, but these cases seem to be very infrequent (e.g., in a random sample of 50 singular count noun \textit{some}-NPs, only three were cases of coercion, and they all occurred in the partitive, as in (9)).

\begin{itemize}
\item (7) a. She stuck my name on some list.
   \hspace{1cm} b. *She stuck my name on some, but not all, list.
\item (8) a. John kicked some cat off the street.
   \hspace{1cm} b. John kicked some, but not all, cat off the street.
\item (9) Well, I had some of that problem.
\end{itemize}

A further 26 cases where the \textit{some}-NP consisted only of \textit{some} were also excluded:

\begin{itemize}
\item (10) Some say that coffee is healthy.
\end{itemize}

This was done because for these cases it is not possible to investigate the effects of the discourse accessibility cues tested in Section 2.3, which assumes that \textit{some} occurs with an embedded NP. However, it is worth noting that in these cases the implicature seems to generally go through.

After the exclusion, 1363 cases of utterances containing \textit{some}-NPs remained. For these cases, implicature ratings were collected in a web-based study, which is described in the following.

### 2.2 Collecting implicature ratings: a test of STRENGTH INVARIANCE

Gradient implicature strength ratings were collected using Amazon’s Mechanical Turk service.\footnote{The experiment can be viewed at \url{https://www.hlp.rochester.edu/mturk/jdegen/7_qpsome/output/qp.html?assignmentId=foo&list=1}. Different lists can be viewed by changing the list parameter to any number between 1 and 67. The data are available upon request.}

#### 2.2.1 Methods

**Participants** 243 participants were recruited on Amazon’s Mechanical Turk and paid $0.80 for each block of 20 items. Participants who completed at least three
blocks received a one-time bonus of $0.20.

**Procedure and materials**  On each trial, participants saw an utterance\(^9\) containing a *some*-NP (the *target utterance*) together with ten utterances from the immediately preceding discourse context (or until the beginning of the dialogue if there were fewer than ten utterances in the previous context). The target utterance was presented in red. Below the target utterance, an almost identical utterance (the *comparison utterance*) was presented which differed only in that the implicature was made explicit by inserting *but not all after some*. The comparison utterance was presented in green font. Two example pairs of (a) target and (b) comparison utterances are shown in (11) and (12).

|(11) a. | I like, I like to read some of the philosophy stuff.  
|      b. | I like, I like to read some, but not all, of the philosophy stuff.  
| (12) a. | And I’ll take some time and do that with her.  
|     b. | And I’ll take some, but not all, time and do that with her.  |

Participants were then asked: “How similar is the statement with ‘some, but not all’ (green) to the statement with ‘some’ (red)?” They provided similarity judgments on a seven point Likert scale with endpoints labeled as “very different meaning” and “same meaning” and individual points labeled as 1, 2, . . . , 7. The more clearly the implicature is part of the speaker’s originally intended meaning, the less of a difference explicitly encoding the content of the implicature should make, and the higher the similarity judgments are expected to be. Conversely, if the content of the implicature was not part of the speaker’s originally intended meaning, making it explicit should lead to a larger perceived shift in meaning and the two utterances should be rated as very dissimilar.

This paraphrase task is a novel measure of scalar implicature and as such deserves further consideration. I briefly discuss three task-related considerations.

First, the effect of locally inserting *but not all* should have different effects on the interpretation of *some* in upward-entailing vs. non-upward-entailing contexts (Chierchia 2004). While this is an important point that deserves further investigation in future work,\(^10\) see Section 2.3.2 for a demonstration that in the dataset reported here, monotonicity properties of the context likely affected only a negligibly small number of cases.

---

\(^9\) An utterance corresponds to a unit of speech that has been transcribed as a sentence in the Switchboard corpus. This includes sentence fragments.

\(^10\) To the best of my knowledge, there exists no large-scale empirical assessment of the rate at which various scalar items occur in upward-entailing vs. non-upward-entailing contexts in naturally occurring language.
A further consideration is that inserting *but not all* may shift the salient Question Under Discussion (QUD) Roberts (1996, 2012) in some cases but not others. That is, participants’ judgments may in some cases reflect not implicature strength, but the difficulty of making reparatory inferences to accommodate the shifted QUD. While it is quite likely that judgments reflect QUD accommodation in some cases (e.g., in cases like the ones listed in (15)), this is not at odds with the notion that these judgments reflect implicature strength. In fact, relevance of the stronger utterance with *all* to a salient contextual QUD is a crucial ingredient in scalar implicature computation (Grice 1975, Matsumoto 1995, Russell 2012, Zondervan 2010). Under the probabilistic view of scalar implicatures, relevance of the stronger alternative to a contextual QUD is one of many factors involved in scalar implicature. Thus, reduced implicature strength due to failed QUD accommodation is perfectly compatible with the view taken in this paper. It is nevertheless an interesting question for future work whether QUD accommodation processes can be teased apart experimentally from ‘core’ implicature computation.

A further consideration regards participants’ potentially varying interpretation and resulting use of the provided Likert scale. In order to effect similar scale interpretations, participants were first familiarized with the task and scale range by completing two practice trials before completing the experimental trials. One of the practice trials was a clear case of a scalar implicature, shown in (13), while the other one, shown in (14), clearly could not give rise to the relevant implicature. Each practice utterance was presented in context (see Appendix A).

(13)  a. I had some of the banana yogurt.
     b. I had some, but not all, of the banana yogurt.

(14)  a. There are probably some peanuts in the pantry.
     b. There are probably some, but not all, peanuts in the pantry.

Participants were told that cases like (13) should receive a high rating and cases like (14) should receive a low rating, but were not instructed on which particular value to assign.

Items were divided into blocks of 20 items each. Each block was rated by ten participants. 11 items appeared in two different blocks in order to ensure that each block consisted of 20 items. Because of this, most items received 10 ratings each and 11 items received 20 ratings each.

2.2.2 Results and discussion

The distribution of participants over number of rated blocks of items is shown in Table 1. Mean number of completed blocks per participant was 5.72 (median: 2).
Table 1  Distribution of participants (bottom rows) over completed number of blocks (top rows).

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>120</td>
<td>3</td>
<td>12</td>
<td>2</td>
<td>35</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>46</td>
<td>48</td>
<td>54</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean by−item implicature strength rating count

Figure 1  Distribution of mean per-item ratings.

Mean overall similarity rating was 3.9 (median: 4). The distribution of raw ratings and (aggregated) mean by-item ratings is shown in Figure 1. Under the Homogeneity Assumption –in particular the STRENGTH INVARIANCE sub-assumption– there should be more high than low ratings. Indeed, ratings should be clustered at the upper end of the scale, reflecting overall strong support for the implicature. However, only 44.7% of ratings were higher than the midpoint of the scale, while 46.6% of ratings were lower than 4. Looking only at the endpoints of the scale, only 14.7% of the data were highest ratings while 19% were lowest ratings. Thus, contrary to STRENGTH INVARIANCE, there is a substantial amount of variation in implicature strength across items.

Examples from the lower, medium, and upper end of the scale are shown in (15) - (17). Numbers on the right indicate mean similarity rating.

(15)  Low similarity rating (little support for implicature)
   a. That would take some planning.\textsuperscript{11}  1.4

\textsuperscript{11}Throughout the paper, where examples are taken from the corpus, the some-NP is highlighted in boldface.
Distribution of scalar implicatures

b. And this would give them a chance to have some positive self-esteem. 1.4

c. You sound like you’ve got some small ones in the background. 1.5

(16) Medium similarity rating (medium support for scalar implicature)

a. And some ways, it might be kind of scary. 4
b. I’d love to have, have some animals. 4

c. It would certainly help them to appreciate some of the things that we have here 4

(17) High similarity rating (much support for implicature)

a. But I think that at some times it can be the right thing to do. 6.7
b. I sold some of them. 6.8

c. I like some country music. 6.9

This amount of variation in implicature strength is quite unexpected from the perspective of the previous literature, which overwhelmingly makes the Homogeneity Assumption. These results constitute a good example of how empirically studying a large group of linguistically untrained language users’ pragmatic judgments about naturally occurring language can yield very different results from received wisdom based on individual researchers’ intuitions about artificial examples.

By the same token, as this is the first study of its kind,\(^\text{12}\) it is important to address potential effects of various methodological choices on the outcome of the study. One potential concern is that when given a gradient scale on which to provide judgments, participants will use the entire scale even if they don’t perceive great meaning differences between items. That is, it is possible that in fact, participants strongly got the upper-bound reading in every case, but distributed their judgments over the scale in order to avoid ‘wasting’ the scale. If control items that differed more strongly in meaning had been included in the items, perhaps participants’ judgments would have actually been clustered at the upper end of the scale, as predicted by the Homogeneity Assumption.

This raises a more general issue for experimental pragmatics: the susceptibility of participants’ pragmatic judgments to both a) the other items included in the experimental stimuli and b) the dependent measure used to collect judgments. I suspend discussion of the more general issue here\(^\text{13}\) and instead focus on how one could test whether participants distributed their judgments over the scale for the

\(^{12}\) Though see van Tiel et al. (2013) and Doran et al. (2012) for attempts to quantify between-scale variation in scalar implicature strength and between-implicature-type variation in implicature strength, respectively.

\(^{13}\) But see, e.g., Degen & Goodman (2014) for an investigation of different dependent measures’ varying sensitivity to context effects in the domain of scalar implicature.
uninteresting and theoretically misleading reason that they wanted to use the entire scale, or for the interesting reason that they perceived the implicature with varying strength.

Assume that participants simply wanted to use the entire scale. In this case, there are no systematic reasons for giving an item a high or a low rating; scale use should be random. If this is so, there should be no systematicity to the strength of participants’ judgments. Each item should have received wildly different (random) ratings from different participants, resulting in by-item means clustered around the midpoint of the scale. This is not borne out in the data: some items received very low means, some items very high ones. Furthermore, it is encouraging for the validity of the paraphrase measure that for many cases (e.g., the examples listed in (15)-(17)), the empirically obtained results are in line with intuition. However, future work should investigate the consequences of using different dependent measures to collect implicature judgments for these items.

A second prediction that emerges if participants did not use the scale systematically is that the variation in item means should not be predictable from contextual features. This prediction is in alignment with the prediction made by the Homogeneity Assumption’s sub-assumption of CONTEXT INDEPENDENCE, which will be tested in Section 2.3.

This Section reported a test of the STRENGTH INVARIANCE sub-assumption of the Homogeneity Assumption. The results revealed a much greater degree of variation in implicature strength for implicatures from some to not all than expected under STRENGTH INVARIANCE, suggesting the assumption is not warranted. The next Section tests the second sub-assumption, that of CONTEXT INDEPENDENCE.

2.3 Analyzing the role of contextual cues in implicature strength: a test of CONTEXT INDEPENDENCE

I next turn to investigations of the individual and joint effects of different contextual cues on scalar implicature strength, none of which are predicted to have an effect if implicatures from some to not all constitute a homogeneous class of context-independent implicatures. The investigated cues are a) the partitive form, b) determiner strength, and c) discourse accessibility as quantified by linguistic mention, topicality, and modification. I discuss each of these in turn.

14 That this is so can easily be shown by a simple simulation treating each item mean as the result of 10 random samples drawn from a 7-point Likert scale. For 1363 items, this yields a Gaussian distribution with a global mean of 4 and standard deviation of 0.6, which is very different from the observed distribution. See Appendix B for details.
2.3.1 Cue 1: the partitive form

Consider the difference between (18) and (19).

(18) Alex ate some of the cashews.
(19) Alex ate some cashews.

Intuitively, there is a clear difference in how strongly each of these utterances gives rise to the implicature that Alex did not eat all the cashews. In the example with the overt partitive form \textit{of}, intuition strongly suggests that Alex did not eat all the cashews, while in the example without the partitive this intuition is much weaker. Before addressing whether these intuitions are substantiated in the empirical data, it is worth discussing reasons for why the partitive has this effect.

First, it is well-known that there are additional constraints on using the partitive structure that are not at play for non-partitive quantifiers. Jackendoff (1977) originally formulated the constraint as one of definiteness of the NP embedded under \textit{some (of)}:

(20) Partitive Constraint I
The complement NP in a partitive must be definite.

Subsequently, this formulation of the constraint was shown to be too strong: there are well-documented cases of indefinite, but specific partitives, as in \textit{one of many people who saw the accident} or \textit{half of a cookie} (Ladusaw 1982).

Reed (1991) re-formulated the constraint as one of discourse accessibility. She proposed that the embedded NP must refer to a discourse accessible group; rather than \textit{evoking} a discourse group, the embedded NP must \textit{refer back to} an already mentioned (or inferable) discourse group. The function of the partitive structure is to evoke a subgroup of that discourse group. Under a discourse accessibility account like Reed’s, the strong preference for the embedded NP to be syntactically definite is explained by the embedded NP’s discourse function: “the need to access a discourse group creates a preference for, but not a restriction to, definite NPs in the embedded position” (Reed 1991: p. 216).

Whence, then, the intuition that partitive \textit{some} more strongly gives rise to the implicature that Alex did not eat all of the cashews than non-partitive \textit{some}? Consider what drawing the implicature requires: in order to infer that the speaker intended to convey that \textit{X} is the case of \textit{some, but not all, Y}, there must be some group \textit{Y}, mutually known by both interlocutors, that can be partitioned. Such groups are precisely Reed’s discourse accessible groups. That is, the partitive’s intuitively high propensity to give rise to scalar implicatures is a consequence of the discourse accessibility constraint on NPs embedded under partitives. It is only with discourse accessible NP referents that scalar implicatures should be able to arise.
Note that this does not prevent utterances with non-partitive some from giving rise to scalar implicatures, i.e. using the partitive is not necessary to get scalar implicatures from utterances with some. As long as the embedded NP is discourse accessible, the scalar inference is possible, whether or not the some-NP is overtly partitive. For example, it seems that if (19) was uttered in a context with a contextually given set of cashews, the speaker should more strongly be taken to mean that Alex did not eat all the cashews than if such a set was not given.

The a priori difference between partitive and non-partitive some in how strongly they are associated with a scalar implicature can be summarized as follows: scalar implicatures can only arise with discourse accessible embedded NP referents. The partitive structure can only be used with discourse accessible embedded NP referents, while non-partitive some can be used with both accessible and inaccessible referents. Thus, the a priori probability of a scalar implicature is higher for partitive some (which always occurs with accessible embedded NP referents) than for non-partitive some (which only sometimes occurs with accessible embedded NP referents).

However, the occurrence of the partitive itself is not sufficient for a scalar implicature to arise, either. Recent evidence from experiments on the processing of scalar implicatures provides some support for this claim. Degen & Tanenhaus (in press) found higher implicature rates for statements with partitive than non-partitive some in a truth-value judgment task. However, implicature rates were not at 100% for either construction, suggesting that the partitive does not categorically force the proper part reading, as has often been noted in the literature (e.g., Horn 1997).

Thus, the presence of the partitive should be a strong, but nevertheless probabilistic, cue that increases implicature strength, but does not fully determine it, compared to cases where the partitive form is absent.

**Data analysis** Here and in the following, I report the results of linear mixed-effects regression models (Baayen et al. 2008) to test the effect of different cues on implicature ratings while simultaneously accounting for conditional dependencies between data points from the same rater. These dependencies are captured in so-called random effects, which offer a convenient way to account for violations of the assumption of independence of each data point (for an introduction directed at language researchers, see Jaeger 2008). This kind of independence is not granted in datasets in which different participants contribute multiple data points; in our case, different participants may have a systematically different perception of how large the shift in meaning is when the implicature is made explicit. Thus a forgiving

---

16 Performing ordinal regression, which accounts for the fact that the obtained data were discrete judgments from a Likert scale, yields the same qualitative results (in terms of significance of effects) as the linear mixed effects model reported here.
participant may have given systematically higher similarity ratings than another, less forgiving, participant. More generally, there may be differences in how different participants use the rating scale. Random effects allow for accounting for this individual participant variability and thus crystallizing the effects of the cues under investigation.

All statistical analyses used mixed-effects linear regression models predicting implicature strength rating from fixed effects of interest (the cues under investigation) and random by-participant intercepts. All fixed effect predictors were centered before entering the analysis. Reported p-values were obtained by MCMC sampling using the pvals.fnc() function in R (Baayen et al. 2008). The partitive and the determiner strength predictor were allowed to interact, as were the three discourse accessibility predictors. I report the main effect of each cue individually. The interaction between partitive and determiner strength is discussed in Section 2.3.2. The interaction between the different discourse accessibility predictors is discussed in Section 2.3.3. The full model is summarized in Appendix D.

**Results**  The dataset of 1363 cases contained 368 (27%) partitives, of which in turn 26.8% were headed by pronouns or demonstratives as in (21) and (22).

(21) Uh, **some of that** unfortunately is legal.
(22) And for **some of them** it was just kind of, I don’t know, not so much a holiday.

As can be seen in Figure 2, the overtly partitive cases received higher implicature ratings than the non-partitive cases ($\beta = 1.01, SE = 0.05, t = 22.05, p < .0001$). Compared to the global mean rating of 3.9, the partitive mean was higher at 5, while the non-partitive mean was lower at 3.5. Similarly, the median rating for partitive cases was 5, while the non-partitive median was 3.

Compared to the 44.7% of cases that globally received ratings above the midpoint of the scale, conditioning on overt partitivity (i.e., excluding non-partitive cases) increases that number to 67.8%. This suggests two things: a) the Homogeneity Assumption seems to be more warranted when the *some*-NP is overtly partitive and b) the partitive is nevertheless not sufficient for unambiguously generating a scalar

---

17 Barr et al. (2013) recommend using the maximal random effects structure whenever possible. In this case, only by-participant intercepts were included because it was not guaranteed that each participant observed items in each of the conditions captured by the fixed effects. In consequence, variability in by-participant random slopes for the fixed effects would not be able to be estimated reliably. Therefore, in accordance with the recommendations made by Barr et al. (2013), the model included only random intercepts and no slopes.

18 Here and throughout the rest of the paper, error bars indicate bootstrapped 95% confidence intervals and numbers in bars indicate number of contributing cases.
implicature: only 25% of ratings were 7s, and 23% of ratings were still below the midpoint of the scale. Examples of partitive cases that received low similarity ratings are shown in (23) - (25) alongside their mean similarity rating.

(23) I wish my mother had had some of those opportunities, because, I think she would have really, she rea-, would have succeeded in a lot of ways, that men, that women were not able to succeed in her generation. 2.4

(24) But when you get into some of these health clubs where you just stand around and wait... 2.9

(25) I just go to be entertained and am not really interested in some of the, like, the Terminator or some of the Schwarzenegger stuff. 2.9

In all three cases, the implicature is not licensed (or only very weakly so) despite the presence of the partitive.

2.3.2 Cue 2: determiner strength

The word some is ambiguous between a weak, indefinite, or non-presuppositional reading, often written as sm because it tends to be unstressed, and a strong, quantificational, or presuppositional, reading (Milsark 1974, 1977, Barwise & Cooper 1981, Ladusaw 1994, Israel 1999). Consider the example in (26).

(26) Some prospectors got the plague.
Distribution of scalar implicatures

The sentence in (26) can mean either that there is an indefinite number of prospectors who got the plague (weak, sometimes also called cardinal interpretation) or that some prospectors got the plague but others presumably did not (strong, sometimes also called partitive or proportional interpretation). In general, determiners can either be unambiguously weak (e.g., a/an and no) or strong (e.g., all and most), or ambiguous between the two readings (e.g., some).

The distinction between weak and strong determiners is central to the distribution of scalar implicatures from some to not all because it has been noted that the use of strong, but not weak determiners, gives rise to scalar implicatures (Ladusaw 1994). Indeed, the partitive form (which, as noted in the previous section, is associated with higher implicature rates than non-partitives) tends to only occur with strong determiners (e.g., Horn 1997, Ladusaw 1994).

However, the weak/strong distinction has been notoriously difficult to pin down (e.g., Horn 1997). The goal here is not to give an exhaustive review of the rich literature on weak and strong determiners, but rather to identify an operationalization of the weak/strong distinction that will facilitate a quantitative test of whether strong some is more likely to give rise to scalar inferences than weak some. To foreshadow, the presuppositionality difference between weak and strong some-NPs (e.g., Lumsden 1988) will be employed to arrive at empirical ratings of the strength of each use of some in the database. I begin by elaborating on some of the properties that have been observed to correlate with the distinction.

Table 2 summarizes the diagnostic tests relevant for our purposes, provided in a review by Horn (1997). The property that we crucially depend on in collecting strength ratings from participants is one made by e.g. de Jong & Verkuyl (1985) and Lumsden (1988). They propose that strong determiners introduce the presupposition that their restriction not be empty and their domain of quantification be part of the domain of discourse. That is, under the strong interpretation of (26), there needs to be some set of prospectors in the domain of discourse of whom it is being predicated that they got the plague. Under the weak reading, the domain of discourse need not contain a set of prospectors - the set is introduced (the discourse group evoked, in Reed’s (1991)’s terms) by the some-NP.

The weak/strong distinction correlates with other properties which are not directly relevant to our purpose (of finding an empirical operationalization of the weak/strong distinction), e.g., the propensity to occur in existential there constructions (Milsark 1974, McNally & Geenhoven 1998) and the ability to occur with individual-level predicates (Carlson 1977, Milsark 1977). Importantly, the literature provides counterexamples to each of these diagnostics (see e.g., Horn 1997, McNally & Geenhoven 1998). Rather than being strict constraints or part of the definition of strong determiners, it seems that these properties are approximate diagnostics and I will treat them as such.
In particular, to arrive at an estimate of the strength of *some* for each of the cases in the database, the presuppositionality difference was exploited in a web-based study collecting participants’ judgments about the use of *some*. To quantify determiner strength, participants rated the similarity of each original utterance from the dataset to the same utterance without *some (of)* on a seven-point Likert scale. The reasoning behind this choice was built on the presuppositional nature of strong NPs: the weak use of *some* does not have a non-empty restriction presupposition associated with it, while the strong one does. Thus, in removing *some (of)*, the change in meaning should be greater for strong than for weak *some*-NPs. Consider examples (27) and (28).

(27) Weak use
  a. But my son needed sm money.
  b. But my son needed money.

(28) Strong use, partitive
  a. And some of the people in our church use birth control.
  b. And the people in our church use birth control.

(29) Strong use, non-partitive
  a. Some history books are pretty scary.
  b. History books are pretty scary.

Mutual entailment holds between the a and b sentences in (27) but not in (28) and (29), i.e., all else being equal, the difference in meaning between the a and b forms in (28) and (29) is greater than in (27). Thus, the higher the similarity rating given for a particular case, the weaker the use of *some* in this case. Conversely, the lower the rating the stronger the use.

---

19 Details of this study can be found in Degen (2013) and Degen & Jaeger (in prep.).
Results

The distribution of mean by-item strength ratings is shown in the left panel of Figure 3. Determiner strength ratings correlated strongly with the diagnostics proposed in the literature (Degen 2013, Degen & Jaeger in prep.). For example, the right panel of Figure 3 demonstrates that partitive cases received on average much lower similarity ratings (i.e., higher strength ratings) than non-partitive cases. However, no bimodal distribution indicating two categorically distinct uses – weak vs. strong – was observed.

Implicature ratings were lower with decreasing determiner strength ($\beta = -0.54$, $SE = 0.03$, $t = -21.10$, $p < .0001$). This is shown in Figure 4. The stronger the use of some, the stronger the support for a scalar inference. Conversely, the weaker the use, the weaker the implicature. This is compatible with the general observation in the literature that strong uses of some can give rise to the implicature, but it is important to note that this is not a perfect correlation (Pearson’s $r = -0.51$). That is, some uses of the determiner were judged as strong but did not strongly support the implicature, whereas others were judged to be weak but nevertheless received high implicature strength ratings. Examples of each of these cases are given in (30) and (33).

(30) Strong determiner, low implicature rating (first number: mean determiner strength rating; second number: mean implicature rating)
   a. I’d like to go to Sundance and Park City and some of those. 2.6; 3.6
   b. What are some of the things they don’t recycle. 4.1; 3.8
   c. Maybe this would be a way to get that feeling back, if we’ve lost some of that. 4.1; 3.9

Cases of strong some that nevertheless give rise to scalar implicatures only weakly, if at all, are not in principle surprising: standard lower-bound interpreta-

20 Note that in this study, strength ratings were not collected for the 99 cases where the head of the embedded NP was a deictic expression like a pronoun or a demonstrative. Thus, strength ratings were not independently available for these cases, but were instead simulated in a principled way. See Appendix C for details of the procedure.

21 There was also an interaction of determiner strength and partitivity of the some-NP ($\beta = 0.43$, $SE = 0.05$, $t = 8.4$, $p < .0001$). One potential reason for this interaction is that determiner strength has an effect on implicature strength for non-partitive (potentially weak or strong) cases, but not for partitive (i.e., by definition strong) cases. However, inspecting the simple effects model reveals that there is an effect of determiner strength on both partitive and non-partitive cases; the significance of the interaction term arises from the effect of determiner strength being weaker for partitive than non-partitive cases. This provides further evidence that even within partitives, there are stronger and weaker cases of some.

22 Note that the correlation is negative because higher ratings in the determiner strength rating task corresponded to weaker uses of the determiner and conversely, lower ratings indicated stronger determiner use. Thus, high implicature ratings should be correlated with low determiner strength ratings, resulting in a negative correlation - which is what we observe.
Figure 3  Distribution of mean by-item strength ratings overall (left) and conditioned on whether or not the some-NP was overtly partitive (right). Higher ratings indicate weaker determiner uses.

Figure 4  Mean by-item implicature rating as a function of decreasing determiner strength. Opacity of each point indicates the contributing number of data points (i.e. darker dots indicate more contributing cases).
Distribution of scalar implicatures

tions, where the implicature does not arise because the stronger alternative is not contextually relevant, should give rise to just this pattern. The example in (30a) seems to be of this type. In contrast, the weak implicature support in (30b) and (30c) seems to have a different source: in (30b), the some-NP is embedded in a wh-question, while in (30c) it is in the antecedent of a conditional. Both of these are instances of non-upward-entailing environments, which have been known to cancel and even flip implicatures (Atlas & Levinson 1981, Chierchia 2004, Chierchia et al. 2008, Horn 1989, Levinson 2000).

Context monotonicity annotation of a random sample of 50 cases yielded only two cases where the some-NP occurred in non-upward-entailing contexts; both of these were polar interrogatives. If this is a good estimate of the rate at which some occurs in these contexts, roughly 4% of some-NPs occur in non-upward-entailing environments. In these cases, implicature ratings should be low. The following two are the polar interrogative cases with their mean implicature strength rating.

(31) Or do some of them play the same song? 4.7
(32) But is it a legal, uh, solution for some companies? 5.4

Both of these mean ratings are higher than the global mean in the dataset, suggesting that at least in polar interrogatives, the implicature is not categorically ruled out. However, a complete test of the effect of non-upward-entailing (and especially downward-entailing) contexts on ratings in our dataset remains to be conducted.

I turn next to examples of cases where determiner use was judged as weak but implicature ratings were nevertheless high.

(33) Weak determiner, high implicature rating (first number: mean determiner strength rating; second number: mean implicature rating)

a. It’s hurting, you know, it’s hurting Germany, for example, too, and some other parts of Europe where they, where they have high industry. 6.4 5.7
b. And, after I, I graduated, I read some of the old classics that I just bluffed my way through and have found that I enjoy them quite a bit, too. 6.2 6
c. But I think that at some times it can be the right thing to do. 6.2 6.7
d. And then on the other hand, I’ve seen some people go into the nursing home and just so happy you know. 5.8 5.7

There seem to be two different things going on here. In the a) and b) cases, use of the determiner is weak in that it is introducing two new discourse groups - other parts of Europe and old classics. However, the modifying post-nominal material
introduces a contrast with a (presumably non-empty) complement set - parts of Europe where they don’t have high industry and the old classics that I did not bluff my way through. In these cases, then, the upper-bound interpretation may not arise as a standard implicature, but as a consequence of the non-empty complement set presupposition introduced by the post-nominal modification.

Similarly, in the c) and d) cases the upper-bound interpretation does not seem to be due to the standard Quantity reasoning, but instead the prior probability of the state of the world signaled by the upper-bound interpretation is high: world knowledge tells us that it is more likely that it is not at all times the right thing to do rather than that it is (whatever it may refer to in this case). And it is more likely that not all people go into the nursing home and are happy rather than that they all are.

Thus, while implicature support is strongly correlated with determiner strength, factors like monotonicity properties of the context that the some-NP is embedded in, discourse expectations, and world knowledge affect scalar implicatures.

2.3.3 Cue 3: discourse accessibility

As discussed above, Reed (1991) proposed a discourse accessibility constraint on the partitive - the partitive can only be used with embedded NPs referring to discourse accessible referents. Relatedly, strong uses of some have been argued to be both covertly partitive and have a discourse accessibility presupposition on the embedded NP. In this section I investigate the effect of discourse accessibility on scalar implicatures above and beyond overt partitivity and determiner strength.

Several factors contribute to discourse accessibility: here I investigate a) linguistic mention of the embedded NP referent, b) topicality of the some-NP, and c) modification of the embedded NP.

Several researchers have noted that scalar implicatures seem to be affected by information structure. For example, Breheny et al. (2006) found that more scalar implicatures are generated in Greek for sentences in which the some-NP is in subject position compared to when it is in object position. Their explanation is that scalar implicatures should only arise for sentences that address a contextual QUD that is about the constituent containing the scalar item. Because of the strong tendency of Greek (and weaker tendency of English) for subjects to be more likely to contain old information, i.e. information that a QUD is about, scalar implicatures should be more likely to arise for some-NPs in subject position than in positions that are lower on the obliqueness hierarchy (e.g. objects, adjuncts, etc.).

Note, however, that other accounts make the opposite prediction. For example, van Kuppevelt (1996) proposes that scalar implicatures can only arise if the scalar item occurs in the comment part of the sentence, that which answers the contextual QUD. However, the default comment position in English is the object position.
Distribution of scalar implicatures

Taken together, this predicts effects of both linguistic mention and subjecthood on scalar implicatures: implicature ratings should be higher with both previously mentioned and subject some-NPs. Additionally, adding pre- or post nominal modification to an NP that refers to a new (previously unmentioned) entity or group makes this group accessible (Reed 1991). Consider the following example:

(34) When we arrived at the hotel we didn’t know where to go so we asked the guy at the front desk.

The restrictive modifier at the front desk makes the novel mention of the guy discourse-accessible by providing uniquely identifying information (Reed 1991, Webber 1983).

The combination of these three different markers of discourse accessibility - mention, topicality, and modification - could plausibly affect scalar implicature strength in various ways. First, it is possible that each of the markers has an independent, additive effect. For example, an utterance with a modified subject some-NP may more strongly give rise to the implicature than one with an unmodified subject some-NP. This would constitute evidence for a gradient notion of discourse accessibility: the more discourse-accessible a particular some-NP, the stronger the implicature. Another possibility is that discourse accessibility affects implicature strength in a categorical way, such that as long as the discourse accessibility of the some-NP is guaranteed by at least one of the markers (subjecthood, previous mention, or modification), the presence of another marker has no further strengthening effect on the implicature.

To reflect the potential for complex interactions between discourse accessibility markers, predictors for linguistic mention, subjecthood, and modification were allowed to interact in the regression model. I first discuss the main effects of each of the three factors before turning to the interactions in Section 2.3.3.

**Linguistic mention** Nouns in the Switchboard are annotated for whether they are old (previously mentioned), new (not previously mentioned), or mediated (not previously mentioned but contextually inferable). In the some-database, there were 142 old, 767 mediated, and 454 new cases. Figure 5 shows mean strength ratings for different mention categories. There is a clearly gradient increase in implicature strength with increasing discourse accessibility.

For ease of analysis, old and mediated head nouns were collapsed into one category. As predicted, implicature ratings were higher for old than new NPs ($\beta = 0.32$, $SE = 0.04$, $t = 8.62$, $p < .0001$).

---

24 This is assuming that each of the markers has some effect. It is of course also possible that none of them or only a subset affects implicature strength.

25 Old and inferable information tends to pattern together in discourse (Birner 1997).
One surprising finding is that there were many new NPs that nevertheless received high implicature ratings. I discuss this further in Section 2.3.3.

Subjecthood  In the Switchboard corpus, NPs are annotated for whether they are sentential subjects as in (35) or in topicalized constructions like left-dislocations as in (36). These some-NPs stand in contrast to some-NPs that occur in other positions, e.g. as direct objects or in prepositional adjuncts as in (37) and (38).

(35) Some people are motorboaters, you know, which I think is fine. 5.5
(36) Some of those people, they don’t deserve to be let loose. 4.8
(37) I’ve heard some horrible, horrible stories about high school teachers. 3.1
(38) We actually do some work with some people down at Georgia Tech. 4.5

Because there were only 19 cases of topicalized NPs, these were collapsed into the subject NP category. There were thus 257 subject and 1106 other NPs in the some-database. Figure 6 shows mean implicature strength ratings for subject vs. other NPs: subject some-NPs give rise to stronger implicatures than other NPs ($\beta = 0.40, SE = 0.06, t = 7.25, p < .0001$).

Modification  Finally, each case in the database was coded as either modified or unmodified, depending on whether or not the embedded NP contained pre- or post-nominal modification. For example, the examples in (39) and (40) both fell into the modified category, while the case in (41) was classified as unmodified, resulting in
Distribution of scalar implicatures

667 modified and 696 unmodified cases. In addition, partitive cases with possessive embedded determiners were categorized as modified because in those cases, the determiner provided additional information about the relation between the head noun of the embedded NP and already discourse accessible entities, as in (42) where the possessive provides a link between relatives (new) and the speaker’s family (old). There were 12 of these cases in the database overall.

(39) And then I’ve seen some of the Star Trek movies. 6.5
(40) We’re a little farther removed from like Dallas and some of the areas where they probably have more of the homeless and that type of thing. 5.2
(41) We had some friends over as recently as Saturday night. 3.4
(42) Christmas time, uh, some of our relatives would come up from Alabama. 6.3

Figure 7 shows mean implicature strength ratings for modified and unmodified some-NPs. Somewhat surprisingly, and supported by the analysis, unmodified NPs received higher ratings than modified NPs ($\beta = 0.11, SE = 0.04, t = 3.14, p < .001$). This is due to the interaction with especially linguistic mention, which is discussed in the next section.

**Interactions between discourse accessibility factors** The model coefficients for the two-way and three-way interactions between discourse accessibility predictors are shown in Table 3. All interactions were significant, though the one between modification and mention only marginally so. The three-way interaction is visualized...
in Figure 8. Simple slopes analysis revealed that the two-way interaction between linguistic mention and subjecthood was significant for unmodified, but not for modified NPs; for modified NPs, there was only a main effect of subjecthood, such that modified NPs in subject position received higher implicature ratings than modified NPs in other positions. For unmodified NPs, there was an interaction such that both old and subject NPs received higher ratings, but the difference between subject and other NPs was greater for old than for new NPs.

This suggests a role for discourse accessibility that is intermediate between the two roles sketched in Section 2.3.3. As a reminder, the options were either a) that discourse accessibility has a gradient effect on implicature strength—the more accessible, the greater the implicature strength—or b), that discourse accessibility has a categorical effect on implicature strength—if at least one marker of discourse accessibility is present, implicature ratings should be high, and low otherwise. The results suggest that subjecthood has a special status: subject some-NPs ‘boosts’ implicature strength, no matter the presence of other discourse accessibility markers,
indicating an additive effect of subjecthood, in turn supporting a gradient view of discourse accessibility. The more discourse-accessible the *some*-NP, the stronger the implicature. In contrast, previous mention affects implicature strength for unmodified (less discourse-accessible), but not modified (more discourse-accessible) NPs. For cases where discourse accessibility is guaranteed (or at least increased) through modification, mention does not add an extra boost. For cases of unmodified NPs with reduced discourse accessibility, previous mention *does* provide a boost. That is, there is evidence for both categorical and gradient effects of discourse accessibility on implicature strength.

### 2.4 Model evaluation

In Section 2.3 I reported and discussed the effect of multiple contextual features on scalar implicature strength. In particular, partitivity, determiner strength, and three markers of discourse accessibility (and their interactions) all affected implicature strength. Some readers may wonder at this point how well this model does at capturing the data, and whether all of the predictors included in the model are necessary. The second question is easy to answer: the linear regression model used

26 As a side note, a $\chi^2$ test over the linguistic mention x subjecthood contingency table replicates the well-documented tendency for subjects to favor old over new information (33 new subjects vs. 180 old subjects, $\chi^2(1) = 58.73, p < .0001$).
Figure 9  Scatterplot of empirical vs. predicted mean by-item strength ratings for final model (left panel) and model with only by-participant intercepts (right panel). Blue lines indicate best linear fit.

This guarantees that each of the significant predictors (including interaction terms) has an independent effect on implicature strength.

As for the first question, we can inspect two different measures of model quality. The first is the Bayesian Information Criterion (BIC, Schwarz (1978)), a measure of model quality that takes into account the likelihood of the data, given the model. It includes a penalty for added parameters. Models with lower BIC values are preferred over models with higher BIC values. We can thus compare our final model to a basic model that includes only by-participant random intercepts. The final model has a BIC of 55,938 and the basic model 58,493. Model comparison reveals that the final model is a vast improvement over the basic model ($\chi^2(11) = 2663.4, p < .0001$).

A different, more intuitive way of evaluating the model is to compare the empirical data to the values predicted by each of the two (final and basic) models. This is visualized in Figure 9. The first observation is that the final model clearly predicts a much wider range of values than the basic model that only accounts for participant variability, capturing that participant variability is much lower than the variability due to the contextual factors investigated. The better fit of the final model to the data is also reflected in the correlations: the final model predictions are much more highly correlated with the empirical data ($r = .66$) than the basic model predictions are ($r = .16$).

The final model thus provides a quite reasonable fit to the data. However, it is also clear that there is residual variability that the model does not account for.
I briefly discuss potential sources of this variability, some of which were already touched upon in Section 2.2.1.

One source of variability may be the sensitivity of scalar implicatures to embeddings within polarity affecting contexts. I have argued in Section 2.3.2 that due to the very rare occurrence of such contexts at least in the dataset reported here, this factor may play a much smaller role than assumed by some (Chierchia et al. 2008). More likely causes of residual variability are a) the degree of uncertainty that hearers believe speakers to have about the truth of the stronger alternative, b) hearers’ perceived relevance of the stronger alternative to a contextual QUD, and c) the prior probability of the stronger alternative being true, all of which have been argued to play a role in deriving scalar implicatures (Bergen & Grodner 2012, Breheny et al. 2013, Degen 2013, Franke 2009, Gazdar 1979, Geurts 2010, Goodman & Stuhlmüller 2013, Grice 1975, Horn 1989, Matsumoto 1995, Russell 2012, Sauerland 2004, Zondervan 2010).

**Hearer estimate of speaker knowledge** As has been noted, by default only a weak implicature to the effect that the speaker does not believe that the stronger alternative holds arises upon observing a scalar item like *some* (Gazdar 1979, Horn 1989). To get the stronger implicature that the speaker believes that the stronger alternative does not hold, the hearer must make the assumption that the speaker is knowledgeable with respect to the truth of the stronger alternative (Sauerland 2004). The local insertion of *but not all* into the target utterance may have shifted participants’ estimates of the knowledge state that a speaker must have been in in order to produce the modified utterance, thus leading to a greater perceived dissimilarity between target and comparison utterance.

**Relevance of the stronger alternative to a contextual QUD** Locally inserting *but not all* may have shifted the salient QUD that the utterance is interpreted relative to, as discussed in Section 2.2.2. This may have also lowered participants’ perceived similarity between target and comparison utterance.

**Prior hearer belief in the truth of the stronger alternative** In some cases world knowledge about the relative probability of the stronger alternative being false, independent of any considerations of likely speaker knowledge or the QUD, may have guided participants’ judgments. The following are cases that received strong implicature strength ratings despite the model predicting low ratings. The first value is the model’s predicted rating, the second value the actual mean item rating.

(43) *There are some Kurds* living in Iran. 3.4/5.0
And it’s a brick house, with, uh, some wood. 3.3/5.2

It is clear that no inference about speaker intentions is necessary in order to express a judgment that not all Kurds live in Iran and that a brick house is not embellished with all the wood in the world.

All three of these factors are likely contributing to participants’ final strength ratings in complex ways (see also Russell (2012) for a comprehensive discussion of how prior beliefs, relevance of alternatives, and other factors are expected to interact in giving rise to perceived implicature strength). Including estimates of these three factors for each of the items in the dataset will likely improve model fit. How these factors interact with the the contextual factors discussed in Section 2.3 is an interesting empirical question that will shed light on the relation between surface features (e.g. the partitive), semantic features of surface forms (e.g. discourse accessibility), world knowledge (e.g. prior beliefs about likely states of the world), and top-down expectations about language use (e.g. the relevance of an utterance and its alternatives to a contextual QUD), in scalar implicature computation.

3 General Discussion

Lexicalized scalar implicatures have long been accepted as belonging to the class of GCIs (Horn 1989, Levinson 2000). The main grounds for this classification has been individual researchers’ intuitions regarding the relative regularity and context independence with which scalar implicatures arise compared to more context-dependent PCIs. This paper constitutes an attempt to rigorously test the main assumption underlying the classification of scalar implicatures as GCIs, an assumption I termed the Homogeneity Assumption and spelled out in terms of two sub-assumptions: that of STRENGTH INVARIANCE (that scalar implicatures display no or little variability in the degree to which they arise) and that of CONTEXT INDEPENDENCE (that they arise independently of context).

In Section 2 I reported a test of the Homogeneity Assumption for the ⟨all, some⟩ scale, the most clearly lexicalized of scales, in which a large number (243) of linguistically untrained language users’ implicature strength judgments were collected for 1363 naturally occurring utterances containing some. This allowed for separate tests of STRENGTH INVARIANCE and CONTEXT INDEPENDENCE: first, the overall variation in participants’ judgments was analyzed. In a second step, the effect of three features of context (or, from the hearer’s perspective, contextual cues to the speaker’s intention) on implicature strength was analyzed.

In the following I summarize the main experimental results and discuss the implications of these results for a) the status of scalar implicatures as GCIs, b)
processing theories that rely on scalar implicatures constituting a homogeneous class of implicatures, and c) the status of the GCI-PCI distinction itself.

3.1 Summary and discussion of experimental results

The two main results reported in this paper are a) that scalar implicatures from some to not all display much more variation than expected under the Homogeneity Assumption and b) that implicature strength is probabilistically modulated by various features of context. In particular, implicature strength (or the degree to which a speaker is taken to implicate the negation of the stronger alternative) is greater on average when some occurs with the partitive, when its use is relatively strong, and when the embedded NP referent is relatively discourse accessible (i.e., when it has been previously mentioned or is contextually inferable, when the some-NP is in subject position, or when the embedded head noun is modified).

These results suggest that the Homogeneity Assumption, at least for scalar implicatures from some to not all, is not warranted. Not only is implicature strength highly variable between different occurrences of some (pace STRENGTH INVARIANCE), it is also systematically dependent on context (pace CONTEXT INDEPENDENCE). This result undermines the assumption made in the previous literature, which overwhelmingly treats lexicalized scalar implicatures as highly regularized, context-independent inferences (Horn 1972, Huang & Snedeker 2009, Levinson 2000).27 It will therefore be crucial for follow-up work to establish the robustness and generalizability of the result by testing the Homogeneity Assumption both for different scales and different dependent measures.28 It is possible that implicatures using the ⟨all, some⟩ scale just happen to display more variation and context-dependence than expected, and other scales may indeed display the behavior expected under the Homogeneity Assumption. This is unlikely, given the status of the ⟨all, some⟩ scale as the paradigmatic example of a lexicalized scale, which suggests that any other scale should display more, rather than less, context-dependence. But this is an empirical question that can and should be answered by digging into the increasing amount of and increasingly available corpus and judgment data.

An additional important area for future work arises from the observation that, while the model predictions are reasonably correlated with participants’ actual ratings, there is still substantial residual variability in ratings that the model does not capture. As discussed in Section 2.4, this variability is likely due in large part to

27 But see Ariel (2004) for evidence that the not all implicature is similarly infrequent for most.
28 First attempts at establishing implicature strength for different scales and for different implicature types have been made by van Tiel et al. (2013) and Doran et al. (2012); they find a large degree of variability. However, in-depth studies into the context-dependence of these various implicatures remain to be conducted.
factors not presently included in the model, including a) the degree of uncertainty that hearers believe speakers to have about the truth of the stronger alternative, b) hearers’ perceived relevance of the stronger alternative to a contextual QUd, and c) the prior probability of the stronger alternative being true. Future work should attempt to estimate these quantities and integrate them into the model.

3.2 Implications

The results reported in this paper have implications both for theories of how scalar implicatures arise and for theories of how they are processed, which are discussed in the following.

3.2.1 Scalar implicature and the GCI-PCI distinction

The variability in strength and the context dependence exhibited by scalar implicatures from some to not all is at odds with the status of these implicatures as GCIs. Let’s reconsider Grice (1975)’s formulation of what makes a GCI: “the use of a certain form of words in an utterance would normally (in the absence of special circumstances) carry such-and-such an implicature or type of implicature”. The results reported in this paper clearly indicate that an utterance of some cannot be said to normally carry a not all implicature. But maybe the parenthetical in the absence of special circumstances can help? Others, e.g. Horn (1984) have also remarked that a GCI should go through in unmarked contexts. That is, perhaps the test of the Homogeneity Assumption was unfair because it did not exclude marked contexts?

There are multiple reasons why this is not a satisfying objection. First, it is not clear what would constitute a marked context. The presence of at least one of the features that lower implicature strength, identified in Section 2.3? The presence of all strength-lowering features? Something in between? Excluding cases according to any of these criteria will increase implicature strength - this follows from the statistical analyses reported above. However, a) implicature strength in the remaining cases will nevertheless not be maximal and b) the remaining cases will retain variability in strength; that is, STRENGTH INVARIANCE will remain violated. Moreover, trying to save the GCI status by pointing to the markedness of some of the contexts included in the analysis ignores the gradient and systematic dependence of implicature strength on context. This context dependence is arguably the more interesting finding than that of variation in implicature strength; it suggests that scalar implicatures from some to not all are much more PCI-like than previously suspected.

But if some-not-all implicatures are more like PCIs than like GCIs, this does not bode well for the GCI-PCI distinction. Others have previously questioned the
usefulness of the distinction or argued that it is a matter of degree rather than type (Carston 1998, Hirschberg 1985, van Rooij 2003, Wilson & Sperber 1995). The results reported in this paper provide further evidence that, rather than conceiving of the space of form to inference mapping as either categorically context-dependent or context-independent, different implicature types may be more or less context-dependent. This paper has shown that some-not-all implicatures, the traditionally most context-independent of implicatures, are nevertheless systematically context-dependent. Of course there are many implicatures that are more context-dependent; but if we take some-not-all implicatures to provide a lower bound on the context-dependence of implicatures, categorical context-dependence cannot be used as a diagnostic of whether a particular implicature should be considered a GCI or a PCI. But context-dependence has been the diagnostic for distinguishing GCIs and PCIs. Discarding it leaves the GCI-PCI distinction with no other independent diagnostic, rendering it meaningless.

What might be an alternative view of the role of context-dependence in conversational implicatures? We would like to avoid simply saying that all implicatures are context-dependent and leave it at that. Under the probabilistic pragmatics view sketched in Section 1.2, different implicature types fall onto a continuum of context-dependence. The challenge now lies in quantifying the degree of context-dependence for different implicature types. This paper provides an example of how, with the help of corpora of spontaneous speech as well as the web-based collection of regular language users’ interpretation of utterances, one can begin to probe implicatures’ degree of context-dependence. While a precise suggestion for how to quantify context-dependence in the general case lies outside the scope of this paper, the work reported here opens an exciting avenue for the development of data-driven estimates of the degree and type of context-dependence exhibited by different implicature types.

3.2.2 Processing theories of scalar implicature

Overturning the Homogeneity Assumption also has consequences for theories of how scalar implicatures are processed. In Section 1.1.1 I introduced two such theories: the Default model (Levinson 2000) and the Literal-First hypothesis (Huang & Snedeker 2009). I discuss both in turn.

**The Default model** The Default model assumes that the process of computing lexicalized implicatures does not incur a processing cost; only their cancellation does (Levinson 2000). Levinson sees this as a solution to the articulatory bottleneck problem, the question of how it is that communication can proceed as rapidly as it does, assuming that integration of contextual information is effortful and time-
consuming. The crucial assumption that warrants this solution is precisely the Homogeneity Assumption - only if scalar implicatures from *some* to *not all* are context-independent does removing a processing cost from their computation allow for an overall processing speedup.

The results reported here thus undermine the core assumption of the Default model. But what would constitute an alternative solution to the articulatory bottleneck problem?

Under the probabilistic pragmatics approach, the assumption that the integration of contextual information is generally cognitively costly is relaxed. Listeners are assumed to have acquired rich, probabilistic knowledge of the contexts in which speakers intend to communicate the negation of the stronger alternative. When observing an utterance with a scalar item, listeners can then use the available contextual information to infer the speaker’s intention. If contextual support for the implicature is strong, it should be computed more rapidly than if contextual support is weak. A thorough test of this prediction remains to be conducted. A further interesting question is how listeners learn to track the right kinds of contextual features.

**The Literal-First processing hypothesis** The Literal-First hypothesis is the hypothesis that the lower-bound interpretation of an utterance with a scalar item is computed before the pragmatically enriched upper-bound interpretation. Evidence for this staged process would provide evidence for a clear distinction between semantics and pragmatics, and more importantly, for a modular view whereby the pragmatics operates on the semantics. The modularity view is of course widespread in linguistics – literal meanings of expressions are taken to be basic and can be pragmatically enriched. However, it is not clear that this distinction is psychologically meaningful.

The Literal-First hypothesis is associated with two key predictions: that the pragmatic upper-bound interpretation of *some* should be derived more slowly than a) the literal lower-bound interpretation and b) than other literal controls (e.g., utterances in which *some* is replaced by *all*).

The bulk of the empirical findings – from response times (Bott & Noveck 2004, Bott et al. 2012, Neys & Schaeken 2007), reading times (Breheny et al. 2006), mouse movements (Tomlinson et al. 2013), and eye movements (Huang & Snedeker 2009, 2011)– points towards implicatures being costly, in support of the Literal-First hypothesis.

However, making the Homogeneity Assumption is crucial to the interpretation

---

29 There is also ample evidence from the processing literature that scalar implicatures are not computed by default in the general case (Bott & Noveck 2004, Huang & Snedeker 2009, 2011, but see Grodner et al. 2010, Breheny et al. 2013).
of the ‘costly implicature’ results as reflecting a literal-first process, in the following way. If the Homogeneity Assumption is not warranted, an alternative interpretation of the ‘costly implicature’ results is at least as plausible as a literal-first processing mechanism: this alternative is that the observed delays are due to the low prior support for an implicature. To see why this should be so, a brief foray into frequency and predictability effects in other areas of language processing is required.

From a vast body of literature on frequency and predictability effects in other domains of language processing, it is well-known that more frequent or predictable words or structures are processed more quickly than less frequent or predictable words or structures. For example, more frequent words are recognized more quickly and more accurately than less frequent words (Dahan et al. 2001, Marslen-Wilson 1987, Seidenberg & McClelland 1990). Similarly, more frequent, contextually predictable, and less surprising words and structures are read more rapidly than less frequent and predictable ones (Ehrlich & Rayner 1981, Hale 2001, Levy 2008, Mcdonald & Shillcock 2003).

If predictability affects pragmatic processing just as it does lexical, phonological, or syntactic processing, what are the predictions for scalar implicature processing? Under the Homogeneity Assumption, implicature strength is predicted to be high and context-independent. In other words, scalar implicatures are highly predictable compared to literal interpretations of utterances with scalar items. Thus, the pragmatic interpretation of scalar expressions should be processed more rapidly than the literal one, unless literal information is privileged in processing. Indeed, this is exactly the argument of the Literal-First hypothesis. However, if the Homogeneity Assumption does not hold—that is, if scalar implicatures are in fact not predictable from observing a scalar expression alone—then the pragmatic interpretation should not be arrived at rapidly. And in fact, if the literal interpretation is generally more predictable than the pragmatic interpretation, then scalar implicatures are predicted to be delayed compared to literal content. But this is exactly the pattern predicted by the Literal-First hypothesis.

Thus, if the Homogeneity Assumption is not warranted, there are two alternative explanations for ‘costly implicature’ effects: a) the Literal-First hypothesis, and b) that the less predictable interpretation is arrived at more slowly. The evidence against the Homogeneity Assumption reported here thus calls into question the interpretation of ‘costly implicature’ results as unambiguous evidence for a literal-first psychological processing mechanism.

4 Conclusion

The main contribution of this paper is to provide evidence against the Homogeneity Assumption, an assumption that is central to the view of lexicalized scalar impli-
catures as GCIs. Rather than constituting a homogeneous, context-independent class of implicatures, the results reported in this paper suggest that the strength of scalar implicatures from some to not all is highly variable and systematically context-dependent.

This work demonstrates the feasibility of large-scale experimental studies of pragmatic phenomena in naturally occurring linguistic contexts. In an era where individual researchers’ judgments about hand-selected examples are no longer the only source of linguistic data available, studies of this sort will be of utmost importance in informing pragmatic theory moving forward.
Distribution of scalar implicatures

Figure 10  Distribution of observed (left panel) and simulated (right panel) mean by-item ratings.

A  Practice contexts

(45)  Practice item supporting the implicature
Speaker A: Man, this morning I wanted to have some raspberry yogurt and I checked the fridge and it was all gone.
Speaker B: Mhm.
Speaker A: Did you eat it all?
Speaker B: No I didn’t! I had some of the banana yogurt.

(46)  Practice item not supporting the implicature
Speaker A: Is there any food in the house?
Speaker B: Not sure. There’s probably some peanuts in the pantry.

B  Simulation of implicature strength results under random use of scale

To investigate the expected response distribution if participants were using the Likert scale randomly in the web-based study, and compare it to the actual distribution, a simulation was conducted. Ten (the number of judgments obtained from each participant) independent samples were drawn from a seven-point Likert scale 1363 (the number of items in the dataset) times. The resulting distribution had a mean of 4 and standard deviation 0.6 See Figure 10 for a side-by-side comparison of the observed and simulated distribution of mean by-item ratings. The distributions are very different, suggesting that participants did not use the scale randomly, but systematically. Note that the left panel is repeated from Figure 1. Differences in appearance are due to the difference in the scale on which they are plotted.
C Generating strength ratings for pronominal embedded NPs

Strength ratings for the 99 cases with pronominal embedded NP heads were not available. In order to avoid sacrificing these cases, strength ratings were generated for them in a principled way. To understand the procedure, note first that the partitive is mandatory for pronouns and demonstrative heads (see examples (47) - (49) together with their mean implicature rating).

(47) And some *(of) them fizzled out. 6.6
(48) Some *(of) it sounds more like pop music. 5.9
(49) But some *(of) those are pretty big. 5.6

It is thus plausible that these cases would receive strength ratings similar to those of the other partitive cases. Based on this assumption, ratings were generated by sampling from the strength rating distribution of the remaining 269 partitives. That is, the resulting strength distribution for pronoun/demonstrative cases was approximately the same as that of the other partitive cases. These strength ratings were used in the rest of the analysis. Excluding the pronominal head cases did not qualitatively change the results or the significance of effects.

D Full mixed effects linear regression model

The following table contains model coefficients for the full mixed effects linear regression model predicting implicature ratings from fixed effects for cues of interest and log-transformed sentence length as well as random by-participant intercepts. All fixed effects predictors were centered before entering the analysis.

<table>
<thead>
<tr>
<th></th>
<th>Coef β</th>
<th>SE(β)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.02</td>
<td>0.05</td>
<td>77.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Partitive</td>
<td>1.01</td>
<td>0.05</td>
<td>22.0</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Strength</td>
<td>−0.54</td>
<td>0.03</td>
<td>−21.1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Linguistic mention</td>
<td>0.33</td>
<td>0.04</td>
<td>8.6</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Subjecthood</td>
<td>0.43</td>
<td>0.05</td>
<td>8.2</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Modification</td>
<td>0.11</td>
<td>0.03</td>
<td>3.2</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Sentence length</td>
<td>0.15</td>
<td>0.03</td>
<td>5.3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Partitive:Strength</td>
<td>0.43</td>
<td>0.05</td>
<td>8.4</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Linguistic mention:Subjecthood</td>
<td>0.17</td>
<td>0.12</td>
<td>1.3</td>
<td>&gt;0.18</td>
</tr>
<tr>
<td>Linguistic mention:Modification</td>
<td>0.33</td>
<td>0.07</td>
<td>4.5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Subjecthood:Modification</td>
<td>0.25</td>
<td>0.10</td>
<td>2.5</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Linguistic mention:Subjecthood:Modification</td>
<td>0.54</td>
<td>0.24</td>
<td>2.2</td>
<td>&lt;.05</td>
</tr>
</tbody>
</table>
Distribution of scalar implicatures

References


Degen, Judith & T. Florian Jaeger. in prep. Meaning and processing pressures are some (of the) reasons for choosing the partitive.


Academic Press.


Horn, Laurence. 1997. All John’s children are as bald as the King of France:
Distribution of scalar implicatures


Tanenhaus, Michael K, Michael J Spivey-Knowlton, Kathleen M Eberhard & Julie C


