Proceedings of SALT 34: 110–131, 2024

Embedded scalar diversity*

Eszter Ronai
Northwestern University

Abstract This paper is an experimental investigation of embedded scalar implicatures in the context of scalar diversity. We test whether a sentence such as *Every student read some of the books* leads to the implicature that *No student read all of the books*, and similarly whether *Every soup was warm* leads to the implicature that *No soup was hot*—across 42 different lexical scales. We find 1) that embedded implicatures arise; 2) that there is across-scale variation in embedded implicatures, paralleling scalar diversity among global implicatures; and 3) that properties of alternatives (namely, semantic distance and boundedness) that predict global scalar diversity predict variation at the embedded level too. It is argued that these findings are most compatible with an account of embedded implicatures that builds on alternatives, such as the grammatical theory (i.a., Chierchia 2004; Chierchia, Fox & Spector 2012), a modified neo-Gricean account such as in Sauerland 2004, or the "neo-Gricean uncertainty" version of the Rational Speech Act with lexical uncertainty account (RSA-LU, Potts, Lassiter, Levy & Frank 2015). They are, however, less compatible with the "unconstrained uncertainty" RSA-LU model (Bergen, Levy & Goodman 2016; Potts et al. 2015), which leaves unexplained (without further assumptions) why the same alternative-driven variation should occur both in global and embedded implicatures.

Keywords: scalar implicature, scalar diversity, embedded implicature, alternatives

1 Introduction

The possibility of scalar implicatures occurring in so-called embedded environments, e.g., in the scope of *every*, *no* or *exactly one*, has generated a lot of interest over the years in theoretical and experimental semantics-pragmatics. This is because such embedded enrichments cannot be generated under a standard neo-Gricean account of implicature, necessitating a different theoretical approach, or at least a modification to the neo-Gricean account. The current paper's goal is to experimentally adjudicate among competing theories of embedded implicature, specifically those where the implicature generation process builds on scalar alternatives vs. those that do not.

* For helpful discussion and feedback, I would like to thank the SALT 34 reviewers, Camelia Bleotu, Lucas Fagen, Chris Potts, Florian Schwarz, and Michael Tabatowski. Experimental stimuli, data, and R scripts can be found at: https://osf.io/kx42p/?view_only=bf4a09f6e6ad413894eebc61d91a012d

To do so, we test the availability of embedded implicatures across a large variety of lexical scales, capitalizing on *scalar diversity*, that is, the observed variation in the likelihood of implicature across scales. Prior work has found certain properties of scalar alternatives to be able to predict a scale's likelihood of leading to implicature, and therefore the observed variation. This paper's goal is to probe whether a similar variation arises in embedded implicatures, and crucially, whether the same properties of alternatives play a role in explaining it. Our experimental results suggest that this is indeed the case, which constitutes an argument in favor of theoretical treatments of embedded scalar implicature that build on alternatives. The results, however, are less compatible with accounts that eschew alternatives.

The paper is structured as follows. We begin with a brief background on implicature and embedded implicature (Section 2), followed by the different theoretical treatments of embedded implicature (Section 2.1). Section 2.2 introduces the scalar diversity phenomenon, and the properties of alternatives that play a role in it. Section 2.3 reviews relevant prior experimental work on embedded implicature. Experiment 1 is reported in Section 3 and Experiment 2 in Section 4. Section 5 concludes.

2 Background

Scalar implicature (SI) can enrich a sentence with a scalar term (e.g., *some* in (1)) from its literal meaning (1a) to an upper-bounded interpretation (1b).

(1) Mary read some of the books.

a. Mary read at least some of the books.

b. Mary read some, but not all, of the books.

On a (neo-)Gricean account of SI (i.a., Grice 1967; Horn 1972), upper-bounded interpretations arise via hearers' reasoning about informationally stronger unsaid alternatives. In the case of (1), for example, a stronger alternative is *Mary read all of the books*, which is taken to be stronger than (1) since it asymmetrically entails it. On such an account, speakers are taken to follow the Maxim of Quantity and hence should have said the more informative alternative *Mary read all of the books* if it had been true (and they knew so); because they did not say that, hearers can derive the negation of that alternative via the Maxim of Quality. Crucially, hearers' reasoning concerns utterance-level alternatives.

That neo-Gricean accounts take alternatives to be at the level of utterances is important when we look at possible interpretations of sentences where the weaker scalar term is embedded under an operator or quantifier. In (2), for example, the SI-triggering scalar term *some* appears in the scope of the universal quantifier *every student*. If SI calculation happens globally, at the utterance level, then the alternative

to (2) is *Every student read all of the books*, whose negation leads to the *Not every student...* SI in (2a) (henceforth the *weak inference* or *global SI*[1]).

(2) Every student read some of the books.
   a. Not every student read all of the books. weak inference/global SI
   b. No student read all of the books. strong inference/embedded SI

However, there is another potential inference from (2), namely the *No student...* SI in (2b) (henceforth the *strong inference* or *embedded SI*[2]). On a standard, unmodified neo-Gricean account as described above, there is no alternative to (2) that would generate this inference.

2.1 Theoretical accounts of embedded SI

Though strong inferences cannot be generated under the basic neo-Gricean account sketched above, there exist a variety of theories of SI that can capture such inferences. Here, we briefly review some of them, with special attention to one contrast among different theoretical accounts: whether they derive strong inferences by building on scalar alternatives.

The grammatical theory (i.a., Chierchia 2004; Chierchia et al. 2012) derives embedded SI by allowing for the possibility that SI enrichment applies at the embedded, rather than the utterance level. It posits a silent exhaustification operator **O**, with a meaning akin to *only*. As (3) schematically shows, the application of **O** at the embedded site can derive the strong (embedded) inference from *Every student read some of the books*.

(3) Every student x: **O**(x read some of the books).
   $\equiv$ Every student read some but not all of the books. $\equiv$ No student read all of the books.

Crucially for the present paper, deriving SI via the exhaustification operator still makes reference to alternatives, e.g., *all* for *some*, though these need not be utterance-level. On the grammatical theory, it is also possible for **O** to take scope over the whole utterance, as shown in (4), which results in the weak (global) inference that the standard neo-Gricean account was also able to capture.

(4) **O**(Every student read some of the books). $\equiv$ Not every student read all of the books.

1 Global SI will also be used to refer to unembedded SIs such as (1b).

2 The term *embedded SI* is used here descriptively to refer to strong inferences of sentences where the SI-triggering scalar term is embedded under an operator, without any theoretical commitment. We recognize that there are theoretical proposals that can derive such inferences without generating SI in a local, embedded position (see e.g., the modified neo-Gricean proposal referenced in Section 2.1).

There also exist neo-Gricean accounts that can generate strong inferences by alternative-based reasoning at the utterance level. In order to do this, modifications are made to what can serve as an alternative. Relaxing the requirement that alternatives be stronger than —that is, that they asymmetrically entail —the original utterance, and allowing for alternatives to be constructed by simultaneously replacing several scalar terms of the original utterance[3] (i.a., Sauerland 2004), (5) can be taken to be a relevant alternative to *Every student read all of the books*. Here, not only has *some* been replaced by *all*, but *every* has also been replaced by *some*. Negating such an alternative results in the strong inference *No student read all of the books*.

(5) Some students read all of the books.

A third class of accounts that we discuss here are couched in the Rational Speech Acts framework (RSA, Frank & Goodman 2012), which models pragmatic inferences via Bayesian reasoning. Bergen et al. (2016), and Potts et al. (2015) following them, propose that certain pragmatic inferences, including embedded SI, can arise as a consequence of lexical uncertainty (we use the name *RSA-LU* for this class of accounts, for 'RSA with lexical uncertainty'). *Lexical uncertainty* refers to hearers' uncertainty about the particular lexical interpretations intended by the speaker. That is, rather than assuming that speaker and hearer assign the same denotations to lexical items, the RSA-LU allows that conversational participants may differ in the denotations they assign to lexical items, i.e., that they use different lexica. This amounts to saying that conversational participants can use (in principle) arbitrarily strengthened meanings rather than basic semantic denotations, and one part of conversational reasoning is coordinating on these strengthened meanings.

For instance, suppose that we have a domain of entities $\{a,b,c,d\}$ where a, b, and c are books, but d is a film and not a book. The basic semantic denotation of the quantifier *some book* is a set containing every subset of the domain containing at least one book:

(6) $[\![\text{some book}]\!] = \{\{a\},\{b\},\{c\},\{a,b\},\{a,c\},\{b,c\},\{a,d\},\{b,d\},\{c,d\},$
$\{a,b,c\},\{a,b,d\},\{a,c,d\},\{b,c,d\},\{a,b,c,d\}\}$

But under the RSA-LU, a conversational participant's lexicon can include any refinement (that is, any logical strengthening) of this denotation in its place. For instance, there is a lexicon in which *some book* is strengthened to denote $\{\{a,b,c\},\{a,b,c,d\}\}$, the usual meaning of *all books*, and there is also a lexicon where *some book* is strengthened to denote $\{\{a\},\{a,b,c,d\}\}$, which does not have a natural-language

3 Such an approach can overgenerate meanings in other cases, which has led to a large number of proposals regarding the precise theory of alternatives. These details are not relevant for the present paper, and we refer the reader to Gotzner & Romoli (2022: Section 3.2.) for an overview.

equivalent, etc. It is through recursive Bayesian reasoning about these strengthened interpretations in the RSA framework that Bergen et al. (2016) propose to account for embedded SI. Crucially for our purposes, this reasoning process does not invoke alternatives per se; conversational participants can use any logically possible strengthening, not just those derivable from alternative-based reasoning.

Potts et al. (2015) compare multiple possible RSA-LU models of embedded SI, among them *unconstrained uncertainty*, which as in Bergen et al. 2016 allows conversational participants to operate with any arbitrarily strengthened interpretation; and *neo-Gricean uncertainty*, which restricts refinements to the standard set of alternative-excluding meanings (e.g., on the neo-Gricean uncertainty approach, *some* can either remain unstrengthened or it can be strengthened to *some but not all*). As such, the neo-Gricean uncertainty model builds in a role for alternatives in the embedded SI generation process. Potts et al. (2015) find that the neo-Gricean uncertainty model provides the best fit with experimentally elicited data.

In sum, there exist treatments of embedded SI that utilize scalar alternatives, such as the grammatical theory, a modified neo-Gricean account, or the neo-Gricean uncertainty version of RSA-LU. Other versions of the RSA-LU, however, derive these strengthened meanings via an alternative-free process.

2.2 Scalar diversity

Recent experimental work has shown that the likelihood of (unembedded) SI varies robustly across different scales (van Tiel, Van Miltenburg, Zevakhina & Geurts 2016; Baker, Doran, McNabb, Larson & Ward 2009; Beltrama & Xiang 2013; Doran, Ward, Larson, McNabb & Baker 2012). For instance, the *some but not all* SI arises at higher rates than the *warm but not hot* SI in (7).

(7) The soup is warm. → SI-enriched: The soup is warm, but not hot.

Several factors have been identified as predicting (some of the) observed across-scale variation in SI rates. Certain factors relate to properties of scalar alternatives (*all*, *hot*, etc.), such as their distinctness (van Tiel & Schaeken 2017; Westera & Boleda 2020), expectedness (Ronai & Xiang 2022; Hu, Levy & Schuster 2022; Hu, Levy, Degen & Schuster 2023), adjectival polarity (Gotzner, Solt & Benz 2018), or extremeness (Beltrama & Xiang 2013). Other factors derive variation across SI rates from variation in other factors not directly related to scalar alternatives, for instance scales' propensity for negative strengthening (Gotzner et al. 2018), or the relevance of the SIs themselves (Pankratz & van Tiel 2021). For our purpose, the first group of explanatory factors is most relevant. In particular, if we find that variation akin to unembedded scalar diversity arises across lexical scales in their likelihood of leading to embedded SI, and we further find that the same properties of alternatives explain

variation at both levels, that constitutes an argument in favor of theories of embedded SI that rely on alternatives.

This paper uses two previously identified alternative-based predictors of scalar diversity as probes, both related to the distinctness of scalemates. Van Tiel et al. (2016) argue that the distinctness of scalar terms (*some* vs. *all*; *warm* vs. *hot*) is relevant for SI calculation because the two terms should have been distinct enough for the speaker to have considered uttering the stronger alternative if true. If two scalemates are not sufficiently distinct, then the speaker could have been uncertain about which one to use, and their utterance of the weaker one would not necessarily lead to the calculation of SI. That is, the authors predict a positive relationship between distinctness and the likelihood of SI. They identify two components of distinctness: semantic distance and boundedness (see also Horn 1972). Van Tiel et al. (2016) measured semantic distance experimentally, by providing participants with two statements for each scale, e.g., *That is warm* and *That is hot*. Participants were then asked whether the second statement (with *hot*) was stronger than the first, which they had to rate on a 7-point Likert scale from "equally strong" to "much stronger". Results revealed that scales with larger semantic distance led to more robust SI calculation. Boundedness is based on annotation: scales are defined as bounded if the stronger alternative denotes the endpoint of the scale (e.g., <*some*, *all*>) and non-bounded is the stronger alternative is interval-denoting (e.g., <*warm*, *hot*>). Since the weaker term is always interval-denoting, van Tiel et al. argue that scalemates on bounded scales are more distinct, since the weaker and stronger term can be distinguished "on formal grounds alone" (p. 163). Their empirical results revealed that bounded scales led to significantly more SI than non-bounded ones.

As mentioned, the current paper tests whether semantic distance and boundedness, which both crucially make reference to scalar alternatives, are predictors of embedded SI rates in the same way that they predict unembedded scalar diversity.

2.3 Prior experimental work

Since the existence of strong inferences is a key piece of evidence for adjudicating among competing theoretical accounts of SI, a number of experimental studies have tackled whether they are indeed available to hearers. Evidence for the existence of strong inferences has been provided by i.a., Chemla (2009); Chemla & Spector (2011); Clifton & Dube (2010); Gotzner & Romoli (2018). Compare, however, Geurts & Pouscoulous 2009, which does not find such evidence, as well as van Tiel 2013, which shows that some experimental results can be explained instead as typicality effects. This line of work has tended to focus on the availability of embedded SI from a single scale (typically <*some*, *all*>), while work probing embedded SI across different scales remains more limited. Here, we discuss the two

studies that did look at this.

Most relevant to the goals of the present paper is Sun, Tian & Breheny (2018), who have used experimental data from embedded SI ("upper-bound excluded local enrichment" in their terminology) to adjudicate among different theoretical accounts, and argued that their results support the RSA-LU account. In their experiment, participants had to rate the naturalness of sentences such as (8) on a 7-point Likert scale from "very unnatural" to "very natural".

(8) The student is brilliant so not intelligent.

The authors' idea was that for such sentences to be deemed natural, the second scalar term (here, *intelligent*) needs to be enriched with embedded SI. That is, *intelligent* must be understood as *intelligent but not brilliant*; otherwise, the use of *so not* gives rise to a contradiction, resulting in unnaturalness. The authors tested the 43 lexical scales in this paradigm that had formed the basis of van Tiel et al.'s (2016) scalar diversity study. They found that propensity for embedded SI, i.e., naturalness ratings, varied across scales, and this variation correlated with scalar diversity. However, boundedness and semantic distance did not predict the variation in embedded SI, despite being predictors of across-scale variation in SI in unembedded cases. As mentioned, Sun et al. interpreted their results as evidence against the grammatical theory of SI, since the lack of a relationship between the distinctness of alternatives and robustness of embedded SI is "unexpected if local enrichment relies on alternatives to the same extent as global" (p. 11), as is assumed in the grammatical theory. The authors argue that the results instead favor the RSA-LU account, since that "assumes a general narrowing option for semantic interpretation as one of two routes to account for scalar enrichment, and this does not rely on alternatives" (p. 11) —hence, properties of alternatives such as semantic distance or boundedness are not expected to predict the likelihood of SI.

As argued in Section 2.1, however, a distinction should be made among different instantiations of RSA-LU accounts; in particular, Potts et al.'s (2015) neo-Gricean uncertainty model does ascribe a role to alternatives. As such, Sun et al.'s (2018) results can be taken to be most compatible with Bergen et al. 2016 or Potts et al.'s (2015) unconstrained uncertainty model. More critically, we note that Sun et al. (2018) found quite low ratings for sentences such as (8) across the board; 90% of the 43 scales tested were rated lower than 4 on a 7-point scale, and 75% were rated lower than 3. We speculate that a possible reason for this is that sentences of the form *P so not Q* are most natural in a discourse context where *Q* has been previously asserted and the speaker is correcting that assertion. For example, (8) is most felicitous in the following dialogue:

(9) A: The student is intelligent.
B: (No,) the student is brilliant so not intelligent.

Out of the blue, participants may have found Sun et al.'s stimuli unnatural, leading to overall low ratings across all lexical scales. This compression may have also obscured any effects of semantic distance and boundedness. As a more general point, it is also worth noting that in taking their results to favor the RSA-LU account over the grammatical theory, Sun et al. (2018) are arguing on the basis of a null result —which, as speculated here, could have an explanation other than alternatives not playing a role in embedded SI.

In recent work, Bleotu & Benz (to appear) tested four types of implicatures from van Tiel et al.'s 43 lexical scales in two embedding environments, namely *some* and *possible*. An example item from the <*adequate*, *good*> scale embedded under *some* is shown in (10). The experiment probed the availability of global implicatures from the first, embedding scalar term (here, *some*, (10a)), global implicatures from the second, embedded scalar term (here, *adequate*, (10b)), embedded implicatures (10c), and double implicatures where both scalar terms are enriched (10d).

(10) Mary: *Some meals are adequate.*
Would you infer from this that, according to Mary:

a. some, but not all meals are adequate?	global, 1st scale
b. no meal which is adequate is good?	global, 2nd scale
c. some meals are adequate but not good?	embedded
d. some but not all meals are adequate but not good?	double

Following van Tiel et al.'s (2016) methodology, participants were asked to provide "Yes" vs. "No" answers to the questions in (10), with a "Yes" answer indicating the calculation of the relevant implicature. For each lexical scale, all four implicature types were presented on the same screen in a randomized order. In contrast to Sun et al. (2018), Bleotu & Benz (to appear) found not only across-scale variation in the availability of embedded SI (="Yes" responses to (10c)), but also that semantic distance and boundedness were predictors of that variation. Given the different theoretical focus of their paper, however, they did not try to relate this data to accounts of embedded SI such as those outlined in Section 2.1; instead, they argue that a theory of implicature should take into account scalar distinctness, as well as the contextual availability of alternatives (based on a separate manipulation of the Question Under Discussion). Curiously, semantic distance was not found to be a predictor of global implicatures of the 2nd scale (i.e., variation in "Yes" responses to (10b)) —even though such a result would be expected as a replication of van Tiel et al. 2016, which also tested global implicatures of these terms, albeit in sentences with a single scale. This raises the possibility that the complexity of the experimental design could have impacted the empirical findings, warranting an investigation of the availability of embedded SI in a more narrowly targeted way.

3 Experiment 1

Experiment 1 was conducted to test the availability of strong inferences from a wide range of lexical scales, in order to probe whether the robustness of strong inferences varies by-scale in the same manner as global SI does, and whether the two kinds of observed variation can be explained by the same properties of alternatives.

3.1 Participant

119 native speakers of American English participated in an online experiment, administered on the PCIbex platform (Zehr & Schwarz 2018). Participants were recruited on Prolific and compensated $2. Native speaker status was established via a language background survey, where payment was not conditioned on participants' responses. One participant was excluded for being a bilingual; data from 118 participants is reported below.

3.2 Materials and Procedure

Experiment 1 adopted the methods used by Gotzner & Romoli (2018) to test the availability of strong inferences from the <*some*, *all*> scale (see also Chemla 2009). Participants were shown two sentences on every trial and had to judge to what extent the first sentence suggested the second. They indicated their judgment by picking a point on a sliding scale from 0% (definitely not) to 100% (definitely yes).

In the critical items, the first sentence always contained a weaker scalar term embedded under a universal quantifier of the form *every N*. These sentences were created by adapting van Tiel et al.'s stimuli —specifically, one of the sentence frames from their Experiment 2, for all scales except for <*few*, *none*>. *Every* was added to each sentence, and the tense was changed from present to past to avoid a generic reading; e.g., *The soup is warm* became *Every soup was warm*. The second sentence participants saw was either the potential weak inference (11a) of the first sentence, the potential strong inference (11b), or a true (11c) or false control (11d).

(11) Every soup was warm.

a. Not every soup was hot.	weak
b. No soup was hot.	strong
c. At least one soup was warm.	true
d. Not every soup was warm.	false

Condition was manipulated within-participants in a Latin Square design. Each participant saw 42 different lexical scales as critical items. The experiment started with 3 practice trials to familiarize participants with the task.

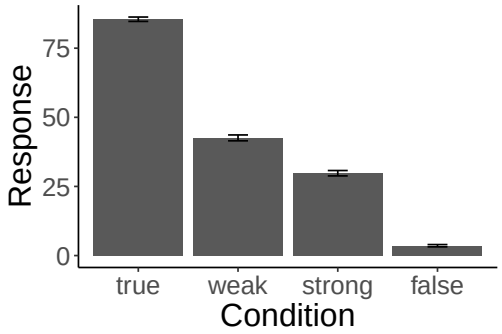

Figure 1 Mean sliding scale response by condition in Experiment 1. Error bars represent standard error.

3.3 Results

Figure 1 shows the results of Experiment 1 averaged over different lexical scales. The statistical analysis of these aggregate results followed that of Gotzner & Romoli 2018. A linear mixed effects regression model was fit using the lme4 package (Bates, Mächler, Bolker & Walker 2015) in R. The model predicted Response (0-100 on the sliding scale) by Condition (true vs. weak vs. strong vs. false). The Condition predictor was treatment coded, with "strong" set as the reference level. The maximal converging random effects structure included by-participant random slopes and intercepts and by-item random intercepts. The analysis revealed that the "true" control condition produced significantly higher Responses than the "strong" condition (Estimate=55.6, Std. Error: 2.75, t=20.19, p<0.001). The "weak" condition was also significantly higher than the "strong" condition (Estimate=12.79, Std. Error: 1.93, t=6.62, p<0.001), while the "false" condition produced significantly lower responses (Estimate=−26.12, Std. Error: 1.47, t=−17.81, p<0.001).

For the next set of analyses, the following data were taken from van Tiel et al. (2016: Table 3) for the 42 scales also tested in the current experiment: SI rates (from their Experiment 2), semantic distance (from their Experiment 4), and boundedness (manually annotated by the authors). Figure 2 shows the by-scale correlation between van Tiel et al.'s SI rates and sliding scale responses in the strong inference condition; a statistical test confirms the existence of a strong positive correlation (Pearson's correlation test: r=0.76, p<0.001). Figure 3 shows the correlation between semantic distance and the strong inference responses. For analyzing the effect of semantic distance, we fit a linear mixed effects regression model to the "strong" data, predicting Response (0-100) by Distance, with by-participant and by-item random intercepts. The model revealed a significant effect such that larger semantic distance leads to higher responses (Estimate=7.28, Std. Error: 3.29, t=2.21, p<0.05).

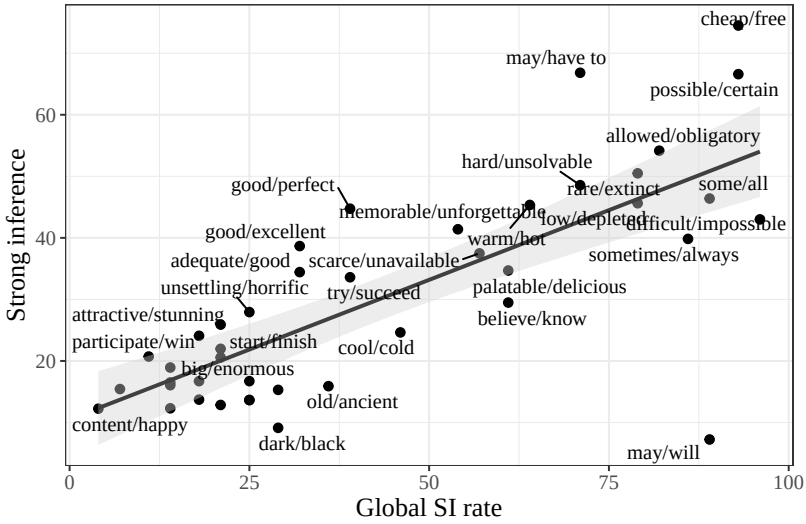

Figure 2 By-scale correlation between van Tiel et al.'s SI rates (*x* axis) and the "strong" condition of Experiment 1 (*y* axis).

Figure 4 shows responses in the strong condition for bounded vs. non-bounded scales. To analyze the effect of boundedness, we again fit a linear mixed effects regression model predicting Response (0-100) by Boundedness in the strong condi-

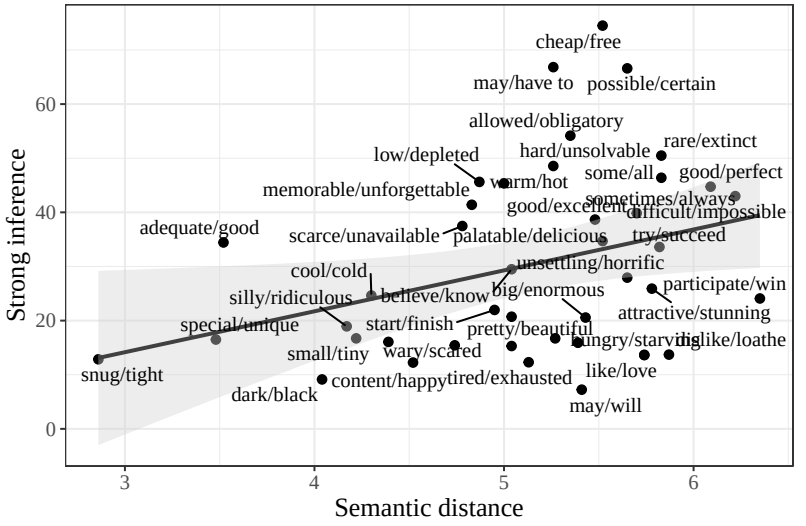

Figure 3 By-scale correlation between semantic distance (*x* axis) and the "strong" condition of Experiment 1 (*y* axis)

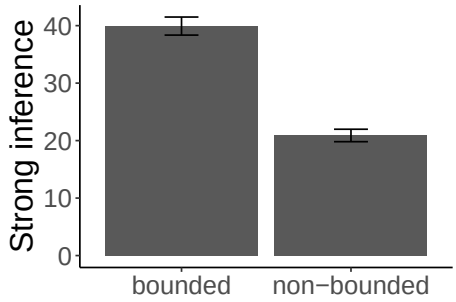

Figure 4 Mean strong inference calculation (i.e., sliding scale response) and standard error in Experiment 1 in bounded vs. non-bounded scales.

tion. Boundedness was scaled sum coded (−0.5: non-bounded; 0.5: bounded), and the random effects structure included by-participant and by-item random intercepts. The analysis revealed significantly higher responses with bounded scales than with unbounded ones (Estimate=18.37, Std. Error: 4.41, t=4.17, p<0.001).

All the aforementioned results are replicated in the weak inference data, namely that we find a positive by-scale correlation with the global SI rates from van Tiel et al. 2016 (Pearson's correlation test: r=0.69, p<0.001), and that greater semantic distance (Estimate=9.11, Std. Error: 2.86, t=3.18, p<0.01) and bounded scales (Estimate=19.41, Std. Error: 3.79, t=5.13, p<0.001) both lead to more robust calculation of the weak inference. However, these results are less theoretically informative, since weak inferences can definitely be derived in the same way as global SI.

3.4 Discussion

The overall pattern obtained in Experiment 1 (Figure 1) closely mirrors that of Gotzner & Romoli 2018, with true>weak>strong>false forming a cline. One difference is that both the weak and the strong inference arose at slightly lower rates in the current experiment (i.e., they received on average lower responses on the sliding scale) than in Gotzner & Romoli 2018. This is expected, since Gotzner & Romoli only tested inferences arising from *some*, which triggers SI more strongly than most other lexical scales; in van Tiel et al. 2016, the *some but not all* (global) SI was calculated at an 89% rate, which only three of the other 42 scales exceeded. Since the current experiment tested 42 different scales, most of which are known to lead to less SI calculation than <*some*, *all*>, it is no surprise that the strong and weak inferences arose less robustly than in Gotzner & Romoli 2018.

Following Gotzner & Romoli (2018), we interpret the strong condition's significant difference from the false control as evidence that strong inferences are

derived by comprehenders and can be detected experimentally. This is in line with findings by i.a., Chemla (2009); Chemla & Spector (2011); Clifton & Dube (2010); Gotzner & Romoli (2018), but contra those by Geurts & Pouscoulous (2009). We further observe that the weak inference is more robust than the strong inference, that is, it received higher sliding scale responses. This can be explained by assuming that sliding scale responses are proportional to the number of readings that make a sentence true (Gotzner & Romoli 2018; Chemla & Spector 2011). The weak inference follows from both embedded and global SI, while the strong inference arises only as embedded SI. To give an example, *No soup was hot* entails *Not every soup was hot*. Hence, if a participant has calculated the strong inference from *Every soup was warm*, they would judge the weak inference to follow too, but without having calculated the global SI. The reader is referred to Gotzner & Romoli 2018; Chemla & Spector 2011 for details of this argument, as well as converging empirical findings.

One may wonder why the sliding scale responses to the true control condition are not at ceiling (85%)[4]. In other words, why is it that participants did not all judge *At least one soup was warm* to be definitely suggested by *Every soup was warm*? There could be multiple possible explanations for this. First, in a context in which there are no soups, *Every soup was warm* could be judged to be vacuously true, whereas *At least one soup was warm* is false. If a participant thought of such a context, that may have lowered their response. Second, it is also conceivable that naive participants imagine a context with multiple soups, e.g., ten, all of which are warm, and then judge *At least one soup was warm* to be too underinformative. Though in this case, *Every soup was warm* still entails *At least one soup was warm*, participants' judgements may have been lowered due to informativity considerations. Third, the way the experimental task was worded could have also contributed to this possibility. If participants interpreted the *P suggests that Q* wording of the task to mean *P could be used to communicate that Q*, or *P means the same as Q*, they may have been looking for paraphrases of *P*, not logical entailments of it.

Going beyond the overall patterns averaged over all lexical scales tested, we found a by-scale correlation between global SI and strong inferences. This means that when a scale is relatively likely to lead to the calculation of global SI, it is also relatively likely to lead to the calculation of the strong inference. In other words, the likelihood of hearers calculating *The soup is not hot* from *The soup is warm* is correlated with the likelihood of them calculating *No soup was hot* from *Every soup*

4 In an additional analysis, we removed data from participants whose responses on the true and false controls were more than 1 standard deviation away from the mean: 12 participants gave on average a <57.79 response to true controls and 2 additional participants gave >18.06 to false controls. Analyzing the remaining 104 participants' data raises the mean true control response to 92%. Importantly, however, the results and conclusions reported in the main analysis of Experiment 1 do not change.

was warm. This finding qualitatively replicates Sun et al. 2018, but finds a much stronger relationship between the two inference types (r=0.76 here vs. r=0.44 in their study). Crucially, and diverging from Sun et al. (2018) (but in line with Bleotu & Benz to appear), we found that two well-established predictors of scalar diversity, semantic distance and boundedness, predicted the variation in strong inferences as well. Fully parallel to findings reported by van Tiel et al. (2016) for global SI (and replicated by, i.a., Gotzner et al. 2018; Sun et al. 2018; Pankratz & van Tiel 2021), we found that the robustness of strong inferences increases with semantic distance, and that bounded scales lead to more strong inference than non-bounded ones.

Importantly from a theoretical viewpoint, these two predictors of (global and embedded) scalar diversity both make reference to stronger alternatives. Boundedness is directly defined as a property of alternatives —whether or not they denote the endpoint of a scale —, while semantic distance is a relation between the weaker scalars and their stronger alternatives, so it also necessarily makes reference to alternatives. Our finding that these properties of alternatives play a role in a scale's likelihood of leading to a strong inference constitutes an argument in favor of theoretical accounts of embedded SI that build on alternatives, such as the grammatical theory (Chierchia 2004; Chierchia et al. 2012), a modified neo-Gricean account like Sauerland (2004), or the neo-Gricean uncertainty model of Potts et al. (2015). On the other hand, a treatment of embedded SI that does not make reference to alternatives, such as the unconstrained uncertainty model of Potts et al. (2015) or Bergen et al.'s (2016) RSA-LU model, would struggle to explain why the same alternative-based across-scale variation arises both in global SIs and strong inferences, when only the former is derived via an alternative-based mechanism. Of course, these accounts could always be supplemented with further assumptions to capture the present data.

More generally, in addition to providing empirical evidence favoring alternative-based treatments of embedded SI, the present study provides novel experimental evidence from scalar diversity for the existence of a single shared mechanism underlying embedded and global SI. Both the finding that the same variation arises in global SI and strong inferences and that the variation can be explained by the same predictors in both cases support such a conclusion. This conclusion is also in line with i.a., Sun & Breheny's (2019) priming study, which showed that the two kinds of SI —global and embedded —can prime each other, which has also been argued to offer support for a shared mechanism.

4 Experiment 2

Experiment 2 was conducted to address a shortcoming of Experiment 1, having to do with using the false condition as baseline. As Gotzner & Romoli (2018), we defined the existence of strong inferences as the strong condition receiving higher sliding

scale responses than the false condition. However, as Gotzner & Romoli (2018) also note, it is possible that the comparison between the strong and false conditions does not adequately allow us to assess the availability of strong inferences. This is because the sentence that participants have to judge in the false condition (e.g., *Not every soup was warm*) is not simply not an inference of the first statement (*Every soup was warm*), but is in fact incompatible with that statement. The strong condition, on the other hand, tests sentences that are compatible with the first statement (*No soup was hot*). Thus, the strong condition might receive higher responses not because it is actually an inference arising from sentences like *Every soup was warm*, but simply because it is compatible with such a sentence —unlike the false condition.

In order to circumvent this problem, Gotzner & Romoli (2018) conducted a second experiment, in which they replaced the false control baseline with a so-called compatible baseline —that is, a sentence that is not an inference of the first statement but is nonetheless compatible with it. For example, participants were asked to judge to what extent *Every student read some of the books* suggested that *Some student read all of the books*; given our running example based on the <*warm*, *hot*> scale, the equivalent compatible control condition would be *Some soup was hot*, given the assertion *Every soup was warm* and the strong inference of interest *No soup was hot*. Gotzner & Romoli (2018) reason that since now both the baseline compatible control condition and the strong condition are compatible with the first sentence, but only the latter is predicted to be an inference by (some) theoretical accounts, higher responses to the strong inference condition can be taken to index its existence —and this is indeed what the authors' findings revealed.

One important property of the compatible control condition, though, is that it is the negation of the target strong inference. This is acknowledged by Gotzner & Romoli (2018) and treated as a benefit; they argue that this ensures that the baseline and the strong conditions are equally complex and relevant. However, one can also think of shortcomings of this experimental design. Since the baseline condition is the negation of the target strong inference, its truth value directly follows from the truth value of that inference. Or looking at it another way, once a participant has calculated the strong inference, the compatible baseline condition becomes incompatible, even though it was not incompatible with the first sentence itself.

For this reason, Experiment 2 takes a different route to the question of how to define the existence of strong inferences. We use the inference task (Geurts & Pouscoulous 2009; van Tiel et al. 2016), where participants see one statement (e.g., *Mary read some of the books*) followed by a question like *Would you conclude from this that Mary didn't read all of the books?*. Participants have to answer with "Yes" (indexing inference calculation) or "No". Using this method, a baseline is no longer strictly necessary. With 0-100 sliding scale responses, it is indeed an important question what responses we can take to suggest that an inference is available. But

with the binary inference task, every time a "Yes" response is given, we can assume that an inference has been calculated, allowing us to sidestep the thorny question of establishing a baseline rating for inference calculation on a 0-100 scale.

4.1 Participants

45 native speakers of American English participated in an online experiment on PCIbex for $1.60 compensation. Participant recruitment and screening was identical to Experiment 1. Data from all participants is reported below.

4.2 Materials and Procedure

Experiment 2 used the inference task to investigate the rates of calculating strong inferences from 42 lexical scales. The materials from the strong condition of Experiment 1 were adapted such that each of the potentially inference-triggering sentences was uttered by a speaker, Mary. The potential inference was embedded in the task question *Would you conclude from this that, according to Mary, <inference>?*. The example in (12) shows the item using the <*warm*, *hot*> scale.

(12) Mary: Every soup was warm.
Would you conclude from this that, according to Mary, no soup was hot?

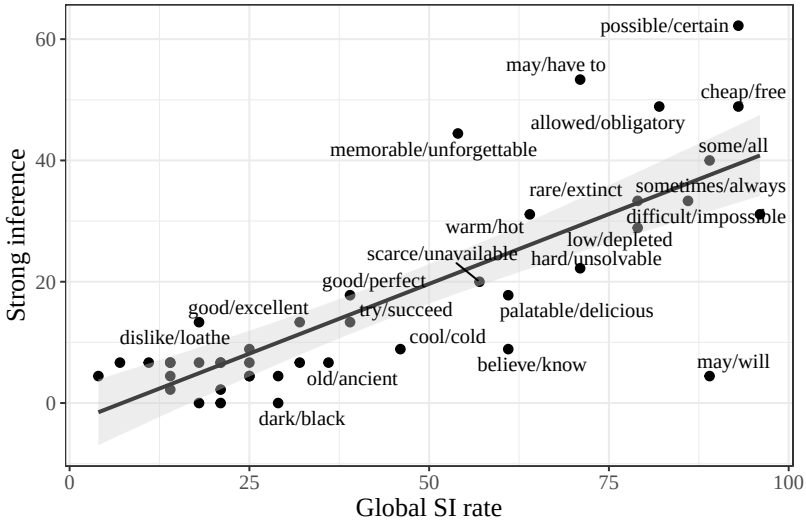

Figure 5 By-scale correlation between van Tiel et al.'s SI rates (*x* axis) and % of strong inference calculation in Experiment 2 (*y* axis).

Participants answered the task questions with either "Yes" (indexing inference calculation) or "No" (suggesting no inference calculation). All participants saw the 42 critical trials, 3 practice trials, as well as 7 catch trials which contained antonyms and hence had an unambiguous "Yes" answer (e.g., *Every street was wide. —Would you conclude from this that, according to Mary, no street was narrow?*).

4.3 Results

Similarly to Experiment 1, as Figure 5 shows, we find a strong positive by-scale correlation between van Tiel et al.'s SI rates and the rate of strong inference calculation (i.e., percentage of "Yes" responses); this relationship is statistically significant (Pearson's correlation test: r=0.8, p<0.001). Next, we fit a logistic mixed effects regression model (lme4) predicting Response (Yes vs. No) by Distance, and found that the rate of strong inference calculation increases with semantic distance (Estimate=0.63, Std. Error: 0.31, z=2.05, p<0.05) —see Figure 6. To check the effect of boundedness, we again fit a logistic mixed effects regression model predicting Response by Boundedness, and found that bounded scales lead to more strong inferences (Estimate=1.54, Std. Error: 0.39, z=3.91, p<0.001) —see Figure 7. (Coding of variables and random effects structures were identical to Experiment 1.)

Experiment 2 thus replicates the crucial finding that properties of alternatives that impact the likelihood of global SI calculation also predict the likelihood of strong

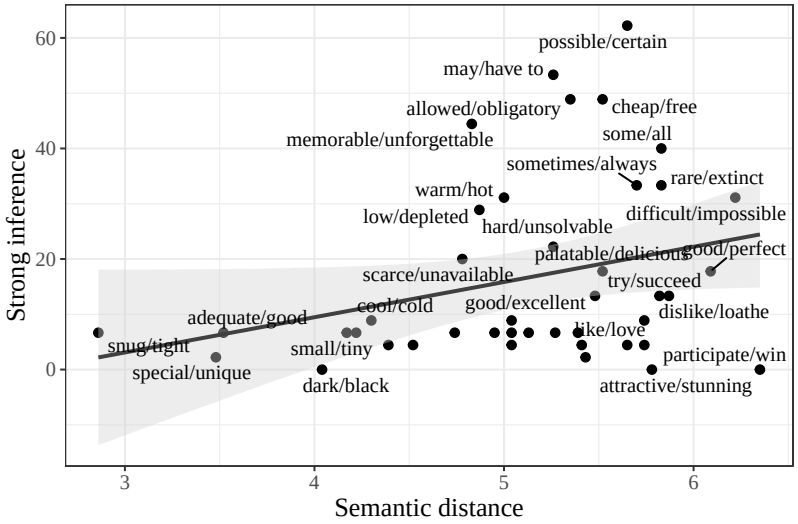

Figure 6 By-scale correlation between semantic distance (x axis) and % of strong inference calculation in Experiment 2 (y axis).

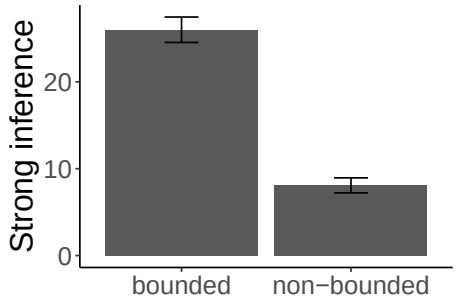

Figure 7 Mean strong inference calculation (=% of "Yes" responses) and standard error in Experiment 2 in bounded vs. non-bounded scales.

inference calculation. This constitutes evidence for alternative-based accounts of embedded SI, now with the concerns about the experimental baseline ruled out.

5 Conclusion

This paper reported on two experiments investigating the availability of embedded SI from a large set of lexical scales, i.e., whether sentences like *Every soup was warm* give rise to the inference that *No soup was hot*. Findings revealed that hearers do calculate such inferences, which successfully extends prior experimental findings which had demonstrated the existence of embedded SI largely on the <*some*, *all*> scale. This cannot be accommodated under an "unmodified" neo-Gricean account of SI, where the only alternative to *Every soup was warm* is *Every soup was hot*, supporting instead accounts that assume additional alternatives, or an entirely different mechanism for SI. The likelihood of embedded SI was found to vary across lexical scales, e.g., an inference of *No soup was hot* was more likely from *Every soup was warm* than *No movie was excellent* was from *Every movie was good*. This variation was strongly correlated with across-scale variation observed in unembedded SI calculation by van Tiel et al. (2016), i.e., scalar diversity. This points to the existence of a shared mechanism underlying the two types of inferences.

Most importantly from a theoretical viewpoint, two properties related to scalar alternatives—semantic distance between scalar terms, and the boundedness of scales—were found to predict variation in embedded SI, paralleling van Tiel et al.'s findings for unembedded SI. This is unexpected under theoretical treatments of embedded SI that do not make reference to alternatives and instead derive these enrichments via unconstrained strengthening of lexical meanings —namely, the RSA-LU account as proposed by Bergen et al. (2016) and Potts et al.'s (2015) unconstrained uncertainty model. Instead, this finding constitutes an argument in

favor of accounts that build in a role for scalar alternatives, such as the grammatical theory (Chierchia 2004; Chierchia et al. 2012), a modified neo-Gricean account like Sauerland (2004), or the neo-Gricean uncertainty RSA-LU model (Potts et al. 2015).

References

Baker, Rachel, Ryan Doran, Yaron McNabb, Meredith Larson & Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1(2). 211–248. doi:10.1163/187730909x12538045489854.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Beltrama, Andrea & Ming Xiang. 2013. Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. In Emmanuel Chemla, Vincent Homer & Grégoire Winterstein (eds.), *Sinn und Bedeutung 17*, 81–98.

Bergen, Leon, Roger Levy & Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9. doi:10.3765/sp.9.20. http://dx.doi.org/10.3765/sp.9.20.

Bleotu, Adina Camelia & Anton Benz. to appear. The role of scalar diversity and question under discussion in deriving implicatures with embedded scales. In *Sinn und Bedeutung 28*, .

Chemla, E. & B. Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28(3). 359–400. doi:10.1093/jos/ffq023. http://dx.doi.org/10.1093/jos/ffq023.

Chemla, Emmanuel. 2009. Universal implicatures and free choice effects: Experimental data. *Semantics and Pragmatics* 2. doi:10.3765/sp.2.2. http://dx.doi.org/10.3765/sp.2.2.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In Adriana Belletti (ed.), *Structures and Beyond*, 39–103. Oxford University Press.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. *87. Scalar implicature as a grammatical phenomenon* 2297–2332. Berlin, Boston: De Gruyter Mouton. doi:doi:10.1515/9783110253382.2297. https://doi.org/10.1515/9783110253382.2297.

Clifton, Charles Jr & Chad Dube. 2010. Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics* 3(7). 1–13. doi:10.3765/sp.3.7.

Doran, Ryan, Gregory Ward, Meredith Larson, Yaron McNabb & Rachel E. Baker. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language* 88(1). 124–154.

Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998. doi:10.1126/science.1218633.

Geurts, Bart & Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and Pragmatics* 2(4). 1–34. doi:10.3765/sp.2.4.

Gotzner, Nicole & Jacopo Romoli. 2018. The Scalar Inferences of Strong Scalar Terms under Negative Quantifiers and Constraints on the Theory of Alternatives. *Journal of Semantics* 35(1). 95–126. doi:10.1093/jos/ffx016. https://doi.org/10.1093/jos/ffx016.

Gotzner, Nicole & Jacopo Romoli. 2022. Meaning and alternatives. *Annual Review of Linguistics* 8(1). 213–234. doi:10.1146/annurev-linguistics-031220-012013.

Gotzner, Nicole, Stephanie Solt & Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. 1659.

Grice, Herbert Paul. 1967. Logic and Conversation. In Paul Grice (ed.), *Studies in the Way of Words*, 41–58. Harvard University Press.

Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*: UCLA PhD dissertation.

Hu, Jennifer, Roger Levy, Judith Degen & Sebastian Schuster. 2023. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *Association for Computational Linguistics* 11. 885–901. doi:10.1162/tacl_a_00579.

Hu, Jennifer, Roger Levy & Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. *Workshop on Cognitive Modeling and Computational Linguistics* 68–74.

Pankratz, Elizabeth & Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition* 13(4). 562–594. doi:10.1017/langcog.2021.13.

Potts, Christopher, Daniel Lassiter, Roger Levy & Michael C. Frank. 2015. Embedded Implicatures as Pragmatic Inferences under Compositional Lexical Uncertainty. *Journal of Semantics* 33(4). 755–802. doi:10.1093/jos/ffv012. https://doi.org/10.1093/jos/ffv012.

Ronai, Eszter & Ming Xiang. 2022. Three factors in explaining scalar diversity. *Sinn und Bedeutung* 26. 716–733.

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391. doi:10.1023/B:LING.0000023378.71748.db.

Sun, Chao & Richard Breheny. 2019. Shared mechanism underlying unembedded and embedded enrichments: Evidence from enrichment priming. *Sinn und Bedeutung* 22(2). 425–441. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/115.

Sun, Chao, Ye Tian & Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology* 9.

van Tiel, Bob. 2013. Embedded Scalars and Typicality. *Journal of Semantics* 31(2). 147–177. doi:10.1093/jos/fft002. https://doi.org/10.1093/jos/fft002.

van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar

diversity. *Journal of Semantics* 33(1). 137–175. doi:10.1093/jos/ffu017.
van Tiel, Bob & Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive Science* 41. 1119–1154.
Westera, Matthijs & Gemma Boleda. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Sinn und Bedeutung* 24(2). 439–454. doi:10.18148/sub/2020.v24i2.908.
Zehr, Jeremy & Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832.

Eszter Ronai
2016 Sheridan Rd
Room 205
Evanston, IL 60208
ronai@northwestern.edu