GLOSSA psycholinguistics

Ronai, E., & Göbel, A. (2024). Watch your tune! On the role of intonation for scalar diversity. *Glossa Psycholinguistics*, 3(1): 26, pp. 1–34. DOI: https://doi.org/10.5070/G60114911

eScholarship University of California

Watch your tune! On the role of intonation for scalar diversity

Eszter Ronai, Northwestern University, US, ronai@northwestern.edu

Alexander Göbel, University of Manchester, UK, alexander.gobel@manchester.ac.uk

Recent research has highlighted that lexical scales vary in their likelihood of giving rise to a scalar inference – a finding labeled scalar diversity. The current paper examines the role of intonation for this phenomenon, which has thus far primarily been studied using written materials. A specific focus in this regard was on the so-called rise-fall-rise contour, which has been argued to (i) convey uncertainty, which could have an influence on scalar inference calculation, and (ii) be sensitive to properties of lexical scales, which could interact with factors driving scalar diversity. Experiment 1 combined production with an inference task to assess the likelihood of different intonational contours, as well as how a given contour affects scalar inference rates. Production of the rise-fall-rise varied across lexical scales, as expected, and led to an increase in scalar inference derivation relative to a fall. The latter finding was further confirmed in Experiment 2, which explicitly manipulated intonational contours in the inference task. The results, thus, show the importance of taking intonation into account when studying scalar diversity and scalar inference more generally, and they also have implications for theories of the rise-fall-rise contour. Additionally, the experiments revealed a contour that is prosodically similar to the so-called Contradiction Contour, but appears to serve a different pragmatic function.

1. Introduction

The investigation of *scalar inferences* (SIs), such as strengthening *some* to *some but not all*, constitutes a well-established testing ground for our understanding of pragmatic reasoning. Due to its ubiquity and tractability, SI is the phenomenon most commonly and comprehensively treated in competing theories of pragmatic mechanisms; these mechanisms, in turn, have been subjected to extensive experimental testing, probing SI's status at the interface of grammar, semantics and pragmatics (Cummins & Katsos, 2019). A recent line of research in this domain has focused on how findings about SIs generalize to other scalar terms beyond the stereotypical cases of *some* and *or* (i.a., Gotzner et al., 2018; Sun et al., 2018; van Tiel et al., 2016). This research has found that lexical scales vary greatly in their potential to give rise to SIs, covering almost the entire spectrum – a finding referred to as *scalar diversity*. A notable commonality among experimental studies conducted in this domain is that they rely on written stimuli: participants read sentences containing scalar terms silently to themselves before providing their response indicating whether they derived an SI. Crucially, there is a considerable amount of evidence supporting the idea that prosodic structure is assigned even during silent reading (Bader, 1998; Fodor, 2002; see also Frazier & Gibson, 2015). As a result, participants are, in principle, able to project whatever intonation they choose onto the stimuli, which may affect their final response.[1]

This issue becomes of particular importance in light of research on the meaning of intonational contours. Specifically, the so-called rise-fall-rise contour (RFR; Ward & Hirschberg, 1985) is often discussed in relation to examples of non-maximal scale items, such as those used in scalar diversity studies (e.g., Constant, 2012), as the contour has been claimed to be infelicitous otherwise (Göbel & Wagner, 2023). This makes it a likely contour choice for these types of contexts. Additionally, different accounts of the meaning of the RFR predict an effect on SI calculation, although the precise nature of this effect depends on the account in question. Finally, there may be additional restrictions on the use of the RFR with respect to properties of the relevant scales (see Göbel & Wagner, 2023). Specifically, the felicity of the contour may vary across the range of scalar terms that are typically used in research on scalar diversity. As a result, participants in scalar diversity studies may be more likely to produce an RFR for certain scales, which may, in turn, affect their likelihood of deriving an SI. Crucially, in this scenario, the properties of lexical scales would only play an indirect role, being mediated by intonation, rather than affecting SI rate directly.

Here we present two experiments investigating this issue in more detail. Experiment 1 uses a combination of a production and an inference task to assess both how likely participants are to produce a certain contour in a given context and how the choice of contour affects the likelihood of drawing an SI. Experiment 2 presents participants with a given contour directly and, again,

[1] We follow Ladd (2008) in defining *intonation* as suprasegmental phonetic features ranging over sentences in a linguistically structured way.

assesses the likelihood of an SI. The results show that the production rate of the RFR varies strongly across lexical scales and that its use – both in production and perception – leads to an increase in SI rate. Thus, the experiments provide strong evidence for the relevance of intonation for the study of SIs generally (in line with, i.a., Gotzner, 2019; Tomlinson et al., 2017) and scalar diversity specifically. Additionally, the findings bear on accounts of the RFR as well as another contour revealed in the production study, which we refer to as the Concession Contour.

The rest of this article is structured as follows. We first provide background on prior research on scalar diversity (2.1), the role of intonation for SI (2.2), accounts of the RFR (2.3), and existing studies of the RFR-SI relationship (2.4). Section 3 details Experiment 1 (production + inference task), and Section 4 details Experiment 2 (perception + inference task). Section 5 offers discussion of our findings in light of the literature on intonational contours and SI. Section 6 concludes.

2. Background

2.1 Scalar inference and scalar diversity

SI represents one of the classic examples of pragmatic enrichment. An utterance containing the quantifier *some* (1), for example, is often enriched from its lower-bounded meaning (1a) to the upper-bounded meaning *some but not all* (1b).

(1) Miriam caught some of the mice.
   a. Miriam caught at least some of the mice. literal
   b. Miriam caught some, but not all, of the mice. SI-enriched

While there are many different theoretical proposals as to how SIs arise, a standard (neo-) Gricean account posits the following. Hearers assume that speakers are following the Maxim of Quantity (Grice, 1967), and are therefore trying to be as informative as is required in the context. A more informative alternative utterance to (1) would have been *Miriam caught all of the mice*. *Informativity* can be defined as asymmetric entailment: *Miriam caught all of the mice* entails *Miriam caught some of the mice*, but not vice versa, hence, the former is more informative (Horn, 1972). Therefore, when a comprehender encounters an utterance like (1), they reason about the speaker's intention behind not uttering the more informative, stronger alternative statement. This may have happened because the stronger alternative is false, and the speaker chose not to utter it in order to avoid violating the Maxim of Quality. This reasoning process leads hearers to derive the negation of the unsaid alternative (*Miriam didn't catch all of the mice*) which, combined with the original utterance's literal meaning (1a), results in the SI-enriched meaning (1b).

While the *some but not all* SI, based on the <*some, all*> lexical scale, is the most widely discussed example, SI can also arise from other pairs of lexical items that form a scale. The example in (2), for instance, is based on the <*happy*, *ecstatic*> scale.

(2) The winner is happy.
   a. The winner is at least happy. literal
   b. The winner is happy, but not ecstatic. SI-enriched

Given (neo-)Gricean assumptions, the *happy but not ecstatic* SI can be derived in the same way as *some but not all*. Hearers of the weaker utterance in (2) can reason that the speaker did not utter the more informative alternative *The winner is ecstatic*, because it would not have been true. The weaker utterance's literal meaning (2a) and the negation of the stronger alternative, then, together give rise to the SI-enriched meaning (2b). But while the mechanism underlying these two different SIs is posited to be the same, they do not arise equally robustly: hearers are much more likely to enrich *some* to mean *not all* than *happy* to mean *not ecstatic* (Ronai & Xiang, 2024). In fact, *scalar diversity* is now a well-replicated finding. This term refers to the substantial variation across different lexical scales in the likelihood that they would lead to SI. In van Tiel et al.'s (2016) highly influential study, for instance, the rate at which participants calculated SIs ranged from 4% to 100%, with the 43 scales tested spanning that full range (see also earlier work by Baker et al., 2009; Beltrama & Xiang, 2013; Doran et al., 2012).

Existing experimental studies of scalar diversity have concentrated on answering the question of what can explain the observed inter-scale variation in SI calculation. How likely a scale is to lead to SI has been related to various properties of the stronger alternative (e.g., *all*, *ecstatic*), or of the relationship between the weaker scalar term (e.g., *some*, *happy*) and that alternative. For example, van Tiel et al. (2016) have shown that the more distinct the weaker and the stronger term are, the more likely they are to lead to SI. This is because the stronger alternative needs to be sufficiently distinct from the weaker term for SI to arise; if the two terms are not distinct enough, the speaker's non-utterance of the stronger term is not necessarily due to its falsity. Here, distinctness was operationalized as semantic distance (as measured in a rating task) and boundedness (whether the stronger alternative is endpoint-denoting). Westera and Boleda (2020) have proposed that semantic relatedness, based on distributional semantics, is another component of distinctness, which they indeed found to be negatively correlated with SI rates. A property of stronger alternatives that has been shown to predict scalar diversity is how expected they are or, in other words, how (un)certain hearers are about the identity of the relevant stronger alternative, given the weaker term uttered – the greater the uncertainty, the less likely SI is to arise (Hu et al., 2022, 2023; see also Ronai & Xiang, 2022). Concentrating on SIs arising from adjectival scales in particular, Gotzner et al. (2018) have related scalar diversity to the underlying scalar semantics of adjectives. Polarity was one relevant predictor, with the authors' results revealing higher SI rates for negative adjectives (e.g., $<$*bad, awful*$>$) than positive ones (e.g., $<$*good, great*$>$); see also Pankratz and van Tiel (2021) for a replication using different diagnostics for polarity. Another factor from adjectival semantics is extremeness:

extreme adjectives (e.g., *excellent*, *huge*) have been shown to lead to lower SI rates (Beltrama & Xiang, 2013; Gotzner et al., 2018). Aside from deriving across-scale variation from properties of the weaker term and its stronger alternative, studies have also suggested that the propensity for SI is linked to another type of semantic process or pragmatic inference that is variable across scales (Gotzner et al., 2018; Sun et al., 2018). Last but not least, the role of context and contextual relevance in explaining scalar diversity has also been investigated, focusing either on discourse or on sentential context (Pankratz & van Tiel, 2021; Ronai & Xiang, 2021a; Simons & Warren, 2018; Sun et al., 2023).

One limitation of this existing body of work that we would like to highlight is that all prior studies of scalar diversity have used exclusively written experimental stimuli, or modeled data from other studies that have done so. This presents a potential issue in light of the fact that – as we will review below in 2.2 – intonation is known to affect SI calculation. Most crucially for our purposes, certain intonational contours are also sensitive to the same factors that have been identified as predicting scalar diversity. As mentioned in Section 1 and further discussed in 2.3, one contour of interest is the RFR, which is predicted to affect the likelihood of drawing an SI. Additionally, the RFR has been argued to be felicitous with positive, but not negative, statements (in negative and positive contexts, respectively; see Göbel, 2019; Göbel & Wagner, 2023). These two factors could, then, conspire to give the appearance that adjective polarity affects SI rates directly. As mentioned, negative scales have been found to lead to SI more robustly than positive ones (Gotzner et al., 2018). Such a finding, however, could in principle be an epiphenomenon arising from the RFR decreasing SI rates and negative adjectives being less likely to be silently read with an RFR. Notably, adjective polarity may be just one factor with the potential to conspire in this way, given that many aspects that constrain the use of the RFR are not yet fully understood. As a result, there is reason to believe that using auditory stimuli, carefully controlling and manipulating the intonation with which SI-triggering utterances are produced, could uncover interesting patterns that written studies on scalar diversity have obscured.

2.2 The role of intonation for SI

As mentioned, there are robust findings in the literature showing that intonation affects how likely SI is to arise for intonation languages like English, French, Dutch and German.[2] We start by reviewing work that has examined the effect of pitch accent placement. Schwarz et al. (2007) investigated SI arising from the $<$*or*, *and*$>$ lexical scale, via sentences such as (3). They varied

[2] Notably, while all languages featured here can be argued to have broadly similar functions for pitch accents – the part of the overall contour perceived as prominent and often marked by a change in pitch – their intonational systems also differ slightly. The discussion should, therefore, not be taken to imply that the phonetic-phonological details of pitch accents are identical across these languages.

the placement of the L+H* accent,[3] which was assumed to mark prosodic focus; it occurred either on the disjunction (3a) or the auxiliary (3b).

(3) a. Mary will invite Fred OR Sam to the barbecue.
b. Mary WILL invite Fred or Sam to the barbecue.

Having been presented with one of the above sentences, participants had to choose between two alternative interpretations: the literal meaning *She will invite Fred or Sam or possibly both* and the SI-enriched meaning *She will invite Fred or Sam but not both*. The authors found a higher rate of SI-enriched, exclusive *not both* interpretations when the L+H* accent was placed on *or* (3a). Using a truth value judgement task, Chevallier et al. (2008) found converging results for French, namely, that prosodic stress on *or* leads to an increase in the exclusive *not both* interpretation. Lastly, Zondervan (2010) conducted a similar manipulation in Dutch, contrasting sentences like those in (4). In (4a), the entire NP containing the disjunction received two H* accents (one on each disjunct), whereas in (4b), the subject received one H* accent.

(4) a. Paola took AN APPLE OR A PEAR from the fruit section.
b. PAOLA took an apple or a pear from the fruit section.

Participants were asked to judge the target sentence – (4a) or (4b) – as true or false, in the context of a story that made it clear that Paola had, in fact, taken both an apple and a pear. If a hearer has calculated the *not both* SI from the disjunction, they would, therefore, judge the target sentence to be false. In line with the other studies discussed, Zondervan (2010) found a significant effect such that more SIs were calculated when *or* was in the accented part of the sentence (4a).

There also exists work manipulating not the placement of the accent, but its type. Several studies in this domain have focused on ad hoc scales (Hirschberg, 1985) giving rise to exhaustive inferences. Gotzner (2019), for instance, tested sentences like (5) in German.

(5) a. Context: The judge and witness followed the argument.
b. Critical sentence: The {judge/JUDGE} believed the defendant.
c. Alternative statement: The witness believed the defendant.

[3] The label for this accent type is part of the widely adopted ToBI annotation system (Beckman et al., 2005), derived from the autosegmental-metrical (AM) theory of intonation (Pierrehumbert, 1980). On this approach, a sentence-level contour consists of a sequence of low (L) and high (H) pitch targets of different accent types (pitch accent, phrasal accent, boundary accent/tone), with the '*' indicating prominence of pitch accents. In the literature presented here and adjacent to it, the L+H* accent is often taken to convey contrastive Focus, but terminology is not always defined and phonetic details may vary or are often missing. For the purposes of this paper, we define *focus* as a semantic-pragmatic correlate of (at least some) pitch accents that evokes alternatives (Krifka, 2008; Rooth, 1992).

Participants were presented with the context sentence, followed by the critical sentence, and then had to make a truth value judgment on the alternative statement. If a participant has calculated the ad hoc inference *The judge, but not the witness, believed the defendant*, then they would judge the alternative statement to be false. Crucially, Gotzner (2019) manipulated the intonation of the critical sentence, which occurred either with an L+H* or an H* accent on the target word (*judge*). The findings revealed that participants computed more exhaustive inferences with an L+H* accent, as indicated by a lower % of True responses.

Using mouse-tracking to investigate online processing, Tomlinson et al. (2017) also compared what effect an L+H* vs. H* pitch accent has on ad hoc SIs in English, and found that the inference is processed earlier under the former contour. Tomlinson and Ronderos (2021), in turn, compared the effect of the L+H* and L*+H contours on the exhaustive interpretation arising from dialogues such as (6).

(6) A: Were Manu and Moni at the party?
B: Manu was there.

The authors were interested in the derivation of the inference that Speaker B believes that Manu was there at the party, but Moni was not (=Speaker B believes that (¬Moni, Manu)). They compared B's utterance when pronounced with the L+H* vs. L*+H contour and found that SIs were more delayed and derived at lower rates with the L+H* contour. Altogether, the studies discussed thus far provide convincing evidence that intonation affects both the likelihood and processing of SIs.

As mentioned above, though the effects of intonation on SI calculation are well established, work on scalar diversity has tended to use written stimuli. Nonetheless, there are two notable exceptions, that is, two studies that manipulated intonation while testing multiple different lexical scales. Cummins and Rohde (2015) tested 20 different English adjectival scales, and presented participants with sentences such as *The view from the hotel room is pretty* in two intonation conditions: neutral vs. with prosodic focus placement on the scalar adjective (here, *pretty*). The authors take the focus manipulation to be a manipulation of the question under discussion (QUD; Roberts, 2012), which they predict would influence SI rates. Indeed, they found that participants were more likely to calculate the SI (e.g., *not gorgeous*) in the focus condition. However, as their by-item results show (Cummins & Rohde, 2015, p. 7, **Figure 1**), scalar terms differ in how susceptible they were to the intonation manipulation. There is substantial variation in effect size – i.e., in how much more likely the SI was to be calculated in the focus condition than in the neutral condition – and 6 scales, in fact, show the opposite pattern to the overall effect. This suggests that it is indeed important to study the effects of intonation on SI calculation across many scales, and to study scalar diversity with auditory stimuli. Crucially, one way in which

our study differs from Cummins and Rohde (2015) is that we are interested in more complex intonational contours over the whole SI-triggering utterance (e.g., the RFR), rather than just manipulating whether the weaker scalar term is focused.

An even more relevant study for the present purposes, de Marneffe and Tonhauser (2019), tested the effect of the RFR on multiple scales. Before discussing this study in more detail, however, we first want to provide sufficient background on prior research on the RFR.

2.3 The rise-fall-rise contour

The use of the RFR is illustrated in the naturally occurring example in (7) on the underlined sentences.

(7) *CK*: If everybody knew everybody, we wouldn't have the problems we have in the world today. You don't rob somebody if you know their name.
JS: You're robbin' ME… (AUDIO)

An early influential account of the RFR comes from Ward and Hirschberg (1985), who propose that it conveys speaker uncertainty with respect to a scale. The authors primarily focus on the RFR in replies to questions as in (8), where its contribution can be intuitively described as a polite hedge. Ward and Hirschberg (1985) capture data like this by proposing that the RFR conveys uncertainty either about whether it is appropriate to evoke a scale (8a), what scale is being evoked (8b), or where a particular value falls on a given scale (8c).

(8) a. A: Are you leaving today?
B: I'm not leaving TODAY… Ward and Hirschberg (1985), (54)
b. A: Are you a doctor?
B: I have a PHD… Ward and Hirschberg (1985), (58)
c. A: Have you ever been West of the Mississippi?
B: I've been to MISSOURI… Ward and Hirschberg (1985), (62)

As an alternative but related proposal, Constant (2012) draws a connection between the RFR and focus particles like *only*. On this view, the RFR quantifies over alternative propositions and indicates that they cannot be safely claimed by the speaker. One – highly relevant – pattern that motivates this account is that the RFR can only occur when the alternatives to the accented element do not resolve all other alternatives (are not "alternative dispelling", in Constant's terms), illustrated in (9). Both maximal scale elements, which either entail the falsity of all stronger alternatives (*no one*) or entail the truth of all weaker alternatives (*all*), are infelicitous, while the element that leaves alternatives open (*most*) is not. This pattern is captured by the assumption that in the cases of *no one* and *all*, the domain of alternatives to the asserted proposition that the RFR quantifies over is empty, and that there is a general ban on this vacuous quantification.

Additionally, the contribution of the RFR is treated as a conventional implicature, by virtue of it being speaker-oriented and independent of at-issue content.

(9) A: Did your friends like the movie?
  a. B: MOST of my friends liked it…
  b. B: #NO ONE liked it…
  c. B: #ALL of my friends liked it… Constant (2012), (33)–(34)

Further related accounts come from Wagner (2012) and Wagner et al. (2013). Wagner differs from Constant in assuming that the RFR operates over alternative speech acts rather than propositions, and that it contributes a presupposition rather than a conventional implicature. That is, the alternatives assumed to be evoked by the RFR are not calculated relative to the focus of the sentence (e.g., {*Most/None/All/…*} *of my friends liked it…* in (9)) but, more broadly, to what else could have been said. This adjustment is meant to capture the RFR's ability to be embedded as its own speech act that is separate from the rest of the sentence, rather than having to take scope over the whole utterance, as shown with the appositive relative clause in (10).

(10) John – who likes sweets – was an obvious suspect.

Wagner and colleagues focus on the incompleteness component of the RFR, stated in (11), and present experimental evidence from contexts like (12) that the RFR is produced more frequently and perceived as more acceptable in partial answers, compared to complete answers.

(11) **RFR** (*p*): The speaker asserts *p* but considers it to be only an incomplete answer to the question under discussion.

(12) a. Partial answer
  Q: Is either Bill or Susan coming to the party?
  A: Bill is coming.
  b. Complete answer
  Q: Is Bill coming to the party?
  A: Bill is coming.

Although the previous three accounts seem closely related, they differ in a small but important detail. While all three accounts are compatible with the RFR providing an incomplete answer when the truth of other alternatives is unknown, Constant additionally allows alternatives to be unclaimable, because they are known to be false. This feature captures the fact that the RFR can be followed up with an answer that fully resolves the relevant question, as in (13), which is incompatible with Wagner's and Wagner et al.'s accounts.

(13) A: Did your friends like the movie?
  B: JOHN liked it… the rest of them hated it. Constant (2012), (16)

A different account comes from Westera (2019), which can be viewed as elaborating on the relevance of the QUD, highlighted by Wagner et al. (2013). Embedded in a Gricean theory of pragmatics (for more details on the framework, see Westera, 2017), Westera proposes that the RFR – assumed to also cover cases of Contrastive Topic (Büring, 2003) – conveys information about whether a conversational maxim is being violated or adhered to, and what a speaker takes the available QUDs to be. More specifically, by using an RFR, the speaker is taken to indicate that there are at least two QUDs that are being addressed, and that with respect to one QUD, a maxim is being violated (or suspended), while for another QUD, a maxim is complied with. To illustrate this account with the case in (9), we can assume that the main QUD is the explicitly provided question (*Did your friends like the movie?*), and using the RFR conveys that the answer given does not fully resolve the question, thereby suspending the Maxim of Quantity. As a result, both *no one* and *all* are infelicitous, because they fully resolve the question, and consequently, no maxim is violated, contrary to what the RFR is assumed to convey. A possible secondary QUD for this case could be something along the lines of *Do you think I should go see the movie?* – this is ultimately left to pragmatic reasoning.

Lastly, Göbel (2019) and Göbel and Wagner (2023) shift their attention to the function of the RFR in argumentative dialogues. The observation they make is that the RFR exhibits an asymmetry in replies to statements, depending on the polarity of the initial statement, which they dub *valence asymmetry*. While the RFR is felicitous when providing a positive counterpoint to a negative statement (14a), it is degraded when the order is reversed and its carrier utterance provides a negative counterpoint to a positive statement (14b).

(14) a. A: The bike ride yesterday was really terrible, the weather was horrific.
B: We had a cocktail… (AUDIO)

b. A: The bike ride yesterday was really great, the weather was perfect.
B: #We had an accident… (AUDIO)

Notably, this pattern is unexplained by previous accounts, since B's replies in (14a) and (14b) do not differ in whether alternatives are left open or not. The authors, hence, propose that the RFR conveys the presence of a stronger alternative on a pragmatically inferred scale. For cases like (14), this scale concerns an evaluation, here, of the quality of the bike ride, where the positive reply implies a stronger – or better – alternative to A's statement, whereas the negative reply implies a weaker – or worse – one. For cases like (9), on the other hand, the scale is one of logical entailment, such that a stronger alternative to *most* would be *all*, capturing the pattern in a similar way as previous accounts.

This account of the RFR and the case of the valence asymmetry directly connect to the notion of adjective polarity in the scalar diversity literature, as mentioned earlier. As an illustration, consider the case of the scale $<$*ugly*, *hideous*$>$, categorized as a pair with negative polarity

by Gotzner et al. (2018). On the assumption that *ugly* and *hideous* are on a measurement scale regarding beauty with other adjectives like *pretty* and *beautiful*, the stronger predicate *hideous* would actually be lower than *ugly* on the scale (see Solt, 2015). As a result, the RFR would be predicted to be unacceptable by Göbel and Wagner (2023). This prediction is borne out intuitively for the item in (15).[4] We would, therefore, expect the RFR to occur less frequently with negative scales than positive scales, and to contribute to the appearance of a polarity effect in scalar diversity, assuming the RFR has its own independent effect on SI rates.

(15) A: Is the wallpaper hideous?
B: ??It is ugly… (AUDIO)

Regarding SI rates, the accounts of the RFR differ in their predictions about its effect on the likelihood of drawing an SI. Ward and Hirschberg (1985), Wagner (2012) and Wagner et al. (2013) can all be argued to predict a decrease in SI rate for the RFR, relative to a Fall: drawing an SI relies on negating a stronger alternative, which is incompatible with having to leave the truth of said alternative open, as required by these accounts.[5] In contrast, Göbel (2019) and Göbel and Wagner (2023) simply treat the RFR as implying the existence of a stronger alternative, while remaining agnostic regarding its truth value. A possible effect of the RFR could, then, be that highlighting the salience of the relevant alternative leads to an increase in SI rate. The idea that the salience of alternatives can affect SI rate in this way is supported by findings from the written domain from Ronai and Xiang (2024; see also Ronai & Xiang, 2021b; Yang et al., 2018; Zondervan et al., 2008), who found that a prior question that mentions the stronger alternative leads to an increase in SI rate, relative to when the SI-triggering sentence occurs without a question context, or following a question that mentions the weaker scalar term itself. An intermediate position is taken by Constant (2012), whose account is compatible with either an increase or a decrease, given that alternatives can be unclaimable either because they are considered false (SI increase) or because they are not known (SI decrease). Similarly, Westera's (2019) account allows some flexibility, such that the exact predictions with respect to SI rate are less definitive. On the one hand, the account includes a prediction that the RFR may not exhaustively resolve the main QUD, which would result in a decrease in SI rate. On the other hand, in the cases we are concerned with, the RFR may also pick up on a secondary QUD – so the explicitly mentioned question may no longer be the main QUD, and the RFR may be compatible with an increase in SI rate.

[4] Note that this judgment is only meant to hold in the absence of further context. For instance, in a situation where A is intentionally looking for an ugly gift for a person they do not like, B's reply seems quite acceptable.

[5] In the case of Ward and Hirschberg (1985), there is an open question about what level the uncertainty could be conveyed at, given the different options provided (8a)–(8c). Following de Marneffe and Tonhauser (2019), we will assume that the most sensible option is one where uncertainty relates to the choice of scalar value rather than the existence or type of scale, given that the target items in studies of SI are inherently scalar.

2.4 Previous work on the effect of RFR on SI calculation

While the relation between the RFR and SIs has not been the main concern of the accounts discussed above, there exist two notable studies that have looked at this relation. The first is de Marneffe and Tonhauser (2019), mentioned above, who tested dialogues such as (16), where the reply contains a weaker scalar term, and the question, a stronger alternative.

(16) *Mike*: Was your hike exhausting?
Julie: It was strenuous.

The authors manipulated whether Julie's answer was pronounced with a neutral fall (H* L-L%) or an RFR (L*+H L-H%), and asked participants to indicate whether Julie "mean[s] that her hike was exhausting" on a 7-point scale (from *definitely no* to *definitely yes*). The results showed that the RFR led to fewer positive responses than a neutral fall, suggesting that it increased the likelihood of drawing a SI. However, while the experiment tested 16 different adjectival scales, the authors' main focus was not on by-scale variation, leaving open the question of how (or whether) the intonation manipulation interacted with scalar diversity.

Moreover, the conclusion drawn by de Marneffe and Tonhauser (2019) has been questioned by Buccola and Goodhue (to appear), the second study on the RFR and SIs to be discussed here. Crucial to their criticism is the distinction between SIs and ignorance inferences: rather than taking the assertion of a weaker alternative as evidence that the speaker believes the stronger alternative to be false, it can also be understood as indicating that the speaker is simply ignorant regarding its truth value. Crucially, replying *no* to the questions posed by de Marneffe and Tonhauser is compatible with either type of inference. Thus, the results can be given an interpretation other than the RFR increasing the likelihood of SI calculation.

To address the mapping of intonation onto pragmatic inferences directly, Buccola and Goodhue (to appear) tested which type of inference – ignorance inference or SI – participants would be more likely to draw after hearing a target sentence either with Fall or RFR, as illustrated in (17).

(17) A: Did all of the guests eat dinner?
B: SOME of them ate dinner.
a. B thinks that not all of the guests ate dinner. SI
b. B isn't sure whether or not all of the guests ate dinner. ignorance inference

In their first experiment, participants were given one contour and had to choose one of the inference options. Both contours led to SI choices overwhelmingly, without a significant difference between them (although the preference was numerically larger for Fall). The second experiment then gave participants both intonation versions at once and asked them to choose between mapping Fall to SI and RFR to ignorance inference, or Fall to ignorance inference and

RFR to SI. Under those conditions, participants showed a preference for the former option, which the authors took as evidence that the RFR conveys uncertainty. While this study did not test the effect of the RFR on the likelihood of SI calculation directly (only via comparison to ignorance inferences) and was restricted to the $<$*some*, *all*$>$ scale, it does add to the body of evidence demonstrating that SI-related interpretations are sensitive to intonational cues.

With this background, we now turn to our own experimental investigation of the role of intonation for scalar diversity.

3. Experiment 1: Production + inference task

In Experiment 1, we investigate the effect of intonation on SI calculation – in the context of scalar diversity – by combining a production task with an inference task. This allows us to see what intonational contours participants produce for various potentially SI-triggering sentences, and whether they calculate SI (as measured by the inference task), given a certain contour.

3.1 Method, materials & design

The stimuli in Experiment 1 consisted of question-answer dialogues containing scalar terms, as in (18). The dialogues featured the following experimental manipulation: the question prompt (Emma's question) and the target sentence (the participant's reply) either contained the same weaker scalar term (18a), or the question contained the chosen stronger alternative (18b). There were 60 lexical scales taken from Ronai and Xiang (2024), in addition to 20 fillers. The SAME vs. STRONG manipulation was administered within-participants, i.e., each participant saw each item only in one condition in a Latin Square design.

(18) Sample Item, Experiment 1

a. *Emma*: Was the winner happy? SAME

b. *Emma*: Was the winner ecstatic? STRONG
You: She was happy.
Given your response, do you think Emma would conclude that the winner was not ecstatic?

Participants first saw the full dialogue on the screen. After pressing a button, they heard an audio recording of the question (*Was the winner happy?* or *Was the winner ecstatic?*) and had to record themselves saying the reply (*She was happy.*). Emma's questions were presented auditorily, in addition to in written form, in order to make the task more natural. Afterwards, they were given the task question *Given your response, do you think…?* (italicized in (18)) and chose between "Yes" and "No" as their answer. In this adapted version of the inference task from van Tiel et al. (2016) (see also, i.a., Pankratz & van Tiel, 2021), if a participant responds with "Yes", that can be taken to indicate SI calculation: that the participant has enriched *happy* to *not ecstatic*. Responding with "No", on the other hand, suggests that the participant has not calculated the SI and takes *happy* to

be compatible with *ecstatic*.[6] Altogether, this method allowed us to gather data on the production rates of relevant contours on the target sentence across conditions and items, as well as examine SI rates in light of a given contour being produced.

Recordings were manually annotated by the second author in terms of the overall contour used by the participant on a given item. The annotator listened to target sentences in Praat. Sentences were presented without the context sentence or knowledge of condition in order to avoid bias. Contours were categorized according to a combination of the visual pitch information available and the auditory impression, given that audio quality was not always sufficient to guarantee accurate pitch tracking. The "a priori" categories originally included five contours:

i. a pitch accent on the scalar item, e.g., *happy* in (18), and a monotonous final fall – (L+)H* L-L% in ToBI labels (=Fall),
ii. a rising pitch accent on the scalar item, followed by a low phrasal accent and a rising final boundary tone – {L*+H/L+H*} L-H% (= RFR),
iii. a pitch accent on the auxiliary (if present), which we take to indicate Verum Focus (see Höhle, 1992; Lohnstein, 2015; as well as 3.5 below), followed by a monotonous final fall (= Verum Focus Fall, following Goodhue et al., 2016),
iv. a monotonous final rise without a preceding rising accent – L* H-H% (= Rising Declarative, see, e.g., Gunlogson, 2001),
v. Other/Unclear.

However, after initial inspection, two changes were made. First, Rising Declaratives were taken out, due to the contour not occurring sufficiently frequently. Second, as mentioned in Section 1, there was a notably frequent use of a contour with an initial and a final high tone that we labeled "Concession Contour", illustrated in (19), so it was added as one of the categories.

(19) (A: Was the winner happy/ecstatic?)
B: She was happy. (AUDIO)

3.2 Procedure

The experiment was implemented through prosodyExperimenter (https://github.com/prosodylab/prosodylabExperimenter). Participants first saw a welcome screen, followed by a chance to adjust their volume and test their microphone, an online consent form, and a language background questionnaire. Afterwards, there was a test in which participants were played three sounds and had to choose which one was the quietest, which required the use of

[6] Buccola and Goodhue (to appear) and Ronai and Xiang (2024) argue that a "No" response is also compatible with ignorance regarding the status of the stronger alternative. We come back to this issue in Section 5.

headphones. For the main part of the experiment, participants provided their production of the target sentence and then answered the question for the inference task, as described above. There were three practice trials after the instructions were received, followed by a total of 80 stimuli. The experiment concluded with a chance to provide feedback. A test version of the experiment can be accessed at https://prosodylab.org/~agobel/conepi/30-scaRFR_Pro2AFC/?SESSION_ID=Glossa&mode=experiment.

3.3 Participants

64 monolingual native speakers of American English were recruited on Prolific and compensated $4 or $5 (depending on time). One participant's response file was not properly saved and, hence, not annotated. During annotation, participants were excluded if their responses were unnatural (N = 5), or if they were monotonous across items, in that they almost exclusively chose either Fall or Verum Focus Fall (N = 21).[7] We were, thus, left with 37 participants for the data analysis reported below.

3.4 Results

3.4.1 Production rates

The counts by condition for each category are shown in **Figure 1**. The first thing to note is that Fall is by far most frequent contour used, comprising about 56% of the total recordings, even after excluding monotonous participants. Next, we can see that the RFR was used almost exclusively in the STRONG condition. The Concession Contour, on the other hand, trended toward occurring more frequently in the SAME condition, but was more evenly distributed. Finally, Verum Focus Fall occurred almost exclusively in the SAME condition. We discuss the implications of these findings below in 3.5.

3.4.2 SI rates

We next looked at the rate of SI calculation, as measured by the inference task, depending on the contour produced by the participant. We restricted this analysis to Fall as a baseline, RFR as the intended contour of interest, and the Concession Contour for exploratory purposes. SI rates, i.e., the proportion of "Yes" responses, for those three contours by condition are shown

[7] While this latter criterion leads to a high exclusion rate, we consider it justified, since we were not interested in how often people use the RFR in general, but in *when* and *how* they use it. We attribute the large number of exclusions mainly to the online setting; participants were, essentially, asked to simulate a natural-sounding conversation while sitting by themselves in front of a computer. As a result, even though all experimental items were dialogues, many participants' productions resembled reading a passage out loud from a book instead of participation in a conversation.

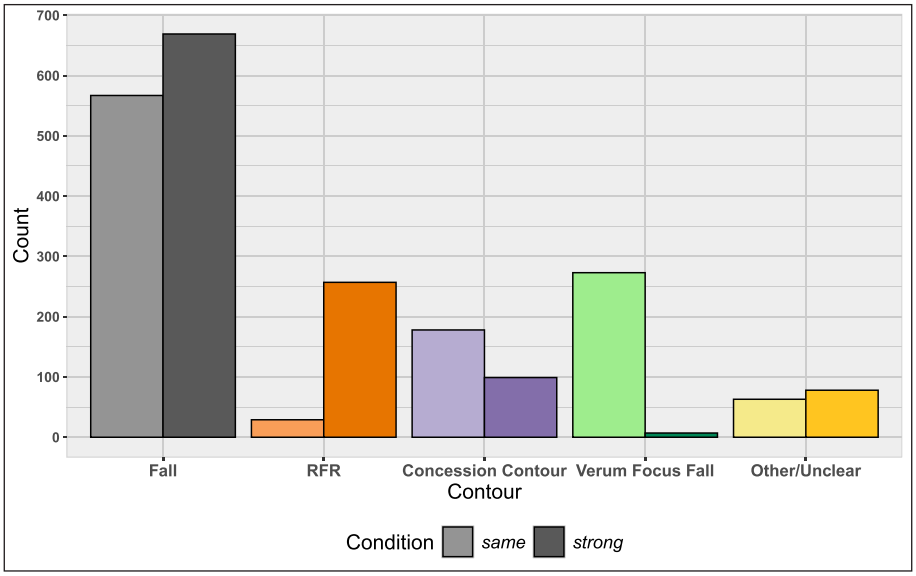

Figure 1: Production rates by contour and condition in Experiment 1. Lighter colors (left) correspond to the SAME condition, and darker colors (right), to the STRONG condition.

in **Figure 2**.[8] For the statistical analysis, we fit a logistic mixed effects regression model using the lme4 package in R (Bates et al., 2015). The model predicted Response in the inference task ("Yes" vs. "No") as a function of Contour (RFR vs. Fall vs. Concession Contour), Condition (SAME vs. STRONG) and their interaction. It included the maximal random effects structure supported by the data (Barr et al., 2013): random by-participant and by-item intercepts and slopes for the Condition predictor. Both fixed effects predictors were treatment-coded: in Contour, the Fall level served as baseline, while in Condition, the STRONG level served as baseline.

The analysis revealed the following results. First, the SAME condition produced lower SI rates than the STRONG condition (Estimate = –1.13, SE = 0.27, z = –4.14, $p < 0.001$). There was no evidence that this effect differed across contours, i.e., there were no significant interactions (Estimate = –0.58, SE = 0.59, z = –0.98, $p = 0.33$; Estimate = 0.04, SE = 0.4, $z = 0.11$, $p = 0.92$). Second, Fall showed the lowest SI rate (33.5% in the SAME condition and 45.3% in the STRONG condition), followed by the Concession Contour (48.3% in the SAME condition and 61.6% in the STRONG condition), which produced a significantly higher rate (Estimate = 0.7, SE = 0.31, $z = 2.25$, $p < 0.05$). Lastly, the RFR produced the highest SI rate (55.2% in the SAME condition and 70% in the STRONG condition), also significantly higher than the baseline Fall (Estimate = 0.89, SE = 0.23, $z = 3.81$, $p < 0.001$).

[8] Note that while individual circles in **Figure 2** correspond to the proportion of "Yes" responses per participant in that condition, these proportions also depend on how many times that participant produced the given contour. That is, if a participant only produced the RFR on one item, then their proportion of "Yes" responses could only be 100% or 0%. This problem will not arise in Experiment 2, when we directly manipulate contours.

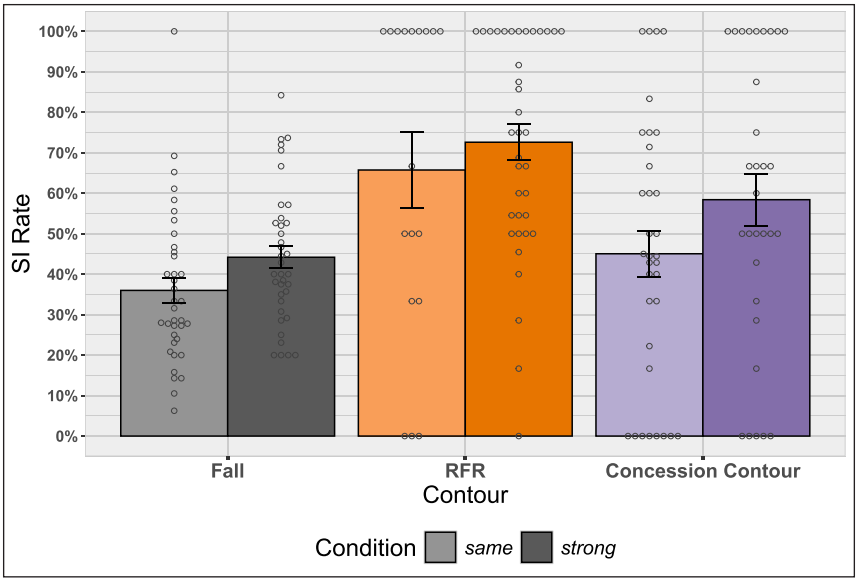

Figure 2: Mean SI rates (and SE) by contour and condition in Experiment 1. Lighter colors (left) correspond to the SAME condition, and darker colors (right), to the STRONG condition. Circles denote individual participants.

3.5 Discussion

3.5.1 Production rates

The experiment provided data from two sources: production rates of contours and inference rates given the production of a certain contour. For production rates, there are four findings to mention. First, participants' primary choice of contour was a Fall, which made up slightly over half of all productions. We attribute this overwhelming preference to the online setting, with participants sitting by themselves in front of a computer, which may make it difficult to voice-act fully naturally. Additionally, the annotation focused solely on the overall pitch contour and did not take into account other acoustic factors, such as duration or intensity, which is worth investigating in future studies (see, for example, Sandberg & Cole, 2022).

Second, we saw that Verum Focus Fall occurred exclusively in the SAME condition, where the scalar terms in the question prompt (Emma's question) and the target sentence (the participant's reply) were identical. This finding serves as a sanity check, since in the SAME condition, the scalar term in the reply (*happy*) is given, and accenting a given word is usually marked. Shifting prominence to the auxiliary prevents such a violation. However, not all items allowed for this pattern, since not all target sentences included auxiliaries (e.g., *Did the train slow? It slowed.*). This explains why the rate of Verum Focus Fall is not as high as one would expect.

Third, the RFR almost exclusively occurred in the STRONG condition, where the question prompt mentioned a stronger alternative. This is in line with Göbel and Wagner's (2023) account of the RFR, which takes it to convey the presence of a stronger alternative. In the STRONG

condition, the requirement for a stronger alternative to be present is explicitly satisfied by the question prompt (*Was the winner **ecstatic**?*). In Section 5, we elaborate more on how other theoretical accounts might capture the observed STRONG-SAME asymmetry.

Finally, the experiment revealed the frequent use of a contour that was not previously considered as a relevant option for the given contexts, which we refer to as the Concession Contour, illustrated in (20) (repeated from (19)) with one of the productions elicited in the experiment. Its prosodic characteristics are an initial high tone followed by a fall up until a concluding rise. This pitch shape exactly parallels that of the so-called Contradiction Contour (Liberman & Sag, 1974), illustrated in (21). However, intuitively the two contours seem to make different contributions, with the reply in (20) sounding much less like a proper contradiction. We will also return to this issue in Section 5.

(20) (A: Was the winner happy/ecstatic?)
B: She was happy. (AUDIO)

(21) *JS*: These balloons aren't gonna stay filled 'til New Year's!
CK: Those aren't for New Year's! Those are my everyday balloons. (AUDIO)

3.5.2 SI rates

Moving on to SI rates, Experiment 1 had three main findings. First, SI rates were higher in the STRONG condition than in the SAME condition. This replicates Ronai and Xiang (2024), who conducted the same manipulation using written stimuli. The STRONG-SAME difference can be explained in at least two ways. In the STRONG condition, the question directly mentions the relevant stronger alternative, thereby increasing its salience and encouraging hearers to reason about it. Additionally, in that condition, only on its SI-enriched meaning does the answer constitute a congruent one, in the sense of Gualmini et al. (2008), Hulsey et al. (2004), and Zondervan et al. (2008). Since this finding is not of primary interest to the current study, we direct the reader to Ronai and Xiang (2024) for further discussion, as well as to, i.a., Degen (2013), Degen and Tanenhaus (2015), Kursat and Degen (2020), and Ronai and Xiang (2021b) for findings regarding the context-sensitivity of SI.

More crucially, we also found that the RFR led to an increase in SI rates, relative to a Fall. As discussed in 2.3, this pattern is unexpected based on several theoretical accounts of the RFR: those that take the contour to correspond to the alternative being left open (Wagner, 2012; Wagner et al., 2013; Ward & Hirschberg, 1985). It is, however, directly in line with predictions of Göbel (2019) and Göbel and Wagner (2023). As for previous experimental results, de Marneffe and Tonhauser (2019) had similarly found the RFR to result in increased SI rates, while Buccola and Goodhue (to appear) found the opposite effect. We discuss how our findings can be reconciled with the latter study in detail in Section 5.

Lastly, Experiment 1 also found the novel Concession Contour to yield a higher SI rate than a Fall, but less so than the RFR, a full discussion of which we will come back to later.

3.5.3 Relation to scalar diversity

While the effect of the RFR on SI rates constitutes an interesting finding for theoretical accounts of the contour, further analyses are needed to more precisely determine how intonation interacts with the phenomenon of scalar diversity. For instance, one important possibility is that those lexical scales that show a high SI rate in written studies might also happen to be the ones produced more often with an RFR in our study, such that RFR rate is a direct correlate of SI rate. If this is the case, then the RFR does not actually encourage SI calculation, and its effect of leading to increased SI rates in Experiment 1 arises, instead, as an epiphenomenon. This hypothetical, whereby the RFR co-occurs with scales that robustly lead to SI, could also receive a different interpretation. Namely, since previous studies of scalar diversity relied on written stimuli, it is conceivable that certain scales were found to lead to higher SI rates than others due to their propensity to be silently read with an RFR intonation. On this view, finding that RFR rates and SI rates are linked would suggest that written studies of scalar diversity suffered from a confound. To investigate these possibilities, we conducted an additional correlational analysis. **Figure 3** shows the by-item (that is, by-scale) correlation between RFR productions in the STRONG condition of Experiment 1 and SI rates from Ronai and Xiang's (2024) written study, which conducted the same dialogue manipulation on the same stimuli as we did.

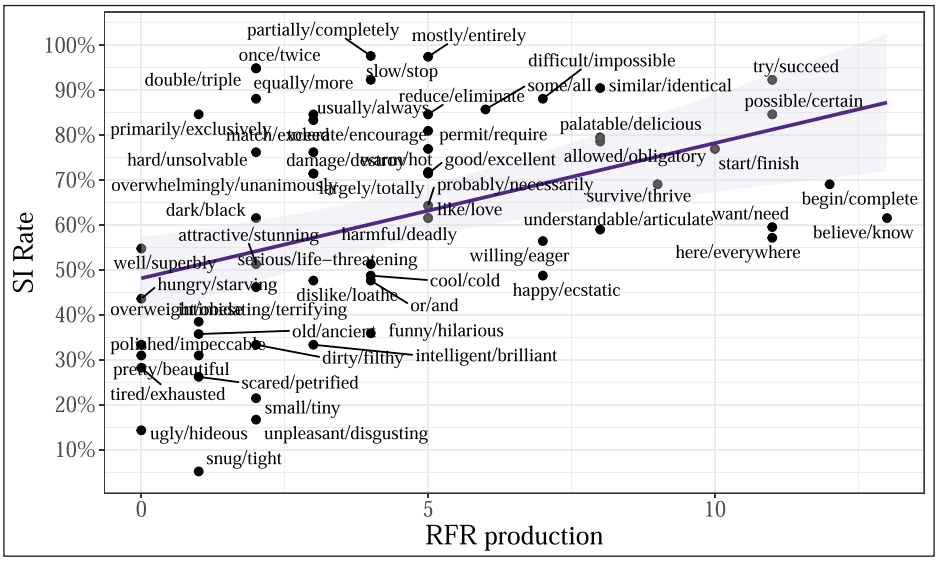

Figure 3: By-item correlation between RFR productions in Experiment 1 and SI rates from Ronai and Xiang's (2024) written study (STRONG condition).

We find a moderate positive correlation (Pearson's correlation test: r = 0.45), indicating that the RFR was indeed produced more frequently with scales that are more likely to lead to SI.[9] This can be interpreted as suggestive evidence that the RFR's effect on SI rates is indeed related to its co-occurrence with lexical scales that are more likely to lead to SI. Or, alternatively, that a scale's likelihood of triggering SI calculation is linked (in part) to its propensity to be silently read with RFR. Before drawing any firmer conclusions, however, we will further investigate these possibilities in Experiment 2, which uses a perception task that allows us to assess the contribution of intonation independently of lexical factors.

In 2.1, we raised the possibility that the RFR may interact with predictors of scalar diversity in ways that constitute possible confounds, unless properly controlled for. One such factor is polarity. Göbel (2019) and Göbel and Wagner (2023) suggest that the RFR is only felicitous with positive statements, while Gotzner et al. (2018) have found higher SI rates with negative adjectival scales than with positive ones. On theories of the RFR that predict it to lower SI rates, e.g., because it indexes uncertainty, this would open up the possibility that what results in the polarity effect in scalar diversity is that positive scales are more likely to be silently read with the RFR. In Experiment 1, we found the RFR to increase SI rates, not decrease them, which suggests that the above confound is not at play. Nonetheless, it is worth briefly looking at the effect of polarity on both RFR productions and SI rates. To do this, we analyzed the subset of our lexical scales that had been annotated for polarity by Gotzner et al. (2018) (see their 2.1.2.4. for details). This included 21 adjectival scales that are fully identical across the two studies, as well as <*pretty*, *beautiful*>, where we adopted Gotzner et al.'s annotation for <*pretty*, *gorgeous*>. Of these 22 scales, the RFR was produced 52 times with positive scales (4.33 average) and 19 times with negative scales (1.9 average). That is, the RFR occurred more than twice as frequently with positive scales, in line with the predictions of Göbel (2019) and Göbel and Wagner (2023). To check the effect of polarity on SI rates, a logistic mixed effects model was fit, predicting Response ("Yes" vs. "No") by Polarity (treatment-coded, with NEGATIVE serving as baseline). The model included random intercepts for participants and items. This analysis revealed that POSITIVE and NEGATIVE scales did not differ significantly from each other in their likelihood of leading to SI (Estimate = 0.12, SE = 0.86, $z = 0.14$, $p = 0.89$).[10] This means that Experiment 1 ultimately did not reveal evidence supporting the hypothetical conspiracy of the RFR's polarity asymmetry and effect on SI rates: we actually found that the RFR increases SI rates, and our set of items happened to include adjectival scales that do not show a polarity effect. But it remains the case that factors

[9] The same correlation is much weaker in the SAME condition (Pearson's correlation test: r = 0.24), which follows from the RFR not being produced very robustly in the SAME condition in the first place.

[10] This would seem to run counter to Gotzner et al.'s findings, but, in fact, if we look at the same 22 scales from their work that were tested in our study, we find that those also did not produce different rates of SI; in the relevant subset of Gotzner et al.'s data, the average SI rate for positive scales is 37.17%, while for negative ones, it is 37.1%.

governing intonational contours overlap with those predicting SI rates, and future work should, therefore, still keep this potential interaction in mind.

4. Experiment 2: Perception + inference task

To address open questions from Experiment 1, Experiment 2 focuses on the independent effect of intonation on SI rates. For this, we combine a perception task with the inference task: participants listen to potentially SI-triggering sentences in different intonational conditions before making an SI judgment.

4.1 Method, materials & design

We used the same materials as in Experiment 1 (60 experimental stimuli + 20 fillers), but restricted to the STRONG condition, since the RFR was rarely produced in the SAME condition. Additionally, both the question prompt (Emma's question) and the target sentence (now Luke's reply) were presented auditorily, without the text being visible on the screen. The target sentence occurred with one of three contours: a FALL, the RFR, or the CONCESSION CONTOUR, in a Latin Square design. After listening to one version of the dialogue, participants were asked the same task question (italicized in (22)) as in Experiment 1 – with the only modification being that the target speaker was no longer referred to as *you* but as *Luke*, i.e., the task question included *Given Luke's response…*. As before, we take a "Yes" response to index SI calculation, and a "No" response to index that the participant has not calculated the SI. A sample item with recordings is shown in (22), and pitch tracks for the three contours are shown in **Figure 4**.

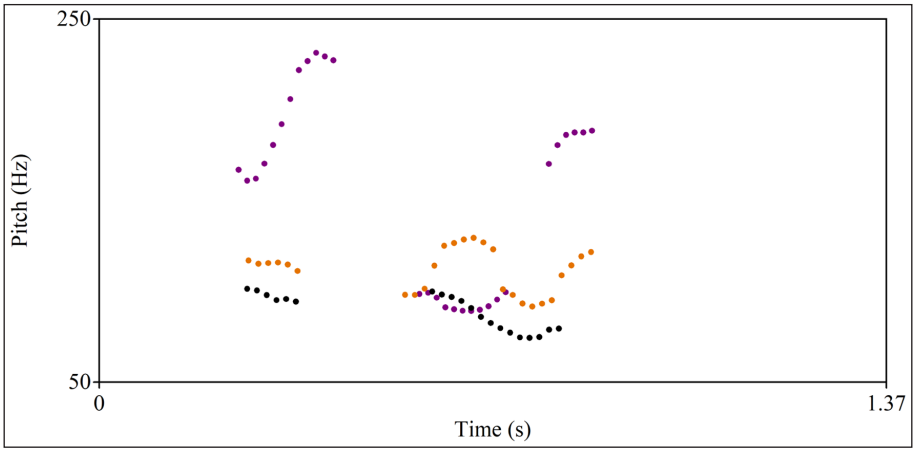

Figure 4: Pitch tracks for the target sentence *It was serious*, from the <*serious*, *life-threatening*> scale, with FALL in black (AUDIO), RFR in orange (AUDIO) and CONCESSION CONTOUR in purple (AUDIO).

(22) Sample Item, Experiment 2
Emma: Was the winner ecstatic?
Luke: She was happy. {[FALL], [RFR], [CONCESSION]}
Given Luke's response, do you think Emma would conclude that the winner was not ecstatic?

4.2 Procedure

The general procedure was largely the same as for Experiment 1, except there was no mic check. A test version can be accessed at https://prosodylab.org/~agobel/conepi/31-scaRFR_Aud2AFC/?SESSION_ID=Glossa&mode=experiment.

4.3 Participants

90 monolingual native speakers of American English were recruited on Prolific and compensated $2.50. 17 participants were excluded for failing the headphone test. Data from the remaining 73 participants is reported below.

4.4 Results

SI rates – that is, the proportion of "Yes" responses – by contour are shown in **Figure 5**. To analyze the results, we fit a logistic mixed effects regression model predicting Response ("Yes" vs. "No") as a function of Contour (Fall vs. RFR vs. Concession Contour). The fixed effects predictor was treatment-coded, with Fall as the reference level. The maximal converging random effects structure included by-participant intercepts and by-item intercepts and slopes. We found significantly higher rates of SI calculation with the RFR than with the Fall (Estimate $= 0.4$, SE $= 0.12$, $z = 3.25$, $p < 0.01$). The difference between Fall and Concession Contour, on the other hand, was not significant (Estimate $= 0.04$, SE $= 0.12$, $z = 0.39$, $p = 0.70$).

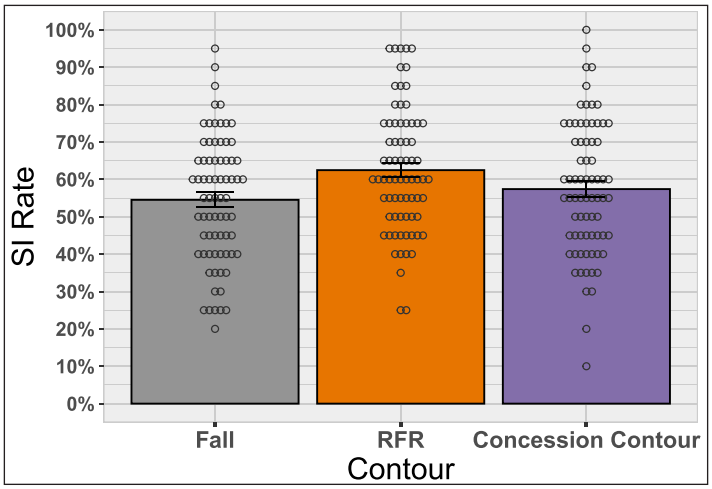

Figure 5: SI rates (and SE) by contour in Experiment 2. Circles correspond to individual participants.

4.5 Discussion

The results largely replicated the findings from Experiment 1. Fall received the lowest SI rate (54.5%), RFR the highest (62.5%), and the Concession Contour was numerically in between the two (57.4%). However, the differences were much smaller than in Experiment 1, such that only the comparison between Fall and RFR reached statistical significance. This compression may be due to the more mediated nature of the task: rather than judging one's own production – and by virtue of that, most likely intention – the perception experiment required not only reasoning about the intention of someone else's choice of intonation, but also how that might affect the hearer. The fact that the experiment was able to replicate the previous data is thus even more notable. As before, the RFR leading to an increase in SI rates supports theoretical accounts where it is analyzed as indicating the presence of a stronger alternative (Göbel, 2019; Göbel & Wagner, 2023), while it is less compatible with those that take the contour to correspond to uncertainty or the alternative being left open (Wagner, 2012; Wagner et al., 2013; Ward & Hirschberg, 1985).

In our discussion of Experiment 1, we raised the possibility that the RFR increasing SI rates is epiphenomenal, i.e., arising as a consequence of it occurring more frequently with items that are more likely to lead to SI to begin with. Experiment 2 was specifically conducted to further probe this possibility by directly manipulating the intonational contour of all items. Since Experiment 2 found the same effect of the RFR as Experiment 1, but this effect now cannot be reduced to by-scale variation in SI rates, as the RFR vs. Fall contrast occurred with all items, we can conclude that the RFR indeed encourages SI calculation.

As mentioned, a different interpretation of the (moderate) by-scale correlation between RFR productions and SI rates is also available. Namely, it could be the case that the reason why certain scales were found to produce higher rates of SI than others in written studies of scalar diversity is that such scales are the ones more likely to be silently read with the RFR. The current Experiment 2 data is also informative with respect to this possibility, as we can check whether the different lexical scales' relative likelihood of leading to SI remained consistent across different intonational contours. To do this, we calculated rank-order correlations using Kendall's τ_B.[11] This analysis finds that SI rates with the RFR are correlated with both the Fall ($\tau_B = 0.63$) and the Concession Contour ($\tau_B = 0.67$), showing that the relative order of different lexical scales remains largely (though not entirely) the same. This, in turn, suggests that what makes a lexical scale a "high SI rate" scale is not simply its propensity to be silently read with RFR in a written study.

The next section turns to further discussion of what the results of Experiments 1 and 2 tell us about SIs and scalar diversity, as well as intonational contours.

[11] A statistic of –1 indicates full reversal of the rankings, while a statistic of 1 indicates the same ranking, i.e., that lexical scales occur in the same order when ranked by the SI rate they produced.

5. General discussion

This article presented data from two experiments investigating the role of intonation for scalar diversity. Experiment 1 used the combination of production and an inference task to assess, first, what contours speakers use in dialogues involving scalar terms, and second, how the choice of contour affects the likelihood of drawing an SI across different scales. We saw that – despite a strong overall preference for Fall, likely due to the online setting – participants frequently used the RFR, as expected, in addition to the unexpected use of a contour we labeled Concession Contour. Crucially, both the RFR and the Concession Contour led to an increase in SI rates relative to a Fall, with the RFR's effect being stronger. Moreover, the rate at which participants produced each of these contours was not uniform, but varied by lexical scale. To ensure that the effect of contour on SI rates was not actually driven by this variation, Experiment 2 presented participants with recordings of the same target sentences, manipulating the type of contour. The results from this experiment replicated the main overall pattern, with both the RFR and the Concession Contour numerically increasing the likelihood of SI, relative to Fall, although only the comparison between RFR and Fall was significant. We now turn to discussing how the combined findings inform the study of scalar diversity specifically, and SI more generally, as well as the theories of the intonational contours involved.

The finding that scales vary in their likelihood of receiving an RFR contour in production has implications for scalar diversity. Namely, it raises the possibility that scales similarly vary in whether they are silently read with RFR in written studies. As a result, when comparing SI rates across different lexical scales using written stimuli, it is not easily discernible if differences are driven by the lexical scales themselves or mediated through the effect of lexical scales on rates of intonational contours, which then affect the likelihood of SI. At the same time, the by-scale correlation between RFR productions and written study-based SI rates is only moderate (**Figure 3**), and the relative ranking of scales remains reasonably consistent across different intonational contours (see 4.5). This cautions against interpreting our results too strongly, as suggesting that intonation being masked in prior studies is such a serious confound that existing scalar diversity findings should, necessarily, be reassessed. Instead, it seems likely that the interaction of scalar terms with intonation, e.g., their propensity to be silently read with the RFR, is one among many other factors – such as semantic distance or boundedness, as discussed in 2.1 – that play a role in scalar diversity.

As things stand, it is not well understood precisely what factors matter for the felicity of a contour such as the RFR, nor is the observed variation in SI calculation fully explained. But, based on existing proposals, some properties of the linguistic signal matter for both. We have investigated one such factor, the polarity of adjectival scales, following up on the possibility that negative scales only lead to higher rates of SI than positive scales due to a difference in their compatibility with the RFR. While we ultimately found this not to be the case in our own data

set, such possibilities should be taken into account in future work. More generally, based on the results of our article (as well as prior work, such as Gotzner, 2019; Tomlinson et al., 2017), future studies of SI should ideally control for the effects of intonation. As mentioned, we do not fully understand what governs the use of the RFR, and our work has found it to increase the rate of SI calculation. Consequently, written studies can never fully rule out the possibility that a participant's SI judgment had been affected by projecting an RFR contour onto the stimuli.

Turning to the implications of the results for accounts of the RFR, there were three relevant sources of evidence. First, the RFR was almost exclusively produced in Experiment 1 when the question prompt contained a stronger alternative to the scalar term in the target sentence. Of the accounts discussed in 2.3, this pattern is most straightforwardly explained by the account of Göbel (2019) and Göbel and Wagner (2023). On their view, the RFR presupposes the presence of a stronger alternative, which is provided in the STRONG condition, but not the SAME condition, thus explicitly licensing its use. The increased production rate could then be attributed to a principle like Maximize Presupposition (Heim, 1991), which encourages speakers to use a presupposition trigger – in this case the intonational contour – whenever possible. The accounts of Constant (2012) and Wagner et al. (2013) capture these results, insofar as they rule out the use of the RFR in the SAME condition, either due to no alternatives being left open (Constant) or the reply providing a complete answer (Wagner et al.). However, these accounts would have to be augmented by a general theory of why people choose to express one meaning over another when there are multiple options, since neither is couched in terms of presupposition, and hence a principle like Maximize Presupposition does not apply.

Westera (2019), in turn, could capture the observed STRONG-SAME difference in RFR productions via the assumption that finding a secondary QUD is easier in the STRONG condition. Such an assumption seems plausible, given that the SAME condition is maximally restricted in how the reply relates to the question, while the lexical mismatch in the STRONG condition leaves it more open whether the question is sufficiently addressed, potentially leading participants to search for other questions that could be addressed instead, or in addition. In contrast to previous accounts, on Ward and Hirschberg's (1985) account, it is less clear how to explain the production patterns; given the uncertainty view, there is no obvious reason why there should be an asymmetry between our two conditions. Finally, the RFR being restricted to the STRONG condition is least compatible with Wagner (2012), since on this account, the RFR is treated as quantifying over alternative speech acts, rather than being restricted by focus. As such, it is not clear why one could not indicate something else that could have been said in the SAME condition, leaving it unexplained why the two conditions should differ.

Secondly, and most importantly, the experimental results showed that the RFR led to an increase in SI rate, relative to a Fall. This finding provides direct evidence against the accounts of Ward and Hirschberg (1985), Wagner (2012) and Wagner et al. (2013), all of which predict

a decrease instead. Constant (2012) and Westera (2019), on the other hand, are able to capture our empirical findings, since these accounts are, in principle, compatible with both uncertainty and strengthening. Finally, similarly to the data from production rates, the SI rate pattern is accounted for by Göbel (2019) and Göbel and Wagner (2023) as an effect of salience: by virtue of the RFR presupposing the presence of a stronger alternative, this alternative can be assumed to be more salient, which, in turn, facilitates drawing the SI. This reasoning is analogous to a salience-based explanation of the finding – replicated here – that mentioning the stronger alternative in a preceding question in a dialogue also increases SI rate.

It is also worth briefly addressing the third source of evidence for accounts of the RFR, namely, the variation in production rate across lexical scales. As discussed in relation to **Figure 3**, the RFR occurred more than twice as often on lexical scales of positive polarity, compared to scales of negative polarity. This pattern is uniquely in line with the accounts of Göbel (2019) and Göbel and Wagner (2023): on the view that the alternative presupposed by the RFR is not simply stronger, but higher on a scale, pairs like <*ugly, hideous*> would be expected to license an SI by virtue of one item being stronger, but not license the use of the RFR on the weaker item, given that *ugly* is a higher degree of beauty relative to *hideous* and hence higher on the scale. However, it is also clear that this account cannot capture the full range of observed variation across scales. While a more detailed discussion of how the various scale properties could relate to accounts of the RFR goes beyond the scope of this article, we believe that it serves as a promising source of evidence for future research.

Let us now turn to the question of how the effect of the RFR on SI rate in our experiments relates to prior studies (see also 2.4). Our findings are in line with de Marneffe and Tonhauser (2019), who similarly found that the RFR increased SI rates. Buccola and Goodhue (to appear), on the other hand, found that participants are more likely to pair the RFR with an ignorance inference interpretation, and a Fall with an SI, than the other way around. The authors took this finding to support an uncertainty view of the RFR. Given such an account, and, indeed, Buccola and Goodhue's empirical results, we would expect the RFR to decrease SI rates, which is the opposite of what we found. Here, we suggest some ways in which this conflict can be reconciled.

Notably, Buccola and Goodhue's (to appear) experiments tested SIs based on only the <*some*, *all*> scale, while de Marneffe and Tonhauser (2019) and the current article focused on a larger number of lexical scales. In fact, in our Experiment 2, the effect of intonation on the <*some*, *all*> scale is in line with Buccola and Goodhue: SI was calculated at a rate of 88.57% with Fall, and 73.33% with the RFR. Since this data point represents only one of our 60 items, it does not lend itself to statistical analysis, but the numerical trend observed is in the direction supported by Buccola and Goodhue's work: the RFR led to fewer SIs. Taking the three relevant studies together, it is conceivable that the *some but not all* SI is affected differently by intonation than other scales. While it is perhaps the paradigmatic example of SI, there are other ways in which

it is not representative of the entire class of lexical scales: for example, it leads to SI calculation more robustly than almost any other scale.

Another important difference between our experiments and Buccola and Goodhue's is that they specifically contrasted SIs with ignorance inferences: both potential meanings were explicitly made available to participants. In contrast, the inference task in our own experiments is primarily aimed at identifying when participants have calculated an SI ("Yes" response), but obscures some other possible interpretations. As noted by Ronai and Xiang (2024), given our dialogue manipulation, three different meanings may underlie Luke's answer *She was happy*. It could correspond to an SI-enriched meaning (23a), or an ignorance meaning (23b), where Luke can only say that the winner was happy, but he does not know whether she was ecstatic, or to a meaning where the weaker term *happy* is used as a (near)-synonym to *ecstatic* (23c).

(23) Emma: Was the winner ecstatic?
Luke: She was happy.
a. She was happy (but not ecstatic). SI
b. (Well,) she was happy. ignorance
c. (Yes,) she was happy. happy ≈ ecstatic

Making the same observation that there are these three possible meanings, Buccola and Goodhue (to appear, p. 11) argue that the RFR is intuitively only compatible with (23a) and (23b), while a Fall is only compatible with (23a) and (23c). They further reason that if participants frequently intended to produce meanings like (23c) in our Experiment 1, and used Fall to do so, that could explain why not many of the Fall productions corresponded to SI calculation. This, coupled with a bias for SI over ignorance inferences affecting the interpretations of the RFR, could have led to the illusion that RFR leads to more SIs than Fall. However, this only explains the findings of our Experiment 1, not our Experiment 2. Moreover, if Fall is indeed inappropriate for conveying a meaning like (23b), that could have influenced the authors' own experiment. Namely, participants may have chosen to pair an ignorance inference with the RFR simply to avoid matching it to a contour it is incompatible with (Fall). We, thus, take the overall evidence to be a challenge for uncertainty accounts of the RFR, and more in favor of accounts such as Göbel (2019) and Göbel and Wagner (2023).

Lastly, the relevance of the present study for research on the meaning of intonational contours goes beyond theories of RFR, namely, by revealing the frequent use of the Concession Contour. In terms of its distribution, we found that the Concession Contour was produced more in the SAME condition, but also present in the STRONG condition. Additionally, in both experiments, the Concession Contour, like the RFR, contributed to an SI rate increase, relative to a Fall, although to a lesser extent. Given that our main focus is on SIs and the finding of the Concession Contour is unexpected, we believe a full account of the Concession Contour

goes beyond the scope of this article. One point we would like to address, however, is how the Concession Contour might relate to the recognized category of the Contradiction Contour, and, specifically, if they should be treated as variants of each other or as distinct categories.

As mentioned in 3.5, the Concession Contour and the Contradiction Contour seem closely related prosodically: both contours have an initial "floating" high tone that is not aligned with lexical stress, as well as a final rise. One notable difference is that the pitch height for the Contradiction Contour seems exaggerated. However, this difference could be attributed to paralinguistic factors, such as emotional arousal. The exaggerated contour of the Contradiction Contour is, thus, not a conclusive argument against treating it as a variant of the Concession Contour (and vice versa).

On the semantic-pragmatic side, the Contradiction Contour has been argued to presuppose that there is contextual evidence for the complement of the prejacent proposition (Goodhue & Wagner, 2018). That is, in the stereotypical contradictory use, the Contradiction Contour on p is licensed by the prior assertion of $\neg p$ (and vice versa). However, the question is if this account could capture the use of the contour in our experimental conditions, i.e., in response to a question, and its effect on SI rate. One possibility could be to adjust the account to conceive of contextual evidence in terms of degrees (cf. Farkas & Roelofsen's (2017) notion of credence levels). Given that asking about a proposition p ($= ?p$) is usually only licensed when the speaker is not committed to either p or $\neg p$, there is a sense in which the question act raises doubt about p and, hence, provides contextual evidence – albeit weak – against p being true, licensing the Contradiction Contour in the guise of what we labeled Concession Contour. This adjustment could, then, easily account for the use of the contour in the SAME condition, and also in the STRONG condition, on the additional assumption that asking about a stronger alternative implies doubt about a weaker alternative as well, although to a lesser degree, in line with the numerical difference in production rates. The account could even allow one to integrate the observation about the difference between Contradiction Contour and Concession Contour regarding pitch exaggeration, by correlating pitch height with the degree of contextual evidence: pitch will be higher overall in a contradictory use, since asserting p constitutes the maximal amount of contextual evidence against $\neg p$, whereas pitch is reduced in reply to a question, because $\neg p$ is, by definition, still at least a possibility.

However, it is unclear how this account could explain the increase in SI rate with the Concession Contour found in the experiments. Using a contour to communicate that there is evidence against p (for instance, *The winner was happy*) is, in principle, independent of the attitude the speaker has toward its strengthened interpretation. Moreover, and maybe more crucially, viewing the use of the Concession Contour in terms of contextual evidence fails to take into account properties of the lexical scales. But as **Figure 6** shows, there was substantial variation observed in production rates of the Concession Contour across scales – as with the case of the RFR. We, thus, conclude that a unified account of Concession Contour and Contradiction

Contour in terms of gradient contextual evidence faces serious challenges, and leave a more in-depth investigation into this issue and alternative possibilities for future research.

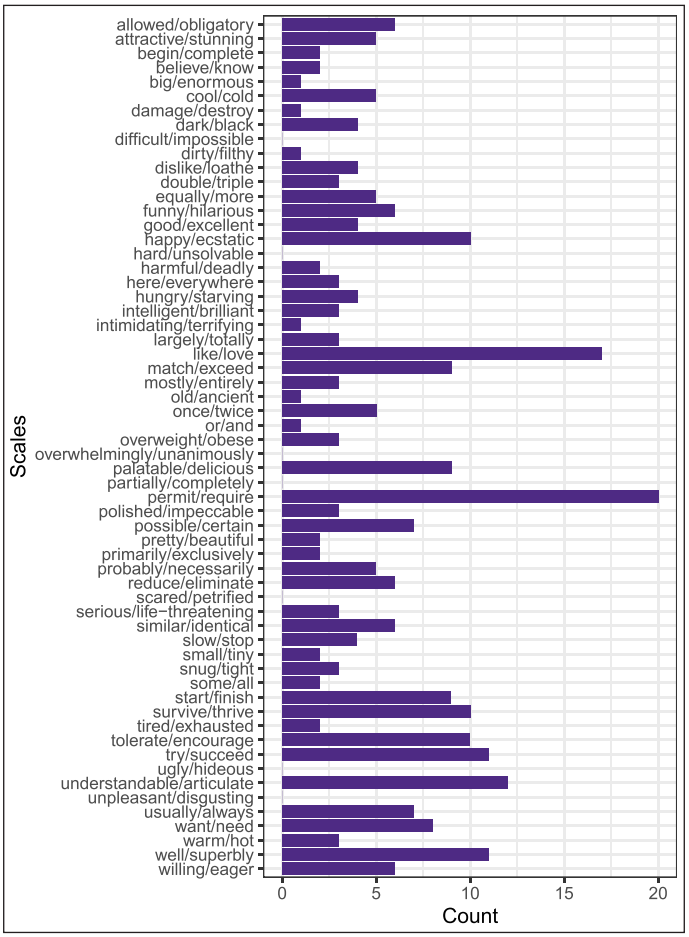

Figure 6: Production rates for Concession Contour by item in Experiment 1.

6. Conclusion

Recent research in experimental pragmatics has focused on the phenomenon of scalar diversity: the inter-scale non-uniformity of SI. While most existing studies in this domain have relied on written stimuli, the present article begins to probe the interplay of intonation and SI calculation across different scales. In two experiments, we tested the effect of different contours (either produced by participants or directly manipulated in the stimuli) on the robustness of SI calculation. We found that SI rates varied by contour, and, in particular, the so-called RFR contour made SIs more likely to arise. The production rate of RFR also varied across lexical scales. These findings point to the importance of considering intonation in studies of SI and scalar diversity. Further, they also inform theoretical treatments of the RFR, being most easily captured by accounts that link its felicity to the presence of a stronger alternative.

Data accessibility statement

Experimental files, results files and analysis code are stored in an OSF repository at https://osf.io/6a9wg/?view_only=e14a3c89b1474918b30c1d59c202fbff.

Ethics and consent

The studies reported in this paper were approved by the Princeton Institutional Review Board (#15015). Informed consent was obtained from all participants.

Acknowledgements

We are indebted to Emma Nguyen and Luke Adamson for providing audio stimuli, to Thomas Sostarics for help with the visualizations, as well as to Dan Goodhue, Sunwoo Jeong, Deniz Rudin, Michael Wagner, the UPenn Experimental Semantics Lab, the SALT 33 audience, and three anonymous reviewers for feedback. This material is partially based upon work supported by the National Science Foundation under Grant No. #BCS-2041312.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

The authors contributed equally to this work and are listed in reverse alphabetical order.

ORCiD IDs

Eszter Ronai: https://orcid.org/0000-0003-1578-0938

Alexander Göbel: https://orcid.org/0000-0002-7920-9071

References

Bader, M. (1998). Prosodic influences on reading syntactically ambiguous sentences. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 1–46). Kluwer. DOI: https://doi.org/10.1007/978-94-015-9070-9_1

Baker, R., Doran, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics*, *1*(2), 211–248. DOI: https://doi.org/10.1163/187730909X12538045489854

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199249633.003.0002

Beltrama, A., & Xiang, M. (2013). Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17* (pp. 81–98).

Buccola, B., & Goodhue, D. (to appear). The effect of intonation on scalar and ignorance inferences. *Proceedings of the 59th Annual Meeting of the Chicago Linguistic Society (CLS 59)*, https://ling.auf.net/lingbuzz/007464.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics and Philosophy*, *26*, 511–545. DOI: https://doi.org/10.1023/A:1025887707652

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology*, *61*(11), 1741–1760. DOI: https://doi.org/10.1080/17470210701712960

Constant, N. (2012). English rise-fall-rise: A study in the semantics and pragmatics of intonation. *Linguistics and Philosophy*, *35*, 407–442. DOI: https://doi.org/10.1007/s10988-012-9121-1

Cummins, C., & Katsos, N. (2019). 1. Introduction. In *The Oxford handbook of experimental semantics and pragmatics*. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780198791768.013.33

Cummins, C., & Rohde, H. (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology*, *6*, 1779. DOI: https://doi.org/10.3389/fpsyg.2015.01779

de Marneffe, M.-C., & Tonhauser, J. (2019). Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. In E. Onea, M. Zimmermann, & K. von Heusinger (Eds.), *Questions in discourse* (pp. 132–163). Brill. DOI: https://doi.org/10.1163/9789004378322_006

Degen, J. (2013). *Alternatives in pragmatic reasoning* [Doctoral dissertation, University of Rochester].

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710. DOI: https://doi.org/10.1111/cogs.12171

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, *88*(1), 124–154. DOI: https://doi.org/10.1353/lan.2012.0008

Farkas, D. F., & Roelofsen, F. (2017). Division of labor in the interpretation of declaratives and interrogatives. *Journal of Semantics*, *34*, 237–289. DOI: https://doi.org/10.1093/jos/ffw012

Fodor, J. D. (2002). Prosodic disambiguation in silent reading. *Proceedings of NELS*, *32*, 113–137.

Frazier, L., & Gibson, E. (2015). *Explicit and implicit prosody in sentence processing: Studies in honor of Janet Dean Fodor*. Springer. DOI: https://doi.org/10.1007/978-3-319-12961-7

Göbel, A. (2019). Additives pitching in: L* + h signals ordered focus alternatives. *Proceedings of SALT 29*, 279–299. DOI: https://doi.org/10.3765/salt.v29i0.4612

Göbel, A., & Wagner, M. (2023). On a concessive reading of the rise-fall-rise contour: Contextual and semantic factors. *Proceedings of ELM 2*, 83–94. DOI: https://doi.org/10.3765/elm.2.5395

Goodhue, D., Harrison, L., Su, Y. T. C., & Wagner, M. (2016). Toward a bestiary of English intonational contours. In B. Prickett & C. Hammerly (Eds.), *Proceedings of the North East Linguistics Society 46* (pp. 311–320).

Goodhue, D., & Wagner, M. (2018). Intonation, 'yes' and 'no'. *Glossa: A Journal of General Linguistics*, *3*, 1–45. DOI: https://doi.org/10.5334/gjgl.210

Gotzner, N. (2019). The role of focus intonation in implicature computation: A comparison with 'only' and 'also'. *Natural Language Semantics*, *27*, 189–226. DOI: https://doi.org/10.1007/s11050-019-09154-7

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*, 1659. DOI: https://doi.org/10.3389/fpsyg.2018.01659

Grice, H. P. (1967). Logic and conversation. In P. Grice (Ed.), *Studies in the way of words* (pp. 41–58). Harvard University Press. DOI: https://doi.org/10.1163/9789004368811_003

Gualmini, A., Hulsey, S., Hacquard, V., & Fox, D. (2008). The Question-Answer Requirement for scope assignment. *Natural Language Semantics*, *16*(3), 205–237. DOI: https://doi.org/10.1007/s11050-008-9029-z

Gunlogson, C. (2001). *True to form: Rising and falling declaratives as questions in English* [Doctoral dissertation, University of California, Santa Cruz].

Heim, I. (1991). Artikel und Definitheit. In A. v. Stechow & D. Wunderlich (Eds.), *Handbuch der Semantik* (pp. 487–535). de Gruyter. DOI: https://doi.org/10.1515/9783110126969.7.487

Hirschberg, J. B. (1985). *A theory of scalar implicature* [Doctoral dissertation, University of Pennsylvania].

Höhle, T. N. (1992). Über Verum-Fokus im Deutschen. In J. Jacobs (Ed.), *Informationsstruktur und Grammatik* (pp. 112–141). Opladen. DOI: https://doi.org/10.1007/978-3-663-12176-3_5

Horn, L. R. (1972). *On the semantic properties of logical operators in English* [Doctoral dissertation, UCLA].

Hu, J., Levy, R., Degen, J., & Schuster, S. (2023). Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, *11*, 885–901. DOI: https://doi.org/10.1162/tacl_a_00579

Hu, J., Levy, R., & Schuster, S. (2022). Predicting scalar diversity with context-driven uncertainty over alternatives. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 68–74. DOI: https://doi.org/10.18653/v1/2022.cmcl-1.8

Hulsey, S., Hacquard, V., Fox, D., & Gualmini, A. (2004). The Question-Answer Requirement and scope assignment. In A. Csirmaz, A. Gualmini, & A. Nevins (Eds.), *MIT working papers in linguistics* (pp. 71–90). MITWPL.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*, 243–276. DOI: https://doi.org/10.1556/ALing.55.2008.3-4.2

Kursat, L., & Degen, J. (2020). Probability and processing speed of scalar inferences is context-dependent. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1236–1242). Cognitive Science Society.

Ladd, R. D. (2008). *Intonational phonology*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511808814

Liberman, M., & Sag, I. (1974). Prosodic form and discourse function. *Proceedings of CLS*, *10*, 416–427.

Lohnstein, H. (2015). Verum focus. In C. Féry & S. Ishihara (Eds.), *The Oxford handbook of information structure* (pp. 290–313). Oxford Academic. DOI: https://doi.org/10.1093/oxfordhb/9780199642670.013.33

Pankratz, E., & van Tiel, B. (2021). The role of relevance for scalar diversity: A usage-based approach. *Language and Cognition*, *13*(4), 562–594. DOI: https://doi.org/10.1017/langcog.2021.13

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* [Doctoral dissertation, MIT].

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics [Earlier version appeared in OSU Working Papers in Linguistics 49 in 1996]. *Semantics and Pragmatics*, *5*, 1–69. DOI: https://doi.org/10.3765/sp.5.6

Ronai, E., & Xiang, M. (2021a). Exploring the connection between Question Under Discussion and scalar diversity. *Proceedings of the Linguistic Society of America*, *6*(1), 649–662. DOI: https://doi.org/10.3765/plsa.v6i1.5001

Ronai, E., & Xiang, M. (2021b). Pragmatic inferences are QUD-sensitive: An experimental study. *Journal of Linguistics*, *57*(4), 841–870. DOI: https://doi.org/10.1017/S0022226720000389

Ronai, E., & Xiang, M. (2022). Three factors in explaining scalar diversity. *Proceedings of Sinn und Bedeutung*, *26*, 716–733.

Ronai, E., & Xiang, M. (2024). What could have been said? Alternatives and variability in pragmatic inferences. *Journal of Memory and Language*, *136*, 104507. DOI: https://doi.org/10.1016/j.jml.2024.104507

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, *1*, 75–116. DOI: https://doi.org/10.1007/BF02342617

Sandberg, K., & Cole, J. (2022). *The role of duration in signaling scalar alternative sets* [Poster presented at the 9th Experimental Pragmatics Conference]. https://sites.northwestern.edu/katesandberg/files/2022/09/XPrag-Pres_final.pptx

Schwarz, F., Clifton, C., Jr., & Frazier, L. (2007). Strengthening 'or': Effects of focus and downward entailing contexts on scalar implicatures. *University of Massachusetts Occasional Papers in Linguistics*, *33*(1), 9.

Simons, M., & Warren, T. (2018). A closer look at strengthened readings of scalars. *Quarterly Journal of Experimental Psychology*, *71*(1), 272–279. DOI: https://doi.org/10.1080/17470218.2017.1314516

Solt, S. (2015). Measurement scales in natural language. *Language and Linguistics Compass*, *9*, 14–32. DOI: https://doi.org/10.1111/lnc3.12101

Sun, C., Tian, Y., & Breheny, R. (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology*, *9*. DOI: https://doi.org/10.3389/fpsyg.2018.02092

Sun, C., Tian, Y., & Breheny, R. (2023). A corpus-based examination of scalar diversity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. DOI: https://doi.org/10.1037/xlm0001278

Tomlinson, J., Gotzner, N., & Bott, L. (2017). Intonation and pragmatic enrichment: How intonation constrains ad hoc scalar inferences [PMID: 28697695]. *Language and Speech*, *60*(2), 200–223. DOI: https://doi.org/10.1177/0023830917716101

Tomlinson, J., & Ronderos, C. R. (2021). Does intonation automatically strengthen scalar implicatures? *Semantics and Pragmatics*, *14*(4), 1–30. DOI: https://doi.org/10.3765/sp.14.4

van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*(1), 137–175. DOI: https://doi.org/10.1093/jos/ffu017

Wagner, M. (2012). Contrastive topics decomposed. *Semantics and Pragmatics*, *5*, 1–54. DOI: https://doi.org/10.3765/sp.5.8

Wagner, M., McClay, E., & Mak, L. (2013). Incomplete answers and the rise-fall-rise contour. In R. Fernández & A. Isard (Eds.), *Proceedings of the 17th workshop on the semantics and pragmatics of dialogue* (pp. 140–149). SEMDIAL.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, *61*, 747–776. DOI: https://doi.org/10.2307/414489

Westera, M. (2017). *Exhaustivity and intonation: A unified theory* [Doctoral dissertation, University of Amsterdam].

Westera, M. (2019). Rise-fall-rise as a marker of secondary QUDs. In D. Gutzmann & K. Turgay (Eds.), *Secondary content: The linguistics of side issues* (pp. 376–404). Brill. DOI: https://doi.org/10.1163/9789004393127_015

Westera, M., & Boleda, G. (2020). A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung*, *24*(2), 439–454. DOI: https://doi.org/10.18148/sub/2020.v24i2.908

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology*, *9*, 1720. DOI: https://doi.org/10.3389/fpsyg.2018.01720

Zondervan, A. (2010). *Scalar implicatures or focus: An experimental approach* [Doctoral dissertation, University of Amsterdam].

Zondervan, A., Meroni, L., & Gualmini, A. (2008). Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In T. Friedman & S. Ito (Eds.), *Proceedings of SALT 28* (pp. 765–777). Cornell University. DOI: https://doi.org/10.3765/salt.v18i0.2486