Bilingualism: Language and Cognition

cambridge.org/bil

Scalar diversity and second-language processing of scalar inferences: A cross-linguistic analysis

Greta Mazzaggio[1,2], Federica Longo[2,3], Penka Stateva[2] and Bob van Tiel[4]

[1]Department of Humanities, University of Florence, Florence, Italy; [2]Center for Cognitive Science of Language, University of Nova Gorica, Nova Gorica, Slovenia; [3]Department of Cognitive Sciences, Psychology, Education, and Cultural Studies, University of Messina, Messina, Italy and [4]Department of Philosophy, Theology, and Religious Studies, Radboud University, Nijmegen, The Netherlands

Research Article

Cite this article: Mazzaggio, G., Longo, F., Stateva, P. and van Tiel, B. (2025). Scalar diversity and second-language processing of scalar inferences: A cross-linguistic analysis. *Bilingualism: Language and Cognition* 1–13. https://doi.org/10.1017/S1366728925000392

Received: 14 November 2023
Revised: 19 March 2025
Accepted: 27 March 2025

Keywords:
scalar inference; conversational implicature; scalar diversity; second language; linguistic transfer; pragmatics

Corresponding author:
Greta Mazzaggio;
Email: greta.mazzaggio@unifi.it

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Abstract

We investigate the processing of scalar inferences in first language (L1) and second language (L2). Expanding beyond the common focus on the scalar inference from 'some' to 'not all', we examine six scalar expressions: 'low', 'scarce', 'might', 'some', 'most' and 'try'. An online sentence-picture verification task was used to measure the frequency and time course of scalar inferences for these expressions. Participants included native English speakers, native Slovenian speakers and Slovenian speakers who spoke English as their L2. The first two groups were tested in their L1, while the third group was tested in their L2. Results showed that the English-L2 group resembled the Slovenian-L1 group more than the English-L1 group in terms of inference frequency. The time course for scalar inference computation was similar across all groups. These findings suggest subtle pragmatic transfer effects from L1 to L2, varying across different scalar expressions.

Highlights

- We experimentally tested scalar inference processing in second language (L2)-English speakers.
- We compared results with first language (L1)-English and L1-Slovenian speakers.
- Cross-linguistic and cross-scalar differences in scalar inference rates were found.
- English-L2 speakers relied on their L1-Slovenian, suggesting pragmatic transfer.

1. Introduction

A speaker who says (1) may imply that not all actors are famous.

(1) Some actors are famous.

This type of inference is called a *scalar inference*. Scalar inferences are called so because they are associated with *lexical scales*. Lexical scales are sets of expressions that are ordered in terms of logical strength, for example, <some, all>. A speaker who utters a positive sentence containing the weaker expression on a scale (e.g., 'some') may imply that the corresponding sentence with the stronger expression (e.g., 'all') is false (e.g., Gazdar, 1979; Geurts, 2010; Horn, 1972).

Although much of the research on scalar inferences has focused on the inference from 'some' to 'not all', the class of lexical scales is highly diverse. To illustrate, Table 1 provides a representative sample.

Scalar inferences are typically explained as a variety of *conversational implicature*, that is, as inferences that can be calculated on the basis of the literal meaning of the utterance and the assumption that the speaker is *cooperative* (Grice, 1975). In the case of (1), the literal meaning can be paraphrased as in (2).

(2) At least some, and possibly all, actors are famous.

So, according to its literal meaning, (1) is compatible with a situation in which all actors are famous. However, if the speaker believed that this situation obtained, it would have been more informative, and hence cooperative, for them to say (3) instead of (1).

(3) All actors are famous.

Given that the speaker did not produce this more informative alternative, it may be inferred that, according to the speaker, not all actors are famous. When we combine this scalar inference

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Table 1. Sample of lexical scales (from Pankratz & van Tiel, 2021)

Adjective	<good, excellent> <big, huge>	<small, tiny> <dry, arid>	<warm, hot> <pale, white>
Noun	<shock, disbelief> <envy, resentment>	<respect, reverence> <discomfort, illness>	<solace, safety> <precision, rigidity>
Verb	<equal, surpass> <seek, obtain>	<spark, ignite> <discomfort, illness>	<request, require> <imply, state>

with the literal meaning of the utterance, we arrive at the *pragmatically enriched* meaning, which is paraphrased in (4).

(4) At least some, but not all, actors are famous.

In this aetiological sketch, the literal meaning precedes the pragmatically enriched meaning, in the sense that the literal meaning functions as a premise in the reasoning process that, ultimately, results in the pragmatically enriched meaning.[1] A key question in experimental pragmatics is whether the theoretical priority of the literal meaning is reflected in human psychology, that is, whether deriving scalar inferences is associated with a cognitive cost (e.g., Geurts & Rubio-Fernández, 2015; Noveck, 2018; Recanati, 1995).

On the one hand, proponents of *relevance theory* argue that, in settings in which there is no strong support for scalar inferences, there should indeed be a parallelism between theory and psychology (e.g., Noveck & Sperber, 2007; Sperber & Wilson, 1995). According to relevance theory, hearers attempt to piece together the speaker's intention based on the literal meaning of an utterance and its surrounding context and based on the expectation that the utterance is optimally relevant. Generally (i.e., unless there is strong contextual evidence to the contrary) the literal meaning of the utterance is a good first cue to the speaker's intention. Only if the hearer is dissatisfied with the relevance of this literal interpretation will they decide to compute the scalar inference. According to relevance theory, this process of meaning construction is cognitively taxing and time-consuming.

By contrast, Levinson (2000) argues that the pragmatically enriched meaning should be easier to retrieve than the literal meaning. Levinson's proposal builds on the observation that human communication has a comparatively slow information transmission rate because of the time needed for phonetic articulation (i.e., we can only talk so fast). One way of reducing this articulatory bottleneck is by integrating generalised conversational implicatures – such as scalar inferences – into the lexical meaning. Thus, according to Levinson, sentences such as (1) receive a scalar inference by default, though this inference can be overridden (e.g., by continuing with 'In fact, all actors are famous').

More recent proposals have tried to find middle ground between these two approaches by arguing that the presence or absence of a processing cost for deriving scalar inferences depends on various factors, including the question under discussion (Ronai & Xiang, 2021), the structural characteristics of the alternatives (Chemla & Bott, 2014; van Tiel & Schaeken, 2017), the naturalness of the utterance (Degen & Tanenhaus, 2015) and the polarity of the scalar expression (e.g., van Tiel & Pankratz, 2021; van Tiel, Pankratz, & Sun, 2019).

[1]Note that other theories of scalar inference exist, such as the grammatical theory attributing their derivation to an exhaustivity operator constrained by scalar alternatives, but these are beyond the scope of the present work (cf. Chierchia et al., 2012; Fox, 2007).

One of the ways in which the predictions of these proposals have been tested is by comparing the derivation of scalar inferences in first language (L1) and second language (L2). It is well-known that processing L2 input draws upon more cognitive resources than processing L1 input (e.g., Clahsen & Felser, 2006; Juffs, 2001; White & Juffs, 1998). Given this observation, one may formulate two predictions about the derivation of scalar inferences in L1 and L2.

i. *Frequency*: If the derivation of scalar inferences is cognitively costly, people may be less likely to derive scalar inferences in L2 when compared to L1, since, in the former case, they may not have sufficient cognitive resources at their disposal to derive the scalar inference and thus may resort to the literal meaning.
ii. *Time course*: If the derivation of scalar inferences is cognitively costly, people may take longer to derive scalar inferences in L2 when compared to L1, since, in the former case, they have fewer cognitive resources at their disposal, which may increase the time needed for deriving scalar inferences.

In the upcoming two sections, we describe previous studies that have tested these two predictions. As we will see, these studies provide tentative evidence in support of both predictions. Afterwards, we discuss an important limitation of the current state of the art, namely that it has concentrated exclusively on the scalar inference from 'some' to 'not all'. Over the past years, it has become increasingly clear that findings for 'some' do not always generalise to other scalar expressions (e.g., van Tiel et al., 2016). Hence, we carried out an experiment in which we tested the derivation and processing of six different types of scalar inferences in L1 and L2, focusing on native speakers of Slovenian who also spoke English as L2. As we will see, the results of our study speak against both aforementioned predictions and suggest effects of pragmatic transfer between L1 and L2 (e.g., Bou-Franch, 1998).

1.1. Prior research on the frequency of scalar inferences in L1 and L2

Slabakova (2010) was the first to study how frequently people derived scalar inferences in L1 and L2. She tested three groups of participants: (i) L1 speakers of English, (ii) L1 speakers of Korean and (iii) L1 speakers of Korean who spoke English as their L2. The first two groups were tested in their L1; the third group in their L2.

To test the frequency of deriving scalar inferences, Slabakova carried out a sentence verification task borrowed from Noveck (2001). In this task, participants had to evaluate whether they agreed or disagreed with certain sentences. The target condition consisted of underinformative sentences with 'some', such as (5).

(5) Some elephants have trunks.

Such underinformative sentences are true according to their literal meaning, since there are elephants that have trunks. However, the corresponding scalar inference, which states that not all elephants have trunks, is false. Hence, the proportion of 'disagree' responses provides a measure of the frequency with which scalar inferences were derived.

Comparing the three language groups, Slabakova found that participants were significantly more likely to reject underinformative sentences with 'some' in L2 than L1. This finding goes counter to the relevance-theoretic idea that the derivation of scalar inferences is cognitively costly. Instead, Slabakova's findings appear to support Levinson's idea that scalar inferences are default inferences and that L2 speakers had fewer cognitive resources at their disposal

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

to overturn these defaults to arrive at the literal meaning of the sentence.

Although Slabakova's findings are compelling, they have not been replicated since. Instead, most studies have found no significant difference in the rates of scalar inferences in L1 and L2, neither for adults (Dupuy et al., 2019; Feng & Cho, 2019; Snape & Hosoi, 2018) nor for children (Antoniou et al., 2020; Antoniou & Katsos, 2017). Mazzaggio et al. (2021) even found that adults are *less* likely to derive scalar inferences in L2 than L1.

In a recent study, Khorsheed and van Tiel (2024) offer an explanation for these conflicting findings. Their explanation rests on the assumption that people indeed have difficulties deriving scalar inferences in L2. However, Khorsheed and van Tiel propose that these difficulties may be masked by the use of experimental tasks that allow participants to process target sentences at their leisure. Thus, Slabakova – who found no evidence that deriving scalar inferences in L2 is cognitively costly – presented the experimental sentences in a self-paced pen-and-paper questionnaire, whereas Mazzaggio and colleagues – who found lower rates of scalar inferences in L2 – presented them auditorily and forced participants to respond within three seconds after sentence offset. Khorsheed and van Tiel argue that pen-and-paper questionnaires, such as the one used by Slabakova, allow L2 speakers to compensate for their difficulties with deriving scalar inferences by taking more time and effort to process the sentences (see also Ellis, 2009; Hopp, 2022, for discussion of methodological effects on research in L2).

In support of their proposal, Khorsheed and van Tiel carried out two sentence verification tasks similar to the one used by Slabakova, testing participants in their L1, Malay, and in their L2, English. The first experiment used a self-paced pen-and-paper questionnaire. In the second experiment, the sentences were flashed on screen, one word at a time, and participants were instructed to respond as quickly as possible. In the first experiment, Khorsheed and van Tiel found that participants were equally likely to derive scalar inferences in L1 and L2. However, in the second experiment, low-proficiency L2 speakers were significantly less likely to derive scalar inferences in L2 than L1.

Taken together, and with the exception of the outlying study by Slabakova, the current experimental record thus suggests that people are less likely to derive scalar inferences in L2 than L1, but that this effect may be mitigated by the use of experimental tasks in which participants can take their time to process and evaluate the relevant sentences. This conclusion ties in with relevance theory, since it suggests that the derivation of scalar inferences is cognitively costly.

1.2. Prior research on the time course of scalar inferences in L1 and L2

A recurrent finding in research on scalar inferences in L1 is that their derivation is *time-consuming* (e.g., Bott & Noveck, 2004; Chemla & Bott, 2014; Cremers & Chemla, 2014; van Tiel, Pankratz, & Sun, 2019; van Tiel & Schaeken, 2017). Thus, Bott and Noveck (2004) carried out a sentence verification task in which participants were presented with underinformative sentences with 'some', such as (6).

(6) Some dogs are mammals.

In their Experiment 3, participants could respond ad libitum. In this experiment, Bott and Noveck found that participants were roughly equally likely to accept or reject such sentences. However, looking at response times, Bott and Noveck found that participants took significantly longer to reject underinformative sentences such as (6) than to accept them. This difference in response times was absent for control sentences that were unambiguously true or false, such as those in (7).

(7) a. Some mammals are dogs. (True)
b. Some dogs are insects. (False)

Bott and Noveck thus observed an interaction effect on response times between condition (target versus control) and response ('true' versus 'false'). Following van Tiel, Pankratz, and Sun (2019), we will call this interaction effect the Bott and Noveck (B&N) effect.

Given that processing L2 input is cognitively demanding, one might expect that the size of the B&N effect is modulated by proficiency. Khorsheed et al. (2022) provide evidence in support of this hypothesis. In particular, Khorsheed and colleagues replicated B&N sentence verification task in English, but with L2 speakers. Khorsheed and colleagues found that the B&N effect emerged for all participants irrespective of their proficiency. However, the effect was significantly more pronounced for less proficient speakers than for more proficient speakers. Recently, Khorsheed and van Tiel (2024) failed to replicate this finding, though they also tested considerably fewer participants (213 against 110 participants).

Taken together, the current experimental record tentatively suggests that proficiency is negatively correlated with the size of the B&N effect. Again, this conclusion ties in with relevance theory, because it suggests that deriving scalar inferences is cognitively costly and that this cognitive cost is amplified when processing input from a language in which one is not particularly fluent.

1.3. Our study

In summary, research on scalar inferences in L2 has broadly confirmed the relevance-theoretic account, according to which the derivation of scalar inferences in out-of-the-blue contexts is cognitively costly. At the same time, the current experimental record has an important limitation, which is that studies on the topic have mainly focused on the scalar inference from 'some' to 'not all', even though they purport to provide information about the mechanism of scalar inferencing in general. Apparently, these studies assume that the <some, all> scale is representative for the entire family of lexical scales.

Recent studies have provided overwhelming evidence against this uniformity assumption (e.g., Hu et al., 2023; Ronai & Xiang, 2022; van Tiel et al., 2016). For example, van Tiel, Pankratz, and Sun (2019) showed that the B&N effect that was observed for 'some' did not consistently generalise to other scalar expressions. This observation of scalar diversity raises an important question: To what extent do the findings about the frequency and time course of scalar inferences in L2 generalise to scalar expressions other than 'some'?

To address this question, we tested three groups of participants: (i) L1 speakers of English, (ii) L1 speakers of Slovenian and (iii) L1 speakers of Slovenian who spoke English as their L2. The first two groups were tested in their L1; the third group in their L2.

To study the derivation of scalar inferences, we carried out a sentence-picture verification task. In this task, participants saw a sentence that was followed by a picture, and they had to indicate whether the sentence was a good or bad description of the corresponding picture. In the control condition, the sentence was unambiguously true or false. In the target condition, the sentence was literally true, but the corresponding inference was false. Crucially,

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

while we thus tested the inference from 'some' to 'not all', we also tested five other types of scalar inferences. Table 1 provides an overview of the sentences and pictures that were tested. These materials were borrowed from van Tiel, Pankratz, and Sun (2019).

The proportion of 'false' responses in the target condition provides a measure of the frequency with which participants derived scalar inferences. In addition, we computed participants' response times to measure the time course of deriving scalar inferences.

Focusing on 'some', we expect to replicate the current state of the art. Recall that, in terms of frequency, it was found that people are less likely to derive scalar inferences in L2 than L1, unless they can take their time to process and evaluate the sentences. In our task, the sentence disappeared before the picture was shown, so that participants had to keep the sentence in memory when they determined its adequacy as a description of the picture. In addition, participants were instructed to respond as quickly as possible. Hence, we expect fewer scalar inferences, that is, fewer 'false' responses in the target condition, in L2 than L1. In terms of time course, we expect that the B&N effect associated with the scalar inference of 'some' is significantly more pronounced in L2 than L1, based upon the results of Khorsheed et al. (2022).

Focusing on the other scalar expressions, we distinguish between two possibilities. On the one hand, it may be the case that the uniformity assumption is correct and that other scalar expressions pattern with 'some' in that they (i) lead to fewer scalar inferences in L2 than L1 and (ii) are associated with a significantly greater B&N effect in L2 than L1. On the other hand, given earlier observations of scalar diversity, it may turn out that the effects of proficiency (i.e., L1 versus L2) are not uniform across scalar expressions.

One potential source of such scalar diversity lies in the role of *pragmatic transfer*. Pragmatic transfer refers to the observation that L2 speakers frequently apply pragmatic regularities from their L1 to their L2, even if these regularities do not correspond to the way that L1 speakers behave (e.g., Bou-Franch, 1998; Kasper, 1992; Kecskes, 2015; Mazzaggio & Stateva, 2024). In the case of scalar inferences, pragmatic transfer effects may emerge when L2 speakers assume that certain varieties of scalar inferences are equally robust – or equally fickle – in their L2 as in their L1.

Of course, such pragmatic transfer effects will only be visible if there are differences in the rates of scalar inferences between languages (e.g., between English and Slovenian in our study). While it has sometimes been claimed that scalar inference rates are universal (e.g., Slabakova, 2010, p. 2446), there is at least some evidence of cross-linguistic variability in the rates of scalar inferences. For example, Katsos et al. (2016) found considerable cross-linguistic variability in the percentage of children that rejected underinformative sentences with 'some', ranging from 13.5% in Malay to 91.0% in Russian (see also Dionne & Coppock, 2022; Stateva et al., 2019).

Our experiment will, first of all, contribute to our relatively limited knowledge of cross-linguistic variability in the rates of scalar inferences. In addition, and perhaps more importantly, if we find such variability, our experiment will indicate to what extent the judgements that L2 speakers of English provide resemble those of L1 speakers of Slovenian (which would suggest effects of pragmatic transfer) or those of L1 speakers of English.

The particular choice of Slovenian-English as our language pair was motivated by prior research suggesting differences in the interpretation of scalar expressions between these two languages. In particular, Stateva et al. (2019) found that the quantifier 'some' and its Slovenian equivalent 'nekaj' exhibit subtle differences in meaning and usage, which makes this pairing an informative test case for investigating whether the derivation of scalar inferences in L2 aligns with native-like patterns in the L2 or exhibits transfer effects from L1. Given that most prior studies on scalar inference derivation in L2 have focused on more widely spoken language pairs, this study also contributes to expanding the empirical scope of research on scalar inferences in L2 acquisition.

In the next section, we describe our experiment and present the results.

2. Methods

2.1. Participants

A total of 181 participants took part in our experiment. Participants were divided into three groups: 61 participants were L1 speakers of English and took the experiment in English (mean age: 30.6, standard deviation: 5.6, 46 females), 60 participants were L1 speakers of Slovenian and took the experiment in Slovenian (mean age: 24.2, standard deviation: 3.9, 27 females), and 60 participants were L1 speakers of Slovenian and took the experiment in English, which was their L2 (mean age: 24.4, standard deviation: 4.5, 27 females). Participants were recruited on prolific and were paid £2.25 for their participation.

Participants in the English-L2 group were asked to indicate their proficiency in English in terms of the Common European Framework for Languages. Participants indicated that their level was A1/A2 (N = 1), B1 (N = 5), B2 (N = 14), C1 (N = 26) and C2 (N = 11).

2.2. Materials

The experiment tested six lexical scales: <low, empty>, <might, will>, <most, all>, <scarce, absent>, <some, all> and <try, succeed>. For each lexical scale, we constructed a positive sentence containing the weaker expression on the scale. Each of these sentences was paired with three pictures (see Figure 1). In one picture, the sentence was unambiguously true (the control-true condition); in one picture, it was unambiguously false (the control-false condition); and in one picture, the sentence was literally true but its scalar inference was false (the target condition). There were three minimally distinct versions of each of the three pictures associated with a sentence, which leads to nine items per sentence. In total, the experiment thus consisted of 6 sentences × 9 pictures = 54 items. The order of items was randomised for each participant.

The sentences and pictures were the same as those tested by van Tiel, Pankratz, and Sun (2019), except for the fact that, in their study, they also tested the <or, and> scale. We decided to omit this scale because the target sentence for this scale, that is, 'Either the apple or the pepper is red', could not straightforwardly be translated into Slovenian. In particular, in Slovenian, the dual form is used when the grammatical subject describes exactly two elements, which was the case in the target condition in which both the apple and the pepper were red. However, the dual form would be syntactically awkward in the control condition in which only one of the objects was red. For this reason, we did not include the <or, and> scale in the experiment.

2.3. Procedure

Each trial started with the presentation of the sentence. After pressing the space bar, the sentence disappeared and was replaced by a picture. Participants were instructed to indicate as quickly as possible whether the sentence was a good or bad description of the

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

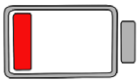
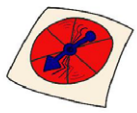
Sentence	Control-True	Control-False	Target
Eng: The battery is low *Slo:* Baterije zmanjkuje			
Eng: Red flowers are scarce *Slo:* Rdeče rože so redke			
Eng: The arrow might land on red *Slo:* Puščica lahko pristane na rdečem			
Eng: Some of the socks are pink *Slo:* Nekatere nogavice so roza			
Eng: Most of the apples are green *Slo:* Ve čina jabolk je zelenih			
Eng: He tried to tie his tie *Slo:* Poskusil si je zavezati kravato			

Figure 1. Pictures tested in our sentence-picture verification task (based on van Tiel, Pankratz, & Sun, 2019) along with the English (*Eng*) and Slovenian (*Slo*) experimental sentences.

corresponding picture. They had to press '1' on the keyboard if they thought the sentence was a good description; otherwise, they had to press '0'. After registering their decision, the picture disappeared and the message 'Press the space bar to continue' was shown. After pressing the space bar, the next trial started.

After finishing the sentence-picture verification task, participants in the English-L2 group were asked to translate the English sentences that were used into their native language, Slovenian. This step served to filter out participants who had not correctly understood the sentences. However, all of the participants in the English-L2 group correctly translated the sentences, so no participants were excluded based on this criterion. The entire experiment took approximately 10-15 minutes to complete.

3. Results

3.1. Data treatment

Six participants (two in the English-L1 group, three in the English-L2 group, and one in the Slovenian-L1 group) were removed from the analysis because they made mistakes in more than 20% of the control items. A total of 175 participants were thus included in the analysis. We also removed trials with a response time below 200 milliseconds or above 15 seconds (22 trials). These exclusion criteria were the same as those used by van Tiel, Pankratz, and Sun (2019).

3.2. Choice proportions

First, we analysed the responses that participants provided. Figure 2 shows the percentage of 'true' responses for each scalar expression, language group and condition. For convenience, Figure 3 also shows the percentage of 'true' responses in the target condition for each scalar expression and language group.

To investigate whether the rates of scalar inferences differed across the three language groups, we analysed the proportion of 'true' responses in the target condition. Thus, we constructed a binomial generalised linear mixed-effects model using the 'glmer' function (Bates et al., 2015) in R (R Core Team, 2023). The model predicted responses in the target condition ('true' or 'false') on the basis of scalar expression ('low', 'scarce', 'might', 'some', 'most' or 'try'), language group (English-L1, English-L2, or Slovenian-L1) and their interaction, including random intercepts for participants. For this analysis, all factors were sum-coded.

To estimate the overall significance of the fixed factors, we carried out likelihood ratio tests using the 'anova' function in R. This analysis showed that scalar expression had a significant effect on responses ($\chi^2(5) = 231.3$, $p < .001$), while language group did not have a significant effect ($\chi^2(2) = 0.2$, $p = .92$). There was also a significant interaction between scalar expression and language group ($\chi^2(10) = 109.1$, $p < .001$).

These results indicate that scalar expressions vary in the probability with which they give rise to scalar inferences, which is in line with earlier observations of scalar diversity (e.g., van Tiel, Pankratz, & Sun, 2019). By contrast, there was no significant effect of language group, indicating that overall, there was no evidence that the language groups varied in the probability of deriving scalar inferences. However, both of these effects should be interpreted with caution, since the effect of scalar expression varied across language groups.

To further investigate the significant interaction between scalar expression and language group, we fitted, for each scalar expression

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

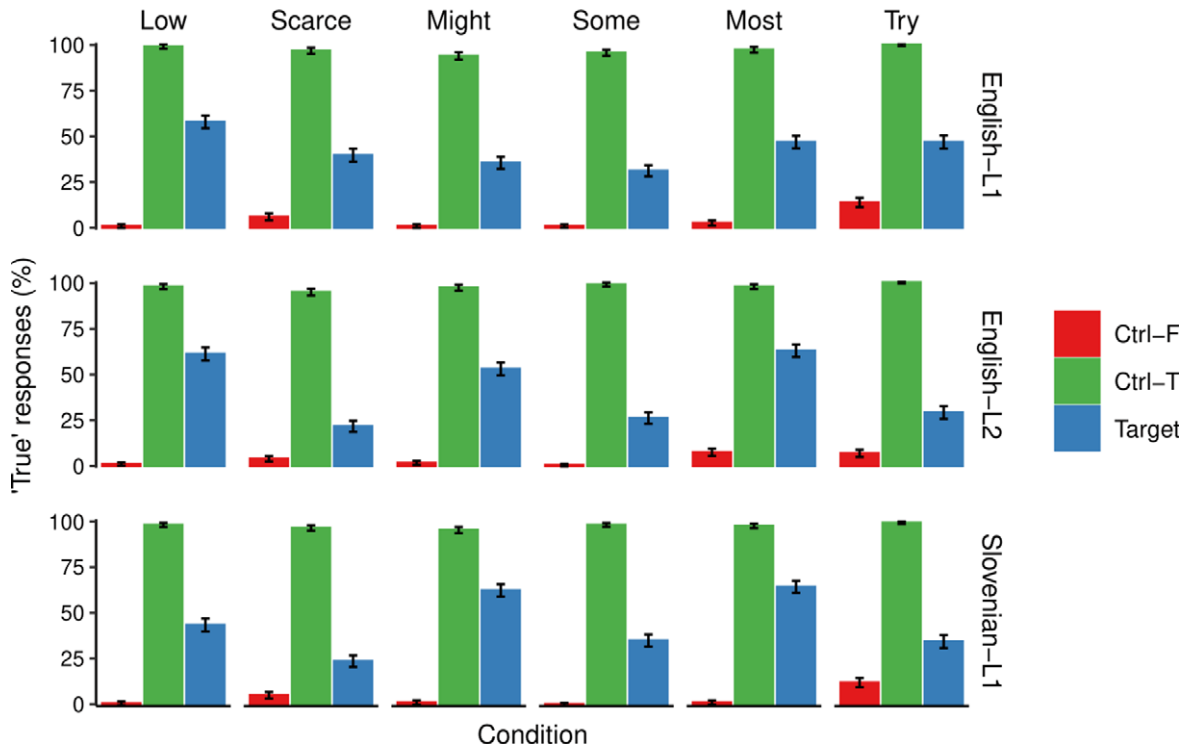


Figure 2. Percentage of 'true' responses for each scalar term, language group and condition. Error bars represent standard errors of the mean.

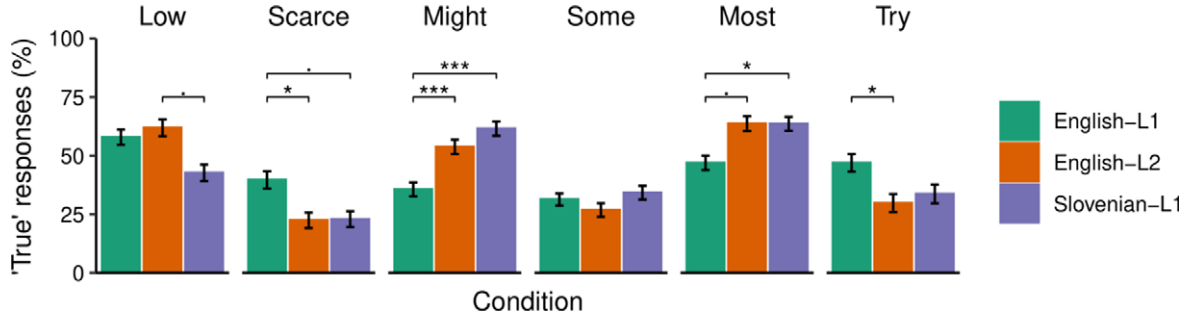


Figure 3. Percentage of 'true' responses for each scalar term and language group. Error bars represent standard errors of the mean. Brackets indicate whether the difference in means was significant at different alpha levels: . = .10, * = .05, ** = .01 and *** = .001.

separately, a binomial generalised linear mixed-effects model. These models predicted responses in the target condition ('true' or 'false') on the basis of language group (English-L1, English-L2 or Slovenian-L1). Using these models, the three language groups were pairwise compared using Tukey's procedure, as implemented in the 'glht()' function of the 'multcomp' package (Hothorn et al., 2008).

First, we compared the proportion of 'true' responses in the target condition in the English-L1 and Slovenian-L1 language groups, to determine whether scalar inferences were equally frequent in the two L1 groups. In contrast to this hypothesis, we observed that the proportion of 'true' responses in the Slovenian-L1 condition was significantly higher than in the English-L1 group for 'might' ($\beta = 16.4$, $SE = 1.9$, $Z = 8.8$, $p < .001$) and 'most' ($\beta = 2.3$, $SE = 1.0$, $Z = 2.4$, $p = .05$). It was marginally lower than in the English-L1 group for 'scarce' ($\beta = -1.6$, $SE = 0.7$, $Z = -2.2$, $p = .07$). All other comparisons were not significant. Hence, there is a significant variability in the frequency with which scalar inferences are derived in English and Slovenian. This raises the question as to whether the English-L2 group patterns with the English-L1 group or with the Slovenian-L1 group. The latter possibility would suggest effects of pragmatic transfer between L1 and L2.

Using the same analyses, we found that the proportion of 'true' responses in the English-L2 condition was marginally higher than in the Slovenian-L1 group for 'low' ($\beta = 2.1$, $SE = 0.9$, $Z = 2.2$, $p = .06$). It was significantly higher than in the English-L1 group for 'might' ($\beta = -15.7$, $SE = 2.0$, $Z = -8.0$, $p < .001$), 'try' ($\beta = 3.6$, $SE = 1.4$, $Z = 2.6$, $p = .02$) and, marginally, 'most' ($\beta = 2.0$, $SE = 1.0$, $Z = 2.1$, $p = .10$), and it was significantly lower than in the English-L1 group for 'scarce' ($\beta = -2.0$, $SE = 0.7$, $Z = -2.7$, $p = .02$). All other comparisons were not significant. Hence, the English-L2 group tended to pattern with the Slovenian-L1 group rather than the English-L1 group, though not universally.

One of our reviewers rightly pointed out that many of the significant differences between the language groups would not survive further (conservative) corrections for multiple comparisons, such as the Bonferroni correction. We concur and therefore

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

caution against drawing strong conclusions from the results for specific scalar expressions. At the same time, the most important conclusion, that is, that the English-L2 group did not consistently pattern with either the English-L1 or Slovenian-L1 group, is strongly corroborated.

3.3. Response times

Next, we analysed participants' response times to determine whether the derivation of scalar inferences caused a delay in response times. Figure 4 shows the mean logarithmised response times for each scalar expression and condition.

To determine whether participants in the English-L2 group were significantly slower in deriving scalar inferences than the two L1 groups, we fitted a linear mixed-effects model predicting logarithmised response times in correct trials on the basis of scalar expression ('low', 'scarce', 'might', 'some', 'most' or 'try'); language group (English-L1, English-L2 or Slovenian-L1); condition (target or control); response ('true' or 'false'); and all of their interactions, including random intercepts for participants. For this analysis, and all of the subsequent ones, all factors were sum-coded. Degrees of freedom and corresponding p-values were estimated using the Satterthwaite procedure, as implemented in the 'lmerTest' package (Kuznetsova et al., 2013).

To estimate the overall significance of the relevant fixed factors, we carried out likelihood ratio tests using the 'anova' function in R. These analyses showed that scalar expression interacted with the interaction between language group, condition and response ($\chi^2(55) = 464.9$, $p < .001$) and that language group interacted with the interaction between scalar expression, condition and response ($\chi^2(46) = 79.9$, $p = .001$). Hence, the relative effect of deriving scalar inferences, that is, the difference in response times between 'true' and 'false' responses in the target condition relative to the control condition, is modulated by both scalar expression and language group.

To further analyse the data, we fitted, for each scalar expression separately, a linear mixed-effects model using the 'lmer' function in R. These models predicted logarithmised response times on the basis of language group (English-L1, English-L2 or Slovenian-L1); condition (target or control); response ('true' or 'false'); and all their interactions, including random intercepts for participants.

First, we inspected whether the interaction between condition and response was significant, in order to determine whether each scalar expression was associated with a B&N effect, that is, whether 'false' responses were slower than 'true' responses in the target condition, relative to the difference between these two types of responses in the control condition. This interaction was significant for 'might' ($\beta = -0.1$, $SE = 0.0$, $Z = -2.6$, $p = .009$), 'some' ($\beta = -0.1$, $SE = 0.0$, $Z = -4.0$, $p < .001$) and 'most' ($\beta = -0.1$, $SE = 0.0$, $Z = -4.1$, $p < .001$). We also observed a significant interaction in the opposite direction (i.e., a reverse B&N effect) for 'scarce' ($\beta = 0.1$, $SE = 4.6$, $p < .001$) and 'try' ($\beta = 0.1$, $SE = 0.0$, $Z = 2.3$, $p = .02$). The interaction was not significant for 'low' ($\beta = -0.0$, $SE = 0.0$, $Z = -0.4$, $p = .66$).

Hence, the derivation of scalar inferences was associated with a significant slowdown in response times for 'might', 'some' and 'most', but not – or even in the reverse direction – for 'low', 'scarce' and 'try'. These findings confirm the results reported by van Tiel, Pankratz, and Sun (2019).

In order to evaluate whether 'false' responses in the target condition were particularly demanding for the English-L2 group, we also looked at the three-way interaction between language group, condition and response. This interaction was only significant, in the expected direction, for the comparison between the English-L1 and English-L2 groups for 'most' ($\beta = 0.1$, $SE = 0.0$, $Z = 2.3$, $p = .02$).

These results do not support the findings of Khorsheed et al. (2022). In their study, the B&N effect associated with the scalar inference of 'some' was significantly more pronounced for less proficient L2 speakers than for more proficient L2 speakers. Here, we observe that, in terms of response times, L2 speakers behave similarly to L1 speakers almost across the board. There was only one exception, namely in the case of 'most' where L2 speakers were indeed significantly more delayed when deriving the scalar inference. However, even in this condition, the L2 speakers were only delayed compared to the L1-English group, but not when compared

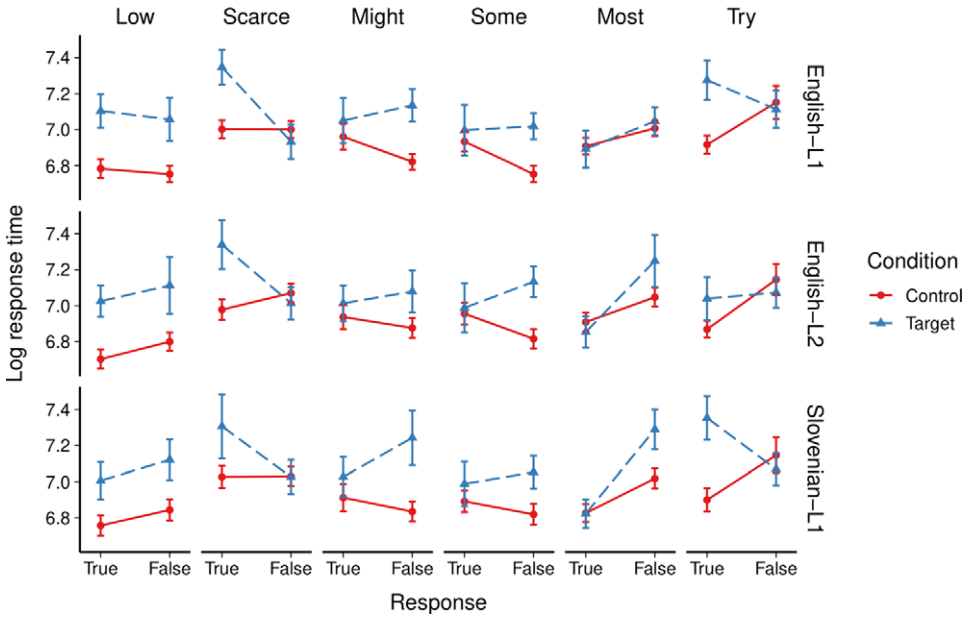


Figure 4. Mean log response times for each scalar term and condition, divided by language group. Error bars represent standard errors of the mean.

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

to the L1-Slovenian group. Hence, taken together, we did not find evidence for the idea that L2 speakers are significantly slower than L1 speakers in deriving scalar inferences.

4. General discussion

This study investigated to what extent current findings about the frequency and time course of scalar inferences in L2 generalise to scalar expressions other than 'some', in light of previous reports that findings for 'some' do not consistently generalise across the entire family of scalar expressions. To this end, we studied the derivation of six types of scalar inferences in three language groups: native speakers of English who were tested in their L1; native speakers of Slovenian who were also tested in their L1; and native speakers of Slovenian who were tested in English, which was their L2. Using a sentence-picture verification task, we thus analysed the frequency and time course of scalar inferences in L1 and L2.

4.1. The frequency of scalar inferences in L1 and L2

First, we discuss the frequency of scalar inferences in L1 and L2. Previous studies that investigated the scalar inference from 'some' to 'not all' have provided mixed results. While Slabakova (2010) observed higher rates of scalar inferences in L2 than L1, later studies consistently failed to confirm this finding, observing instead either no significant difference between L1 and L2 (e.g., Dupuy et al., 2019) or lower rates of scalar inferences in L2 than L1 (Mazzaggio et al., 2021). In a recent study, Khorsheed and van Tiel (2024) argue that L2 speakers have difficulties deriving scalar inferences, but that these difficulties may be masked when using tasks that allow participants to take their time to process and evaluate linguistic input, such as pen-and-paper questionnaires.

Given this claim, we expected to find lower rates of scalar inferences for 'some' in L2 than L1, since, in our experiment, the sentence disappeared before the picture was shown, so that participants had to keep the sentence in memory while evaluating its adequacy as a description of the picture. In contrast with this hypothesis, we observed no significant difference between L1 and L2 in the frequency with which participants interpreted 'some' as implying 'not all'.

A possible explanation for this null result is that the experiment was overall quite easy in terms of the complexity of the target sentences and the corresponding pictures (see Figure 1). Moreover, unlike previous experiments that often tested sentences such as (8), our experiment did not draw upon participants' encyclopedic knowledge.

(8) Some dogs are mammals.

It may be the case that people find it more difficult to evaluate sentences based on their encyclopedic knowledge than to compare sentences against pictures. These factors may explain why we observed no significant difference between L1 and L2 in the frequency with which the scalar inference of 'some' was derived. In support of this explanation, the response times observed in our experiment (1,261msec on average) were substantially faster than those found by Bott and Noveck (2004, Exp. 3), who tested sentences such as (8) and who report mean response times that are consistently above two seconds (see also, e.g., Guasti et al., 2005; Papafragou & Musolino, 2003; Pouscoulous et al., 2007, for further research highlighting the importance of task demands in scalar inference derivation). At the same time, we acknowledge that this explanation is post hoc and that more research is needed to specify the effects of task demands on cognitive resources.

Next, we considered whether the pattern of results for 'some' generalised to other scalar expressions. First, we focused on the two L1 groups to determine whether they were equally likely to derive scalar inferences. In line with previous studies (e.g., Katsos et al., 2016), but in contrast with claims in the literature that scalar inference rates are 'universal' (Slabakova, 2010), we observed considerable variability in the rates of scalar inferences across the two L1 groups. For example, the scalar inferences associated with 'might' (implying 'not necessarily') and 'most' (implying 'not all') were significantly more robust in English than Slovenian, while the reverse held, marginally, for 'low' (implying 'not empty').

An obvious question is how to explain this cross-linguistic scalar diversity. Here, we discuss two possible answers. We note in advance that both of these answers are post hoc in the sense that they are premised in the absence of direct confirmatory evidence. However, we hope that the discussion of these possible answers opens up avenues for further research to uncover the factors that underlie cross-linguistic scalar diversity.

First, it is well-known that the robustness of scalar inferences depends on the *question under discussion* (e.g., Ronai & Xiang, 2021; van Kuppevelt, 1996; Zondervan, 2010; in L2 learners: Starr & Cho, 2022). To illustrate, compare the two dialogues in (9) and (10).

(9) A: How much of the pepperoni pizza did Lucia eat?
B: She ate some of it.

(10) A: Did Lucia try the pepperoni pizza?
B: She ate some of it.

In (9), A's question indicates that they are interested in the amount of pizza that Lucia ate. Hence, the distinction between eating some of the pizza and eating all of it is highly relevant, which makes the scalar inference associated with B's answer robust. By contrast, in (10), A's question suggests that they are not particularly interested in finding out whether or not Lucia ate all of the pizza. Consequently, the scalar inference is intuitively more fickle.

In our experiment, the target sentences were not presented in the context of an explicit question under discussion. Therefore, participants may have inferred such a question to contextualise the sentences. It may be the case that languages, and cultures more broadly, vary in the type of questions (i.e., ones that make the scalar inference relevant or not) that they associate with sentences that are presented without further context. This, in turn, may be connected to the prosodic contours that readers associate with the sentences during reading, since prosody is a reliable cue for the question under discussion (e.g., Ronai & Göbel, 2024; Westera, 2016). If this explanation is on the right track, we predict that much of the cross-linguistic variability in scalar inference rates that we observed would disappear if the target sentences were presented in the context of an explicit question under discussion.

A second possible explanation for the observed variability in the rates of scalar inferences across the two L1 groups is that there are *semantic* differences between the English and Slovenian scalar expressions. Much of the preceding discussion is premised on the assumption that the English and Slovenian scalar expressions are semantically equivalent. In line with this assumption, both groups behaved very similarly in the control conditions (see Figure 2). However, there may be more subtle semantic differences between

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

the scalar expressions. Previous research has shown that such differences are not always straightforwardly detectable even for native speakers. For example, Stateva et al. (2019) showed that native speakers' intuitions about the felicity of English 'some' differ in subtle ways from its equivalents in French ('quelques'), German ('einige') and Slovenian ('nekaj'). In particular, the meaning of English 'some' seems to be more underspecified than that of its equivalents; that is, in English, it is more natural to use 'some' to refer to proportions greater than 50% than in the other languages that Stateva and colleagues investigated.

In their experiment, Stateva and colleagues also compared English 'most' with, inter alia, its Slovenian equivalent 'večina' (we also tested these expressions in our experiment). Here, they did not observe any differences in participants' felicity judgements. However, 'most' and 'večina' differ in their part of speech: while 'most' is a determiner, 'večina' is a noun, which means something along the lines of 'the most' or 'the majority'.

Other scalar expressions differ in their etymology. For example, compare English 'might' with Slovenian 'lahko'. Both express possibility. However, Slovenian 'lahko' developed from the adverb 'easily', 'lightly' and still keeps that meaning in some non-modal uses (Roeder & Hansen, 2006). Moreover, unlike 'might', 'lahko' cannot be used in negative contexts. Hence, even if our English and Slovenian target sentences with 'might' and 'lahko' are truth-conditionally equivalent, the scalar expressions exhibit subtle differences in usage and etymology.

It is well-known that scalar diversity is partly caused by semantic differences (e.g., Gotzner et al., 2017; Sun et al., 2018). For example, consider the notion of *semantic distance*, which refers to the perceived distance in meaning between the scalar expressions on a scale (e.g., Pankratz & van Tiel, 2021; van Tiel et al., 2016). Numerous studies have shown that scalar inferences are more robust if there is greater semantic distance between the expressions on a scale (e.g., van Tiel et al., 2016). A likely explanation for this observation is that, if there is greater semantic distance, it is easier to know whether the stronger term on the scale applies or not and, as a consequence, a hearer is more likely to be justified in concluding that the speaker knows that the stronger expression is inapplicable.

It may be the case that some of the English and Slovenian lexical scales vary in the perceived semantic distance between scalar expressions, which may have impacted the robustness of the scalar inference. For example, 'might' may be used more permissively, that is, in a greater range of situations, in English than its counterpart in Slovenian, so that the semantic distance between 'might' and 'must' is greater in English, leading to a more robust scalar inference in English when compared to Slovenian. Future studies may provide a further insight into such pragmatically relevant microvariation in semantics.

Given that the English-L1 and Slovenian-L1 groups varied in the rates of scalar inferences, we further investigated whether the English-L2 group patterned with the former group or with the latter. Our results do not provide a conclusive answer, but strongly suggest that the English-L2 group patterned with the Slovenian-L1 group rather than the English-L1 group. Thus, we found that the English-L2 group variously provided more or fewer 'false' responses in the target condition than the two L1 groups. In particular, when compared to the English-L1 group, the English-L2 group was more likely to derive the scalar inferences of 'might' (implying 'not necessarily'), 'try' (implying 'not succeed') and, marginally, 'most' (implying 'not all'), but less likely to derive the scalar inference of 'scarce' (implying 'not absent'). Conversely, when compared to the Slovenian-L1 group, the English-L2 group was marginally more likely to derive the scalar inference of 'low' (implying 'not empty').

Hence, with the marginal exception of 'low', we found that the English-L2 group patterned with the Slovenian-L1 group rather than the English-L1 group. There are at least two mutually non-exclusive explanations for this observation.

First, participants in the English-L2 group may have mentally translated the English sentences to their native language Slovenian and answered according to their intuitions about those Slovenian sentences. One argument against this explanation is that, if participants indeed mentally translated the sentences, we would expect that they generally respond more slowly than the L1 groups. However, this is not what we observed.

Second, the observation that the English-L2 group patterned with the Slovenian-L1 group may reflect effects of pragmatic transfer, that is, participants in the English-L2 group may have extended pragmatic regularities from their L1 to their L2 (e.g., Bou-Franch, 1998; Kasper, 1992; Kecskes, 2015). Transfer effects are highly sensitive to whether the L1 and L2 share overlapping pragmatic conventions. If L1 and L2 expressions exhibit semantic or pragmatic differences, learners may experience *negative transfer*, where L1-based inference strategies lead to misinterpretations in L2. Conversely, *positive transfer* can occur when L1 and L2 inferential patterns align, facilitating native-like pragmatic behaviour (Swain, 1972; Volterra & Taeschner, 1978).

In our study, English-L2 speakers may have detected subtle mismatches between English and Slovenian scalar expressions, prompting them to suppress L1-based strategies and rely more on L2 cues. This decision reflects a strategic adjustment rather than a contradiction, as learners actively shift between L1- and L2-based inferencing depending on perceived semantic compatibility (Jarvis & Pavlenko, 2007; Pavlenko, 2009).

Specifically, participants in the English-L2 group may have assumed that English scalar expressions are equally likely to give rise to scalar inferences as their Slovenian counterparts. If correct, this explanation would indicate an important source of difficulty in language learning, namely determining whether or not a scalar expression is to be interpreted with an upper bound, for example, whether someone stating that the party was 'pleasant' is to be interpreted as implying that they had a good time (as it may be in British English) or that it was not a particularly enjoyable party (as it may be in American English). This learnability challenge is consistent with Cummins' threshold hypothesis (Cummins, 1977). L2 learners must reach a certain level of proficiency before they can fully engage with complex pragmatic inferences such as scalar inferences (Bouton, 1992). Below this threshold, processing limitations may prevent the efficient integration of inferential cues, leading learners to default to literal interpretations (Mazzaggio et al., 2021) or rely on L1-based strategies.

Whichever of these two explanations is correct, we hypothesise that there will be effects of language proficiency, in that less proficient speakers are expected to be more likely to show effects of pragmatic transfer and/or mental translation than highly proficient speakers, who may ultimately even behave similarly to L1 speakers.

In our study, we asked participants to self-report their English proficiency, but exploratory analyses did not provide evidence for systematic effects of participants' self-reported proficiency on the rates of scalar inferences. At the same time, self-reports are inherently limited – what counts as B2 level for one participant may count as A2 for another – and our participants did not uniformly cover the full range of possible proficiency levels, since most participants indicated that their English level was at least B2. Future

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

research should study potential effects of L2 proficiency in a more reliable and systematic way.

Before continuing to our second research question, we want to caution against a normative interpretation of our results. In several studies, rejections of underinformative sentences with 'some' are seen as positive evidence of pragmatic competence. For example, Slabakova (2010, p. 2458) interprets her finding that participants are more likely to compute scalar inferences in L2 as suggesting 'superior pragmatic competence'. We find this framing misleading. It would be a mistake to always compute a scalar inference whenever one encounters a scalar expression. After all, the derivation of scalar inferences revolves around the hearer asking themselves why the speaker produced a weaker statement when they could have used a stronger statement instead. One explanation is that the speaker believes the stronger statement to be false. But in some contexts, a more plausible explanation is, for example, that the speaker did not consider this alternative or does not have sufficient evidence. In such contexts, it would be wrong (i.e., not in line with the speaker's intended meaning) to compute the scalar inference.

4.2. The time course of scalar inferences in L1 and L2

Our second research question focused on the time course of scalar inferences. Numerous studies have shown that the inference from 'some' to 'not all' is time-consuming, an effect that has been called the B&N effect (e.g., Bott & Noveck, 2004). In a recent study, Khorsheed et al. (2022) observed that the B&N effect associated with the scalar inference of 'some' was more pronounced for less proficient L2 speakers than for more proficient ones. This finding suggests that the derivation of scalar inferences in L2 is especially difficult for less proficient speakers.

Here, we investigated whether we would find a similar effect of proficiency on response times when comparing L1 and L2 speakers. However, our results did not confirm this hypothesis: L2 speakers were not significantly more delayed when deriving the scalar inference of 'some' than L1 speakers. Several explanations for this apparent discrepancy with the results reported by Khorsheed et al. (2022) may be given.

First, Khorsheed and colleagues compared L2 speakers that varied in their proficiency, whereas we compared L1 speakers and L2 speakers. It may be the case that more proficient L2 speakers pattern with L1 speakers in terms of response times and that such more proficient L2 speakers were overrepresented in our sample. However, as noted earlier, we also asked participants to indicate their proficiency. Exploratory analyses did not confirm the idea that, in terms of response times, less proficient L2 speakers patterned substantially differently from more proficient L2 speakers.

Second, as noted earlier, our sentence-picture verification task is intuitively somewhat easier than the sentence verification task used by Khorsheed and colleagues (and many others, like Cho, 2024; Mazzaggio et al., 2021; Slabakova, 2010), which tested sentences such as (8) that had to be evaluated based on participants' encyclopedic knowledge. As a consequence, it may be the case that the B&N effect was relatively small in our experiment, which may have mitigated the additional difficulty that L2 speakers had with deriving scalar inferences. In line with this suggestion, the mean difference between 'true' and 'false' responses in the target condition of 'some' was substantially smaller in our experiment (88 milliseconds) than in the experiment by Khorsheed and colleagues (~750 milliseconds). Future research should investigate whether differences in the time course of scalar inferences between L1 and L2 speakers emerge when using sentences that draw on encyclopedic knowledge.

Next, we investigated whether other scalar expressions patterned with 'some' in terms of their processing profile. Again, our results provide a negative answer to this question. While some scalar expressions patterned with 'some' in giving rise to a B&N effect, others did not, and still others gave rise to a reverse B&N effect, suggesting that deriving the scalar inference caused participants to respond more quickly. This pattern of results is very similar to what van Tiel et al. (2019, Exp. 1) found.

Van Tiel and colleagues explain their results on the basis of the *polarity* (or *scalarity*) of the scalar expressions. In particular, whereas 'some', 'might', 'most' and 'try' are positive, in that they place a lower bound on their relevant dimension (e.g., 'some' meaning '*at least* one'), 'low' and 'scarce' are negative, in that they place an upper bound (e.g., 'scarce' meaning '*at most* x', where x is determined contextually). The scalar inferences associated with positive scalar expressions express negative information (e.g., 'some' implying '*not* all'), whereas those of negative scalar expressions express positive information (e.g., 'scarce' implying 'existent'). There is a vast literature demonstrating that processing negative information is cognitively costly (e.g., Carpenter & Just, 1975; Clark & Chase, 1972). Hence, according to van Tiel and colleagues, the B&N effect associated with scalar expressions such as 'some' actually reflects participants' cognitive effort with processing and evaluating the negative information, rather than the derivation of scalar inferences, per se.

The outlier in our experiment, but also in Exp. 1 of van Tiel and colleagues, is 'try', which, despite being positive, did not trigger a B&N effect. However, van Tiel and colleagues show that, on other measures of processing effort (e.g., working memory taxation), 'try' did pattern with the other positive scalar expressions (see also Marty et al., 2024; van Tiel, Pankratz, Marty, & Sun, 2019).

Hence, our findings confirm the results reported by van Tiel, Pankratz, and Sun (2019). Interestingly, we find exactly the same pattern of results for both English and Slovenian. This observation suggests that, even though there are marked differences in the frequency with which scalar expressions trigger a scalar inference in different languages, they pattern quite similarly in terms of their processing signature.

Indeed, in terms of time course, the English-L2 group also behaved very similarly to the two L1 groups. Only for one scalar expression did we observe that the derivation of scalar inferences was more time-consuming for the English-L2 group than for the English-L1 group, namely in the case of 'most'. However, even there, we did not observe a significant difference with the Slovenian-L1 group, which indicates that overall, L2 speakers do not take significantly longer to derive scalar inferences than L1 speakers.

4.3. Theoretical and methodological consequences

Returning to the debate between relevance theory and Levinson's defaultism, our results provide evidence against both types of approaches. According to relevance theory, computing scalar inferences in settings in which they are not made contextually relevant is cognitively demanding. Hence, given that processing L2 input draws upon additional cognitive resources compared to L1 input, relevance theory predicts consistently lower rates of scalar inferences in L2 compared to L1. This prediction was not confirmed: depending on the scalar expression, scalar inference rates were variously lower, equal, or higher in the English-L2 group compared to the two L1 groups. By contrast, Levinson (2000) predicts that scalar inferences are default inferences, whose overturning is

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

cognitively demanding. According to this idea, scalar inference rates should be consistently *higher* in L2 compared to L1. We also failed to confirm this prediction. Rather, our results are more in line with intermediate accounts that argue that the presence or absence of a processing cost for scalar inferences is modulated by features of the context or scale (e.g., Degen & Tanenhaus, 2015; van Tiel, Pankratz, & Sun, 2019).

For instance, L2 processing demands may vary depending on the complexity of the scalar expression, the speaker's proficiency and how closely the L2 expressions resemble those in the speaker's L1. While it is unclear which specific factors shaped our results, one possible explanation involves *semantic similarity*. By 'semantic similarity', we refer to the degree of overlap in meaning between scalar expressions in English and their Slovenian equivalents. It may be the case that L2 speakers rely more on their intuitions about their L1 when the English and Slovenian expressions are semantically similar, whereas they may adopt L2-specific conventions when the expressions are less similar. To illustrate, consider the scalar term 'low' in 'The battery is low'. In Slovenian, the equivalent phrase 'Baterije zmanjkuje' (literally 'The battery is running out') emphasises the process of depletion rather than a static low level. This semantic difference may have prompted English-L2 speakers to rely more on their English knowledge, contributing to scalar inference rates similar to those of English-L1 speakers for this expression.

This possibility is supported by studies on L2 acquisition and language transfer, which demonstrate that semantic similarity between L1 and L2 expressions influences how learners process and interpret L2 input. Specifically, L2 learners frequently rely on their L1 when there is overlap in meaning between L1 and L2 expressions, leading to positive transfer (e.g., Jarvis & Pavlenko, 2007). However, as we have seen, transfer can also occur when learners mistakenly apply L1 conventions to L2, especially when subtle differences exist between the languages (i.e., *negative transfer*). This has been observed in pragmatic inferencing, such as in the acquisition of quantifiers (Mazzaggio & Stateva, 2024). While further research is necessary to confirm this idea for scalar expressions, these findings suggest that both positive and negative transfer of lexical meaning may have contributed to the results observed in our study.

4.4. Limitations of the study

A first limitation of our study resides in the methodology. First, while it is typically assumed that scalar inferences are understood in the same way by all language users, emerging studies reveal inter-individual differences in the processing of scalar inferences, as suggested by an anonymous reviewer and as evidenced by Zhang and Wu (2023). Some individuals perceive scalar inferences as pragmatic inferences, others perceive as default meanings, and some maintain flexibility between the two. While we acknowledge the importance of studying such inter-individual differences, doing so systematically would require a much larger sample size – potentially in the hundreds – as well as multiple additional cognitive and psychological assessments, such as executive function tests, working memory assessments and autistic trait measures, to control for potential confounds. Such an endeavour was beyond the scope of our study.

Second, although the sentence verification task is among the most frequently used experimental paradigms to study scalar inferences, several authors have questioned its validity; that is, they have asked whether truth-value judgements are actually reflective of the derivation of scalar inferences (e.g., Degen & Goodman, 2014; Jasbi et al., 2019; Katsos & Bishop, 2011; Kissine & De Brabanter, 2023). Unfortunately, we do not have the space to discuss and address these criticisms. However, we want to emphasise two points.

First, in order to assess the validity of any experimental paradigm, it will be necessary to define what it means to have derived a scalar inference. The central assumption underlying the use of the sentence verification task is that a scalar inference is derived if it becomes a part of what the sentence (or a hypothetical speaker uttering this sentence) means, that is, its falsity becomes a ground for rejecting the entire sentence. In response, Katsos and Bishop (2011) have shown that, often, people who do not reject under-informative sentences with 'some' still notice their pragmatically infelicity. This is an important insight, but whether it should be taken to show that 'false' responses in sentence verification tasks systematically underestimate the frequency of scalar inferences depends on how scalar inferencing is defined. This is a contentious issue that has been largely overlooked in the literature (but see Jasbi et al., 2019; Kissine & De Brabanter, 2023).

Second, even if the sentence verification paradigm can be questioned, the same holds for all of the alternative methodologies that have been proposed in the literature. For example, the inclusion of more than two response options, as suggested by Katsos and Bishop, raises the issue of whether participants who give intermediate responses genuinely derived a scalar inference or whether they simply thought that the sentence was an atypical or odd way of expressing the intended meaning (e.g., van Tiel, 2014). Hence, if there is indeed no single correct way of measuring scalar inferences, our results are pertinent to theorising about pragmatics in L2, even if they call for confirmation using different methodologies, such as visual-world eye-tracking (e.g., Degen & Tanenhaus, 2015).

Finally, as we also noted earlier, while we discussed several potential factors that might explain the variability in scalar inference rates (e.g., question under discussion, cross-linguistic semantic microvariation), our study was not specifically designed to isolate their effects. Indeed, our study was primarily concerned with testing the generalisability of previous findings on the scalar inference from 'some' to 'not all' in L2. However, the unexpected nature of some of the results led us to develop post-hoc explanations. Future studies should investigate the role of these factors more directly by including experimental manipulations that explicitly test under which circumstances L2 learners extend interpretative strategies from their L1 and under which circumstances they become attuned to the pragmatic regularities in the L2.

5. Conclusion

In conclusion, our findings demonstrate cross-linguistic and cross-scalar differences in the computation of scalar inferences. These findings underline the importance of testing a wider range of scalar terms beyond the quantifier scale.

From a cross-linguistic perspective, our study reveals differences between L1 and L2 speakers in the frequency of scalar inferences for specific scalar expressions, suggesting that L2 speakers may rely on their native language when deriving scalar inferences. This finding also raises important questions regarding the extent to which L2 speakers can achieve native-like competence. The underlying reasons for these discrepancies remain uncertain, with potential explanations involving pragmatic transfer, mental translations or subtle differences in the meanings of specific expressions.

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press

Our findings also suggest moving away from the traditional question as to whether the derivation of scalar inferences is cognitively costly and focus instead on uncovering the factors that influence the derivation of scalar inferences in L2. In particular, we have highlighted the potential effects of transfer, as well as cross-linguistic variability in the lexical semantics of scalar expressions. We acknowledge that the exact role of these factors is currently unclear, but encourage future research to look into them in a more controlled way.

Future research should further explore the computational challenges associated with scalar inferences in L2, taking into account a broader range of scalar expressions and larger sample sizes. Additionally, it would be valuable to examine multiple language combinations in order to analyse the underlying semantic structure of the diverse scalar expressions, which could further elucidate the varied patterns of results documented in this study.

Data availability statement. The data that support the findings of this study are openly available on Open Science Framework (OSF) at https://osf.io/jzm3k/.

Acknowledgments. This work has been supported by a grant from the Slovenian Research Agency (ARRS) (Project number: J6-2580) awarded to Penka Stateva. The project has also been supported by funding from the Italian Ministero dell'Istruzione, dell'Università e della Ricerca and the European Union - NextGenerationEU (PNRR – PE05 CHANGES CUP B53C22004010006).

Author contribution. Greta Mazzaggio involved in conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing the original draft, writing review and editing, supervision and project administration; Federica Longo involved in methodology, software, investigation, visualization and writing the review and editing; Penka Stateva involved in writing the review and editing and funding acquisition; and Bob van Tiel involved in methodology, software, formal analysis, data curation, visualization, writing the original draft and writing the review and editing.

Competing interests. The authors declare none.

Ethical approval. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

References

Antoniou, K., & **Katsos, N**. (2017). The effect of childhood multilingualism and bilectalism on implicature understanding. *Applied Psycholinguistics*, **38**, 787–833.

Antoniou, K., **Veenstra, A.**, **Kissine, M.**, & **Katsos, N** (2020). How does childhood bilingualism and bi-diale.ctalism affect the interpretation and processing of pragmatic meanings? *Bilingualism: Language and Cognition*, **23**, 186–203.

Bates, D., **Mächler, M.**, **Bolker, B.**, & **Walker, S**. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Bott, L., & **Noveck, I. A**. (2004). Some utterances are under informative: The onset and time course of scalar inferences. *Journal of Memory and Language*, **51**, 437–457.

Bou-Franch, P. (1998). On pragmatic transfer. *Studies in English Language and Linguistics*, **2**, 5–20.

Bouton, L. F. (1992). The interpretation of implicature in English by NNS: Does it come automatically–without being explicitly taught? *Pragmatics and Language Learning*, **3**, 53–65.

Carpenter, P. A., & **Just, M. A**. (1975). Sentence comprehension: A psycholinguistic model of sentence verification. *Psychological Review*, **82**, 45–73.

Chemla, E., & **Bott, L**. (2014). Processing inferences at the semantics/pragmatics frontier: Disjunctions and free choice. *Cognition*, **130**, 380–396.

Chierchia, G., **Fox, D.**, & **Spector, B**. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Semantics: An international handbook of natural language meaning* (pp. 2297–2332). De Gruyter.

Cho, J. (2024). Scalar implicatures in adult L2 learners: A self-paced reading study. *Second Language Research*, **40**(2), 327–350.

Clahsen, H., & **Felser, C**. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, **27**, 3–42.

Clark, H. H., & **Chase, W. G**. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, **3**, 472–517.

Cremers, A., & **Chemla, E**. (2014). Direct and indirect scalar implicatures share the same processing signature. In S. P. Reda (Ed.), *Pragmatics, semantics and the case of scalar implicatures* (pp. 201–227). Palgrave Macmillan.

Cummins, J. (1977). Cognitive factors associated with the attainment of intermediate levels of bilingual skills. *The Modern Language Journal*, **61**(1/2), 3–12.

Degen, J., & **Goodman, N. D**. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 397–402). Cognitive Science Society.

Degen, J., & **Tanenhaus, M. K**. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, **39**, 667–710.

Dionne, D., & **Coppock, E**. (2022). Complexity vs. salience of alternatives in implicature: A cross-linguistic investigation. *Glossa Psycholinguistics*, **1**, 1–27.

Dupuy, L., **Stateva, P.**, **Andreetta, S.**, **Cheylus, A.**, **Déprez, V.**, **van der Henst, J.-B.**, **Jayez, J.**, **Stepanov, A.**, & **Reboul, A**. (2019). Pragmatic abilities in bilinguals: The case of scalar implicatures. *Linguistic Approaches to Bilingualism*, **9**, 314–340.

Ellis, R. (2009). Measuring implicit and explicit knowledge of a second language. In *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 31–64). Multilingual Matters.

Feng, S., & **Cho, J**. (2019). Asymmetries between direct and indirect scalar implicatures in second language acquisition. *Frontiers in Psychology*, **10**, 877.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics* (pp. 71–120). Palgrave Macmillan.

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition, and logical form*. Academic Press.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Geurts, B., & **Rubio-Fernández, P**. (2015). Pragmatics and processing. *Ratio*, **28**, 446–469.

Gotzner, N., **Solt, S.**, & **Benz, A**. (2017). Scalar diversity, negative strengthening and adjectival semantics. *Frontiers in Psychology*, **9**, 1659.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Guasti, M.T., **Chierchia, G.**, **Crain, S.**, **Foppolo, F.**, **Gualmini, A.**, & **Meroni, L**. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, **20**(5), 667–696.

Hopp, H. (2022). Second language sentence processing. *Annual Review of Linguistics*, **8**, 235–256.

Horn, L. R. (1972). *On the semantic properties of logical operators in English* [Unpublished doctoral dissertation]. Los Angeles: University of California.

Hothorn, T., **Bretz, F.**, & **Westfall, P**. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, **50**, 346–363.

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.

Jarvis, S., & **Pavlenko, A**. (2007). *Crosslinguistic influence in language and cognition* (1st ed.). Routledge.

Jasbi, M., **Waldon, B.**, & **Degen, J**. (2019). Linking hypotheses and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology*, **10**, 189.

Juffs, A. (2001). Psycholinguistically-oriented second language research. *Annual Review of Applied Linguistics*, **21**, 207–223.

Kasper, G. (1992). Pragmatic transfer. *Second Language Research*, **8**, 203–231.

Katsos, N., & **Bishop, D. V. M**. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, **120**, 67–81.

Katsos, N., **Cummins, C.**, **Ezeizabarrena, M. J.**, **Gavarró, A.**, **Kuvač Kraljević, J.**, **Hrzica, G.**, … & **Noveck, I**. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, **113**, 9244–9249.

Kecskes, I. (2015). How does pragmatic competence develop in bilinguals? *International Journal of Multilingualism*, **12**, 419–434.

Khorsheed, A., **Md. Rashid, S.**, **Nimehchisalem, V.**, **Geok Imm, L.**, **Price, J.**, & **Ronderos, C. R**. (2022). What second-language speakers can tell us about pragmatic processing. *Plos One*, **17**, e0263724.

Khorsheed, A., & **van Tiel, B**. (2024). Why second-language speakers sometimes, but not always, derive scalar inferences like first-language speakers: Effects of task demands. *Language Acquisition*, 1–19. https://doi.org/10.1080/10489223.2024.2383574

Kissine, M., & **De Brabanter, P**. (2023). Pragmatic responses to under-informative some-statements are not scalar implicatures. *Cognition*, **237**, 105463.

Kuznetsova, A., **Brockhoff, P. B.**, & **Christensen, R. H. B**. (2013). *LmerTest: Tests for random and fixed effects for linear mixed effect models (LMER objects of lme4 package)* [R package]. https://cran.r-project.org/package=lmerTest

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Marty, P., **Romoli, J.**, **Sudo, Y.**, **van Tiel, B.**, & **Breheny, R**. (2024). Scalar inferencing, polarity and cognitive load. *Glossa Psycholinguistics*, **3**, 15.

Mazzaggio, G., **Panizza, D.**, & **Surian, L**. (2021). On the interpretation of scalar implicatures in first and second language. *Journal of Pragmatics*, **171**, 62–75.

Mazzaggio, G., & **Stateva, P**. (2024). Negative pragmatic transfer in bilinguals: Cross-linguistic influence in the acquisition of quantifiers. *Journal of Psycholinguistic Research*, **53**, 1–16.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicatures. *Cognition*, **78**, 165–188.

Noveck, I. A. (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.

Noveck, I. A., & **Sperber, D**. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In N. Burton-Roberts (Ed.), *Advances in pragmatics* (pp. 184–212). Palgrave Macmillan.

Pankratz, E., & **van Tiel, B**. (2021). The role of relevance for scalar diversity: A usage-based approach. *Language and Cognition*, **13**, 562–594.

Papafragou, A., & **Musolino, J**. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, **86**(3), 253–282.

Pavlenko, A. (Ed.). (2009). *The bilingual mental lexicon: Interdisciplinary approaches* (Vol. **70**). Multilingual Matters.

Pouscoulous, N., **Noveck, I. A.**, **Politzer, G.**, & **Bastide, A**. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, **14**(4), 347–375.

Recanati, F. (1995). The alleged priority of literal interpretation. *Cognitive Science*, **19**, 207–232.

Roeder, C. F., & **Hansen, B**. (2006). Modals in contemporary Slovene. *Wiener Slavistisches Jahrbuch*, **52**, 153–170.

Ronai, E., & **Göbel, A**. (2024). Watch your tune! On the role of intonation for scalar diversity. *Glossa Psycholinguistics*, **3**, 26.

Ronai, E., & **Xiang, M**. (2021). Pragmatic inferences are QUD-sensitive: An experimental study. *Journal of Linguistics*, **57**, 841–870.

Ronai, E., & **Xiang, M**. (2022). Three factors in explaining scalar diversity. In D. Gutzmann & S. Repp (Eds.), *Proceedings of Sinn und Bedeutung* (Vol. **26**, pp. 716–733). University of Cologne.

Slabakova, R. (2010). Scalar implicatures in second language acquisition. *Lingua*, **120**, 2444–2462.

Snape, N., & **Hosoi, H**. (2018). Acquisition of scalar implicatures. Evidence from adult Japanese L2 learners of English. *Linguistic Approaches to Bilingualism*, **8**, 163–192.

Sperber, D., & **Wilson, D**. (1995). *Relevance: Communication and cognition* (2nd ed.). Basil Blackwell.

Starr, G., & **Cho, J.** (2022). QUD sensitivity in the computation of scalar implicatures in second language acquisition. *Language Acquisition*, **29**(2), 182–197.

Stateva, P., **Stepanov, A.**, **Déprez, V.**, **Dupuy, L. E.**, & **Reboul, A. C**. (2019). Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology*, **10**, 957.

Sun, C., **Tian, Y.**, & **Breheny, R**. (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology*, **9**, 2092.

Swain, M. K. (1972). *Bilingualism as a first language*. University of California.

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software manual]. R Foundation for Statistical Computing. https://www.r-project.org/

van Kuppevelt, J. (1996). Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy*, **19**, 393–443.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, **31**, 147–177.

van Tiel, B., & **Pankratz, E**. (2021). Adjectival polarity and the processing of scalar inferences. *Glossa*, **6**, 32.

van Tiel, B., **Pankratz, E.**, **Marty, P.**, & **Sun, C**. (2019). Scalar inferences and cognitive load. In M. T. Espinal, E. Castroviejo, M. Leonetti, L. McNally, & C. Real-Puigdollers (Eds.), *Proceedings of Sinn und Bedeutung 23* (pp. 429–443). Universitat Autònoma de Barcelona.

van Tiel, B., **Pankratz, E.**, & **Sun, C.** (2019). Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, **105**, 427–441.

van Tiel, B., & **Schaeken, W**. (2017). Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science*, **41**, 1–36.

van Tiel, B., **van Miltenburg, E.**, **Zevakhina, N.**, & **Geurts, B.** (2016). Scalar diversity. *Journal of Semantics*, **33**, 137–175.

Volterra, V., & **Taeschner, T**. (1978). The acquisition and development of language by bilingual children. *Journal of Child Language*, **5**(2), 311–326.

Westera, M. (2016). *Exhaustivity and intonation: A unified theory*. [Unpublished doctoral dissertation]. The Netherlands: University of Amsterdam.

White, L., & **Juffs, A**. (1998). Constraints on wh-movement in two different contexts of non-native language acquisition: Competence and processing. In S. Flynn, G. Martohardjono, & W. O'Neill (Eds.), *The generative study of second language acquisition* (pp. 111–130). Erlbaum.

Zhang, J., & **Wu, Y**. (2023). Epistemic reasoning in pragmatic inferencing by non-native speakers: The case of scalar implicatures. *Second Language Research*, **39**, 697–729.

Zondervan, A. (2010). *Scalar implicatures or focus: An experimental approach*. [Unpublished doctoral dissertation]. University of Utrecht, The Netherlands.

https://doi.org/10.1017/S1366728925000392 Published online by Cambridge University Press