

On the interpretation of scalar implicatures in first and
second language

Greta Mazzaggio ^a, , Daniele Panizza ^b, Luca Surian ^a

^a Department of Psychology and Cognitive Science, University of Trento, Corso Bettini N.
31, 38068, Rovereto, TN, Italy

^b Seminar for English Philology, University of Gottingen, Kate-Hamburger-Weg 3, 37073,
Gottingen, Germany

Abstract

We investigated the effect of presenting items in a foreign language (L2) on scalar-implicatures computation. To ensure that L2 processing was more effortful than the processing of the native language (L1), participants were late learners of L2 immersed in an L1 environment and they were presented with oral stimuli under time constraints. If scalar-implicatures computation requires cognitive effort one should find that people are more likely to compute scalar implicatures in L1 than in L2. In two experiments, participants were asked to perform a Sentence Evaluation Task either Italian, their native language, or in a foreign language (English or Spanish). The task included underinformative statements such as “Some dogs are animals” that, if interpreted in a pragmatic way (i.e., “Some but not all dogs are animals”) should be rejected as false. In both experiments, we found more rejections in the native language condition than in the foreign language conditions. These results provide support for models that maintain that scalar implicature computation is effortful.

Keywords

Scalar implicatures; Pragmatics; Default models; Non-default models; Second-language processing

1. Introduction

According to the influential theory set out by Paul Grice (1975, 1989), communication is a co-operative exchange governed by rational expectations about how a conversation should be conducted. Along this line, Grice proposed that participants in a conversation expect each other to obey a set of conversational maxims. These maxims constrain the quantity and quality of the information to be conveyed, and determine how it should be encoded in an utterance. For example, the first maxim of Quantity requires speakers to provide only necessary and sufficient information given the purpose of the exchange. This maxim is violated by the use of (1a) instead of (1c) in a context in which the speaker knows that all students got an A.

- (1) a. Some students got an A.
- b. Not all students got an A.
- c. All students got an A.

Inferring (1b) from (1a) is known in the literature as ‘scalar implicature’ (Horn, 1972) and arises from *some* belonging to a set of alternative quantifiers that are semantically (logically) more informative. The set creates a quantificational scale <some, most, all> that ranges from weak to strong. The quantifier *all* is stronger/more informative than *some*, because $all \sqsubseteq some$ (*all* logically entails *some*). The logical interpretation of *some*,

namely *some and possibly all*, is the lower-bound interpretation that is exemplified in the logic form in (2b), whereas the pragmatic interpretation, that is *some but not all*, is the upper-bound interpretation exemplified in (2c). The latter arises from listeners assuming that a speaker chose the most informative quantifier from the scale.

(2) a. Some of the children danced.

b. $\exists x[\text{child}(x) \wedge \text{danced}(x)]$ = some (and possibly all) of the children danced c.

$\exists x[\text{child}(x) \wedge \text{danced}(x) \wedge \neg \forall x [\text{child}(x) \rightarrow \text{danced}(x)]]$ = some but not all of the children danced

Furthermore, under certain circumstances, the pragmatic interpretation can be cancelled without logical contradiction, as in ‘Some of the children danced, indeed all of them did’.

Apart from quantifiers, among these scales Horn identified connectives (<or, and>), adverbs (<sometimes, often, always>), verbs (<to think, to believe, to know>), modals (<may, must>), numerals (<zero, one, two, etc.>), where the use of the weaker term in the scale invites the listener to infer that the stronger one does not hold; for example, in (3) we can assume (b) from (a), and (d) from (c).

(3) a. I will bring salty or sweet food at the party.

b. I will not bring both salty and sweet food at the party.

c. I think I left my mobile at home.

d. I do not know for sure if my mobile is at home.

How we compute scalar implicatures and whether this process is costly in terms of cognitive resources is under debate, with two main approaches making different predictions: the default models and the non-default models. According to the default models, such as those proposed by Levinson (2000) within the neo-Gricean framework (cf. Horn 1972; Gazdar 1979), the pragmatic interpretation is automatic and represents the ‘default’ meaning: “one that captures our intuitions about a preferred or normal interpretation” (Levinson, 2000: 11). A scalar inference, however, might be cancelled in a subsequent stage. In other words, in line with the default models, when we bump into a scalar term such as *some* we immediately and always interpret it with its upper-bound interpretation *some but not all*. In Levinson’s (2000) terms, such interpretations is automatically driven by the Q heuristic “what isn’t said, isn’t” and if the more informative *all* had not been said, it does not hold. The *some and possibly all* interpretation requires a two-stage derivation process since one needs to cancel the first stage of *some but not all* automatic interpretation.

In contrast, the non-default models, such as those proposed within the framework of the Relevance Theory by Carston (1998) and by Sperber and Wilson (1986, 1995) claim that scalar inferences are pragmatic in nature and the speaker’s

expectations govern the hearer willingness to draw an interpretation based on relevance. Indeed, “the central claim of relevance theory is that the expectations of relevance raised by an utterance are precise and predictable enough to guide the hearer toward the speaker’s meaning” (Wilson & Sperber, 2002: 607). The hearer’s interpretation is reached balancing between a positive cognitive effect and a cognitive effort: the goal of the hearer is to obtain the most relevant interpretation with a minimum effort. With ‘effort’ they refer “to the ease with which the information can be integrated by the processor. If two stimuli provide the same effect, but one requires more effort to process than the other, the easier-to-process one will be higher in relevance” (Noveck, 2018: 27). In other word, the logical interpretation sometimes can perfectly satisfy the hearer in terms of sentence interpretation and this without particular effort; on the other hand, under specific, context-bound situation the hearer might require a more informative interpretation: this pragmatic enrichment may be achieved by means of an effortful cognitive process. Indeed, Wilson and Sperber’s Relevance-Theoretic comprehension procedure is described as follow: “a. Follow a path of least effort in computing cognitive effects: Test interpretive hypotheses (disambiguations, reference resolutions, implicatures, etc.) in order of accessibility. b. Stop when your expectations of relevance are satisfied (or abandoned)” (2002: 613).

A different account that tried to explain difficulties related to the computation of scalar implicatures in a developmental perspective is the Lexical Account (Barner et al., 2011; Foppolo et al. 2012). This account had been proposed in order to explain children’s difficulties with scalar implicatures in spite of preserved pragmatic abilities

in other contexts (e.g., numerals, non-generalized ad-hoc implicatures). In this framework, the problems with scales - such as quantifiers - are a consequence of limitations in representing lexical items as members of a scale and/or accessing the scale.

Psycholinguistic studies realized that those two models clearly predict testable behaviors in terms of processing cost related to the computation of a scalar implicature. Moving from Sperber and Wilson (1986, 1995), some scholars maintain that scalar computation is costly (Bott et al. 2012; Breheny et al., 2006; Panizza et al., 2009) or delayed in that it requires prior access to the literal meaning (Huang & Snedeker, 2009a; Tomlinson et al., 2013). In opposition to these approaches and moving from Levinson (2000), other proponents advance the idea that the access to pragmatic interpretations can be immediate and cost-free (Breheny et al., 2013; Grodner et al., 2010; Degen & Tanenhaus, 2011).

For instance, many studies addressing the scalar-implicature debate (among them, Bott & Noveck, 2004; De Neys & Schaeken, 2007; Guasti et al., 2005; Marty & Chemla, 2013; Noveck, 2001; Papafragou & Musolino, 2003; Pouscoulous et al., 2007) have investigated whether deriving scalar implicatures is cognitively demanding. In particular, the focus has been on looking at reaction times during scalar-implicatures computation and considering whether resource-demanding contexts and/or a paucity of cognitive resources (e.g., working memory load) prevent or reduce pragmatic interpretations. A cognitive cost might be revealed by a variety of psycholinguistic or neurolinguistic measures such as percentages of answers, reaction times, reading

times, patterns of cortical activation and neuronal activity. The idea is that implicatures need processing resources because they require a number of computations that go beyond accessing the lexical meaning of a word from the lexicon. One can expect an increase in the rate of scalar implicatures by either providing enough time to participants' disposal in order to do all the necessary computational steps, and/or having sufficient cognitive skills that allow them to do the computations. Several studies have investigated these effects using populations for which linguistic competence is deemed not fully developed, i.e. young children and second language (L2) non-proficient learners.

Studies on young children demonstrated that they, more often than adults, accept the logical (weaker) term in a context where the stronger term would be more appropriate, supporting the conclusion that the pragmatic interpretation is not automatic (Braine & Romain, 1981; Chierchia et al., 2001; Huang & Snedeker, 2009b; Noveck, 2001; Smith, 1980). However, children's pragmatic interpretations increase under particular task conditions and within clearer contexts (Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004). Their difficulties with scalar implicatures could result from an immature pragmatic competence (Noveck, 2001), from more tolerance towards pragmatic violations (Katsos & Bishop, 2011), from domain-related general cognitive limitations (Reinhart, 1999), from the complexity of quantifiers themselves (Horowitz et al., 2018), from limitations in their lexical knowledge, preventing the access to relevant lexical scales (Lexical Account: Barner et al., 2011; Chierchia et al., 2001; Foppolo, 2007; Foppolo et al., 2012).

Studies on adults' performance on scalar implicatures focused on the cognitive cost of their derivation and on whether scalar-implicatures interpretations are processed in a longer time. Many studies found that pragmatic interpretations are indeed associated with longer processing time due to increased cognitive effort (Bott & Noveck, 2004; Breheny et al., 2006; Degen & Tanenhaus, 2011; Dieussaert et al., 2011; Huang & Snedeker, 2009a; Noveck & Posada, 2003; Politzer-Ahles & Gwilliams, 2015; Tomlinson et al., 2013). In a study of Bott and Noveck (2004), when participants were explicitly instructed to interpret *some* in a pragmatic way they encountered more difficulties compared to participants who were told to interpret it in a logical way, with the difficulty reflected in slower as well as fewer successful responses. This latter study also tested reaction times, predicting that the manifestation of a cognitive effect (e.g. an implicature) depends on the cognitive resources available. They manipulated the resources available to the participants (3000 versus 900 milliseconds to respond). The prediction was that there should be more pragmatic responses in the long condition compared to the short condition. Data confirmed the prediction and found a reliable increase of pragmatic answers when more time was available. When participants had fewer cognitive resources available, fewer scalar implicatures were computed. By contrast, responses to the control sentences did not significantly vary between conditions. Their results seem to support the idea that pragmatic answers rely on cognitive resources. Convergent findings were also reported by Marty and Chemla (2013) and by De Neys and Schaeken (2007), even if conflicting results have been found where online reference resolution

requiring scalar computation (e.g. *some but not all*) was as rapid as with semantic controls (e.g. *all*) (Grodner et al., 2010; Breheny et al., 2013). Indeed, another study showed that the reason why scalar implicatures are quickly processed by adults is the predictability of the event or situation described in the utterance; in other words, participants are computing the implicature before they even hear the critical utterances (Huang & Snedeker, 2018).

Recently, a new stream of research on the cost of scalar-implicatures computation focused on the performance of bilinguals. L2 processing might be a useful experimental ground for the theoretical debate, for two main reasons. First, L2 learners might be slower when they have to process their L2 and this processing is more effortful if they are not balanced bilinguals (Cummins, 1977). This leads to the predictions that L2 processing will be linked to fewer pragmatic interpretations. Second, balanced bilinguals might show cognitive strengthening (i.e., stronger executive functions) and would facilitate switching from the logic to the pragmatic interpretation, or taking into consideration both interpretations to evaluate the more appropriate one, compared to monolinguals (Bialystok et al., 2009; but see, also, Sorace, 2011; 2016).

While most of the available data from both children and adults appear to show that pragmatic implicatures are costly to make, the evidence on bilinguals is more mixed. Some recent studies found no differences between a native language (L1) and L2 processing in pragmatic answers for scalar implicatures (Antoniou & Katsos, 2017; Antoniou et al., 2018; Dupuy et al., 2018; Syrett et al., 2016; Syrett et al., 2017). On the

contrary, other studies found an increase in pragmatic answers by testing sensitivity to conversational violations (Siegal et al., 2009) and scalar implicatures (Siegal et al., 2007) in both early bilingual children and bilingual adults (Slabakova, 2010; Snape & Hosoi, 2018). Explanations for such results on scalar implicatures were attributed, on one side, to decreased processing resources: bilinguals give less logical answers because implicatures are the default answers and they don't have enough resources to cancel them, in line with the default models (e.g., Slabakova, 2010). On the other side, results were attributed to increased cognitive skills, that is bilinguals give more pragmatic answers because to compute implicatures is costly and they have more cognitive resources, in line with the non-default models (e.g., Siegal et al., 2009).

In one of the first adult studies on scalar-implicatures computation and L2 processing, Slabakova (2010) asked English monolinguals and Korean-English bilinguals that were living in the USA to judge the acceptability of underinformative English sentences that included *some*. In addition, a group of native Korean speakers performed the judgment task with materials translated into Korean. In the first experiment, participants were presented with 40 sentences without context (8 true with *all*, 8 false with *all*, 8 felicitous with *some*, 8 infelicitous with *some*, and 8 fillers) and were asked to decide whether they agreed or disagreed with each sentence. Target sentences were of the form of 'Some Xs have Ys', like in 'Some elephants have trunks'. In the second experiment, the author provided participants with a context to make their decision. In both experiments, bilinguals chose the pragmatic interpretation more often than English monolinguals and more often than the Korean

speakers performing the task in Korean. According to Slabakova, these findings support the default models: since, by hypothesis, bilinguals have less cognitive resources at disposal to perform the task, an increase of pragmatic responses suggests they are automatic and easily available. However, this explanation may not be viable. On the one hand, the bilingual participants were categorized as having intermediate to high English proficiency by their TOEFL scores upon admission to a U.S. university, all were living in the U.S., and they used English daily at the time of the study. As Bouton (1992) demonstrated, non-native speakers' computation of implicatures improves, reaching the native-speakers' competence after 4 ½ years of living in the L2 foreign country. This improving after immersion can be related to a faster access to the pragmatic interpretation since "students develop the knowledge and skills that they need to interpret the implicatures appropriately" (Bouton, 1992: 64). Such skills might be a reduced cognitive effort in order to process sentences and higher morpho-syntactic skills that – according to recent studies – seem to correlate with performances on scalar implicatures (Mazzaggio et al., 2017??? Non si capisce a quale pubblicazione ci si riferisca - vedi bibliografia – direi invece che potrebbe essere utile aggiungere, dove pertinente, un riferimento a: Mazzaggio & Surian, 2018); fewer experience and exposition to language may affect participants' confidence on their meta-linguistic judgments (Katsos & Bishop, 2011: 77).

Dupuy and colleagues (2018) tested scalar-implicature processing in French adults learning English or Spanish as their L2 by adopting both a within- and a between-subject design. Participants performed a written Truth-Value Judgment Task

on twelve control items (true all, false all and felicitous some) and on eight target items in which *some* was used in an infelicitous way. In the within-subject design L2 learners saw both the L1 and the L2 sentences. 90 French students, divided into a first group of monolinguals, a second group of learners of English and a third group of learners of Spanish, participated at the experiment. The L2 learners were all upper-intermediate learners (B2 level of the Common European Framework of Reference for Languages). Even if results replicated Slabakova's (2010) results (i.e., L2 learners were significantly more pragmatic than the controls), authors suggested that their overall results did not support Slabakova's conclusion since L2 learners maintained the same strategy to answer in both their L1 and L2, giving the same proportion of pragmatic answers. In their opinion "L2 learning induces a pragmatic bias. [...] Since the participants knew they would be tested in two languages, we can legitimately wonder whether L2 learning results in enhanced pragmatic abilities or if being tested in two languages makes participants more aware of pragmatic cues" (2018: 13-14). In the between-subject condition, 46 French students of an English Studies Degree participated at the experiment. Stimuli were the same of the within-subject condition; half bilinguals saw only the L2 sentences and half saw only the L1 sentences. Results showed that L2 learners did not answer more pragmatically in their L2 than in their L1. In conclusion, both in the within- and in the between-subject designs, rates of pragmatic answers in the L2 condition were similar compared to the L1 condition.

The present study aimed at providing a more stringent test of the competing models (i.e., default and non-default; cost-free vs. costly/delayed) and depart from

other studies since it tests oral processing of scalar implicatures in L2 learners. All participants were Italian native speakers, living in Italy and learning either English or Spanish as their L2. We decided to test late L2 learners that were living in their mother tongue country because this made their L2 processing more effortful than L1 processing (Andreou & Karapetsas, 2004; Cummins, 1977; Sampath, 2005). Our study employed both an L2 with a quantifier-system similar to L1 and an L2 with a different system. The Italian language (L1) has two different existential quantifiers used with countable nouns: *qualche* that must be used followed by a singular noun and *alcuni* (masculine form) or *alcune* (feminine form) that must be followed by plural nouns. Since there are no principled reason to choose one form over the other, in our experiment we tested the form *alcuni/e*, already used in Guasti et al. (2005). As L2 we tested both English (Experiments 1 and 2), a language with only one quantifier (i.e. *some*), and Spanish (Experiment 2), a language that, like Italian, has two different terms, *unos* and *algunos*, even in – differently from Italian – in non-partitive contexts just *algunos* trigger the implicature; *unos* is perfectly fine if used to refer to all members of a set (for a detailed explanation of differences, see Gutiérrez-Rexach, 2001).

In addition, differently from the other studies on bilinguals, our procedure imposed a time limit for interpreting sentences, thereby adding to the resource demands of the task. Since, as we have seen, implicatures need processing resources (i.e., they require a number of computations that go beyond accessing the lexical meaning of a word from the lexicon) we might expect a decrease of the rate of SIs by

overburdening cognitive skills that allow participants to do the computations really quickly (i.e., working memory resources). Moreover, we assumed that participants in L2 conditions should be under a greater cognitive load, due to their weaker linguistic competence paired with time constraints and oral processing, compared to the participants in the L1 condition. Under this assumption, we aim to investigate whether the frequency of pragmatic answers, defined as rejection of underinformative statements, differs in the two groups. Therefore, if the pragmatic interpretations are the non-default interpretations of underinformative utterances and they require cognitive effort, we expect more frequent pragmatic answers in the low-cognitive load, L1 condition than in the high-cognitive, L2 condition.

2. Experiment 1

2.1. Method

2.1.1. Participants

Participants were 86 Italian university students (69 women, mean age 22.0 years, $SD = 4.35$). They were divided into two groups: the L1 group ($N = 31$, 6 men, mean age 23.4 years, $SD = 6.78$) has been tested in their native language (Italian) and the L2 group ($N = 55$, 25 women, mean age 21.1 years, $SD = 1.53$) has been tested in a non-native language (English). Based on an assessment of level of English proficiency by the University Language Centers (according to the Common European Framework of Reference for Languages), the L2 group was consisted of people with low

proficiency (N = 8) at the A2 level, with intermediate-low proficiency (N = 24) at the B1 level, with intermediate-high proficiency (N = 17) at the B2 level and with high proficiency (N = 6) at the C1 level. In order to have more homogenous groups for the analyses we created two groups, a low-proficiency group with people at A2 and B1 levels (N = 32) and a high-proficiency group with people at the B2 and C1 level (N = 23). Participants were not simultaneous bilinguals.

2.1.2. Materials and procedure

The materials consisted of 32 English sentences and their translated Italian equivalents (Appendix 1). Half of the sentences began with the quantifier *some* and half began with the quantifier *all*. Eight of the sentences with *some* were true (e.g., Some dogs are Labradors) and 8 were underinformative (logically true but pragmatically false, e.g., Some children are humans). Eight of the sentences with *all* were universally true (e.g., All snakes are reptiles) and 8 were universally false or absurd (e.g., All animals are carnivorous). A proficient Italian-English bilingual digitally recorded the English and Italian sentences.

The recorded sentences were presented in a Sentence Evaluation Task using PowerPoint software running on a laptop computer. On each trial, a sentence was played and participants indicated whether they agreed or disagreed with it by marking "Yes" or "No", respectively, on the corresponding number on a printed form. The participants had three seconds to produce their response before the recording on

the next trial would be played.¹ The English sentences were presented to the L2 group, and the Italian sentences were presented to the L1 group. Participants were tested in groups at the beginning of language lessons. Participants with an L1 different from Italian were excluded.

2.2. Results and discussion

"No" responses to underinformative sentences with *some* indicated a pragmatic interpretation whereas "Yes" responses indicated a logical interpretation. Figure 1 shows the mean numbers of accurate responses for all conditions (All True, All False, Some Underinformative and Some True); in the Some Underinformative condition we considered the pragmatic response as the correct one (i.e. True for "No" responses and False for "Yes" responses).

The accuracy for the All True Condition was 99.6% for participants tested with their L1, while it was 91.6% for participants tested with their L2. For the All False Condition the accuracy was 100% for participants tested with their L1 and 88% for participants tested with their L2. For the Some True Condition the accuracy was 100% for participants tested with their L1 and 90.1% for participants tested with their L2. Finally, for the Some Underinformative Condition, the pragmatic accuracy was 66.1%

¹ We are aware that for L1 processing three seconds are used as 'long time condition' (e.g. Bott & Noveck, 2004) but for L2 processing this could be a short time. Indeed, also in Bott and Noveck's work, participants with three seconds to answer (tested in their L1) were not at ceiling at control sentences. We believe that both assessing oral processing and imposing a time limit to answer added a cognitive load.

for participants tested with their L1 and 48.7% for participants tested with their L2 (see Figure 1).

We conducted a statistical analysis using a Generalized Linear Mixed Model (GLMM)² considering accuracy as a dichotomous dependent variable, allowing the slopes and intercepts for the within-participants factor Type to change across participants and without the correlations of random effects. The model reported below that best fits the data included subjects and items as random factors³, the Quantifier (All vs. Some), the Type (True vs. False/Underinformative) and the Language (L1 vs L2) as fixed effects, as well as their interactions. The main effects of Quantifier ($\beta = -1.48$; $z = -6.13$; $p < .001$), Type ($\beta = 2.23$; $z = 5.35$; $p < .001$) and Language ($\beta = -4.70$; $z = -3.41$; $p < .001$) were significant, as well as the two-way interaction Quantifier-Type ($\beta = 6.39$; $z = 3.11$; $p = .002$) and the three-way interaction Quantifier-Type-Language ($\beta = -11.53$; $z = -2.12$; $p = .03$). It is important to underline that a two-way interaction between Quantifier and Type represents the effect of the Underinformative sentences vs. the controls whereas a three-way interaction represents how this effect is different across the groups of participants.

² All the GLMM models presented in this paper have been conducted using R and the LmerTest package (Kuznetsova et al., 2017), with maximal random factor matrix (Barr et al., 2013). The correlations of random effects were removed for convergence issues in some of the models, as specified in the text, and all the categorical factors centered prior to analysis.

³ The GLMM that includes the full matrix of 2-way and 3-way interactions did not converge. Thus, we computed the three simpler models with each combination of a 2-way interaction, which resulted significant in every model (Type-Language: $\beta = -3.2$; $z = -3.09$; $p < .01$; Quantifier-Language: $\beta = 3.46$; $z = 3.33$; $p < .001$; Quantifier-Type: $\beta = 2.75$; $z = 5.87$; $p < .001$). We report in the text the model that significantly provides the best fit.

This shows that, overall, participants do not answer to the Some Underinformative condition like for the control All True, All False and Some True conditions. Moreover, the language seems to have an effect on the answers in every condition, with fewer accurate answers in the control conditions and fewer pragmatic answers in the Some Underinformative condition for participants tested with their L2. Yet, this decrement in accuracy is greater in the Some Underinformative condition as indicated by the means. The significant three-way interaction between Quantifier, Type and Language (see Figure 1) supports this conclusion.

To further investigate the different judgments of L1 vs. L2 participants in the critical conditions we conducted a GLMM considering the acceptance rate (dichotomous: "Yes" vs. "No" answer) in the Some (True and Underinformative) conditions. Given that neither the GLMM nor a simpler Generalized Linear Model provided a good fit for the data, we investigated the effect of Language in the single conditions separately, including Language as fixed factor and item and subjects as random factors. In GLMM computed in the Some Underinformative condition the main effect of Language was significant ($\beta = 1.65$; $z = 2.41$; $p = .02$) whereas in the one computed in the Some True condition it wasn't ($p > .1$). This suggests that the effect of Language was stronger in the Some Underinformative condition.

Similar results were obtained by investigating consistency in response patterns: individual participants were classified as consistent pragmatic or logical responders if they rejected as false or accepted as true, respectively, 6 or more underinformative sentences (out of 8). All other participants were classified as non-consistent

responders (L1 = 2/31; L2 = 10/54). Pragmatic responders were more frequent in the L1 (20/29) than in L2 group (24/44), whereas logical responders were more frequent in the L2 group than in the L1 group (9/29 and 20/44, respectively). We conducted a GLMM to test whether this difference was significant, and the main effect of Language approached significance (beta -0.9808, z: -1.951, p = .051)

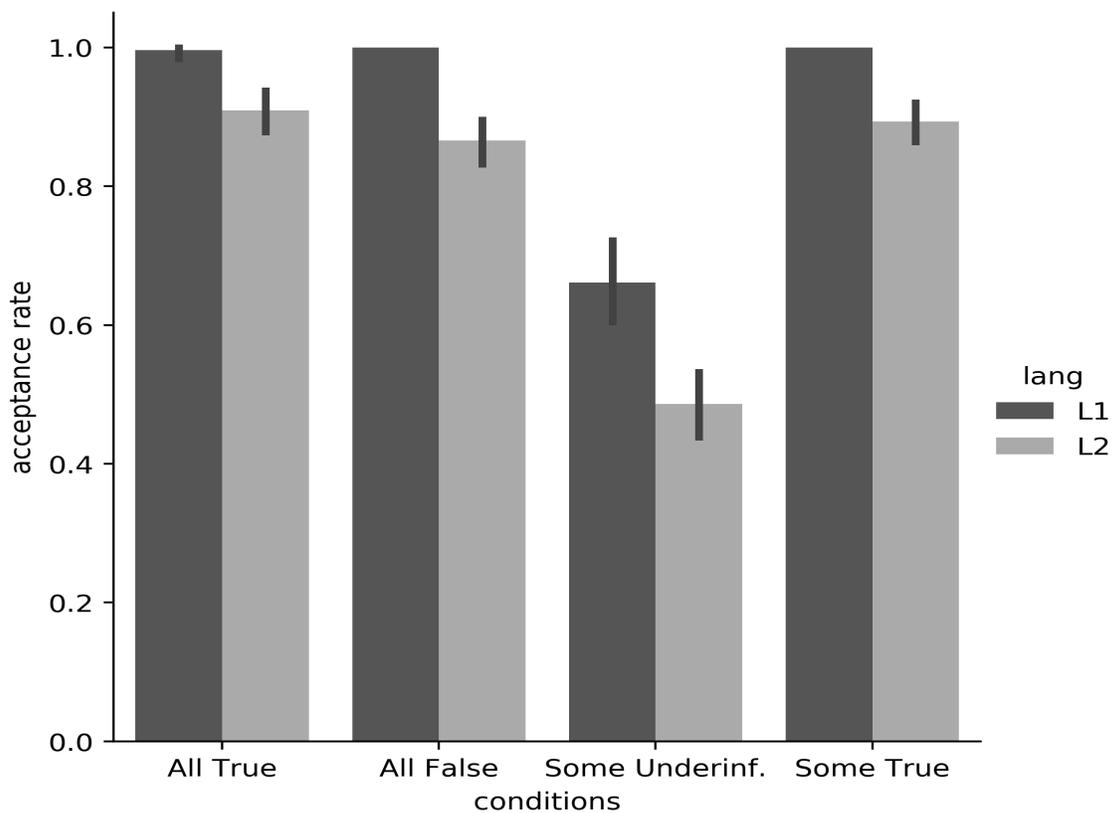


Fig 1. Mean number of accurate responses for all conditions (All True, All False, Some Underinformative and Some True) in the L1 group and the L2 group (English) in Experiments 1. In the Some Underinformative condition we considered the pragmatic response as the correct one Error bars are standard error of the mean.

3. Experiment 2

Experiment 1 showed a greater tendency in the L1 group compared to the L2 group to choose pragmatic answers in the underinformative items. The aim of Experiment 2 was to replicate and extend the results of Experiment 1 by testing not only English as L2, but also Spanish. Moreover, the L2 group participants' familiarity with some of the nouns in the experimental items was assessed.

3.1. Method

3.1.1. Participants

Participants were 393 Italian university students (305 women, mean age 21.9 years, SD = 2.52). They consisted of a L1 group (N = 246, 63 men, mean age 20.5 years, SD = 2.67), a L2 group tested in English (N = 61, 46 women, mean age 22.5 years, SD = 2.59) and a L2 group tested in Spanish (N = 86, 10 men, mean age 21.4 years, SD = 1.14). None of these students had participated in Experiment 1. The L2 proficiency was assessed by the University Language Centers according to the Common European Framework of Reference for Languages. The Italian-English bilinguals group was formed by students with a low-proficiency level (A2 level: N = 8), students with a low-intermediate proficiency level (B1: N = 29), students with a high-intermediate proficiency level (B2: N = 12) and students with an advanced proficiency level (C1-C2: N = 12). In order to have more homogenous groups for the analyses we created two groups, a low-proficiency group with people at A2 and B1 levels (N = 37) and a high-

proficiency group with people at the B2, C1 and C2 levels (N = 24). The Italian-Spanish bilinguals group was formed by students with a low-intermediate proficiency level (B1: N = 31), students with a high-intermediate proficiency level (B2: N = 50) and students with an advanced proficiency level (C1: N = 5). In order to have more homogenous groups for the analyses we created two groups, a low-proficiency group with people at B1 level (N = 31) and a high-proficiency group with people at the B2 and C1 levels (N = 55). Participants were not simultaneous bilinguals.

3.1.2. Material and procedure

The sentences were the same as in Experiment 1 except that they included Spanish translation equivalents, which were digitally recorded by a highly proficient Italian-Spanish bilingual. The procedure was the same as in Experiment 1. In addition, after the task, both groups of bilinguals received a list of some of the nouns from the experimental sentences and were asked to write their translation equivalents in Italian. Participants had been tested in groups at the beginning of language lessons and participants with a L1 different from Italian had been excluded.

3.2. Results

The results of the translation task showed that both bilingual groups were familiar with the nouns. The English L2 group correctly translated an average of 16.6 of the 17 nouns (SD = 0.71) and the Spanish L2 group correctly translated an average

of 14.6 of the 15 nouns (SD = 0.67).

For Experiment 2, the L1 group and the two L2 groups' mean numbers of accurate responses for all conditions are given in Figure 2; in the Some Underinformative condition we considered the pragmatic response as the correct one.

The accuracy for the All True Condition was 97.5% for participants tested with their L1, while it was 88.5% for participants tested with English as their L2 and 91.4% for participants tested with Spanish as their L2. For the All False Condition the accuracy was 98.6% for participants tested with their L1, 92.6% for participants tested with English as their L2 and 96.2% for participants tested with Spanish as their L2. For the Some True Condition the accuracy was 99.2% for participants tested with their L1, 94% for participants tested with English as their L2 and 96.1% for participants tested with Spanish as their L2. Finally, for the Some Underinformative Condition the pragmatic accuracy was 81.1% for participants tested with their L1, 59.8% for participants tested with English as their L2 and 64% for participants tested with Spanish as their L2.

We conducted a statistical analysis using a GLMM considering accuracy as the dependent variable allowing the slopes and intercepts for the within-participants factor Type to change across participants and without the correlations of random effects. The model that best fits the data included subjects and items as random

factors⁴, the Quantifier (All vs. Some), the Type (True vs. False) and the Languages (L1 vs English L2 vs Spanish L2) as fixed effects, as well as their interactions. The main effects of Quantifier ($\beta = -1.09$; $z = -4.33$; $p < .001$), Type ($\beta = 1.41$; $z = 4.91$; $p < .001$) and Language ($\beta = -1.74$; $z = -14.40$; $p < .001$) were significant, as well as the two-way interaction Quantifier-Type ($\beta = 4.27$; $z = 8.17$; $p < .001$) and the three-way interaction Quantifier-Type-Language ($\beta = -0.97$; $z = -2.3$; $p = .02$). This shows that, overall, participants were generally more accurate with *all* than with *some*, and in True conditions than in False/Underinformative ones. Moreover, the two significant interactions suggest that participants do not answer to the Some Underinformative condition like for the control All True, All False and Some True conditions, as indicated by the average ratings where the difference between L1 and L2 is greater in the Some Underinformative condition. This pattern replicates the results of Experiment 1. Moreover, again the languages seem to have an effect on the answers in all Conditions, with less accurate answers in the control conditions and less pragmatic answers in the Some Underinformative condition for participants tested with their L2s. As in Experiment 1, this difference was significantly greater in the Some Underinformative condition as compared to the other conditions.

⁴The GLMM that includes the full matrix of 2-way and 3-way interactions did not converge. Thus, we computed the three simpler models with each combination of a 2-way interaction, which resulted significant in every model (Type-Language: $\beta = -.7$; $z = -4.04$; $p < .001$; Quantifier-Language: $\beta = 0.53$; $z = 3.09$; $p < .01$; Quantifier-Type: $\beta = 4.06$; $z = 7.9$; $p < .001$). We report in the text the model that significantly provides the best fit.

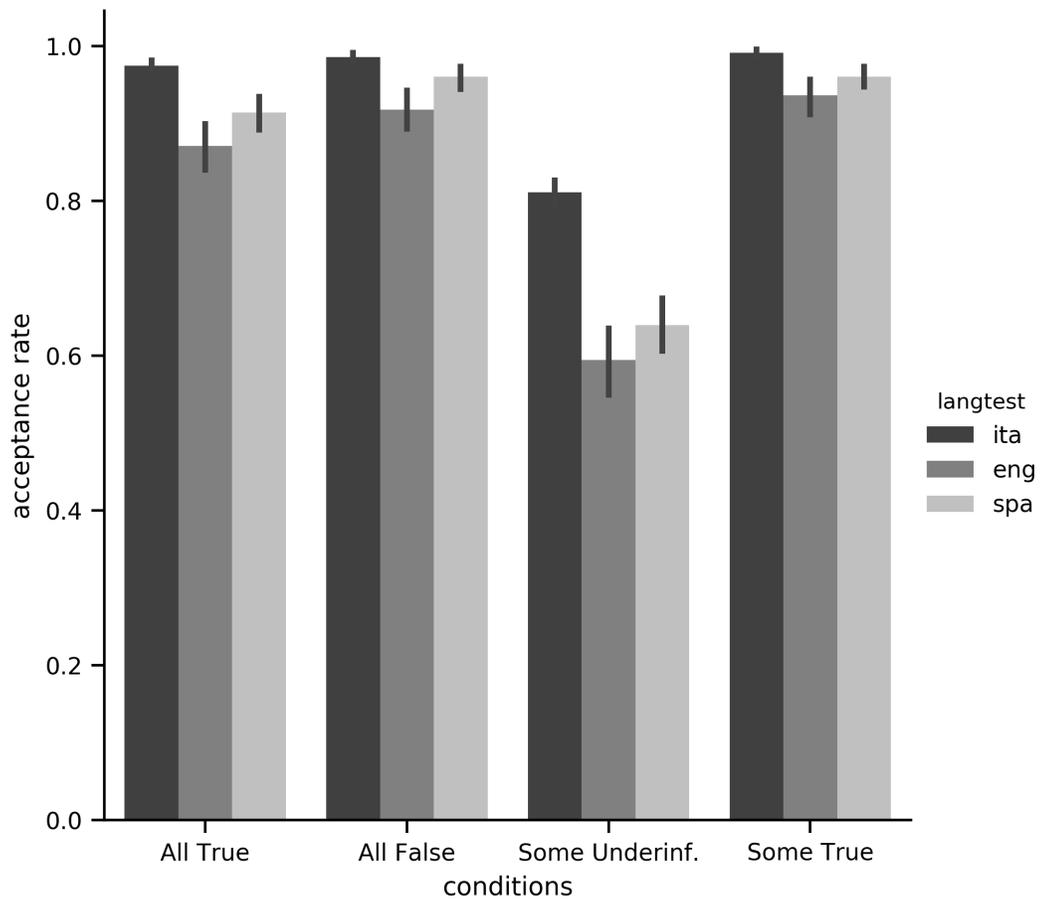


Fig 2. Mean number of accurate responses for all conditions (All True, All False, Some Underinformative and Some True) in the L1 group and the L2 groups (English and Spanish) in Experiments 2. In the Some Underinformative condition we considered the pragmatic response as the correct one. Error bars are standard error of the mean.

Like for Experiment 1, we wanted to further investigate participants' judgments in the critical conditions with Some. For this reason, we conducted a GLMM considering the acceptance rate (dichotomous) in the Some (True and Underinformative) condition, with free random intercepts and slopes for the factor

Type. The model reported below that best fits the data included subjects and items as random factors, the Condition (True vs. False/Underinformative) and the Language (L1 vs L2s) as fixed effects, as well as their interactions. We found an effect of Condition ($\beta = 7.58$; $z = 14.66$; $p < .001$) and also the two-way interaction Condition-Language was significant ($\beta = -3.96$; $z = -7.41$; $p < .001$). These results confirm that the effect of Language is properly on the Some Underinformative condition.

In order to understand whether participants that answered consistently might have contributed to the effect, we decided to run analysis just on consistent participants. Individual participants were classified as consistent pragmatic or logical responders if they rejected as false or accepted as true, respectively, 6 or more underinformative sentences (out of 8). All other participants were classified as non-consistent responders (L1 = 46/246; L2 = 40/147). Pragmatic responders were more frequent in the L1 (176/246) than in L2 group (75/147), whereas logical responders were more frequent in the L2 group than in the L1 group (32/147 and 24/246, respectively). We conducted a GLMM to test whether this difference was significant, and the main effect of Language was significant (beta -1.1407, z: -3.762, $p < .001$)

4. An Overall Analysis on Proficiency

In order to control for an effect of proficiency levels on the accuracy levels in the L2 groups, we decided to combine the L2 English participants of Experiment 1 and 2. We created two proficiency groups, a low-proficiency group with people at A2 and B1 levels and a high-proficiency group with people at the B2 and C1 level. For the

English group we had 69 participants at the low-proficiency level and 47 participants at the high-proficiency level. For the Spanish group we had 31 participants at the low-proficiency level and 55 participants at the high-proficiency level.

The mean numbers of accurate responses for all conditions in the English group, divided by proficiency levels are given in Figure 3, while the mean numbers of accurate responses for all conditions in the Spanish group are given in Figure 4; in the Some Underinformative condition we considered the pragmatic correctness.

Considering the English L2 group, the accuracy for the All True Condition was 88.1% for low-proficient participants, while it was 94.2% for high-proficient participants. For the All False Condition the accuracy was 91.2% for low-proficient participants, while it was 94.5% for high-proficient participants. For the Some True Condition the accuracy was 90.8% for low-proficient participants, while it was 96.1% for high-proficient participants. Finally, for the Some Underinformative Condition the pragmatic accuracy was 52.4% for low-proficient participants, while it was 62.6% for high-proficient participants. Considering the Spanish L2 group, the accuracy for the All True Condition was 89.4% for low-proficient participants, while it was 87.2% for high-proficient participants. For the All False Condition the accuracy was 92.2% for low-proficient participants, while it was 93.1% for high-proficient participants. For the Some True Condition the accuracy was 91.2% for low-proficient participants, while it was 97.4% for high-proficient participants. Finally, for the Some Underinformative Condition the pragmatic accuracy was 63.5% for low-proficient participants, while it was 54% for high-proficient participants.

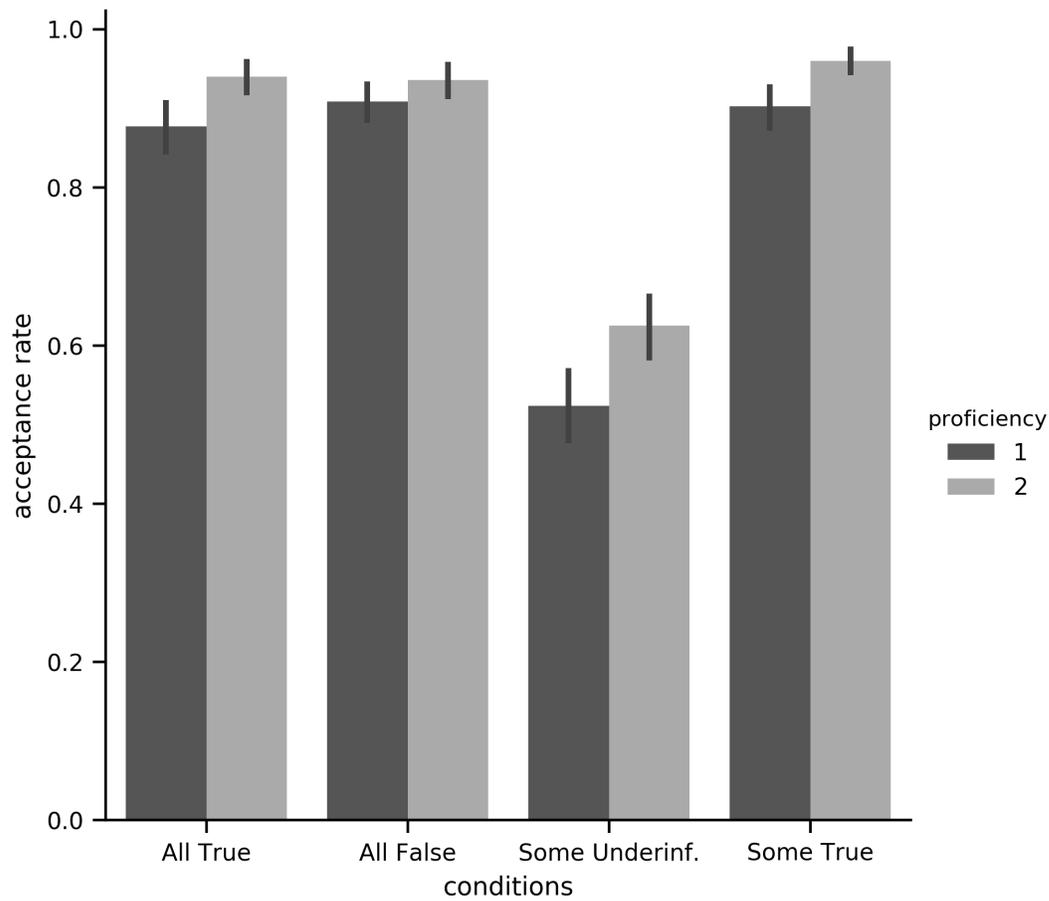


Fig 3. Mean number of accurate responses for all conditions (All True, All False, Some Underinformative and Some True) in the L2 English group (Experiment 1 and 2), divided by proficiency level; 1 (dark-grey bar) represents the low-proficiency group and 2 (light-grey bar) represents the high-proficiency group. In the Some Underinformative condition we considered the pragmatic response as the correct one. Error bars are standard error of the mean.

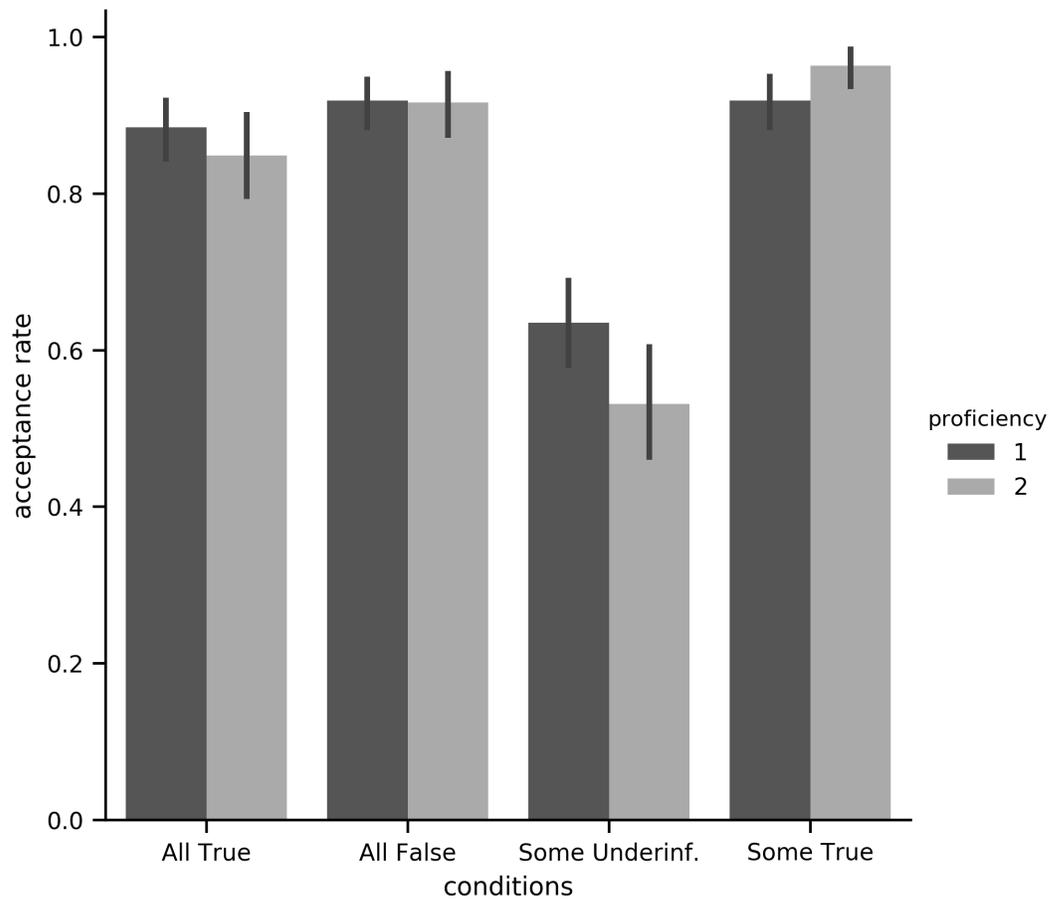


Fig 4. Mean number of accurate responses for all conditions (All True, All False, Some Underinformative and Some True) in the L2 Spanish group, divided by proficiency level; 1 (dark-grey bar) represents the low-proficiency group and 2 (light-grey bar) represents the high-proficiency group. In the Some Underinformative condition we considered the pragmatic response as the correct one. Error bars are standard error of the mean.

We conducted GLMM considering accuracy as the dependent variable, including subjects and items as random factors, random slopes and intercepts for the factor

Type. We also included the Quantifier (All vs. Some), the Type (True vs. False), the Proficiency group (low vs high) and the Languages (L1 vs English L2 vs Spanish L2) as fixed effects, as well as their interactions. The main effects of Quantifier ($\beta = -1.13$; $z = -4.41$; $p < .001$), Type ($\beta = 1.36$; $z = 5.3$; $p < .001$) and Proficiency ($\beta = 0.79$; $z = -4.41$; $p < .001$) were significant, but not the Language ($\beta = 0.48$; $z = 0.26$; $p = .79$). Also the two-way interaction Quantifier-Type ($\beta = 3.26$; $z = 6.33$; $p < .001$), the three-way interaction Quantifier-Type-Proficiency ($\beta = -0.92$; $z = 2.54$; $p = .01$) and the four-way interaction Quantifier-Type-Proficiency-Language ($\beta = 2.01$; $z = 2.62$; $p = .009$) were significant, but not the three-way interaction Language-Type-Proficiency ($p > .1$) or Quantifier-Type-Language ($p > .1$). This analysis overall replicates the results of Experiment 1. Participants were sensitive to the difference between truth vs. falsity and truth vs. underinformativeness. Yet, these results also show that the proficiency level has an influence on the judgments of the Some Underinformative condition compared to the control conditions, which is different with respect to the L2. Participants with English as L2 with a lower proficiency level, gave less pragmatic answers than participants with high proficiency level. The pattern is reversed when considering Spanish as a L2, with less pragmatic answers in the high proficiency group. This behavior results in the four-way interaction reported above. This is an unanticipated finding and it will be discussed in detail in the discussion section.

5. Discussion

Building on prior research on pragmatic processing in L2 computation, for the

first time we assessed the oral processing of scalar implicature in L2 learners, tested either in their L1 (Italian) or in their L2s (English or Spanish). We found that, participants tested with materials in L2 were less likely to derive a pragmatic interpretation of underinformative sentences compared to participants tested with materials in L1. On the assumption that L2 oral processing, under time constraints, imposes a higher cognitive load than L1 processing, the decrease in pragmatic interpretations of underinformative sentences can be taken as evidence that deriving such pragmatic interpretations is costly and non-automatic.

Therefore, the current pattern of results is consistent with the non-default models' view of scalar implicatures as well as those approaches maintaining that scalar computation is not cost-free. In contrast, our results run against the default models and the cost-free models of scalar computation. Indeed, according to the former class of approaches, in order to compute a scalar implicature, the listener should execute several steps. When interpreting a sentence like "Some Xs are Ys", first we consider the literal meaning of the sentence; then, we generate the more informative-alternative sentence "All Xs are Ys"; finally, we negate the more informative alternative in order to strengthen the meaning of the sentence and to obtain the pragmatic interpretation "Some but not all Xs are Ys". If we do not have enough time and cognitive resources to go through all those steps, we might be limited to a semantic interpretation (i.e. "Some and possibly all Xs are Ys").

Whilst our results found that participants tested in their L2 performed more poorly (with less pragmatic answers) than when they were tested in their L1, Dupuy

et al., (2018) found that participants performed similarly in the two language conditions. Several reasons may account for the discrepancy between our results and Dupuy et al.'s. First, participants in our study had time constraints since they were asked to perform their task as quickly as possible, whereas participants in Dupuy et al.'s study did not. Second, materials in our study were presented orally whereas in their study they used written items. Thus, processing auditory information and using L2 under time constraint might have increased the difficulty of our task, requiring more resources for performing the item evaluations.

However, time constraints were absent both in Dupuy et al.'s and in Slabakova's studies, that also both used written materials. Therefore, those factors cannot account for the discrepancies between those two studies. As we have previously discussed in the Introduction, bilinguals in Slabakova's study were immersed in an L2 environment and their L2 processing was probably as automatic as L1 processing. Notice that other studies that did not impose time constraints failed to replicate Slabakova's results, reporting no difference in pragmatic interpretation between L1 and L2 speakers (Antoniou & Katsos, 2017; Antoniou et al., 2018; Dupuy et al., 2018; Syrett et al., 2016; Syrett et al., 2017). These inconsistencies may derive from a third factor.

Considering our study and the studies that found no differences between L1 and L2 processing, we might speculate that immersion probably played a more important role than participants' proficiency (Fortune, 2012). Accordingly, when Bouton (1992) tested bilingual students on conversational implicatures immediately after their arrival in the USA and then after 4 years and a half, he found a great improvement in

their performance. Indeed, Cummins' Threshold Hypothesis (1977, 1978; see also Ardasheva et al., 2012; Farrell, 2011; Green, 1986; Karapetsas & Andreou, 2004; Ricciardelli, 1992; Sampath, 2005) shows how only balanced bilinguals display the cognitive advantages firstly theorized by Peal and Lambert (1962).

The fact that Slabakova's bilinguals were more likely to derive a pragmatic interpretation in their L2 than in their L1 may reflect such general metacognitive advantage that proficient bilingualism bestows (Adesope et al., 2010; Bialystok & Senman, 2004; Bialystok & Shapero, 2005; Kushalnagar et al., 2010; Mezzacappa, 2004; Pelham & Abrams, 2014). Following this line, when bilingual children – exposed to two languages every day – have been tested on the detection of violations of Gricean maxims, they performed better than monolinguals (Siegal et al., 2010). Current results are consistent with findings showing that bilingual adults perform better on a logical reasoning task when they are tested using a foreign language than when they are tested using their native language (Costa et al., 2014, pp. 4-5). In other words, we suggest that the metacognitive advantages only show up in the case of high proficiency in L2.

Another aspect that we might take into consideration is that there is a methodological difference between Slabakova' study and the present one. As previously pointed out, Slabakova used a Truth-Value Judgment Task (TVJT), while we used a Sentence Evaluation Task. In many works (e.g., Guasti et al., 2005) it has been demonstrated that different methodologies might produce a different rate of pragmatic answers in the same population. However, we believe that such aspect

cannot fully account for the difference between our study and Slabakova's one since we should have expected an overall difference (more or less pragmatic answers, independently of the group, like in Guasti et al.'s work) and not the opposite pattern as we have found.

Now, we should consider our results in comparison with both the Dupuy et al.'s ones and with Slabakova's ones, highlighting the inversion in the pattern of results between our study and Slabakova's, whose design we followed quite closely. On the one hand, if just immersion played a crucial role, we should have found no differences between L1 and L2 participants, like for Dupuy et al. On the other hand, if the use of time constraints and the oral processing were the only factors in play, it is hard to explain the differences between Slabakova's results and the results of other studies that found no differences between L1 and L2. Hence, we suggest that the three factors that we mentioned (i.e., the immersion in the L2 environment, the use of time constraints and the presentation of items in the oral modality) might have jointly played a role in yielding our results. Within our account's framework, we may hint at a continuum from immersed bilinguals with high cognitive resources (who show an increase in pragmatic responses to underinformative sentences, since they have the resources to easily compute scalar implicatures), to bilinguals tested in their L1 environment (who show no difference in pragmatic responses with respect to L1, because they have the time to derive the scalar implicatures), to bilinguals tested in their L1 environment under time constraints (who show a decrease in pragmatic responses to underinformative sentences, since their limited resources allow for more

automatic responses to be given). The available data suggest that these factors are involved but do not allow to tell whether their contribution is additive or they interact.

Furthermore, it is of interest to consider both the L1 and the L2 tested, since our results on the proficiency level seems to suggest that languages may affect the pattern of behavior, which is something not entirely new in the literature (Katsos et al., 2016). Interestingly, in our data, we saw that if English and Spanish as L2s behave similarly in opposition to L1 (i.e., with less pragmatic answers), the proficiency levels of participants seem to affect differently the two languages groups. Specifically, participants with English as L2 gave fewer pragmatic answers in the low proficiency group than in the high proficiency one, whereas participants with Spanish as L2 gave fewer pragmatic answers in the high proficiency group compared to the low proficiency one. Our predictions go in the direction of English participants' behavior: we expect that, since listening to a sentence in L2 is cognitively more demanding, participants with less morpho-syntactic skills will have greater difficulties in dealing with underinformative sentences, hence producing fewer pragmatic responses. Furthermore, the different behavior of participants with Spanish as L2 is unexpected because in the analysis of responses the two languages behaved in a similar way, as attested by the lack of significant interaction between language, type and quantifier. Does this result constitute a problem for the idea that a more logical behavior in L2 is due to higher cognitive demands? Two considerations suggest that this might not be a problem. First of all, we clearly demonstrated within two experiments that the

cognitive operations involved in processing L1 are different, qualitatively and quantitatively, from those involved in processing L2. Second, in our study the proficiency level is computed on participants' reports, without considering other factors such as the duration of the immersion in a linguistic community, thus this index should be taken with caution (while the difference between a L1 and a L2 is a fact).

Having stated this proviso, the effect of proficiency in Spanish might be due to different factors that boil down to specific properties of this language. For instance, Spanish grammar is very similar to Italian grammar, unlike English one. Moreover, Spanish and Italian have commonalities at the lexical/phonological level. Indeed, Spanish includes a quantifier, *algunos*, which is lexically very similar to the Italian *alcuni* that constitutes the target word in our experiments. For this reason, non-proficient participants might have adopted a strategy counting on lexical similarities between the two languages (transfer mechanism) to reason similarly to their L1. Differently, high proficient participants are probably more used to process Spanish without any sort of transfer and with all the cognitive difficulties that this implies. Non-proficient participants in the English L2 condition cannot take advantage of lexical similarities because Italian and English are very different. As a second hypothesis, the effect of similarity between Italian and Spanish might have played out the opposite way: high proficient participants might have automatically activated their L1 while processing Spanish sentences, and this co-activation resulted in interference and processing strain that caused the rate of pragmatic answers to drop.

Further investigation on a wider range of quantifiers and typologically different languages needs to be conducted to assess the validity of these possibilities.

The present results also do not allow deciding between competing models that claim that scalar implicatures are costly. For example, our results can be explained for by the Lexical Account on the computation of scalar implicatures (Barner et al., 2011; Foppolo et al., 2012). As we have seen, in this account, the problems with implicatures are a consequence of limitations in representing/accessing lexical items in scale. Thus, one may propose that the logical interpretation of underinformative sentences results from difficulty with accessing the <some, all> scale. Applying this account to our results, one needs to assume either (or, possibly, both) of two viewpoints. For the majority of our participants fluency in L2 was not very high and we may thus assume that their mastering of the <some, all> scale was not optimal: non-ceiling performances in control sentences point to this possibility. Thus, access to the <some, all> scale might have been affected by limitations on the representations of the comprising items. The second interpretation rests on the time-constraints factor. So, it might be the case that reduced time in our study might have prevented L2 participants from accessing the scale in an optimal way, i.e. fully exploiting their knowledge about the scale. If we further assume that the items more directly accessible were also the more easily represented and available, this may be seen as a special case of the non-default model. Further research is needed to adjudicate between different non-default models or those based on costly implicature computation (e.g. Relevance Theory and Lexical accounts); a main focus should be on

oral processing of scalar implicatures with participants immersed in a L2 environment, with or without time constraints.

The present study had some limitations. First, we checked whether our L2 participants were not simultaneous bilinguals and we asked them their proficiency level; however, also other factors related to their L2 skills would be worth considering, such as when they started learning the L2, whether they have any L2 experience outside the formal teaching or whether they spent a period abroad. Second, while we know that, in Italy, English is studied from the 3rd grade we did not collect more detailed information in participants' language learning experiences. Third, it might be interesting to replicate our data with a within-subject design. It would be desirable to address such limitations in future studies.

6. Conclusions

In conclusion, our study brings new data to the field of study of scalar implicature computation by testing a different type of L2 processing (i.e. oral processing) with time constraints. We assessed the oral processing of scalar implicatures in L2 learners and found that participants tested in L2 were less likely to derive a pragmatic interpretation of underinformative sentences than participants tested in L1. Since L2 oral processing, under time constraints, is more resource demanding than L1 processing (Borghini & Hazan, 2018), the present results provide evidence that deriving such pragmatic interpretations is costly and non-automatic. Results of the present study and previous evidence on bilinguals bring supportive data for non-default models and the approaches maintaining that scalar implicature

computation is not cognitively free of cost.

Appendix 1

Items

All true		
English (L2)	Spanish (L2)	Italian (L1)
All snakes are reptiles	Todas las serpientes son reptiles	Tutti i serpenti sono rettili
1. All cats are animals	Todos los gatos son animales	Tutti i gatti sono animali
2. All men are humans	Todos los hombres son personas	Tutti gli uomini sono persone
3. All birds are animals	Todos los pájaros son animales	Tutti gli uccelli sono animali
All cobras are snakes	Todas las cobras son serpientes	Tutti i cobra sono serpenti
All dogs are animals	Todos los perros son animales	Tutti i cani sono animali
All horses are mammals	Todos los caballos son mamíferos	Tutti i cavalli sono mammiferi
All sunflowers are flowers	Todos los girasoles son flores	Tutti i girasoli sono fiori

All false

English (L2)	Spanish (L2)	Italian (L1)
All animals are carnivorous	Todos los animales son carnívoros	Tutti gli animali sono carnivori
4. All cats are dogs	Todos los gatos son perros	Tutti i gatti sono cani
All stones are singers	Todas las piedras son cantantes	Tutte le pietre sono cantanti
All flowers are professors	Todas las flores son profesoras	Tutti i fiori sono professori
All pens are animals	Todos los lápices son animales	Tutte le matite sono animali
All children are grandmothers	Todas las niñas son abuelas	Tutte le bambine sono nonne
All televisions are cars	Todos los televisores son coches	Tutte le televisioni sono automobili
All books are drinks	Todos los libros son bebidas	Tutti i libri sono bevande

Some true

English (L2)	Spanish (L2)	Italian (L1)
1. Some dogs are Labrador	Algunos perros son Labrador	Alcuni cani sono Labrador
Some children are blonde	Algunos niños son rubios	Alcuni bambini sono biondi
Some flowers are red	Algunas flores son rojas	Alcuni fiori sono rossi
Some cats are Persians	Algunos gatos son Siameses	Alcuni gatti sono Persiani
Some houses are rented	Algunas casas son altas	Alcune case sono affittate
Some mobiles are iPhones	Algunos teléfonos son iPhones	Alcuni cellulari sono iPhone
Some dresses are blue	Algunos vestidos son azules	1. Alcun Alcuni vestiti sono blu
Some lakes are big	Algunos lagos son grandes	Alcuni laghi sono grandi

Some underinformative

English (L2)	Spanish (L2)	Italian (L1)
Some children are humans	Algunos niños son personas	Alcuni bambini sono persone
5. Some salmons are fish	Algunos salmones son peces	Alcuni salmonei sono pesci

Some horses are animals	Algunos caballos son animales	Alcuni cavalli sono animali
Some tulips are flowers	Algunos tulipanes son flores	Alcuni tulipani sono fiori
Some dogs are animals	Algunos perros son animales	Alcuni cani sono animali
Some women are humans	Algunas mujeres son personas	Alcune donne sono persone
Some giraffes are animals	Algunas jirafas son animales	Alcune giraffe sono animali
Some roses are flowers	Algunas rosas son flores	Alcune rose sono fiori

Acknowledgements. We are grateful to proff. Remo Job and Anne Reboul for insightful discussions. This work has been supported by grants from the Fondazione ONLUS Marica De Vincenzi.

References

- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207-245.
- Andreou, G., & Karapetsas, A. (2004). Verbal abilities in low and highly proficient bilinguals. *Journal of Psycholinguistic Research, 33*(5), 357-364.
- Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilingualism on implicatures understanding. *Applied Psycholinguistics, 1-47*.
- Antoniou, K., Veenstra, A., Kissine, M. & Katsos, N. (2018). The impact of childhood bilingualism and bi-dialectalism on pragmatic interpretation and processing. In *Proceedings of the 42nd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.
- Ardasheva, Y., Tretter, T. R., & Kinny, M. (2012). English language learners and academic achievement: Revisiting the threshold hypothesis. *Language Learning, 62*(3), 769-812.
- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition, 118*(1), 84-93.
- 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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bialystok, E., & Senman, L. (2004). Executive processes in appearance–reality tasks: The role of inhibition of attention and symbolic representation. *Child Development*, 75(2), 562-579.

Bialystok, E., & Shapero, D. (2005). Ambiguous benefits: The effect of bilingualism on reversing ambiguous figures. *Developmental Science*, 8(6), 595-604.

Bialystok, E., Craik, F. I., Green, D. W., & Gollan, T. H. (2009). Bilingual minds. *Psychological Science in the Public Interest*, 10(3), 89-129.

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: a pupillometry study. *Frontiers in Neuroscience*, 12, 152.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, 66(1), 123-142.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3), 437-457.

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.

Braine, M. D., & Rumain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, 31(1), 46-70.

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, 126(3), 423-440.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3), 434-463.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: Applications and implications*. Amsterdam: Benjamins.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the 25th Boston University Conference on Language Development*, pp. 157-168. Somerville, MA: Cascadilla Press.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B.

(2014). Your morals depend on language. *PloS One*, 9(4), e94842.

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

Cummins, J. (1978). Metalinguistic development of children in bilingual education programs: Data from Irish and Canadian Ukrainian-English programs. M. Paradis (Ed.) *Aspects of Bilingualism*. Columbia, S. C.: Hornbeam Press.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54(2), 128-133.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In *Proceedings of the Cognitive Science Society*, 33(33).

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, 64(12), 2352-2367.

Dupuy, L., Stateva, P., Andreetta, S., Cheylus, A., Déprez, V., Henst, J. B. V. D., Jayez J., Stepanov A., & Reboul, A. (2018). Pragmatic abilities in bilinguals: The case of scalar implicatures. *Linguistic Approaches to Bilingualism*. [⟨](#)

[10.1075/lab.17017.dup](https://doi.org/10.1075/lab.17017.dup) . [⟨hal-01803048⟩](#)

- Eberhard, K. M., Shin, W.-J., Gundersen, S., Che, C., & Boldt, B. (2013). Mental Simulation of Action Sentences in One's Native vs. Second Language. Paper presented at the 6th Annual Embodied and Situated Language Processing. University of Potsdam, Potsdam, Germany.
- Farrell, M. P. (2011). Bilingual competence and students' achievement in physics and mathematics. *International Journal of Bilingual Education and Bilingualism*, 14(3), 335-345.
- Foppolo, F. (2007). *The logic of pragmatics. An experimental investigation with children and adults.* LAP Lambert Academic Publishing
- Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language Learning and Development*, 8(4), 365-394.
- Fortune, T. W. (2012). What the research says about immersion. *Chinese language learning in the early grades: A handbook of resources and best practices for Mandarin immersion*, 9-13.
- Gazdar, G. (1979). Pragmatics: Implicature, Presupposition, and Logical Form, 37-62.
- Green, A. (1986). A time sharing cross-sectional study of monolinguals and bilinguals at different levels of second language acquisition. *Brain and Cognition*, 5(4), 477-497.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and*

Semantics, Vol. 3. New York, NY: Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1), 42-55.

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.

Gutiérrez-Rexach, J. (2001). The semantics of Spanish plural existential determiners and the dynamics of judgment types. *Probus*, 13, 113-154.

Horn, L. (1972). *On the semantic properties of the logical operators in English*. Ph.D. dissertation, UCLA, Los Angeles, CA.

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2018). The trouble with quantifiers: exploring children's deficits in scalar implicature. *Child Development*, 89(6), e572-e593.

Huang, Y. T., & Snedeker, J. (2009a). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58(3), 376-415.

- Huang, Y. T., & Snedeker, J. (2009b). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, 45, 1723-1729.
- Huang, Y. T., & Snedeker, J. (2018). Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, 102, 105-126.
- Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition*, 94(3), B79-B89.
- Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67-81.
- Katsos, N., Cummins, C., Ezeizabarrena, M. J., Gavarró, A., Kraljević, J. K., Hrzica, G., ... & Van Hout, A. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, 113(33), 9244-9249.
- Kushalnagar, P., Hannay, H. J., & Hernandez, A. E. (2010). Bilingualism and attention: A study of balanced and unbalanced bilingual deaf users of American Sign Language and English. *Journal of Deaf Studies and Deaf Education*, 15(3), 263-273.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests

in linear mixed effects models. *Journal of Statistical Software*, 82(13).

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

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology*, 4.

Mazzaggio, G., Surian, L., 2018. A diminished propensity to compute scalar implicatures is linked to autistic traits. *Acta Linguist. Acad.* 65 (4), 651e668.

Mezzacappa, E. (2004). Alerting, orienting, and executive attention: Developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. *Child Development*, 75(5), 1373-1386.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165-188.

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

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2), 203-210.

Panizza, D., Chierchia, G., & Clifton, C. (2009). On the role of entailment pattern and scalar implicatures in the processing of numerals. *Journal of Memory and Language*, 61, 503-518.

- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3), 253-282.
- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12(1), 71-82.
- Peal, E., & Lambert, W. E. (1962). The relation of bilingualism to intelligence. *Psychological Monographs: General and Applied*, 76(27), 1.
- Pelham, S. D., & Abrams, L. (2014). Cognitive advantages and disadvantages in early and late bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2), 313.
- Politzer-Ahles, S., & Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography. *Language, Cognition and Neuroscience*, 30(7), 853-866.
- 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.
- Reinhart, T. (1999). The processing cost of reference-set computation: Guess patterns in acquisition. *OTS Working Papers in Linguistics*.
- Ricciardelli, L. A. (1992). Bilingualism and cognitive development in relation to threshold theory. *Journal of Psycholinguistic Research*, 21(4), 301-316.

- Sampath, K. K. (2005). Effect of bilingualism on intelligence. In Proceedings of the 4th International Symposium on Bilingualism, pp. 2048-2056.
- Siegal, M., Iozzi, L., & Surian, L. (2009). Bilingualism and conversational understanding in young children. *Cognition*, 110(1), 115-122.
- Siegal, M., Matsuo, A., Pond, C., & Otsu, Y. (2007). Bilingualism and cognitive development: Evidence from scalar implicatures. In Proceedings of the Eighth Tokyo Conference on Psycholinguistics (pp. 265-280). Tokyo: Hituzi Syobo.
- Siegal, M., Surian, L., Matsuo, A., Geraci, A., Iozzi, L., Okumura, Y., & Itakura, S. (2010). Bilingualism accentuates children's conversational understanding. *PloS One*, 5(2).
- Slabakova, R. (2010). Scalar implicatures in second language acquisition. *Lingua*, 120(10), 2444-2462.
- Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, 30(2), 191-205.
- Snape, N., & Hosoi, H. (2018). Acquisition of scalar implicatures. Evidence from adult Japanese L2 learners of English. *Linguistic Approaches to Bilingualism*, 8(2), 163-192.
- Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, 1(1), 1-33.

- Sorace, A. (2016). Referring expressions and executive functions in bilingualism. *Linguistic Approaches to Bilingualism*, 6(5), 669-684.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition* (2nd ed.) Cambridge, MA: Harvard University Press.
- Surian L., & Job R. (1987). Children's use of conversational rules in a referential communication task. *Journal of Psycholinguistic Research*, 16, 369-382.
- Syrett, K., Austin, J., Sanchez, L., Germak, C., Lingwall, A., Perez-Cortes, S., Arias-Amaya, A., & Baker, H. (2016). The influence of conversational context and the developing lexicon on the calculation of scalar implicatures. *Linguistic Approaches to Bilingualism*, 7(2), 230-264.
- Syrett, K., Lingwall, A., Perez-Cortes, S., Austin, J., Sánchez, L., Baker, H., Germak, C., & Arias-Amaya, A. (2017). Differences between Spanish monolingual and Spanish-English bilingual children in their calculation of entailment-based scalar implicatures. *Glossa: A Journal of General Linguistics*, 2(1).
- Tomlinson Jr, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, 69(1), 18-35.
- Wilson, D., Sperber, D., 2002. *Relevance theory*. In: Horn, L.R., Ward, G. (Eds.), *Handbook of Pragmatics*. Blackwell, Oxford.

Greta Mazzaggio is a postdoc researcher at the University of Neuchâtel (Switzerland) with a Swiss Government Excellence Scholarships. After an MA in Linguistics at the University of Verona (Italy), she obtained a PhD in Cognitive Sciences at University of Trento (Italy) with a thesis on the processing of conversational implicatures. She worked for 2 years as postdoc researcher at the Humanities Department of the University of Florence (Italy). Her main areas of interest are in the field of experimental pragmatics, with research on different topics (processing of implicatures, irony comprehension, pronouns mastering) and populations (typical children and adults, atypical populations). She is also interested in cognitive aspects related to second-language acquisition.

Daniele Panizza is a postdoc researcher at the English Linguistics Department of the University of Göttingen, Germany. He received a PhD in Cognitive Sciences at University of Trento (Italy) with a thesis on the interpretation and processing of numerals and conversational implicatures. His main areas of interests are in the fields of theoretical linguistics, experimental pragmatics, sentence processing and language acquisition. He has investigated the comprehension and processing of scalar quantifiers and semantic/pragmatic inferences in children and adults by the means of psycholinguistic (eye-tracking with reading and visual world paradigms) and neurolinguistic (event related potentials) methodologies.

Luca Surian is a Full Professor of Developmental Psychology at the University of Trento, Italy. He also held appointments at the Medical Research Council - Cognitive Development Unit in London and at the Departments of Psychology of the Universities of Padua, Trieste and Greensboro (North Carolina), where he was supported by a Fulbright Scholarship. For many years, Prof. Surian has been carrying out research on the development of language and cognitive processes, with a particular interest in the conceptual development and the acquisition of communicative competence and social cognition. Prof. Surian's research was reported in more than 100 scientific publications.