UNIVERSITY OF TRENTO
DEPARTMENT OF PSYCHOLOGY AND COGNITIVE SCIENCES

DOCTORAL SCHOOL IN
PSYCHOLOGICAL SCIENCES AND EDUCATION

Doctoral Dissertation
in the subject of Cognitive Sciences

Reading Between the Lines:

Conversational Implicature Processing

in Typical and Atypical Populations

Advisors
Chiar.mo Prof. Luca Surian
Chiar.mo Prof. Remo Job

PhD candidate
Dott.ssa Greta Mazzaggio

CICLO XXXI

August 2018

This thesis is dedicated to my beloved grandmother, Teresa.

Table of Contents

Page

INTRODUCTION……....1

1. Conversational Implicatures: a Definition……6

2. The Computation of Conversational Implicatures: a Theoretical Framework…….10

2.1. The Neo-Gricean Approaches and the 'Strong defaultism'…….……10

2.2. The Relevance-Theoretic Approaches…….12

2.3. The Grammar-driven Approaches and the 'Weak defaultism'……..…..…14

3. The Experimental Turn and This Thesis Proposal……17

3.1. The The Experimental Turn: Children's Processing……19

3.1.1.The Lexicalist Account…….…..20

3.1.2 The Processing Account……..…23

3.1.3. The Pragmatic Account……25

3.2. The The Experimental Turn: Adult's Processing…….……27

3.2.1. The Default Model……...……..27

3.2.2. The Literal-First Model…….…….……..28

3.2.3. The Constraint-Based Model……..……..29

3.3. This Thesis' Outline...……..……31

PART 1 – TYPICAL DEVELOPMENT:

THE COMPREHENSION OF CONVERSATIONAL IMPLICATURES……...37

Chapter 1. Scalar- and Ad-hoc-Implicature Processing in Typically Developing Children..39

PART 2 – ATYPICAL DEVELOPMENT:

THE COMPREHENSION OF CONVERSATIONAL IMPLICATURES ……..79

Chapter 2. Scalar- and Ad-hoc-Implicature Processing in Children with Autism Spectrum Disorders…….……..81

PART 3 – ADULT COMPETENCE WITH SCALAR IMPLICATURES…….119

Chapter 3. Scalar Implicature Computation and the Role of the Autistic Quotient……..…121

Chapter 4. Scalar-Implicature Computation in Second-Language Oral Processing…..……145

Chapter 5. The cost of scalar implicature: inference or infelicity? A Reaction-Time and Eye-Tracking Study…….171

GENERAL DISCUSSION AND PERSPECTIVES……....221

1. Main Findings…….…….223

1.1 The Scalar/Ad-hoc Implicatures Debate …….….223

1.2 The Debate around the Acquisition of Scalar Implicatures in Typically Developing Children …….…….…..224

1.3 The Default-Non-Default Debate of Scalar-Implicatures Computation…...226

1.4 The Role of Theory of Mind in the Computation…….230

2. Conclusion……233

REFERENCES…….…237

ACKNOWLEDGMENTS……..……..253

RINGRAZIAMENTI……..…….256

INTRODUCTION

> *'When I use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means just what I use it to mean—neither more nor less.'*
>
> Lewis Carroll, *Through the Looking Glass.*

Is it true that when we utter a word, our interlocutor will interpret the same meaning that we were willing to attribute to that word? You are probably thinking that, no, this is not always the case. Sometimes, indeed, you may have heard the sayings 'what you don't say speaks volumes' or 'you should read between the lines'. Furthermore, rhetoricians recognized how *minus dicimus et plus significamus* (Hoffmann, 1987, p. 21). The idea that the unsaid matters as much as spoken words may sound like a commonplace but it has been originally pointed out by the British philosopher John Stuart Mill, who wrote:

> "If I say to any one, "I saw some of your children to-day", he might be justified in inferring that I did not see them all, not because the words mean it, but because, if I had seen them all, it is most likely that I should have said so: though even this cannot be presumed unless it is presupposed that I must have known whether the children I saw were all or not" (1867, p. 501).

A century later, in 1967, Herbert Paul Grice − in his famous *William James Lectures* delivered at Harvard University – developed and formalized those ideas into what is known today as one of the most influential theories regarding the logic behind a conversation, elegantly set out in his seminal works *Logic and Conversation* (1975), *Further Notes on*

Logic and Conversation (1978) and *Studies in the Way of Words* (1989).[1] Grice (1975), for the first time, described the phenomenon of *implicature*.

An "IMPLICATURE is a component of speaker meaning that constitutes an aspect of what is *meant* in a speaker's utterance without being part of what is *said*" (Horn & Ward, 2006, p. 3). We must then distinguish the semantic meaning of a word or of a sentence from the speaker's meaning. Such interpretation of the meaning of a sentence based on extra linguistic aspects is part of what we can define Pragmatics. A clear example of the divergence between Semantics and Pragmatics can be found in Orwell's *Down and Out in Paris and London:*

> "Words used as insults seem to be governed by the same paradox as swear words. A word becomes an insult, one would suppose, because it means something bad; but in practice its insult-value has little to do with its actual meaning. For example, the most bitter insult one can offer to a Londoner is 'bastard'—which, taken for what it means, is hardly an insult at all. And the worst insult to a woman, either in London or Paris, is 'cow'; a name which might even be a compliment, for cows are among the most likeable of animals. Evidently a word is an insult simply because it is meant as an insult, without reference to its dictionary meaning; words, especially swear words, being what public opinion chooses to make them. In this connexion it is interesting to see how a swear word can change character by crossing a frontier" (1933, p. 179).

Some swear words can be analyzed in pragmatic terms. Taking Orwell's example, if I say that a woman is a 'cow', I'm literally saying something false, or – in Gricean terms – I'm not adhering to a *Cooperative Principle:*

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (Grice, 1975, p. 45).

[1] It is worth mentioning that some theorists (see Reboul, 2005) consider John Austin as the father of the field of Pragmatics, particularly with his William James Lectures given at Harvard (1955) and the related publication *How to do things with words* (Austin, 1962).

In particular, in affirming that a person is an animal, I'm not respecting (cfr. flouting) what Grice defined the first *Maxim of Quality*: "Do not say what you believe to be false". Thus, the listener who moves from the basic assumption that I'm cooperative will enrich the literal meaning of 'cow' with the metaphorical meaning of 'an unpleasant woman'. Grice (1975, pp. 45-46) did not described just this maxim, but a series of maxims and sub maxims that we must obey in order to be cooperative:

> *"Quantity*
> 1. Make your contribution as informative as is required (for the current purposes of the exchange).
> 2. Do not make your contribution more informative than is required.
> *Quality* (Supermaxim - Try to make your contribution one that is true):
> 1. Do not say what you believe to be false.
> 2. Do not say that for which you lack adequate evidence.
> *Relation*: Be relevant.
> *Manner*: (Supermaxim - Be perspicuous):
> 1. Avoid obscurity of expression.
> 2. Avoid ambiguity.
> 3. Be brief (avoid unnecessary prolixity).
> 4. Be orderly."

The specific type of implicature that is mainly addressed in this thesis is the results of the violation of the first Maxim of Quantity and it is the so-called "Scalar Implicature". First and foremost, we shall provide a brief outline of Conversational Implicatures.

1. Conversational Implicatures: a Definition

Scalar implicatures are part of what Grice defined CONVERSATIONAL IMPLICATURES (specifically, GENERALIZED CONVERSATIONAL IMPLICATURES), as opposed to another type of implicatures, the CONVENTIONAL IMPLICATURES. These latter implicatures are defined 'conventional' because they are not computable through the adherence to the conversational maxims but they always arise when a specific term is used. An example of a Conventional Implicature is given in (1).

(1) a. She is a linguist and she is *even* funny.

b. IMPL: Linguists are not generally funny.

When presented with a sentence like in (1a) we always implicate the meaning in (1b), because it is the adverb *even* – with its conventional meaning – that automatically triggers the implicature. And this is the case for other terms such as *therefore* or *but* (for a better description of the phenomenon see Potts, 2005).

Scalar implicatures are generated thanks to the use of specific lexical items as well, but their computation is not automatic. According to Horn (1972), scalar implicatures are generated because those specific lexical items are considered to be part of a scale in which they are ordered from the less informative (L_i) to the more informative (M_i). Thus, given the lexical scale $<L_i, M_i>$, when the speaker decides to use the L_i item, s/he automatically implicates the negation of the M_i item, since $M_i \subseteq L_i$ (M_i logically entails L_i). In other words, according to Horn, the use of one element of the scale implicates the negation of the terms at its rights in the same scale. For example, considering the quantifiers *some* (2a) and *all* (2b), they can be seen as part of a scale <some, all>. Thus, there are two possible interpretations of (3a): a literal (lower-bound) interpretation, like in (3b), and a pragmatic (upper-bound)

interpretation, like in (3c). According to Grice, in listening (3a) we take into consideration the existence of the more informative statement in (3d), but since the speaker should be cooperative and did not utter (3d), we implicated the negation of the stronger alternative and got the meaning in (3c).

(2) a. $[\![some]\!] = \lambda P.\lambda Q.\, P \cap Q \neq \emptyset$

b. $[\![all]\!] = \lambda P.\lambda Q.\, P \subseteq Q$

(3) a. *Some* children got a biscuit.

b. *Some and possibily all* children got a biscuit.

c. *Some but not all* children got a biscuit.

d. *All* children got a biscuit.

This kind of Gricean reasoning can be summarized as follows:

i. S said "Some Xs are Ys"

ii. I know that exist a more informative statement that is 'All Xs are Ys' ('All Xs are Ys' ⊆ "Some Xs are Ys")

iii. S should obey at the Conversational Maxims or at least at the Cooperative Principle

iv. If S would have known that 'All Xs are Ys' he would have said so, to obey the Cooperative Principle and the first Maxim of Quantity

v. Since S said "Some Xs are Ys" it means that either S knows that it is not the case that 'All Xs are Ys' or that he does not know whether 'All Xs are Ys'

Apart from quantifiers scale, there are other lexical terms that can be seen as part of a scale and that can give rise to scalar implicatures, such as connectives <or, and>, numerals <one,

two, three, …>, verbs <to begin, to finish>, adverbs <sometimes, often, always>, modals <may, should, must>, adjectives <warm, hot> and so on (Horn, 1972; Levinson, 1983, p. 134).

Grice (1975, pp. 57-58) also stated specific properties for conversational implicatures. According to the cancellability property, an implicature can be cancelled explicitly asserting that it is not the case for such implicature (e.g., *Some children got a biscuit, indeed all of them did*) or if the context makes clear that the speaker is not adhering to the Cooperative Principle. Moreover, a proposition – even if paraphrased – containing the implicature will always tend to convey such implicature (non-detachability property) and the hearer must follow an inferential path to calculate it (calculability property).

In this thesis, we will study also a different type of CONVERSATIONAL IMPLICATURES, namely PARTICULARIZED CONVERSATIONAL IMPLICATURES or Ad-hoc implicatures. Contrary to scalar implicatures, ad-hoc implicatures arise from contextual features, rather than lexical ones. In other words, ad-hoc implicatures need a specific context in order to be computed. Let us consider the example in (4). If I say (4b) in a general conversation (i.e., outside the specific contextual features in (4)), I'm simply asserting that you have a sister, that your sister went out and that she was elegant. On the other hand, if I say (4b) in a context in which it is an answer for (4a), you will probably infer that your sister used the missing dress.

(4) a. I cannot find my new dress.

b. Your sister was pretty elegant when she left home.

The kind of reasoning for ad-hoc implicatures computation might be seen as similar to the one for SIs. Since (4b) is a blatant violation of the maxim of Relevance, the listener will find a possible meaning that might explain why the speaker said something irrelevant. As Levinson

(1983, p. 127) outlined, “all implicatures that arise from observing the maxim of Relevance are particularized, since utterances are relevant only with respect to the particular topic or issue at hand”. Also, there are ad-hoc implicatures that might arise from observing – like for SIs – the first maxim of Quantity and these are the ones that we are dealing with our works in this thesis (Chapter 1 and 2).

(5) a. My child is the one with glasses.

b. My child is the one with glasses and a hat.

c. My child is the one with (only) glasses.

In a playground where there is a child with only a hat, a child with only glasses, a child with a hat *and* glasses and a child with no hat nor glasses, upon uttering (5a) the hearer will probably derive the implicature in (5c) considering that an alternative sentence, like the one in (5b), might have been uttered. Indeed, we might think of a scale <no hat/glasses, only hat/glasses, both hat/glasses> that is created contextually, instead of lexically (Stiller, Goodman, & Frank, 2015). Similar to scalar implicatures, ad-hoc implicatures are cancellable: indeed, considering the example in (5a) ‘My child is the one with glasses’, I can continue saying ‘My child is the one with glasses, indeed he has glasses and a hat’. To sum up:

> “ a. An implicature *i* from utterance U is *particularized* iff U implicates *i* only by virtue of specific contextual assumptions that would not invariably or even normally obtain
>
> b. An implicature *i* is *generalized* iff U implicates *i* unless there are unusual specific contextual assumptions that defeat it” (Levinson, 2000, p. 16).

2. The Computation of Conversational Implicatures: a Theoretical Framework

Since Grice (1975), the debate on the computation of conversational implicatures, and particularly of scalar implicatures, has been broadly fostered with an important contribution of the new field of Experimental Pragmatics (Noveck, 2001; Noveck & Sperber, 2004). The two main areas of interest are related to how implicatures are generated (i.e., linguistically or pragmatically) and if such processing is demanding in terms of cognitive resources. As we will partially find out trough this thesis, experimental investigation considered different populations with various techniques in order to shed light on how specific implicatures are computed.

When considering how scalar implicatures are generated, we may consider three principal theoretical accounts: the Neo-Gricean approaches, Relevance-Theoretic approaches and Grammar-driven approaches (Foppolo, 2012). In this section we shall hint – with no claim of completeness – at the major points of the different approaches. Particularly, we will focus on aspects that we will run across during this thesis, that are:

- distinction between generalized and particularized implicatures
- global vs. local interpretation of scalar alternatives
- default vs. non-default computation of implicatures
- diverse mechanisms of computation (linguistic vs. extra-linguistic)

2.1. The Neo-Gricean Approaches and the 'Strong defaultism'

According to Neo-Gricean approaches (particularly, Levinson, 2000), upper-bound interpretations are the default interpretation when a sentence containing an underinformative term is encountered. Contextual factors might sometime lead to a lower-bound interpretation. The prediction is that a generalized conversational implicature is a "default inference, one that captures our intuitions about a preferred or normal interpretation" (Levinson, 2000, p. 11) that

is computed costless, automatically, and its cancellation can be obtained with a cost, as a second step. Levinson, argued that:

> "According to the standard line (more often presupposed than justified), there are just two levels to a theory of communication: a level of sentence-meaning (to be explicated by the theory of grammar in the large sense) and a level of speaker-meaning (to be explicated by a theory of pragmatics, perhaps centrally employing Grice's notion of meaning$_{NN}$). […] This view, although parsimonious, is surely inadequate, indeed potentially pernicious, because it underestimates the regularity, recurrence, and systematicity of many kinds of pragmatic inferences. What it omits is a third layer, what we may call the level of *statement*- or *utterance-meaning* […] or, as I will prefer below, *utterance-type-meaning.* This third layer is a level of systematic pragmatic inference based *not* on direct computations about speaker-intentions but rather on general expectations about how language is normally used. These expectations give rise to presumptions, default inferences, about both content and force; and it is at this level (if at all) that we can sensibly talk about speech acts, presuppositions, conventional implicatures, felicity conditions, conversational presequences, preference organization, and so on and of special concern to us, generalized conversational implicatures" (2000, pp. 22-23).

In his work, Levinson also reformulated the Gricean maxims in what he defined heuristics; if, for example, for Grice scalar implicatures derive from the adherence to the first Maxim of Quantity, for Levinson scalar implicatures are generated trough the following of the Q-heuristic, that is formulated as "What isn't said, isn't" (2000, p. 31). Levinson proposed also the I Heuristic and the M Heuristic: the former asserts that "What is simply described is stereotypically exemplified" (2000, p. 32), a statement in line with the Gricean second Maxim of Quantity ("Do not make your contribution more informative than is required"); the latter asserts "What's said in an abnormal way, isn't normal; or Marked message indicates marked situation" (2000, p. 33) and it is in line with Gricean Maxims of Manner ("avoid obscurity of expression, avoid ambiguity, be brief and be orderly"). If Levinson's proposal (i.e. inferences

are computed costless and automatically) seems elegant and intuitive, experimental research, as we will see in this thesis, goes to the opposite direction.

2.2. The Relevance-Theoretic Approaches

"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, p. 607). Accordingly to the Relevance-Theoretic approaches, firstly set out by Dan Sperber and Deirdre Wilson (1986; 1995), our communication is not governed by the adherence to a Cooperative Principle or to specific maxims but it is oriented toward the *Cognitive Principle of Relevance*: "Human cognition tends to be geared to the maximization of relevance" (2002, p. 609). And what is relevant? According to Wilson and Sperber (2002, p. 610), an input is relevant to us when it gives positive cognitive effects and is presented and processed in a specific context, that is when it brings something useful for our world representation. Thus, according to the authors, positive cognitive effects and processing efforts are related to relevance in a way that we can affirm that something is relevant to an individual under the following conditions:

> " a. Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.
>
> b. Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time" (2002, p. 609).

Not only did Wilson and Sperber outline the Cognitive Principle of Relevance, but they also enunciated the *Communicative Principle of Relevance*, according to which "every ostensive stimulus conveys a presumption of its own optimal relevance" (2002, p. 612). In an ostensive-inferential communication, in which the goal is to inform the interlocutor regarding our

willingness to communicate something, ostensive stimuli are oriented to capture the attention of the hearer on our desire to convey a specific message. Such stimuli should be relevant enough to capture the listener's attention and to convey the idea that the effort to process the input will be worth it (i.e., there is a presumption of optimal relevance). How can this reasoning be applied to implicatures? First of all, Wilson and Sperber distinguished between explicatures and implicatures, as follows:

> " (I) An assumption communicated by an utterance U is explicit [hence an "explicature"] if and only if it is a development of a logical form encoded by U.
> [Note: in cases of ambiguity, a surface form encodes more than one logical form, hence the use of the indefinite here, "a logical form encoded by U."]
>
> (II) An assumption communicated by U which is not explicit is implicit [hence an "implicature"]" (2002, p. 635).

Accordingly, there are no distinctions between particularized and generalized conversational implicatures; indeed we could say that:

> "all the implicatures that are called "generalized" […] are assimilated in Relevance Theory to explicatures and derived from the linguistically encoded logical form by pragmatical enrichment […] i.e. as content considered being explicitly communicated and not simply inferred" (Foppolo, 2012, p. 24).

Apart from these descriptive aspects, these theories make specific predictions regarding the defaultness of interpreting underinformative items in a scale. According to Relevance-Theoretic approaches, scalar inferences are pragmatic in nature and are governed by speaker's expectations on the hearer willingness to draw an interpretation based on relevance. How far the hearer may go is all about the balance between 'gain' (positive cognitive effect) and 'pain' (cognitive effort): the hearer will try to obtain the most suitable interpretation with a

minimum effort. A logical interpretation of a scalar term "could very well lead to a satisfying interpretation of this term in an utterance" (Bott & Noveck, 2004, p. 439). The further enriching step might be done but with a cost. 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, p. 613).

It is worth noting that, according to Chemla and Singh (2014, p. 376), Relevance-Theoretic approaches do "not offer a theory of scalar implicatures"; rather, they provide "a general framework and terminology to draw and describe expectations about how theories of scalar implicatures and almost just any other phenomena can lead to processing predictions" and that they are "not at all a competitor of the Gricean and grammatical theories". We won't analyze the assumptions of Relevance-Theoretic approaches in depth because what has been said is enough for the purpose of this thesis.

2.3. The Grammar-driven Approaches and the 'Weak defaultism'

From what we have discussed so far, we can make some specific predictions. First, scalar implicatures should be computed when a sentence containing the less informative term in a scale is uttered in a context in which the more informative alternative in the scale is relevant; indeed the hearer compares what have been said with what could have been said but had not. Second, scalar implicatures are computed globally and post-compositionally, i.e. after the computation of the semantic and of the truth-conditional content of the sentence (Chemla & Singh, 2014a, p. 374; Foppolo, 2012, p. 28).

We will now consider different approaches according to which scalar implicatures are not pragmatically driven but linguistically driven (i.e., grammatically) and, thus, according to which scalar implicatures are computed locally (e.g., Chierchia, 2006; 2013; Chierchia et al. 2008; Fox, 2007). Chierchia (2013) proposed the *Extended Standard Gricean Theory* according to which scalar implicatures are computed by means of a covert counterpart of *only*, called exhaustivity operator (O). This 'exhaustification process' will occur just when it is advantageous (i.e., relevant) in a particular context, that is only when "lexically […] activated alternatives are relevant to the conversational goals" (Chierchia, 2013, p. 123). Indeed, lexical alternatives are always accessible but the exhaustification process is optional.

Let us see how this exhaustification process works. For example, considering the scale <or, and>, a sentence like (6a) can be logically interpreted as (6b), formally represented in (7b) where B_s stands for "the speaker believes that", or pragmatically as (6c), formally represented in (7a). If we apply the silent operator (6d) the result of the operation is the strengthened meaning, formally represented in (7c) (Chierchia, Fox, & Spector, 2008, p. 4).

(6) a. Mary *or* John will come.

b. Mary will come *and* John will come.

c. Only [Mary *or* John will come].

c. *O* [Mary *or* John will come].

(7) a. B_s (come(m) ∨ come(j))

b. B_s (come(m) ∧ come(j))

c. B_s (O_{ALT} (come(m) ∨ come(j))) = B_s (come(m)∨ come(j) ∧ ¬ (come(m)∧ come(j)))

In (7c) we bumped into O_{ALT}, that is when the silent operator is applied to scalar alternatives. Indeed, according to Chierchia's proposal, for an implicature to be derived, "it is not sufficient to activate the representation of the basic meaning; it is also necessary to have and retrieve the scale component" (Foppolo, 2012, p. 195). In other words, any expression ||α|| has plain semantic value and a strong value $||\alpha||^{S}$ assigned by the grammar and that is the negation of the stronger alternative in the scale of α:

> "Such a scalar value is computed by exploiting the stronger alternatives to the plain value, which constitute $||\alpha||^{S\text{-}ALT}$, i.e. the set of the stronger alternatives that are lexically determined given the scalar lexical entry" (Foppolo, 2012, p. 30).

Considering the quantificational scale, we might now consider that the quantifier *some* has:

- a plain value: $||some|| = \lambda P \lambda Q\ \exists x[P(x) \wedge Q(x)]$
- a set of alternatives: $||some||^{S\text{-}ALT} = \{some < many < most < all\}$
- a scalar value: $||some||^{S} = \lambda P \lambda Q\ \exists x[P(x) \wedge Q(x)] \wedge \neg \forall x\ [P(x) \rightarrow Q(x)]$
  (Foppolo, 2012, p. 30)

Building on the idea that the exhaustification process is syntactic in nature, the natural consequence is to think that it might apply to embedded clauses. Indeed, Chierchia also proposed that a globalist view of implicature computation cannot account for examples like in (8). Considering a sentence like (8a), the alternative sentence should be (8b) and the computed implicature (8c). However, (8c) is too weak to express the meaning conveyed by (8a), indeed the meaning of (8a) seems more compatible with (8d) (Foppolo, 2012, p. 28).

(8) a. Daniel believes that some students like semantics.

b. Daniel believes that all students like semantics.

c. Daniel *does not* believe that all students like semantics.

d. Daniel believes that *not all* students like semantics.

To sum up, according to Grammar-driven approaches, scalar implicatures are computed through linguistic mechanisms that work on lexical alternatives. Implicatures are computed locally. Particularized and generalized implicatures are assumed to be computed differently. Regarding the processing cost associated with the derivation of the implicature, it is less clear whether this approach predicts it or not. As Bart Geurts (2010, p. 94) suggested, if on the one side we have Levinson and his 'strong defaultism' claims, on the other side grammatical approaches can be seen as proponent of a 'weak defaultism' in which they make "no claims about processing", indeed they merely assert "that upper-bounding inferences are the norm". We will not examine further those proposals, since it is beyond the scope of this thesis.

3. The Experimental Turn and This Thesis' Proposal

In the previous paragraphs, we have seen that different theoretical accounts on the computation of implicatures make different predictions. We have "Gricean accounts" according to which we must consider the speaker's intentions when we derive scalar implicatures, considering both linguistic and extra-linguistic information. Specifically, we have to consider the semantic meaning of a proposition, the speaker's willingness to respect the purposes of a cooperative exchange (i.e., for scalar implicatures, we must consider some respect for the Maxim of Quantity) and the interlocutor's epistemic state, acknowledging a

certain acquaintance with the involved topic of the exchange. This seems to predict a role of Theory of Mind (ToM), that is the cognitive capacity to attribute mental states to us and to others and to recognize that those mental states can differ one from the others (Premack & Woodruff, 1978); ToM plays an important role when we have to interpret other people's behaviour and act accordingly. Furthermore, Relevance-Theoretic accounts (e.g., Sperber, & Wilson, 1986) and Grammatical accounts (e.g., Chierchia, 2006; 2013) may assume a certain role of both the interlocutor's epistemic state and of the context but they predict different computational mechanisms (i.e., either based on relevance and a series of inferential processes or grammatically driven by a silent operator *O*). One of the strikingly different accounts, mainly theorized by Levinson (2000), proposed that implicatures are computed as a default regardless of context, as soon as they are encountered, and subsequently cancelled if required by the context. In Levinson's view (2000, p. 4) we usually interpret them "without too much calculation of such matters as speakers' intentions, encyclopedic knowledge of the domain being talked about, or calculations of others' mental processes."

Insomuch as the predictions made by the different accounts might be seen as easily testable (e.g., cost vs non-cost, a role for ToM vs. no role for ToM, ecc.), groundbreaking works of Noveck (2001) for children's processing and Bott and Noveck (2004) for adults' processing have inspired a series of experiments which tried to better our understanding of the process of implicature computation.

3.1 The Experimental Turn: Children's Processing

Noveck's (2001) aforementioned essay brought new research interest on the acquisition of scalar implicatures in children. However, his work moved from three previous studies that showed how children tend to interpret logically weak scalable lexical items: Smith (1980), Braine and Rumain (1981) and Paris (1983).

Smith (1980) assessed 60 children (4- to 7-year-olds) on their knowledge and interpretation of quantifiers. Although children showed a good general comprehension of the quantifiers, they tended to accept more than adults sentences such as "Do some birds have wings?". Braine and Rumain (1981) in four tasks tested children (5- to 10- year-olds) as well as adults with the comprehension of the disjunction 'or'. Again, children accepted more than adults an inclusive interpretation of 'or' (with both the two disjunctions true) instead of an exclusive pragmatic interpretation (with just one element true). Similar results had been obtained in Paris (1973).

Through three experiments, Noveck's work (2001) had the goal to developmentally assess whether children initially interpret logically – instead of pragmatically – a weak term in a scale. In the first experiments he tested 68 children (5- to 9- year-olds) and 20 adults – English-speakers – with modals. Participants had to judge an underinformative sentence like 'There might be a parrot in the box' in a context in which it *must* be a parrot in the box. Results showed that children give logical answers significantly more than adults and they begin to reason pragmatically with age. In the second experiment, 35 children (5- to 7- year-olds) and 16 adults – speakers of English – had been tested with the same task of the first experiment but with after pre-training. Results showed an effect of training (i.e., more logical answers) just for adults. In the third experiment, 61 children (8- to 10- year-olds) and 15 adults – speakers of French – had been assessed with the existential quantifier *certains*. Participants had to evaluate underinformative sentences such as 'Some giraffes have long necks'. Children did not reject such sentences like adults.

After Noveck (2001), many other studies found that children tend to answer logically more often than adults and many hypotheses have been formulated to explain this pattern of results. We shall now consider some of the most influential proposals.

3.1.1 The Lexicalist Account

According to the *lexicalist* account, the fact that children tend to interpret scalar implicatures logically is due to problems in retrieving the scalar alternatives or to a non-mature lexicalization of the scale. Considering the quantificational scale, it might be that they know the meaning of *some* and *all*: indeed, when tested they never show difficulties in the control sentences, but they might not know that *some* and *all* are part of a scale (<some, all>), or they are probably building the links amongst quantifiers belonging to the scale. According to Barner, Brook and Bale:

> "A failure to represent lexical items as members of psychological scales could explain numerous results in the literature, and also the apparent discrepancy between children's difficulty with implicatures and their relatively sophisticated use of pragmatic cues elsewhere in language acquisition" (2011, pp. 86-87).

They tested 60 4-year-old native speakers of English with both sentences with *some* (and *only some*) in which scale members are context-independently specified and with sentences in which scale members are context-dependently specified. For example, they presented a card in which there were a cow, a dog and a cat, all sleeping. Then they asked questions like "Are (only) some of the animals sleeping?" (context-independent) or ''Are (only) the cat and the cow sleeping?" (context-dependent). The *only* was never prosodically emphasized. First of all, the use of *only* had a significant effect just for sentences with context-dependent alternatives, with higher pragmatic interpretations when *only* was used. In the case of sentences with context-independent alternatives (e.g., some/all), children gave more logical interpretations in both sentences with *only* and without *only*. According to authors, "Since *only* forces strengthening grammatically (and clearly did so for contextual alternatives), no pragmatic inference was required. Thus, children's failure to derive strengthened

interpretations for some is consistent with a hypothesized failure to generate relevant scalar alternatives" (2011, pp. 90-91). Furthermore, they proposed that children, in order to compute implicatures, must both acquire the "core meanings" and the "syntactic properties" of scale-items and also "perform additional learning in order to rapidly and automatically access lexical items as scalar alternatives" (2011, p. 91).

Similar conclusions are reached in Foppolo, Guasti and Chierchia (2012) who suggest that a factor that might play a role in children's computation of scalar implicatures is "the maturation of the lexicon that, in the case of scalar items, involves two layers of lexical representation, the basic meaning and the scale: the link between these two needs to be acquired and automatized" (p. 391). They reached this conclusion through 6 experiments in which they manipulated different factors:

- In Experiment 1 they tested 63 Italian children of different age (4- to 7- year-olds) and 12 adults with a Truth Value Judgment Task in order to analyze the developmental effect of scalar-implicatures computation. They found out that 6 is the critical age: afterwards children behave like adults, whereas they split beforehand (i.e., half of them answer pragmatically and half of them logically).
- In Experiment 2 authors manipulated the Italian quantifier used. In Experiment 1 they used 'qualche' (*some*) while in Experiment 2 they used 'alcuni dei' (*some of*); the goal was to assess whether the use of the partitive might help children with the computation. Twenty-four children of 5-year-olds and 12 adults had been tested with a Truth Value Judgment Task similar to Experiment 1's one. Results showed that the use of the partitive did not help children.
- Experiment 3 had the goal of priming the scale presenting a correct *all*-statement before a *some*-statement. Results of 12 5-year-olds children showed no facilitation-effects.
- In Experiment 4 the same children tested in Experiment 2 were presented with a

Conversational Violation Test to check whether children that do not compute implicatures are generally non-mature in terms of pragmatic skills. This was not the case since children could detect violations of the conversational norms tested.

- In Experiment 5, 17 children that previously failed to compute implicatures in other Truth Value Judgment Task experiments had been assessed again with a Felicity Judgment Task. They had to select the most appropriate sentence between an *all*-sentence and a *some*-sentence when describing a picture in which the use of *all* is more informative. Children performed incredibly well (95% of correct answers) showing that they are able to distinguish the more informative element in the scale.
- In Experiment 6, 47 5-year-olds and 40 adults had been tested with a Truth Value Judgment Task similar to the ones used for the other experiments. The only difference was that both a correct and an incorrect use of *all* were presented to participants before the *some*-underinformative sentence. Whilst presenting just the correct use of *all* before the evaluation of the undeinformative sentence did not change the rate of pragmatic answers (Experiment 4), in this case the percentage of pragmatic answers rose from 40% to 72.5%.

Foppolo et al. explained the increase of pragmatic answers in the last experiment asserting that:

> "we can think of an under- informative statement as ambiguous between a basic and a strengthened meaning; we know that children are not automatically aware of ambiguities, but they are able to shift their perspectives when informed of the alternatives. We think that it is exactly this information that they got in our Experiment 6, which helped them to recognize the two layers of the meaning of *some* and retrieve the scale (provided that this is already available in the lexicon) so that they could choose the most appropriate, strengthened meaning and reject the underinformative sentence" (2012, p. 389).

In conclusion, according to a *lexicalist* account it seems that children's difficulties in deriving

the pragmatic interpretation might be scale-specific and that in order to compute pragmatic interpretations children (a) need a certain command of the lexical items, (b) they need to know that those lexical items are part of a scale and (c) they need scale-items retrieval abilities.

3.1.2 The Processing Account

The *processing* account considers children's difficulty with scalar implicatures as related with the processing cost of the computation. Pouscoulous, Noveck, Politzer and Bastide (2007) might be seen as the credited authors of the *processing* account. In their first experiment they tested 23 children (9- to 10-year-olds) and 19 adults with a Truth Value Judgment Task based on an artificial scenario, instead of encyclopedic knowledge. Participants saw four cardboard boxes and a variety of plastic animals that could be inside or outside the boxes. For example, subjects had to evaluate a sentence like "Some turtles are in the boxes" when actually *all* turtles were. Authors tested also for negative sentences with *some*. Results showed that children were more likely than adults to answer logically to underinformative sentences, as was expected. Higher percentages of logical answers have been found for negative sentences like "Some elephants are not in the boxes" in a case in which there were no elephants inside the boxes and two elephants outside the boxes.

Experiment 2 had the goal of manipulating the task in order to give more spare cognitive resources to children and thus make them more able to compute implicatures. First, they asked questions on items inside the box only, to make the task easier. Second, instead of asking to evaluate a sentence, participants had to actively act to answer at a puppet's request. For example, if the puppet said "I would like some boxes to contain a token" when there were five boxes with a token inside each box, participants might act pragmatically removing at least one token or they might act logically leaving the scenario as it was. Third, they used the French quantifier *quelques* (instead of *certains*) that the authors considered easier to process

for children. Again, negative sentences were presented, such as “I would like some boxes to not contain a token”. Sixty-six children from a nursery class, 30 kindergarten-children, 54 second-graders children and 21 adults participated at the experiment. Results showed an increase in pragmatic answers compared to Experiment 1, thus the task turned out to be easier, and rates of pragmatic answers significantly increased with age. Again, negative sentences appeared to be more difficult compared to positive sentences.

Experiment 3 was similar to Experiment 2 but participants had been divided in 2 groups, one group had been tested with the French word *quelques* and the other with the French word *certains*. Twenty-five children (9- to 10-year-olds) and 28 adults had been tested. If for adults the quantifier term did not have an effect on results, children derived more implicatures when *quelques* was used.

In conclusion, Pouscoulous et al. (2007) demonstrated that young children can produce implicatures (more than what was previoulsy asserted) but they are not at adults’ level; children reach an adult performance after the age of seven. Moreover, the fact that the use of an indefinite expression (i.e., *quelques* vs *certains*) led to more implicatures derived in children but not in adults represents a demonstration – in the authors’ view – of the role of lexical complexity. Indeed, they state that “children clearly understand the meaning of both expressions, but the added processing cost of *certains* makes the task harder, thus reducing children’s rate of implicature production” (2007, p. 371). The view that children’s problems with scalar implicatures may be related with the processing costs required to compute them can be reconciled with the Relevance-Theoretic approach proposed by Sperber and Wilson (1995).

3.1.3 The Pragmatic Account

Children difficulties with scalar implicatures have also been explained considering their pragmatic system. For example, Katsos and Bishop (2011) consider children as more tolerant of pragmatic violations than adults. In their paper, together with a classical binary judgment task in which they replicated previous results with 5- to 6-year-old children (i.e., children rejected less underinformative sentences compared to adults), they had a ternary judgment tasks. In this latter task participants (18 5- to 6-year-old children and 10 adults) could evaluate how much a sentence was correct assigning a 'small strawberry' (false), a 'big strawberry' (underinformative) or a 'huge strawberry' (optimal), instead of answering 'right' or 'wrong'. With such design children showed certain sensitivity to underinformative sentences, with a preference to assign the midsize strawberries. With a third experiment they also assessed 15 5-year-old children and 10 adults with a sentence-to-picture matching paradigm, finding a performance similar to that of adults. As previously mentioned, authors consider the findings as consistent with the idea that children are able to compute implicatures, but they are simply more tolerant than adults in binary judgment tasks. They did not exclude that the factors described by other accounts (e.g., processing costs, task complexities, informativeness) might be involved in children's behavior but the authors' view is that their role is linked with what is defined as 'pragmatic tolerance'. An alternative explanation that authors considered is that children have less meta-linguistic skills than adults and they tend to accept underinformative sentences in order to avoid explaining why the sentence is wrong. Indeed authors wrote: "Children may not be as competent as adults in expressing complex judgments such as a 'yes, but. . .' or 'half right, half wrong' as opposed to simple 'yes' or 'no'. In this case, young children may default to a simple 'yes'" (2011, p. 77).

Another recent proposal has been pointed out by Skordos and Papafragou (2016): authors focused on the role of conversational relevance and considered children's problem with the computation of scalar implicatures as related with their difficulties in considering alternative scalar items as relevant in particular contexts. In their first experiment they assessed 90 5-year-old children and 36 adults with a classical Acceptability Judgment Task. They manipulated it with the stronger scale member (i.e. *all*) presented before or after the weaker alternative (i.e., *some*) in order to assess whether the presence of the more informative item could enhance the relevance of the alternative, thus leading to more scalar implicatures derived. This is what they found. In Experiment 2, they tested 50 5-year-old children and 24 adults again with an Acceptability Judgment Task to check whether the simple presence of the *all*-alternative before the *some*-alternative helps the computation of the implicature or the alternative in the context must be relevant in order to be useful. Results showed that "even in contexts where the stronger scalar term (*all*) was explicitly mentioned, children did not benefit from its presence unless the scalar term was seen as a relevant stronger alternative. Thus, the accessibility of the stronger scalar term is a necessary but not sufficient condition for the generation of SIs in children" (2016, p. 13). Finally, in their third and last experiment, 60 5-year-old children and 24 adults have been assessed to demonstrate that even with the presence of another relevant quantifier (*none*) children could generate scalar implicatures like adults. In conclusion, Skordos and Papafragou suggested that their results seem to support neither Noveck's (2001) proposal nor Katsos and Bishop's (2011) one and that it is just the accessibility of scalar alternatives that helps children with the computation of scalar implicatures.

3.2 The Experimental Turn: Adult's Processing

If studies on children help on what concerns the acquisition and development of pragmatic interpretation of scalar items, studies on adults focus on the cost of such process. It is again Noveck (with Bott) who carried out a pioneering work in 2004, described and replicated in Chapter 5. In their work, Bott & Noveck showed with four experiments that pragmatic interpretations are computed with more effort, concluding that their data can be explained by a Relevance-Theory and not by a Neo-Gricean account. This result has been confirmed "with various methodologies, such as the truth-value judgment task (Bott and Noveck 2004; Degen and Tanenhaus 2011), self-paced reading (Breheny et al. 2006; Chemla et al. 2013), and eye-tracking studies (Huang and Snedeker 2009a; Storto and Tanenhaus 2005)" (Chemla & Singh, 2014b, pp. 387-388). The debate has now turned into an analysis of the cost and three main models are often discussed in the experimental literature: Default model, Literal-first model and Constraint-Based model.

3.2.1. The Default Model

The Default model' assumption moves from the Neo-Gricean theoretical approaches (see Chapter *2.1*) and it considers the pragmatic interpretation of scalar terms as a default interpretation: this is due to a principle of communicative efficiency. Specifically,

> Rapid, effortless inference processes for deriving GCIs are proposed as a solution to the problem of the articulatory bottleneck: while humans can only produce a highly limited number of phonemes per second, communication nevertheless proceeds remarkably quickly. The inferences that allow listeners to derive context-dependent interpretations, such as those presumably involved in computing particularized implicatures, are assumed to be slow and resource-intensive, in the sense of classic two-process models of attention that distinguish between *automatic* processes, which are fast, require few resources and arise independent of context and *controlled processes*, which are slow, strategic and resource demanding […]. In contrast, a scale containing a small set of lexical alternatives – like <all, some> - could be pre-compiled and thus automatically accessed, regardless of context. (Degen & Tanenhaus, 2015, p. 670).

Thus, according to this view, every time a weak scalar term is encountered it is automatically interpreted with its upper-bound meaning (i.e., the pragmatic meaning). However, there are no studies that clearly go in this direction, even when the goal was precisely to assess this model (Bezuidenhout & Morris, 2004), and this is why we are going to focus more on the other models.

3.2.2. The Literal-First Model

The Literal-First model (Huang & Snedeker, 2009) considers a two-step processing of scalar implicatures: every time a weak scalar term is encountered, the semantic interpretation is always accessed before the pragmatic one. Huang and Snedeker (2009) reached this conclusion after three experiments on online implicatures comprehension with the use of eye-tracking technologies. They created stories in which there were four different characters (two boys and two girls) in 4 different areas of the screen; characters received two types of objects (e.g., socks and soccer balls). After looking the objects divided between the characters on a display, participants heard instructions of the form "Point to the girl that has [QUANTIFIER] of the socks" and eye movements were recorded. The first two experiments were similar and only the objects' distribution was manipulated for control sentences. "During the Quantifier phase" they "found increased fixations to the Target for the *two*, *three*, and *all* trials suggesting that listeners were able to quickly use the lexical semantics of the number words and the strong scalar quantifier to disambiguate the referent" and "a delayed disambiguation for the *some* trials, suggesting that initial processing was limited to the lower-bounded lexical semantics of this weak quantifier" (Huang & Snedeker, 2009, p. 395).

The third experiment was similar but the number of distributed objects was equated (e.g., one girl with 3 objects X, one boy with 3 objects X, the second girl with 6 objects Y and the second boy with 0 objects Y). They also compared a ''2-referent trials" (i.e., where two

referents are compatible with the semantics interpretation of *some*) with a ‘‘1-referent trials” (i.e., where just one referent is compatible with the semantic interpretation of *some* because the other one has no objects). In Experiment 3, they saw again “delays in looks to the Target for critical trials that contrasted some with a total set (2-referent trials). However, similar delays were not seen in trials that contrasted some with an empty set (1-referent trials). […] These results suggest that resolution of the Target is quicker via semantic analysis than pragmatic inference” (Huang & Snedeker, 2009, p. 403). In conclusion, Huang and Snedeker’s result goes in the direction of other offline studies in which a cost for pragmatic interpretation has been found.

Contradictory results have been found by Grodner, Klein, Carbary and Tanenhaus, (2010), in which a similar Picture Selection Task showed no delayed looking times for pragmatic interpretations of *some*. However, differently from Huang and Snedeker, they had no numerical quantifiers in the filler items. Thus, the cost found in Huang and Snedeker’s work might be related to the difficulty difference between the filler items with *two* and *three* and the undefined quantity of the sentences with *some* (Tomlinson, Bailey, & Bott, 2013, p. 21).

3.2.3. The Constraint-Based Model

Constraint-Based model is part of the more general Context-Driven framework in which the role of context is central to determine whether there is a cost in computing an implicature.

> For example, Relevance Theory (Sperber & Wilson, 1995), like the Literal-First hypothesis, assumes that the semantic interpretation is basic. In Relevance Theory the upper-bound meaning is only computed if required to reach a certain threshold of relevance in context. In contrast to the Literal-First hypothesis, however, Relevance Theory does not necessarily assume a processing cost for the pragmatic inference. If the context provides sufficient support for the upper-bound interpretation, it may well be computed without incurring additional processing cost. However, a processing cost will

> be incurred if the upper-bound interpretation is relevant but the context provides less support (Degen & Tanenhaus, 2015, p. 671).

The Constraint-Based model moves from three considerations (Degen & Tanenhaus, 2015, p. 671):

a) Utterance comprehension is probabilistic and constraint-based;
b) Listeners generate expectations of multiple types about the future (e.g., phonetic and syntactic properties of utterances, the domain of reference, the meaning conveyed by the speaker);
c) Interlocutors can rapidly adapt their expectations to different speakers, situations, etc.

Based on these assumptions, Degen and Tanenhaus' (2015) have taken the view that implicatures are computed without particular effort in any circumstance in which there is a "support from multiple cues"; however, if those cues are not present implicatures might be derived with more effort and they will not be as strong. Under this view, the goal of their 2015 study was to determine such cues. Specifically, the cues under investigation were the use of the partitive *of* (*some* vs. *some of*) and the listener's assumption on the scalar alternatives available to the speaker. Authors developed a gumball paradigm "in order to investigate whether listeners are sensitive to the partitive and to the naturalness and availability of number descriptions as lexical alternatives to *some* in scalar implicature processing using a range of different set sizes" (Degen & Tanenhaus, 2015, p. 678). In the experimental set there is a gumball machine formed by an upper chamber (where there are all the gumballs at the beginning) and a lower chamber (where the gumballs drop); through three experiments, participants, after listening to a sentence like 'You got [QUANTIFIER] gumballs', had to judge its naturalness or to evaluate whether they agree with it. Response times have

also been colleted (Experiment 3). Main results showed that a) the naturalness of *some* was lesser for small sets (1 – 3) and for unpartitioned set (all gumballs) compared to the intermediate ones (6–8); b) *some of* was less natural than *some* ; c) *some* was less natural compared to exact numbers when the two are intermixed, particularly for the smallest set sizes; d) *all* was more natural than *some* for unpartitioned set; e) naturalness scores predict response times. All in all, results led authors to conclude that "the speed and robustness of an implicature is determined by the probabilistic support it receives from multiple cues available in the linguistic and discourse context, including the task/goal relevant information" (Degen & Tanenhaus, 2015, p. 702), in line with Constraint-Based models' prediction. In the authors' view the distinction between scalar and ad-hoc implicatures does not exist.

3.3. This Thesis' Outline

This thesis will take a closer look at the experimental side of the topic under debate and a series of data collected in the last three years will be presented: the aim was to add some pieces to the complex puzzle just introduced on the mechanism behind the comprehension of conversational implicatures. To do so, in a series of experiment we manipulated both the type of implicatures (scalar vs. ad-hoc) and the population under investigation (typical vs. atypical; children vs. adults). Some of the papers presented herewith are submitted to scientific journals, others are in preparation.

In the first chapter of this dissertation I will present a work on ad-hoc and scalar-implicatures computation in typically developing children. Since several studies in the past years documented children's failure in deriving scalar implicatures but not in deriving ad-hoc implicatures, looking at children's acquisition can help shed light on the theoretical frameworks that better account for the derivation of these strengthened meanings. The debate on children's acquisition of implicatures is focused on the reason behind their difficulties compared to adults. As we have seen, some authors attribute children's difficulties with the

computation of scalar implicatures to their not yet fully developed pragmatic abilities: children have been considered either as more tolerant of pragmatic violations than adults ('tolerance account', Katsos & Bishop, 2011) or as unable to recognize what is conversationally relevant ('relevance account', Skordos & Papafragou, 2016). Other authors suggested that the problem in deriving scalar implicatures might be due not to a lack of general pragmatic reasoning, but to their immature knowledge of lexical scales ('lexicalist account', Barner, Brooks, & Bale, 2011; Foppolo, Guasti, & Chierchia, 2012; Tieu, Romoli, Zhou, & Crain, 2015). In one experiment, we compared children's derivation of both scalar and ad-hoc implicatures in order to identify the role of general pragmatic reasoning and of their knowledge of lexical scales. To obtain a direct comparison between these two types of implicatures, we used the same Picture Selection Task for both of them. We also assessed children's IQ, morpho-syntatictic abilities and ToM abilities. In another experiment, we also compared the results of a Picture Selection Task and of a Truth-Value Judgment Task for scalar implicatures only, in order to evaluate the role of visual alternatives in their computation. In this latter experiment, children's performance with the scalar item 'some' did not improve in the Picture Selection Task compared to the classical Truth-Value Judgment Task. In the ad-hoc- scalar comparison, we confirmed that children have more difficulties with scalar- than ad-hoc implicatures, suggesting that children's problem in the computation does not lie in a general lack of pragmatic comprehension. The fact that children correctly selected the target for ad-hoc implicatures indicates that children are able to take salient alternatives into consideration. The roles of mentalizing and linguistic skills are discussed.

In the second chapter, we will once again focus on the distinction between scalar and ad-hoc implicatures but testing children with Autism Spectrum Disorders (ASD). Pragmatic abilities of people with ASD are generally considered impaired but previous studies on scalar implicatures found no difficulties in ASD population compared to typical population, even if

they have ToM deficits. These studies, however, focused just on scalar implicatures, while we aim to extend such research to different types of implicatures. Since previous studies found a correlation in the ASD groups between verbal abilities and scalar-implicature computation, we decided to assess also ad-hoc implicatures, which arise on the basis of contextual instead of linguistic features. Moreover we tested younger population (from 4 to 9 y.o.), at an age in which also typically-developing children begin to compute implicatures. In this way, we aimed at checking whether there is a delay in ASD children compared to typically-developing children in the ability to derive pragmatic meanings. Furthermore, we decided to look more at possible correlations between general intelligence, ToM and linguistic skills and implicatures computation. We used the same Picture Selection Task of the study presented in Chapter 1. Results showed that, again, scalar implicatures are more difficult than ad-hoc implicatures in the group of control children. However, unlike what was found in previous studies, ASD children displayed greater difficulties than typically developing children of the same age for both type of implicatures.

Overall, these two first studies on typically developing children and on children with ASD had the goal to better disentangle the debate around differences between generalized and particularized implicatures, with both the use of a same methodology (Picture Selection Task) and searching for possible roles of cognitive, mentalizing and linguistic skills. We confirmed that scalar implicatures are more difficult to compute than adhoc implicatures and that ToM skills might play a role for scalar implicatures.

From the third chapter on, we moved to adult population. In Chapter 3, we decided to assess the association between autistic traits in the broader phenotype and performance in tasks requiring the computation of scalar implicatures. We specifically selected a population that was previously shown to possess, overall, higher autistic traits: students of scientific disciplines. Our goal was to investigate whether specific cognitive traits in our participants

might be related with less pragmatic answers in the task. In two experiments we checked the degree of acceptance of underinformative scalar items in students enrolled either in a scientific or in a humanistic curriculum, assessing their autistic traits using the Autism-Spectrum Quotient questionnaire. We found that students enrolled in Science curricula provided less pragmatic answers compared to students enrolled in Humanities curricula. Moreover, autistic traits, and specifically ToM skills, were negatively associated with the number of pragmatic answers.

In the fourth chapter, we will assess the effect of a second language on the oral processing of scalar implicatures: the goal was to check whether the cognitive effort caused by the processing of a language different from the mother tongue one, might interfere with their derivation. We asked bilingual students whose second language was either English or Spanish to perform a *Sentence Evaluation Task*. We also included a group of bilinguals that performed the same task in their first language. We found more pragmatic answers in the first-language condition than in the second-language condition, suggesting that deriving a scalar implicature is effortful. In our analysis, this study provides data against the default models of scalar-implicatures computation.

After finding strong evidence against the default hypothesis, the fifth chapter attempts to focus on the debate regarding the cost of scalar-implicatures computation with the use of different experimental techniques. First of all, in one experiment we offered a replication of the first study that did investigate such cost, increasing the interest on this topic: Bott & Noveck (2004). Specifically, we decided to replicate their third experiment that was a Sentence Evaluation Task with the use of categorical sentences, in which they found that more time is needed to answer pragmatically than logically in underinformative sentences. Similarly, we found longer reaction times (RTs) when participants answered pragmatically to underinformative categorical sentences. With the second experiment (and through a Sentence

Evaluation Task) we excluded the possibility that the cost of the implicature computation may be due to an experimental artefact, i.e. that it is easier for participants to move down in the conceptual hierarchy than to move up. We used categorical sentences with artificial categories (using pseudo-words): there were two different sets of images, one that validated and one that did not validate the sentence. We tested two groups, one group saw the validating condition, and one group saw the other one. Results confirmed again a cost when participants gave pragmatic answers compared to logical answers. Finally, through a third experiment we measured not only RTs but also reading times and eye-movements. With the use of a variety of images (with two characters and three areas of interest) and objects, we asked participants to run a self-paced reading and evaluate underinformative sentences, as well as control sentences. Surprisingly, we did not find differences in reading times between pragmatic and logical answers in the infelicitous condition. Our results suggest that having a visual context reduced the effort. Eye-tracker data did not bring significant evidence in the expected direction, probably due to the fact that participants could look at the picture before reading the sentences; however, we propose some future development of our study.

To conclude, this thesis' goal is to bring new data into the complex and diverse field of study of implicature computations, testing different population with different techniques. We report new findings and propose future developments of the field (see *General Discussion*).

PART 1 –

TYPICAL DEVELOPMENT:

THE COMPREHENSION OF CONVERSATIONAL IMPLICATURES

CHAPTER 1

Scalar- and Ad-hoc-Implicature Processing in Typically Developing Children

This chapter is based on the following original proceedings:

Foppolo F., Mazzaggio G., Panzeri F. and Surian L. (2018). What's behind some (but not all) implicatures. *Front. Psychol. Conference Abstract:* XPRAG.it 2018 -Second Experimental Pragmatics in Italy Conference. doi: 10.3389/conf.fpsyg.2018.73.00041

The article is in preparation

Abstract

Children's ability to cope with conversational inferences has been the matter of a lively debate both in the linguistic and the psychological literature of the past decade. Several studies in the past years showed difficulties in preschoolers when computing the scalar implicature associated to the quantifier *some,* which typically gives rise to the *some but not all* inference, while better performance has been documented with ad-hoc implicatures, i.e. pragmatic inferences that ensue from context-driven scales. Children's behavior has been accounted for in terms of pragmatic failure (i.e., tolerance of pragmatic violations, detection of conversational relevance), processing costs and delay in the lexicalization of the scale for scalar quantifiers. In Experiment 1 (N = 58) we make a methodological contribution, showing that the Truth Value Judgment Task, traditionally employed to investigate children's pragmatic ability and recently criticized, prompts a rate of pragmatic responses comparable to the Picture Selection Task in which alternatives are presented visually and no metalinguistic judgment is provided. In Experiment 2 (N =141) we charted the developmental trajectory of scalar and ad-hoc implicatures, and we found that preschoolers performed better with ad-hoc pragmatic inferences than with scalar implicatures. We further contributed by linking children's performances on implicatures with cognitive and linguistic measurements. A role of morphosyntactic competence has been found for both kinds of implicatures, while performance on Theory of Mind tasks was more positively associated with success on the scalar implicature task. An explanation is given in terms of a lexicalist approach to scalar implicatures.

1. Introduction

Children's ability to cope with conversational inferences has been the matter of a lively debate both in the linguistic and the psychological literature of the past decade. A new trend of experimental investigation emerged from pioneering studies by Chierchia, Crain, Guasti, Gualmini, and Meroni (2001) on children's interpretation of disjunction and by Noveck (2001) on children's interpretation of the quantifier *some* and the modal *might*. Since then, a thriving body of research has been devoted to experimental pragmatics, with a special focus on children's (and adults') interpretation of the quantifier *some* (cf. Chemla & Singh, 2014a,b and Skordos & Papafragou, 2016 for an overview).

To introduce the topic of our investigation, imagine a typical situation in which two hungry boys are looking for a snack in the kitchen cupboard. If boy A utters sentence (1) to boy B after opening the cookie box, then boy B will infer that there are some cookies left in the box, i.e. (2). This inference is known as a Scalar Inference (scalar implicature). When the strictly literal meaning of (1) is enriched with the scalar implicature in (2), we get the strengthened meaning (3).

(1) Mommy ate *some* of the cookies.

(2) Mommy did not eat *all* of the cookies.

(3) Mommy ate *some but not all* of the cookies.

The scalar implicature in (2) arises by virtue of the fact that the speaker chose to utter (1) instead of a possible - equally plausible - alternative, namely (4), which would be compatible with a situation in which the box is empty.

(4) Mommy ate *all* of the cookies.

Assuming that the speaker is cooperative, the hearer will thus infer (3) on the basis of the speaker's choice to utter (1) over (4), following a standard Gricean reasoning about the maxims of conversations that rule our conversational exchanges (Grice, 1975). In particular, the Gricean Maxim of Quantity I urges the speaker to provide as much information as required by the goal of the exchange: in the example at hand, the quantity of cookies left in a box is relevant to the hearer, and the use of a sentence like (1) in the case Mommy ate *all* of the cookies would be *underinformative* given that a more informative alternative, i.e. (4), could be uttered.

Pragmatic inferences that ensue from lexical scales like <*some*, *all*> are one kind of generalized (conversational) implicature and depend on the scalar ordering of items on a scale of informativity in which the stronger element in the scale (in this case, *all*) entails the weaker term (*some*). In such kind of implicatures, alternatives are linguistically determined by their position on this scale (Horn, 1972).

There's also another kind of conversational implicature, one in which alternatives are not linguistically pre-determined but are made available due to special features of the context. Sticking to our example above, suppose our hungry kids find two cookie boxes in the cupboard, one with a ribbon and one with a ribbon and a flag. Then boy A peeps in the boxes and utters (5).

(5) Mommy ate all the cookies in the box with the ribbon!

We bet no one thinks that boy B would be in trouble in choosing the right box in this context: in fact, he would probably not even consider picking up the box with the ribbon and the flag, despite the fact that this box, too, has a ribbon on it. Why so? For a reasoning

analogous to the one made before: sentence (5) is one of the possible alternatives that the speaker could have used. In particular, he could have uttered (6), which is optimally informative in this context; the fact that he didn't, entitled the hearer to infer that (7) is what the speaker meant, which guides him towards the correct choice.

(6) Mommy ate all the cookies in the box with the ribbon and the flag!

(7) Mommy ate all the cookies in the box with (only) the ribbon!

While the steps beyond the derivation of the pragmatic inferences in (3) and (7) are the same, the way alternatives are brought to salience is radically different: while in the case of generalized scalar implicatures alternatives are linguistically encoded in a scale, in the case of particularized implicatures alternatives are construed ad-hoc in the context of utterance. A difference along this dimension makes a crucial difference for some theoretical accounts, but not for others, as we will discuss.

As for children's performance in the derivation of pragmatic inferences, previous works in the acquisition literature show that pre-school aged children have difficulties in deriving the scalar implicature associated to *some*, and that difficulties remain for some of the kids even in the most facilitating and ecological settings (cf. Foppolo, Guasti, & Chierchia, 2012 and Skordos & Papafragou, 2016 for an overview). On the other hand, a recent study by Stiller, Goodman and Frank (2015) showed that three and a half year-old children are already adult-like in interpreting *Ad-hoc* (henceforth, ad-hoc implicature) implicatures (see also Jackson & Jacobs, 1982 and Surian & Job, 1987 for related findings).

Different hypotheses have been formulated to account for children's difficulty with scalar implicatures. Under one approach, that we might label the *lexicalist* approach, children's failure with the *some but not all* implicature stems from the fact that children have

not lexicalized the scale yet or have problems in retrieving the scalar alternatives to *some*: though knowing the meaning of *some* and *all*, they might not know that the two are part of a scale, or they might still be in the process of drawing stable links amongst quantifiers belonging to the same scale. Under Foppolo et al.'s (2012) proposal, children's failure with the scalar implicature associated to *some* reflects children's immaturity at the lexical level, in which two layers of meaning should be associated to the quantifier: the existential meaning (according to which "some P Q" instantiates the existence of at least one P that Q) - and the scalar or enriched meaning (according to which this quantifier is ordered with other alternatives on a scale so that "some P Q" leads to the negation of the stronger scalar alternatives "most P Q" and "all P Q"). While almost all children aged 6 and 7 are adult-like in the derivation of the enriched meaning of *some*, 5-year-old children split, and only some of them derive the scalar implicature, despite the fact that all of them recognize that *all* is a better description than underinformative *some* (Foppolo et al., 2012: Exp. 1 and Exp. 5). Similarly, Barner, Brook and Bale (2011) argued that children have difficulty in accessing the relevant scalar alternatives in the lexicon, and showed that such difficulty is not grounded in memory limitation, as shown by children succeeding in accessing relevant contextual alternatives. Thus, under the *lexicalist* hypothesis, there is a developmental stage in which children have not completed the additional learning step that links scalar quantifiers in a scale and consequently fail in rapidly and automatically accessing lexical items as scalar alternatives (see also Tieu, Romoli, Zhou, & Crain, 2015. for an explanation along these lines for the scalar implicature associated with the scale <*or, and*>).

A different approach is what we might label the *processing* account: within a Relevance-Theoretic framework, Pouscoulous, Noveck, Politzer and Bastide (2007) hypothesized that children's problems with scalar implicatures are due to difficulties in the optimization process between the cognitive gains of uttering the most informative utterance

and the processing costs required to access it. Their argument is that, in most cases, the non-enriched interpretation of a scalar term will often suffice as a relevant-enough interpretation of the utterance in which it occurs; in the light of a balance between effect and effort, children might fail to access the enriched meaning in more complex scenarios and tasks.

A more general *pragmatic* approach, however, associates children's failure with scalar inferencing with their yet immature pragmatic system. This general idea is set out differently in a variety of accounts. Under Katsos and Bishop's hypothesis (Katsos & Bishop, 2011), children fail to reject underinformative *some* because they are more tolerant of pragmatic anomalies than adults: though recognizing that underinformative *some* is not optimal, they tolerate it, as if they were conforming to non-adult-like pragmatic norms. More recently, Skordos and Papafragou (2016) argued in favor of an important role for conversational relevance in the derivation of scalar implicatures and proposed that children's problem with scalar implicatures might lie in their failure to recognize that the scalar terms constitute relevant alternatives in certain contexts. They tested children's performance in three different experiments. In Experiment 1 and 3, they modulated the availability of scalar alternatives, testing whether children may be encouraged to generate a scalar implicature from the use of a weak alternative (*some*) by the mere presence of the stronger lexical member of the quantifier scale (*all*) or another scalar quantifier (*none*) during the experiment. The hypothesis that the relevance of lexical alternatives plays a role in scalar-implicature generation was tested in Experiment 2, by manipulating the degree to which the stronger lexical item could be easily recognized as a relevant alternative by children in a given context. Their results show that children were more prone to reject underinformative uses of *some* when alternatives were made accessible and relevant in the course of the experiment (see also Foppolo et al., 2012: Experiment 6 for a similar finding).

With respect to the distinction introduced above between particularized (such as ad-hoc implicatures) and generalized implicatures (such as scalar implicatures), these three general accounts make different predictions. In principle, no difference should be expected between types of implicatures within a *pragmatic* or a *processing* account. According to these approaches, children's non-adult-like behavior relies on a principle of pragmatic tolerance or on a failure in accessing or recognizing relevant alternatives (because of processing costs, as in Pouscolous et al.'s account, or because of low saliency, as in Skordos and Papafragou's account). If this is the case, then these difficulties should affect all kinds of implicatures, independently of their status. In fact, under some theoretical accounts, like the Relevance Theory (Sperber & Wilson, 1986; 1995), the derivation of all pragmatic inferences rests on the same underlying mechanism, one in which context, together with the evaluation of costs and benefits, plays a key role in determining when enrichment is relevant or not.

Under *lexicalist* approaches, on the other hand, a difference between ad-hoc and scalar implicatures is expected: while in the case of ad-hoc implicatures the alternatives that are activated depend solely on context, in all generalized implicatures including scalar implicatures, the set of alternatives is a feature of the language relying on the lexical representation of the scalar item itself. It is worth noting that also in this approach context might intervene in favoring (or suspending) the inference, in obedience to Gricean's maxims of conversations. The crucial difference is in the access to the alternatives, which depends on a linguistic representation and a lexical retrieval mechanism in the case of scalar quantifiers, while it is purely context-driven in the case of ad-hoc scales.

As we said, previous results indicate that children as young as 3,5 are successful in deriving ad-hoc implicatures (Stiller et al., 2015), while a more complex picture emerges in the case of scalar implicatures, for which children's success has been shown to vary

considerably across ages, materials and tasks; in general, not all children at age 5 have reached an adult-like stage yet.

In a recent study, Horowitz, Schneider and Frank (2017) compared ad-hoc and scalar implicatures directly by means of a picture selection task modelled after Stiller et al. (2015; see also, Jackson & Jacobs, 1982 and Surian & Job, 1987 for a preliminary version of this task), in which the child had to select a target (among different pictures) by following oral instructions. The authors tested children aged 4 and 5, reporting a better performance with ad-hoc than scalar implicatures. They also found a correlation between children's rate of interpretation of the scalar quantifier *some* as *some but not all* and their performance on the negative quantifier *none*. In order to account for these findings, they suggested that part of children's problems with scalar implicatures may be rooted in some difficulties at the semantic rather than at the pragmatic level or in a lack of general processing resources.

In this paper we aim at contributing to this debate by presenting two experimental studies. In a first study (Experiment 1), we tested children's performance on scalar quantifiers and scalar implicatures in two tasks, administered within subjects in different sessions: a power point version of the classical Truth Value Judgment Task (similar to the one used by Katsos & Bishop, 2011) and a referential Picture Selection Task similar to the one used by Horowitz et al. (2017). The aim of this preliminary study was twofold. First, we wanted to compare the two tasks directly. In the Picture Selection Task, the child is asked to select a picture that corresponds to a sentence. Target sentences, i.e., "Some S are P", are semantically true descriptions of two pictures: one fits the pragmatic interpretation of the sentences, that is only some S are P, and the other the logically weaker reading, in which all S are P. If a child is computing the pragmatic inference, she will select the picture that corresponds to the pragmatic interpretation. But if the child does not derive the implicature, sticking to the semantic reading "at least some S are P", in principle she might select both pictures. The

Truth Value Judgment Task, on the other hand, allows to discriminate pragmatic and logical answers: when presented with a context in which all S are P, that is described with the target sentence "Some S are P", if the child accepts the description, she can only have the logical reading – at least if the task comprises control items to check whether she knows the correct use of *all*. When the child rejects the description in the underinformative scenario, she must provide a justification, and thus those children who argue that the sentence is not acceptable because all S are P in the scenario are indisputably deriving the implicature. Administering the two tasks at the same group of children enables us to check what are the choices in the Picture Selection Task of the "pragmatic" and of the "logical" responders of the Truth Value Judgment Task. No study so far compared the rate of pragmatic responses in the same group of children tested with a Truth Value Judgment Task and a Picture Selection Task task, analyzing Task as a within subject variable.

The second aim of the preliminary experiment was to validate the novel Picture Selection Task for scalar implicatures in Experiment 1 for using it in our second study (Experiment 2), in which we compared scalar implicatures and ad-hoc implicatures directly by means of the same task. Although Horowitz et al. (2017) carried out this comparison in a recent contribution, we independently designed a Picture Selection Task apt to compare ad-hoc and scalar implicatures. Differently from them, we did not include the negative quantifier *none* in our experiment. Indeed, these authors found an interesting correlation between the rate of derivation of the scalar implicature and children's correct answers on *none*. However, we think that this result deserves further investigation, that goes beyond the purposes of this paper. One crucial thing to mention, though, is that *none* does not belong to the scale <some, all>, and its presence might affect children's computation of scalar inference with *some*, favoring a lower-bound interpretation of this quantifier. Indeed, in a visual world eye-tracking study on adults conducted by Foppolo and Marelli (2017) they found that listening to *some*

after a sentence with *none* induced slower convergence towards the pragmatic target compared to the case in which *some* followed a sentence with *all.* For these reasons, we only include *all* as control in our study. In addition, we also correlated the rate of implicatures generated by each child with other standardized measures of cognitive and linguistic development (such as non-verbal IQ, lexical and grammatical abilities and Theory of Mind). No study so far has systematically investigated the correlation between these factors and scalar-implicature computation in typically developing monolingual children. These analyses had two purposes. From a purely developmental perspective, we aimed at understanding the developmental factors beyond children's general performance in pragmatic tasks. Furthermore, our study also aimed at comparing ad-hoc and scalar implicatures with respect to these factors: from a theoretical perspective, this analysis can foster the debate about the nature (and alleged differences) between types of inferences, and the steps involved in their computation.

Experiment 1 revealed no difference between the Truth Value Judgment Task and the Picture Selection Task in the derivation of scalar implicatures, despite the fact that in the Picture Selection Task the scalar alternatives to *some* were also presented as a visual alternative with no metalinguistic judgment required to select the target: in both tasks, children's rate of derivation of the scalar implicature was around 55%, and children were bimodally distributed (Guasti et al., 2005). Experiment 2 revealed a difference in the rate of derivation of pragmatic inferences between ad-hoc and scalar implicatures, with children succeeding more with the former than with the latter, in line with what found by Horowitz et al. (2017). Correlating children's performance across types of implicatures with their performance in linguistic and other cognitive tasks, we found evidence of a significant contribution of morphosyntactic abilities in the derivation of pragmatic inferences, and a positive correlation between ToM abilities and scalar but not ad-hoc implicatures. In the final

session of this paper, we will discuss the impact of these findings for theories of acquisition of pragmatics, and for theoretical approaches to implicatures.

2. Experiment 1

This experiment consists of two scalar implicature tasks administered to the same group of participants in different experimental sessions, a classical Truth Value Judgment Task and a novel Picture Selection Task. It serves two main goals: first of all, we wanted to verify whether the pragmatic responders in the Truth Value Judgment Task (who reject underinformative *some* in an all-scenario) will then select the picture in which only some S are P in the Picture Selection Task. If this is the case, this sets a baseline for Experiment 2 in order to verify the Picture Selection Task as a sensitive task for scalar implicature derivation.

2.1. Participants

Fifty-eight children aged 4 and 5 were tested (4;2-6;0; M = 63 months). Kids were recruited from two kindergarten schools in the North of Italy and were tested after both parents signed a consent form for participation.

2.2. Materials and procedure

With this first experiment our goal was to compare a Truth Value Judgment Task with a novel Picture Selection Task for scalar implicature computation. For this reason, we decided to maintain similar structural characteristics: both tasks were presented in a ludic fashion and were administered through a PPT presentation on a laptop PC computer in a quiet room of the kindergarten. All the target sentences were pre-recorded to control for prosody. Children's responses were annotated by the experimenter on an answer sheet. The Ethical Committee of Trento University approved this experiment. The tasks were administered one after the other

and the order of presentation was randomized. In the case in which children were tired or distracted, the session was stopped.

The two tasks are detailed in the following sessions.

2.2.1. Truth Value Judgment Task for Scalar Implicatures

This task was adapted from Foppolo et al., (2012) and Katsos and Bishop (2011). Children were presented with a box and two characters that appeared on the screen: Davide, a boy, and Lucy, a foreign girl who is learning to speak Italian. The children's task is to help Lucy to improve her Italian by judging her descriptions of different scenarios. For each trial, the scenario consisted of an array of six objects of the same kind (e.g., apples, dolls, hats); by means of his magic wand, Davide put either none, some (ie. 4 out of 6) or all of the six objects inside the box. The child saw the objects moving from their initial position to the box, while hearing a "magic" sound produced by Davide's wand triggering the movement. At the final stage, Lucy described what happened by saying "Davide put some of/all the Xs in the box" and children were asked to judge whether Lucy said things "right" or "wrong", and to correct her if she said something wrong.

To familiarize children with the task, the session started with two warm-ups in which a clearly true or a clearly false description of the scene was provided. The test phase comprised four target sentences with *some* used in an underinformative way (UI) to describe a scenario in which Davide put all six objects in the box (Figure 1), and eight control items: four with the quantifier *all* (two true and two false), and four with the quantifier *some* (two true and informative, and two false). The experimental conditions are summarized in Table 1.

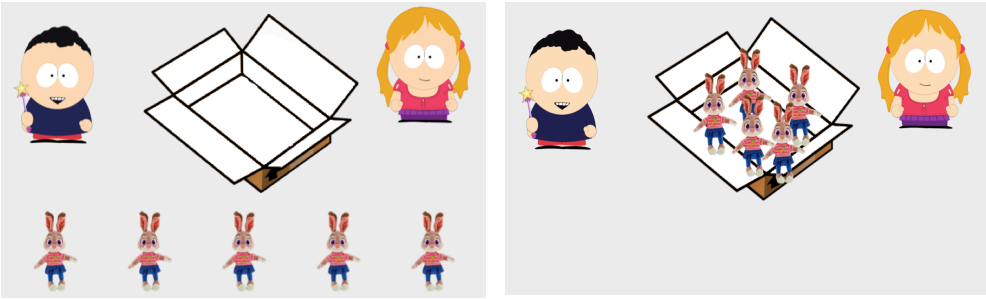


Figure 1. Example of an underinformative item for *some* for the Truth Value Judgment Task. The left panel shows the initial array of object; the right panel shows the final outcome of Davide's action, after he moved the objects from the array to the box by means of his magic wand. Lucy's sentence in this trial was: "Davide put some of the bunnies in the box".

ITEM TYPE	**SENTENCE**	**OBJECTS IN BOX**	**EXPECTED ANSWER**
SOME-TRUE	Davide put *some of* the X in the box	4 / 6	Yes
SOME-FALSE	Davide put *some of* the X in the box	0 / 6	No
SOME-UNDERINFORMATIVE	Davide put *some of* the X in the box	6 / 6	No (= scalar implicature) Yes (= no scalar implicature)
ALL-TRUE	Davide put *all* the X in the box	6 / 6	Yes
ALL-FALSE	Davide put *all* the X in the box	4 / 6	No

Table 1. Description of the different types of items in the Truth Value Judgement Task.

2.2.2. Picture Selection Task for Scalar Implicatures

This task is a classical picture selection in which participants had to find the correct target − among 4 pictures − by exploiting a sentential clue; this task is modelled after Surian and Job (1987) and Stiller et al. (2015). In this task, children were introduced to a boy (Daniele) and were told that Daniele wanted to play a game with the children by giving them

a clue to find specific objects or individuals on the screen. For example, children saw a scenario with 4 birthday cakes (Figure 2) and had to find the one that Daniele is addressing by hearing this clue: “Guess which one is my cake, I give you a clue: On my birthday cake, some of the candles are burning”.

All the test sentences were previously pre-recorded to control for prosody. After a warm-up trial in which children were corrected in case of no answer or wrong answer, the test phase started. In this phase, children were presented with four sentences containing *some*, two sentences containing *all* and one control sentence with no quantifier. Items were shown in a pseudo-randomized order so that the quantifier *all* always appeared first in the sequence.

The structure of the scenario was kept constant across trials of the same kind. In the *all*-scenario, there was a target (for example, a garden with 5/5 red flowers), two competitors (for example, a garden with 3/5 red flowers and a garden with 3/5 blue flowers) and one distractor (a garden with no flowers). In the *some*-scenario, there were two types of distractors: for example, a cake with no candles and a cake with no burning candles. There were also two types of possible targets: a pragmatic target, in which the array of objects was only compatible with the pragmatic meaning of *some,* namely *some but not all* (i.e. a cake with 3/5 burning candles) and an underinformative-competitor, in which the array of objects was also compatible with the more informative quantifier (*all*) (i.e. a cake with 5/5 burning candles). Note that participants who derived the scalar implicature should converge on the pragmatic target (and should do so consistently across trials); participants who stick to the logical interpretation (at least some) could opt for the underinformative competitor, but they could in principle.also select the pragmatic target.

Figure 2. Example of a *some* item in the Picture Selection Task: "Guess which one is my cake, I give you a clue: on my birthday cake, some of the candles are burning".

As discussed in Horowitz et al. (2017), this task is designed to enhance contextual relevance of the *all*-alternative: in fact, this is presented as a visual alternative in the scenario, and it is also given as a linguistic alternative during the experiment. Moreover, the task is in principle simpler than the Truth Value Judgment Task in that no metalinguistic judgment is required, thus lowering the computational resources required to solve the task. As highlighted before, on the other hand, when the Picture Selection Task is presented in isolation the results are not easily interpretable: if the child selects the underinformative competitor, this means that she is assessing the logical meaning of *some*, without deriving the implicature. In contrast, selections of the pragmatic target could be due to the derivation of the implicature (pragmatic responders) or, in principle, to the random choice between target and competitor for a logical responder. A direct comparison of the two tasks within the same group of children can shed some light on children's choices.

2.3. Results

In the Truth Value Judgment Task, the responses on controls were coded as "correct" if the child correctly accepted or rejected the target true and false statements respectively. Responses on test statements were coded as "correct" if the child rejected the underinformative statement with *some* and mentioned *all* in their justification for rejection. In the Picture Selection Task, the responses were coded as "correct" if the child selected the target picture for *all* and the pragmatic-target for *some*.

The two tasks yielded similar results: while children's accuracy on controls was above 94% in both tasks (94.8% in the Picture Selection Task and 94.2% in the Truth Value Judgment Task), children were not adult-like in deriving the scalar implicatures, regardless of the task: they rejected the underinformative-*some* sentences in the Truth Value Judgment Task 55.6% of the times, and they selected the pragmatic target in the Picture Selection Task 57.6% of the times (Figure 3). Performances in the Picture Selection Task and performances in the Truth Value Judgment Task highly correlate (r = .48, t = 4.0496, df = 56, $p < 0.001$).

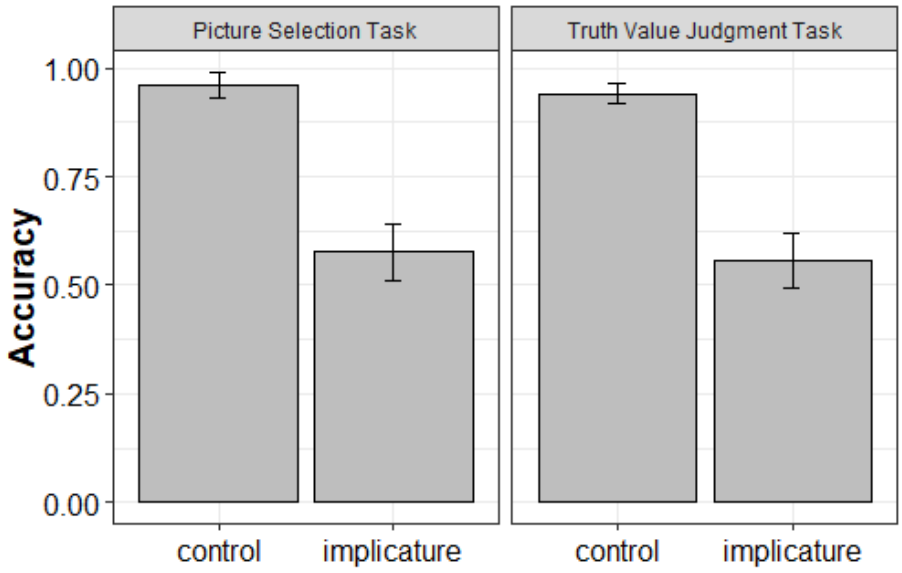


Figure 3. Children's accuracy on controls and underinformative-*some* across tasks.

Note that, as found in other studies (Guasti et al., 2005), children displayed a bimodal distribution in the Truth Value Judgment Task, consistently rejecting or accepting the underinformative-*some* sentences. The distribution of children's responses in the two tasks depending on the number of times (out of 4) they selected the pragmatic target in the Picture Selection Task or rejected the underinformative-some sentence in the Truth Value Judgment Task is plotted in Figure 4. As the graph shows, only 2 (3%) children in the Truth Value Judgment Task and 10 (17%) children in the Picture Selection Task did not provide consistent responses in the underinformative trials.

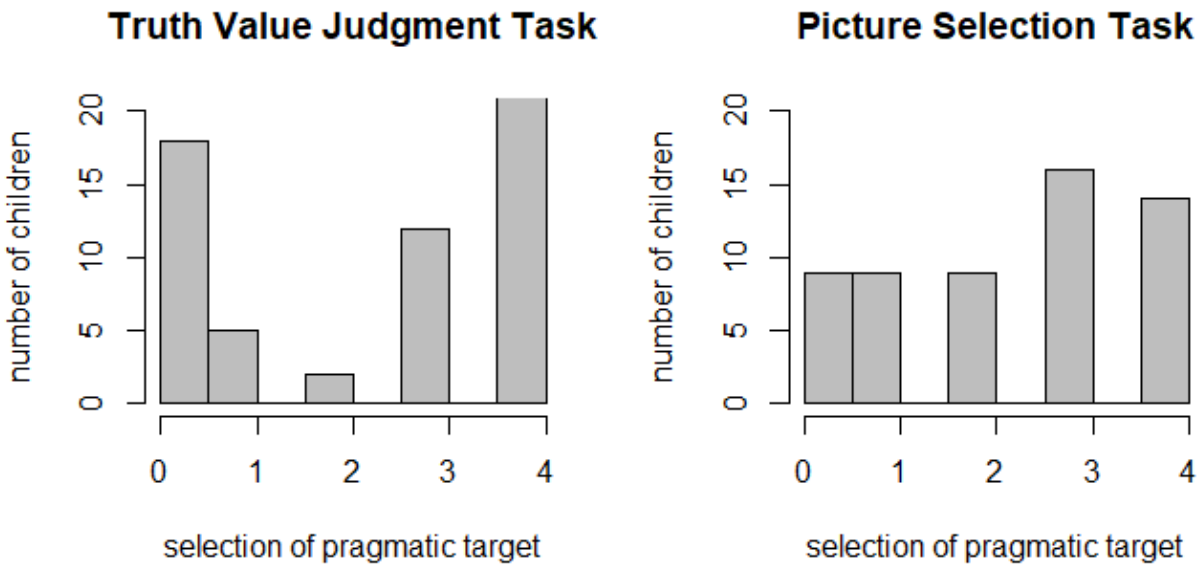


Figure 4. Distribution of children's responses with respect to the number of times (up to 4) they rejected the underinformative-some sentence in the Truth Value Judgment Task (left panel) or selected the pragmatic target in the Picture Selection Task (right panel).

By inspecting children's wrong responses in the Picture Selection Task, we found that they selected the underinformative competitor (for example, the cake on which all the candles were burning) more than 92% of the times.

We run a logistic regression analysis considering accuracy as the dependent variable, subjects and items as random factors, and Task (Picture Selection Task vs. Truth Value Judgment Task), Type (implicature vs. controls) and Age in months as fixed effects. For a better fit of the model, all the variables were centered prior to analysis. A significant effect of Type (Est = -4.102, *SE* = .625, z = -6.567, $p < .001$) and Age (Est = .167, *SE* = .043, z = 3.865, $p < .001$) is revealed: accuracy on implicatures is significantly lower than controls and performance is worse for younger kids. A marginal interaction across Task, Type and Age (Est = 0.154, *SE* = .081, z = 1.905, p = .057) is also observed. No significant effect is revealed for Task (Est = .000, *SE* = .611, z = .001, p = .999) instead.

2.4. Discussion

Children showed excellent knowledge of the semantics of the quantifiers involved in the scale: they were at ceiling in the Truth Value Judgment Task in judging sentences with *some* and *all* in which these quantifiers applied to true and false situations (cf. Foppolo et al., 2012: Experiment 6 for a similar result). Nonetheless, children's performance was less than optimal in judging underinformative-*some* sentences. Also, children displayed a bimodal distribution, always consistently accepting or rejecting underinformative-*some* sentences in the Truth Value Judgment Task, or consistently selecting the underinformative or the pragmatic target pictures in the Picture Selection Task (Guasti et al., 2005).

Our study is the first one to compare the Truth Value Judgment Task and the Picture Selection Task directly on the same group of children: thus, the lack of a significant difference is also revealing from a methodological point of view. Children's selection of the pragmatic target with a *some*-sentence could in principle due to the computation of the implicature or to the random choice of a logical interpretation of the quantifier (at least some). The high correlation of children's performances on the two tasks ensures that the tasks tap the

same pragmatic ability: the pragmatic kids who rejected the underinformative-*some* in the Truth Value Judgment task were those who selected the pragmatic target in the Picture Selection Task; and the logical children who accepted the underinformative-*some* in the Truth Value Judgment Task were those who selected the underinformative-competitor in the Picture Selection Task.

Having confirmed that the "correct answers" of the two tasks identify pragmatic children who derive the implicature, the fact that children's performance on scalar implicatures was the same in the two tasks can shed light on the role of alternatives in the computation of pragmatic inferences. Our results show that the Truth Value Judgment Task, in which children have to listen to a story and have to provide a metalinguistic (binary) judgment, does not prove more difficult than a Picture Selection Task. Crucially, no difference was revealed between the Picture Selection Task and the Truth Value Judgment Task in the rate of derivation of scalar implicature, despite the fact that in the Picture Selection Task the child simply had to select the image that represents the pragmatic interpretation of *some* in a scenario in which the scalar alternative *all* was provided as a visual alternative. *Contra* Skordos and Papafragou (2016), this factor doesn't seem to boost scalar-implicature derivation: we show that, regardless of the task, children's performance with scalar implicatures was far from optimal.

As set out in the introduction, the main goal of this first experiment was that of assessing the validity of using the Picture Selection Task to test scalar implicatures in order to compare children's performance with scalar and ad-hoc implicature within the same group of children by means of this task. This has been done in Experiment 2.

3. Experiment 2

The aim of the second experiment was to compare children's derivation of scalar and ad-hoc implicatures using the same Picture Selection Task paradigm and to link children's performances in these pragmatic tasks with other measures of cognitive and linguistic development.

3.1. Participants

One hundred and forty-one children aged 3 to and 9 were tested. Seventy-five kids were enrolled in the kindergarten (3;10-6;0, M = 60;9 months), and 66 in the primary school (grade 1st to 3rd; 6;10-9;2, M=89;7 months). Children were recruited in different kindergartens and primary schools in Northern Italy. All children were tested after both parents signed a consent form for participation.

3.2. Materials and procedure

Children were administered two Picture Selection Tasks: one was the same Picture Selection Task used for scalar implicatures in Experiment 1. The second was a novel Picture Selection Task designed for ad-hoc implicatures, modeled after Surian and Job, 1987 and adapted from Stiller et al. (2015). Like the Picture Selection Task for scalar implicatures, children had to point at the correct target − among 4 pictures − by exploiting a sentential clue. The story is similar to that for scalar implicatures: participants should follow Daniele's hint to find the correct referent of his expression. For example, Daniele says "Guess which one is my bed, I give you a clue"; then the four pictures appear and Daniele says: "On my bed there is a teddy bear". On the basis of this sentence the child has to find the correct target among the four (Figure 5). As for scalar implicatures, the scenario displays two distractors (i.e. an empty bed and a bed with a penguin on it) and two potential targets: the pragmatic target (i.e. the bed with only the teddy bear) and the underinformative target (i.e. the bed with a teddy bear and a

penguin). As for scalar implicatures, if the ad-hoc implicature is computed the children should select the bed with the teddy bear alone, under the reasoning that, if Daniele wanted them to point to the other bed (the one with the teddy bear *and* the penguin) he should have referred to that bed by explicitly mentioning both things. The fact that he didn't, should entitle the listener to derive the inference that the referent of the request is the bed with only the teddy bear on it.

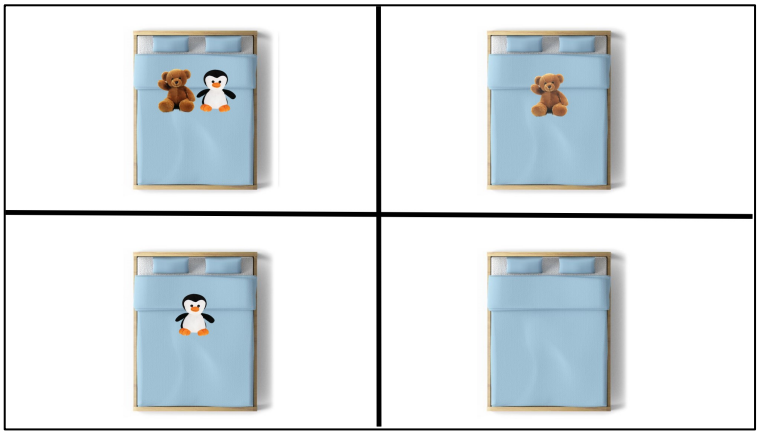


Figure 5. Example of an underinformative item in the Picture Selection Task for ad-hoc implicatures: "Guess which one is my bed, I give you a clue: on my bed there is a teddy bear".

After a warm-up item, children were presented with four ad-hoc implicatures and a control sentence in a pseudo-randomized order. In addition to the two Picture Selection Tasks for testing implicatures, we wanted to control for possible factors that could be correlated with the ability to derive pragmatic inferences of different kinds. To this purpose, children were also tested with a battery of standardized tests: the Raven Coloured Progressive Matrices to test for non-verbal IQ (Italian standardization by Belacchi, Scalisi, Cannoni, & Cornoldi, 2008); the test for Lexical comprehension and the test for Grammatical comprehension taken

from the Batteria di Valutazione del Linguaggio (Marini, Marotta, Bulgheroni, & Fabbro, 2015) to test for receptive vocabulary and morphosyntactic abilities; a series of Theory of Mind tasks adapted in Italian from the first four tasks of Wellman and Liu (2004) to test 1st order ToM (*Appendix 1*, for a description). Our aim was to verify whether cognitive skills (as measured by Raven CPM and by ToM tasks) and/or linguistic (lexical and morphosyntactic) abilities were correlated with none, only one or both types of implicatures, with the goal of identifying possible predictors of pragmatic inferencing on the one hand, and of shading light on which underlying mechanisms scalar and ad-hoc implicatures have in common, and in which they differ.

The tasks were administered by an experiment in a quiet room of the kindergarten or of the school, after children were familiarized with the experimenter and with the laptop PC computer. In the case children were tired, the tasks have been administered in different days. This study is part of a more extensive research on pragmatic comprehension in typically and atypically developing children that has been approved by the University of Trento's Ethical Committee.

3.3. Results

As in the previous Picture Selection Task study, the responses were coded as "correct" if the child selected the pragmatic target; in the case of *some*, this corresponded to the picture in which only some of the objects were affected (for example, the bottom right picture in Figure 2); in the case of ad-hoc implicatures, this corresponded to the picture in which only the object mentioned was present (for example, the top right picture in Figure 5). In all the other cases, the answer was coded as "incorrect".

Children's accuracy on controls was above 95% in both tasks (99% in the Ad-hoc implicature Task and 95% in the Scalar Implicature Task; Figure 6, left panel); overall, the

rate of derivation of ad-hoc implicatures was higher than scalar implicatures (86% vs. 74% respectively; Figure 6, right panel). As observed in the previous study, the overall majority of the incorrect responses corresponded to the selection of the underinformative target (around 90% of the times) and the majority of children provided consistent responses in the underinformative trials (87% in the ad-hoc task and 84% in the scalar implicature tasks), either always selecting (or always failing to select) the pragmatic target, as shown in Figure 7.

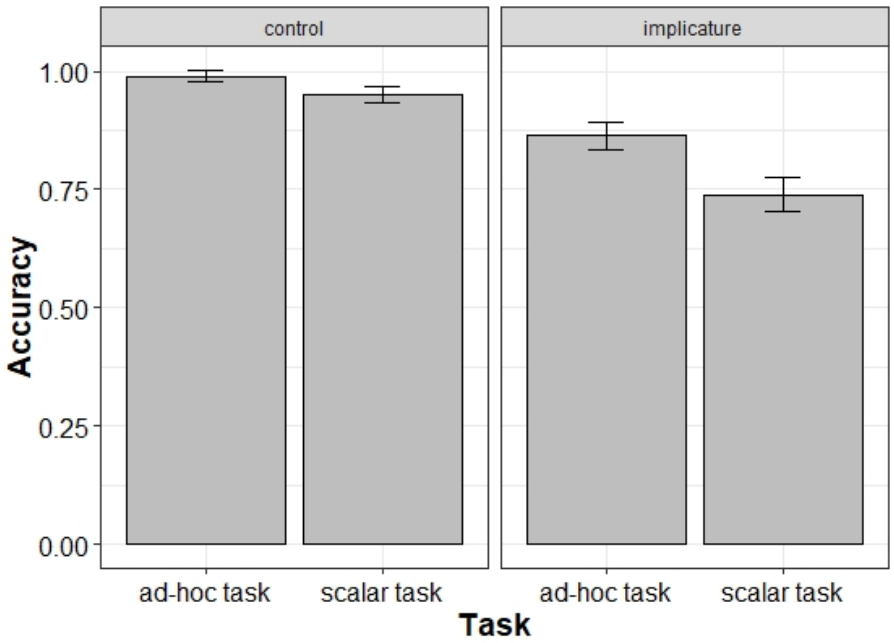


Figure 6. Children's accuracy on controls (left panel) and implicatures (right panel) in the two Tasks (ad-hoc vs. scalar implicature Picture Selection Task).

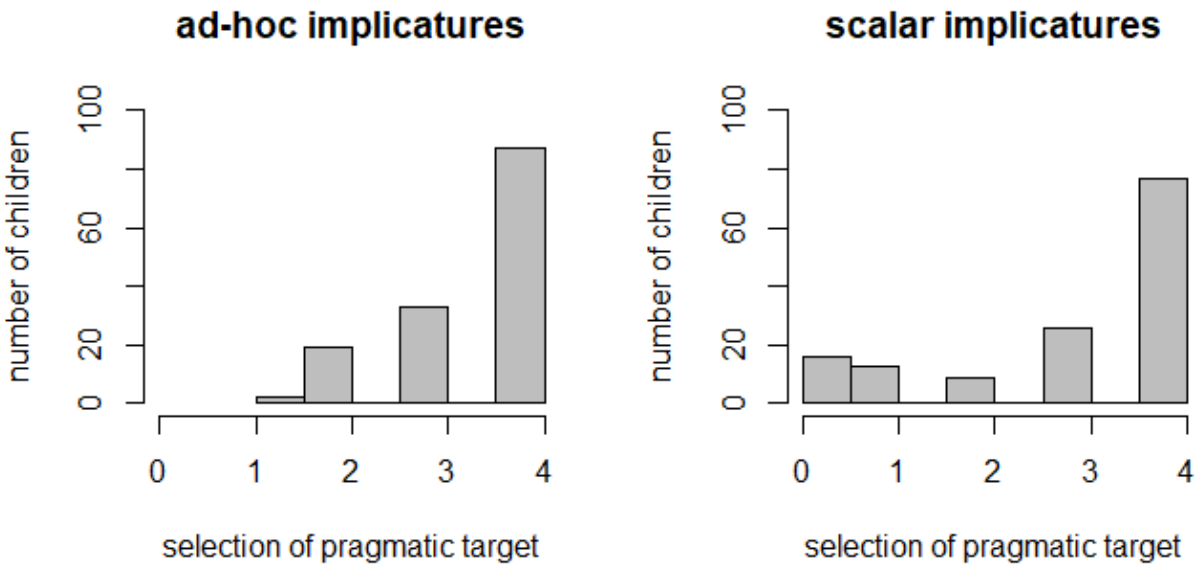


Figure 7. Distribution of children's responses with respect to the number of times (up to 4) they selected the pragmatic target in the ad-hoc (left panel) and the scalar task (right panel).

We ran a logistic regression analysis considering accuracy as the dependent variable, subjects and items as random factors, and Condition (implicatures vs. controls), Task (scalar vs. ad-hoc task), Age (in months), Grammatical comprehension (henceforth, Morphosyntax), Lexicon, Raven and 1st order ToM scores, as well as their interactions, as fixed effects. For a better fit of the model, all the variables were centered prior to analysis; following a backward stepwise model selection-procedure, all the factors and interactions that resulted as non-significantly improving the model-fit were removed. The model that best fits the data reveals: a significant effect of Condition (accuracy on controls is significantly higher than accuracy on implicatures; Est = -2.648, *SE* = 0.547, *z* = -4.836, *p* < .001), a significant effect of Age (accuracy is significantly higher for older children; Est = .049, *SE* = .011, *z* = 4.304, *p* < .001), a significant effect of Task (accuracy is significantly higher for the ad-hoc than the scalar task; Est = -1.540, *SE* = .512, *z* = -3.008, *p* = .003) and a significant effect of Morphosyntax

(accuracy is significantly higher for children with higher scores in the Grammatical comprehension test; Est = .062, *SE* = .027, *z* = 2.288, *p* = .02).

Considering the fact that controls were at ceiling, we next ran a logistic regression analysis considering accuracy on the implicature condition only as the dependent variable, subjects and items as random factors, and type of Implicature, Age (in months), Morphosyntax, Lexicon, Raven and 1st order ToM as fixed effects, as well as their interactions. For a better fit of the model, all the variables were centered prior to analysis; as before, following a backward stepwise elimination procedure, all the factors and interactions that resulted as non-significantly contributing to the best fit of the model were removed. As before, the model revealed a significant effect of type of Implicature (accuracy on scalar implicature was lower than accuracy on ad-hoc implicatures: Est = -1.482, *SE* = .670, *z* = -2.211, *p* = .03), Age (younger children's accuracy was lower than older children's: Est = .053, *SE* = .012, *z* = 4.325, *p* < .001) and Morphosyntax is found (accuracy is significantly higher for children with higher scores in the Morphosyntax test; Est = .078, *SE* = .021, *z* = 3.797, *p* < .001). Focusing on implicatures only, the model also showed a significant interaction between type of Implicature and ToM: albeit not being significant as a factor, ToM resulted to significantly modulate accuracy on scalar (Est = .391, *SE* = .184, *z* = 2.121, *p* = .03), but not ad-hoc implicatures (Est = .029, *SE* = .215, *z* = .136, *p* = .892).

The gap between children's performance with ad-hoc and scalar implicature is even more evident if we split children in two age groups, as in Figure 8. Clearly, the younger children struggle more with scalar than ad-hoc implicatures (57% vs. 78%), while the older kids show parallel performance in both (94% for scalar implicature and 96% for ad-hoc implicature). No difference is observed for controls across age groups and tasks (in all cases, performance is above 90%). It is worth noting that we fully replicated the results from Experiment 1 for scalar implicature for the younger kids.

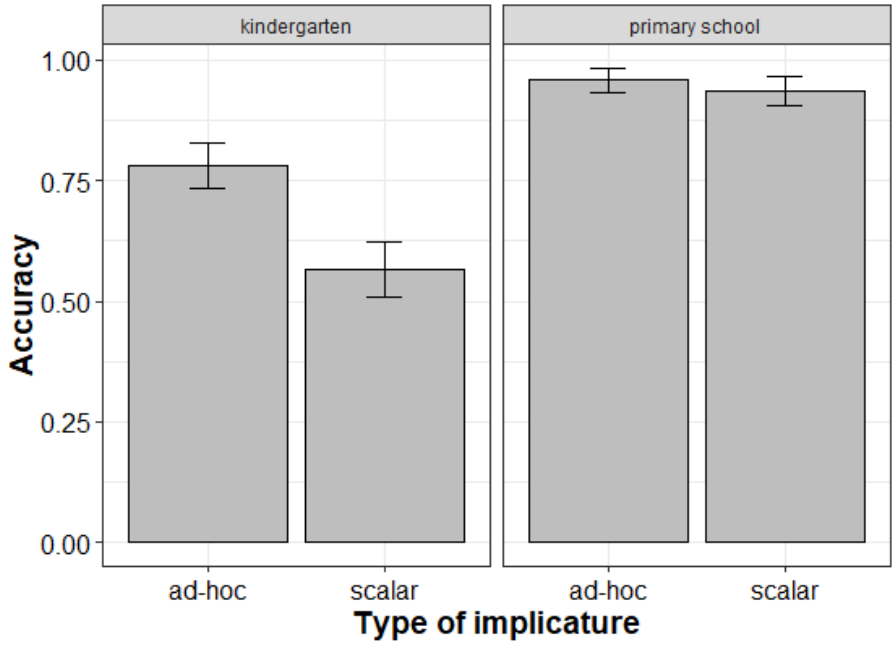


Figure 8. Plot (for descriptive purposes only) of children's accuracy on the two types of implicatures (ad-hoc vs. scalar implicature) across school groups (kindergarten children, N = 75, left panel, and primary school children, N = 66, right panel).

Since we found that the difference between ad-hoc implicatures and scalar implicatures is only revealed in the younger group (Figure 8), to further inspect the contribution of the different factors we focused on the kindergarten children only (N = 75). We performed an analysis considering accuracy on implicatures as the dependent variable, subjects and items as random factors, and type of Implicature, Morphosyntax, Lexicon, Raven and ToM, as well as their interactions, as fixed effect. We removed all the factors and interactions that resulted as non-significant for the best model-fit. As before, the two types of implicatures resulted to be significantly different, with scalar implicatures being more difficult than ad-hoc (Est = -1.926, SE = .702, z = -2.745, p = .006). For these younger kids, however, Morphosyntax is the only measure that turned out to be a significant predictor for implicature computation,

independently of type (Est = .076, *SE* = .026, *z* = 2.890, *p* = .004), while a significant interaction was again found between ToM and type of Implicature (Est = .823, *SE* = .276, *z* = 2.987, *p* = .003): while performance on ad-hoc implicatures does not depend on levels of ToM (even kids with low ToM scores have a good performance with ad-hoc implicatures), performance with scalar implicatures does positively correlate with ToM scores, as shown in Figure 9.

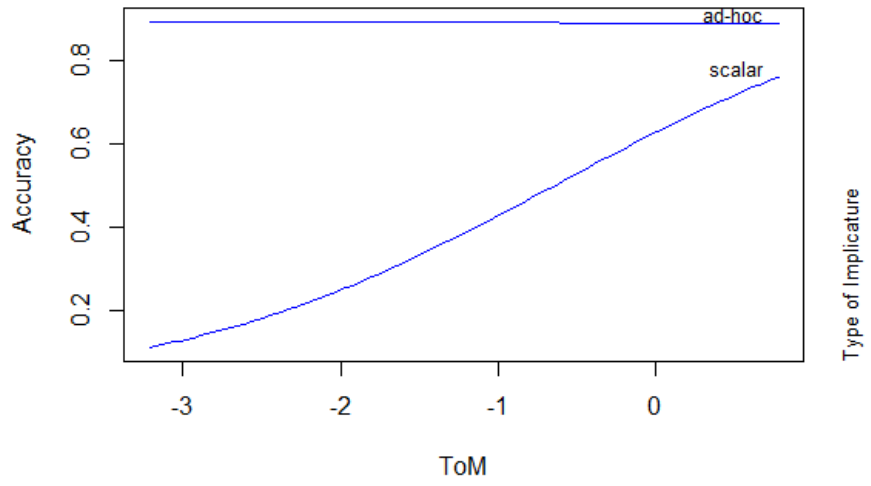


Figure 9. Correlation effects between levels of ToM and type of Implicature.

Our data suggest that the abilities required to derive the two types of implicatures might possibly rely on different modules. To further inspect these results, we performed a Focused Principal Component Analysis (Figure 10). This analysis revealed some interesting patterns. First of all, ToM, as measured in our tasks, seems to be independent from Linguistic abilities and Raven, being it located orthogonally from these factors. Second, the factor that has a higher correlation with scalar implicatures is ToM (*r* = .04), and the factor that correlates less is Raven (*r* < .02). Third, this pattern is different from ad-hoc implicatures, given that in this

case both Raven and Linguistic abilities seem to be the most correlated factors (r = .04 for Raven and Lexicon; r = .06 for Morphosyntax). These descriptive analyses reveal an interesting difference between the two types of implicatures, which goes beyond the fact that children performed differently with them. Such analyses also reveal that the underlying factors of these two kinds of implicatures are different.

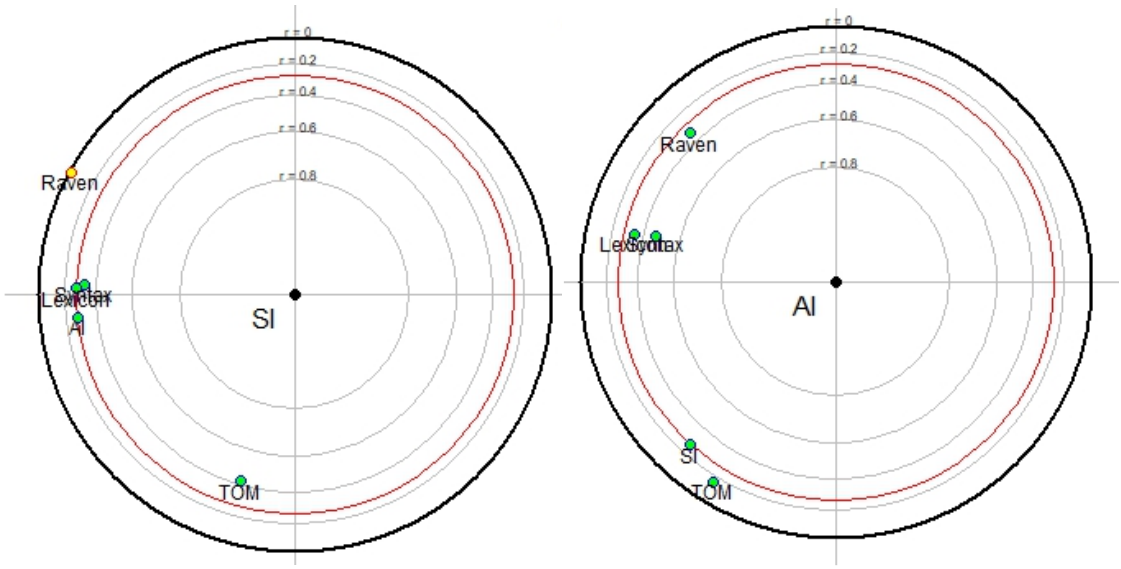


Figure 10. Output of the Focused Principal Component Analysis for scalar implicatures (SI, left panel) and ad-hoc implicatures (AI, right panel).

3.4. Discussion

Children showed an overall excellent performance on controls, and an overall good performance on ad-hoc implicatures. In fact, the group of older kids showed an optimal performance on both types of implicatures. The younger kids, instead, behaved differently in the two tasks: on the one hand, they had not much trouble in detecting the pragmatic target in the case of inferences that were built on a contextual basis, like it is the case for ad-hoc implicature. On the other, they struggled more to detect the pragmatic target for the scalar quantifier, and almost half of the kids consistently failed this task, pointing to the

underinformative target for *some*. These results are in line with those found by Horowitz et al. (2017) in a similar comparison between scalar and ad-hoc implicature.

Beyond their results, we investigated this difference in a developmental perspective, showing that children up to age 6 have more difficulties with scalar implicatures than ad-hoc implicatures, while this difference disappears after age 6, when their performance becomes optimal in both kinds of implicatures. This result is also in line with previous developmental findings about the *some but not all* implicature (Foppolo et al., 2012: Exp. 1).

The correlation analyses showed a difference between scalar and ad-hoc implicature in the factors underlying children's success. Specifically, children's morphosyntactic competence positively predicted their ability to generate pragmatic inferences. More specifically, the measures of ToM abilities were found to modulate children's success with scalar inference, while no impact of ToM on the ability to generate ad-hoc inferences was found.

4. General Discussion

In this paper we tested pre-school and school-aged children's ability to compute generalized (scalar) implicatures (i) in different tasks (Experiment 1); (ii) in comparison with particularized (ad-hoc) implicatures (Experiment 2); (iii) in correlation with other cognitive and linguistic measures.

First, we replicated previous findings in showing that pre-school aged children have difficulties in deriving pragmatic inferences, especially those involving a scalar quantifier, and that their performance is better with those inferences in which alternatives are construed ad-hoc, on the basis of the context (Horowitz et al., 2017). Extending previous findings, we also characterized the developmental trajectory of the two types of implicatures: we showed that preschoolers perform better with ad-hoc than scalar implicatures; we also show that the

delay of the latter over the former kind of implicature is confined to a specific stage in development, namely pre-school age, as scalar implicatures reach an optimal performance at the same age as ad-hoc implicatures, namely in school-age, as it is evident in Figure 8.

We also made a methodological contribution in this paper, showing that the Truth Value Judgment Task, traditionally employed to investigate children's pragmatic ability and recently criticized, prompts a rate of pragmatic responses comparable to the (allegedly) less demanding Picture Selection Task in which alternatives are presented visually and no metalinguistic judgment is provided.

We finally analyzed the impact of different linguistic (lexicon and morphosyntax) and cognitive measures (non-verbal IQ and 1st order ToM) on the rate of pragmatic responses and showed a different pattern of dependencies in the two types of inferences investigated. In particular, we showed that 1st order ToM correlates with the rate of pragmatic answers in the case of scalar quantifiers, but not of ad-hoc scales.

As we outlined in the introduction, children's difficulty with pragmatic inferences is a well attested fact in the acquisition literature. What is still debated is the source or the nature of this difficulty: under some accounts this difficulty is believed to be more grounded in a yet immature linguistic representation of the scale (Barner et al., 2011; Foppolo et al., 2012); in others, it is associated with yet immature pragmatic or processing systems that make children more tolerant than adults when facing pragmatic anomalies (Katsos & Bishop, 2011), or less prone to detect what is contextually relevant or optimal in a given exchange (Skordos & Papafragou, 2016).

We believe that the results of Experiment 2 can shed some light on the disagreements between *lexicalist*, *pragmatic* and *processing* accounts of children's failure with pragmatic inferences. Certainly, it seems hard to reconcile the difference shown between ad-hoc implicatures and scalar implicatures within a *processing* account. If the mechanisms involved

in the generation of all kinds of implicatures rest on the same basis, they should require the same processing costs, and thus children's asymmetrical competence is unexpected. It seems difficult to provide a sensible explanation of the reason why scalar alternatives are costlier than contextually retrieved alternatives without resorting to the linguistic nature of scalar, but not ad-hoc implicatures. Similarly, such difference is not readily predicted or easily explained by a *pragmatic* account, or a *tolerance* account either: if children's pragmatic system were yet immature, the consequence would be a failure in all kinds of pragmatic inferences, unless one assumes that the mechanisms involved in the generation of the inference, or in the retrieval of the alternatives, are not equivalent across types of triggers. In all cases, it seems hard to reconcile these alleged differences in the processes or mechanisms underlying pragmatic inference without letting the nature of the scale involved in the computation entering the picture.

Conversely, this difference is straightforwardly accounted for within a *lexicalist* approach to scalar implicatures: according to the latter, some children might have reached the stage in which they *can* compute all the steps involved in the derivation of pragmatic inferences, but they might have problems in accessing the scalar alternatives in some cases. In particular, problems are predicted in those cases in which the scale requires an *a priori* lexicalization, a process that might take some time. This account would explain why children show a good performance with ad-hoc implicatures but not with scalar implicatures: only the former does not require a preliminary lexicalization of the scale, as alternatives are directly retrieved from the context. The scale <some, all>, instead, needs to be in place before enriching the meaning of *some* as *some but not all*.

The analyses of the factors involved in the two types of inferences also reveal a difference between the two types of inferences, which is straightforwardly accounted for by a *lexicalist* account, while it requires additional speculations by other accounts. The fact that

ToM does not modulate children's accuracy with ad-hoc implicatures seems to suggest that ToM is not a relevant factor in the derivation of pragmatic inferences; thus, it looks as if the fact that it does modulate children's derivation of scalar implicatures might be interpreted as a consequence of the fact that scalar implicatures, like ToM abilities, take time to be acquired. We believe that the correlation observed between ToM and scalar-implicature computation might simply reflects the maturational stages of both abilities around the age of 5: those children that are in a certain maturational stage (that is best captured by ToM tasks, although not strictly related to domain specific ToM abilities) are also able to derive the scalar inference related to *some.* Those children that are still in a more immature stage are not able to derive scalar implicatures. According to our interpretation, ToM, at least as it is captured by the task used in our study, seems not to be a prerequisite for implicature computation; if it was, one would have to explain why it is not necessary in the case of ad-hoc implicatures. We are well aware of the fact that a measure of ToM ability strictly depends on the task used, and how it is difficult to disentangle ToM scores in a specific task from other abilities that might interfere in the task, such as verbal intelligence. Even if it is not always easy to look at the interface between language and ToM (De Villiers, 2007), we suggest that ToM scores might simply capture the maturational level that is required to access scalar implicatures, which is higher than that required to access ad-hoc implicatures. As before, a developmental effect across implicatures is best captured by a *lexicalist* approach to scalar implicatures: children know how to play the game of being cooperative interlocutors, provided that they know the *terms* of the game.

Acknowledgments

This work has been supported by grants from the Fondazione ONLUS Marica De Vincenzi. We are grateful to Istituto Comprensivo Statale G. Ciscato (Malo), Istituto Comprensivo Trento 6 (Trento) and Scuola Infanzia Sacra Famiglia (Peschiera del Garda) for their support, and particularly to Dott. Bruno Sandri, Dott.ssa Manuela Segata, Dott.ssa Paola Pasqualin and Dott.ssa Sabrina Pallavicini. We also thank Sophia Marlene Bonatti, Petra Rossato and Antonio Di Soccio for data collection.

Appendix 1 (Wellman & Liu, 2004, p. 538).

Diverse Desires

Children see a toy figure of an adult and a sheet of paper with a carrot and a cookie drawn on it. ‘‘Here’s Mr. Jones. It’s snack time, so, Mr. Jones wants a snack to eat. Here are two different snacks: a carrot and a cookie. Which snack would you like best? Would you like a carrot or a cookie best?’’ This is the owndesire question. If the child chooses the carrot: ‘‘Well, that’s a good choice, but Mr. Jones really likes cookies. He doesn’t like carrots. What he likes best are cookies.’’ (Or, if the child chooses the cookie, he or she is told Mr. Jones likes carrots.) Then the child is asked the target question: ‘‘So, now it’s time to eat. Mr. Jones can only choose one snack, just one. Which snack will Mr. Jones choose? A carrot or a cookie?’’ To be scored as correct, or to pass this task, the child must answer the target question opposite from his or her answer to the own-desire question. This task was derived from those used by Wellman and Woolley (1990) and Repacholi and Gopnik (1997).

Diverse Beliefs

Children see a toy figure of a girl and a sheet of paper with bushes and a garage drawn on it. ‘‘Here’s Linda. Linda wants to find her cat. Her cat might be hiding in the bushes or it might be hiding in the garage. Where do you think the cat is? In the bushes or in the garage?’’ This is the own-belief question. If the child chooses the bushes: ‘‘Well, that’s a good idea, but Linda thinks her cat is in the garage. She thinks her cat is in the garage.’’ (Or, if the child chooses the garage, he or she is told Linda thinks her cat is in the bushes.) Then the child is asked the target question: ‘‘So where will Linda look for her cat? In the bushes or in the garage?’’ To be correct the child must answer the target question opposite from his or her

answer to the ownbelief question. This task was derived from those used by Wellman and Bartsch (1989) and Wellman et al. (1996).

Knowledge Access

Children see a nondescript plastic box with a drawer containing a small plastic toy dog inside the closed drawer. ‘‘Here’s a drawer. What do you think is inside the drawer?’’ (The child can give any answer he or she likes or indicate that he or she does not know). Next, the drawer is opened and the child is shown the content of the drawer: ‘‘Let’s see y it’s really a dog inside!’’ Close the drawer: ‘‘Okay, what is in the drawer?’’ Then a toy figure of a girl is produced: ‘‘Polly has never ever seen inside this drawer. Now here comes Polly. So, does Polly know what is in the drawer? (the target question) ‘‘Did Polly see inside this drawer?’’ (the memory question). To be correct the child must answer the target question ‘‘no’’ and answer the memory control question ‘‘no.’’ This task was derived from those used by Pratt and Bryant (1990) and Pillow (1989), although it was modified so that the format was more parallel to the contents False-Belief task.

Contents False Belief

The child sees a clearly identifiable Band-Aid box with a plastic toy pig inside the closed Band-Aid box. ‘‘Here’s a Band-Aid box. What do you think is inside the Band-Aid box?’’ Next, the Band-Aid box is opened: ‘‘Let’s see y it’s really a pig inside!’’ The Band-Aid box is closed: ‘‘Okay, what is in the BandAid box?’’ Then a toy figure of a boy is produced: ‘‘Peter has never ever seen inside this Band-Aid box. Now here comes Peter. So, what does Peter think is in the box? Band-Aids or a pig? (the target question) ‘‘Did Peter see inside this box?’’ (the memory question). To be correct the child must answer the target question ‘‘Band-Aids’’ and answer the memory question ‘‘no.’’ This task was derived from one used

initially by Perner, Leekam, and Wimmer (1987) and widely modified and used since then (see Wellman et al., 2001).

PART 2 –

ATYPICAL CHILDHOOD:

THE COMPREHENSION OF CONVERSATIONAL IMPLICATURES

Chapter 2.

Scalar- and Ad-hoc-Implicature Processing in Children with Autism Spectrum Disorders

This chapter is based on the following original article:

Mazzaggio, G., Foppolo, F. Job, R., & Surian, L. (2018). Ad-hoc and Scalar Implicatures in Children with Autism Spectrum Disorder.

Manuscript under review in Journal of Autism and Developmental Disorders.

Abstract

Pragmatic abilities of people with Autism Spectrum Disorder (ASD) are generally considered impaired. Nonetheless, previous studies demonstrated a good competence of ASD people in some areas of pragmatics, like the derivation of the scalar implicatures related to the use of a linguistic scale, such as the scale of quantifiers that includes "some" and "all". Our study extends previous research to younger children with ASD (4-9 years old) and compares the derivation of scalar implicature to a different type of implicature, the so-called *ad-hoc* implicatures, which are based on a contextual – rather than a linguistic – scale. Although more than 50% of the children with ASD performed well on both kinds of implicatures, as a group their performance with both kinds of implicatures remains significantly lower than age-matched typically developed peers. By assessing the contribution of morphosyntactic skills, IQ and Theory of Mind (ToM) skills, we found that general cognitive abilities (IQ) seem to play a role in ASD children's pragmatic performance, while ToM skills seem to modulate their performance with scalar, but not ad-hoc, implicatures.

1. Introduction

When we speak, sometimes what we literally say is different from what we implicitly communicate. Indeed, by uttering a proposition in a particular context or by means of specific linguistic expressions, we might *implicate* something more than the simple combination of the words in that sentence. This module of language use is known as 'Pragmatics', and notoriously constitutes an area of weakness for Autism Spectrum Disorders (ASD). People within the spectrum are characterized by impairment in social interactions with, for example, failure of to- and-fro conversations with others (DSM-5: American Psychiatric Association, 2013). Researchers, indeed, often found difficulties in ASD in some areas of pragmatics, particularly with irony processing, turn taking, inappropriate topic change, difficulties with the use of an appropriate register, pronouns and deictic terms use, metaphors and humor comprehension (e.g., Baltaxe, 1977; Chin & Bernard-Optiz, 2000; Clifford & Dissanayake, 2008; Loukusa, Leinonen, Kuusikko, Jussila, Mattila, Ryder, Ebeling, & Moilanen, 2007; Mazzaggio, Panzeri, Giustolisi, & Surian, 2018; Naigles, Cheng, Rattanasone, Tek, Khetrapal, Fein, & Demuth, 2016; Ozonoff & Miller, 1996; Reddy, Williams, & Vaughan, 2002; Rundblad & Annaz, 2010; Surian, Baron-Cohen, & Van der Lely, 1996).

In this paper we will focus on a specific aspect of pragmatics, namely conversational implicatures. These are pragmatic inferences that arise when a speaker utters certain items, like *some*, that might be pragmatically enriched so as to mean *some but not all*. These will be introduced in section 1.1. Some previous researches tested these pragmatic inferences in ASD individuals; we will discuss these in section 1.2. Beyond previous findings, we extended our investigation to different kinds of implicatures; we also tested children at an age in which the ability to compute implicatures is known to be under development in typically developing populations; most crucially, we also correlated ASD children's performance in conversational inference with measures of general cognitive ability, as well as linguistic and theory of mind

skills. Our final aim was to understand how ASD children and TD children differ in the development of their ability to draw pragmatic inferences, and which factors modulate this ability. Our findings will contribute to a more thorough understanding of ASD individuals' difficulty in pragmatic tasks, and it will also contribute to the current theoretical debate about the nature of pragmatic inferences.

1.1. Conversational Implicatures

Implicatures arise because, according to the philosopher Paul Grice, people communicate in accordance to a general *Cooperative Principle*:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (Grice, 1975, p. 45)

More specifically, Grice outlined a series of maxims and sub maxims that must be respected to act cooperatively in the exchange, and that can be divided in four categories (Grice, 1975, pp. 45-46). According to the category of *Quantity*, our contribution should be as informative as it is required by the purposes of the exchange. According to the category of *Quality* we should try to make our contribution one that is true, and we should not say what we believe to be false or that for which we lack adequate evidence. We should also be relevant (*Relation*) and we should be perspicuous, that is we should avoid obscurity of expression or ambiguity and we should be brief and orderly (*Manner*).

From the listener's basic assumption that the speaker will obey these maxims, a certain kind of conversational implicatures may arise. Let us consider examples in (1).

1) a. My grandmother took *some* of her pills.

   b. My grandmother took *some but not all* of her pills.

   c. My grandmother took *some and possibly all* of her pills.

d. My grandmother took *all* of her pills.

In hearing (1a) the listener might derive the implicature in (1b); such kind of implicature is one example of what Grice (1975) called *generalized implicature,* also defined *scalar implicature* (Horn, 1972). According to Horn, scalar implicatures are generated thanks to the use of specific lexical items that are part of a scale in which they are ordered with respect of their informativeness, from the less informative to the most informative. The quantifiers *some* and *all* are an example of scalar quantifiers linked in such scales, that take the form of <some, all>, in which *all* is the most informative element and *some* the less informative one, provided that *all* logically entails *some* and thus *some* is true in larger set of circumstances, including all situations in which the more informative quantifier *all* applies. Accordingly, given the lexical scale <some, all>, when the speaker utters (1a) over the more informative (1d), by virtue of the maxim of Quantity stated above, s/he automatically implicate the negation of the stronger alternatives in the scale, and this is how the scalar implicature in (1b) arises. This interpretation, however, is not part of the lexical content of *some,* which is logically compatible with the lower-bound interpretation in (1c) where *some* is interpreted as *some and possibly all* (logical interpretation); it constitutes, instead, the upper-bound interpretation in (1b), where *some* is interpreted as *some but not all* (pragmatic interpretation). Being an implicature, the inference in (1b) is cancellable: one could say 'my grandmother took *some* of her pills, indeed she took *all* of them', without contradiction. In a Gricean view, scalar implicatures are derived in three steps: the first step requires the computation of the literal meaning of the sentence, like in (1c); the second step requires the access to the relevant alternatives, such as in (1d); the third step requires the negation of the relevant stronger alternative, to obtain the reading in (1b). Whether this third step requires considering the speaker's ability to infer other people's mental state, such as Theory of Mind (ToM) skills, it

is still under debate. In Gricean terms, in order to compute the third step, we should consider our interlocutor's willingness of being cooperative, thus we need to consider that if s/he did not utter the more informative statement it is because s/he does not know or believe that statement to be true, obeying the Maxim of Quantity. Other views (e.g., Chierchia, 2006 & 2013), even if not explicitly excluding the role of mentalizing reasoning, also consider the intervention of grammatical operators in the derivation of generalized implicatures, and in the activation of the alternatives.

Together with those generalized implicatures, Grice described also the so-called *particularized implicatures*, henceforth *ad-hoc implicatures*. We can define ad-hoc implicatures all those inferences that arise by uttering a particular proposition in a specific context. Let us consider a situation in which a person should find the correct t-shirt (e.g., a t-shirt with dots) in a chest with a t-shirt with dots and stripes and a t-shirt with only dots, after listening to the sentence in (2a). That person should not have particular problems in choosing the t-shirt with *only* dots, since s/he knows that there was an alternative instruction, (2b), that could have been uttered instead of (2a), but it was not: this led to the inference in (2c), which is an example of ad-hoc implicature.

(2) a. Bring me my t-shirt: it is the one with dots.

b. Bring me my t-shirt: it is the one with dots and stripes.

c. Bring me my t-shirt: it is the one with (only) dots.

Similarly to scalar implicatures, ad-hoc implicatures require some steps in order to be computed: in the first step, the literal meaning of the sentence is computed (2a); the second step requires the access to the relevant alternatives, such as in (2b); the third step requires the negation of the relevant stronger alternative, to obtain the reading in (2c). Despite the fact that the steps required are similar in the two kinds of implicature, a crucial difference exists at the

second step: if for scalar implicatures the relevant alternatives are linguistically accessed through the activation of the linguistic scale <some, all>, for ad-hoc implicatures the relevant alternatives are built *ad-hoc* through the context.

In the past years, the psycholinguistic literature mainly focused on scalar-implicatures computation in children and adults (for a review, Chemla & Singh, 2014a,b). It has been extensively demonstrated that children compute less scalar implicatures than adults, especially at age 4 and 5 (e.g., Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Chierchia, Guasti, Gualmini, Meroni, Crain, & Foppolo, 2004; Foppolo, Guasti, & Chierchia, 2012; Guasti, Chierchia, Crain, Foppolo, Gualmini, & Meroni, 2005; Huang & Snedeker, 2009; Noveck, 2001; Papafragou & Musolino, 2003) even if when manipulating features of the context or of the experimental design they can reach a pragmatic interpretation more easily (Foppolo, Guasti ,& Chierchia, 2012; Papafragou & Tantalou, 2004; Pouscoulous, Noveck, Politzer, & Bastide, 2007). Children difficulties have been ascribed to different factors, and the debate is still open. According to what we can define the 'tolerance account' (Katsos & Bishop, 2011), children do not generally lack the competence to compute scalar implicatures, but they are more tolerant regarding pragmatic violations compared to adults. Differently, the 'relevance account' (Skordos & Papafragou, 2016) predicts that accessibility of the stronger scalar term is crucial for children and that failure in pragmatic processing is due to a difficulty in recognizing what is relevant in the conversation. Finally, according to the 'lexicalist account' (Barner & Bachrach, 2010; Barner, Brooks, & Bale, 2011; Foppolo et al., 2012; Tieu, Romoli, Zhou, & Crain, 2015) children fail with scalar implicatures because they do not have access to scalar alternatives (yet); more specifically, children are not in the developmental stage in which they have lexicalized the scale and in which they can easily retrieve the scalar alternatives. Thus, considering the scale <some, all>, children might know

the meaning of *some* and the meaning of *all*, but they do not consider them as one the alternative of the other, with one being less informative and the other one more informative.

Interestingly, when we consider the computation of ad-hoc implicatures, children do not show the same difficulties than for scalar implicatures and they can compute them at four – and even three – years old (Stiller, Goodman, & Frank, 2015). Moreover, when children are tested with the same task for ad-hoc implicatures and scalar implicatures, they perform significantly better with the former (Foppolo, Mazzaggio, Panzeri, & Surian, 2018; Horowitz, Schneider, & Frank, 2017). All in all, these results suggest that different kinds of implicatures possibly require different mechanisms, and might rely on different abilities in order to be derived: for example, provided that the scale is lexically built for scalar implicatures, these kind of implicatures might rely more on linguistic abilities; on the contrary, particularized implicatures such as ad-hoc implicatures might rely more on general cognitive or ToM abilities in which the alternative statements are evaluated with respect to a shared conversational context.

Testing populations with ASD, specifically with High Functioning Autism (HFA), might be of particular interest to assess the role of intervening factors in implicature computation because people with such disorders are characterized by normal intelligence and good verbal abilities, but impairments in social relation and in ToM.

1.2. Scalar implicatures in autism

With respect to the derivation of conversational implicatures, a series of studies assessed the ability to compute scalar implicatures in populations with HFA. Results are not consistent; some researchers did not find a difference with typical population (Chevallier, Wilson, Happé, & Noveck, 2010; Hochstein, Bale, & Barner, 2017; Pijnacker, Hagoort, Buitelaar, Teunisse, & Geurts, 2009; Su & Su, 2015) while other recent studies did (Pastor-

Cerezuela, Tordera Yllescas, González-Sala, Montagut-Asunción, & Fernández-Andrés, 2018; Shaeken, Van Haeren, & Bambini, 2018).

In their seminal study, Pijnacker et al. (2009) tested 56 Dutch adults, 28 with ASD and 28 matched controls; the ASD group included 11 participants with High-Functioning Autism (HFA) and 17 with Asperger syndrome.[2] Participants had to judge sentences as true or false, and were exposed to underinformative sentences containing the scalar quantifier 'some' (e.g., *Some sparrows are birds*) and 'or' (e.g., *Zebras have black or white stripes*), that are logically true but pragmatically infelicitous if a scalar implicature is derived (namely, *Some but not all sparrows are birds* and *Zebras have black or white stripes but they do not have black and white stripes*). Due to their ToM deficits, the ASD group was expected to provide less pragmatic answers compared to matched typical adults. Unexpectedly, results showed a similar pattern of responses in the two groups. Overall, the ASD group's pragmatic answers were not significantly different from those observed in the control group, when considering the scalar terms. However, when considering separately HFA and Asperger participants within the ASD group, the former gave less pragmatic answers than the latter in the underinformative condition with *some* and – marginally – also in the underinformative condition with *or*. Moreover, researchers found a correlation between pragmatic answers and verbal intelligence in the HFA group; the higher the verbal intelligence, the more frequent the pragmatic answers. On one hand, these findings seem to support the view that linguistic abilities play a role in the computation of scalar implicatures; according to the authors, it might be that HFA participants leverage their verbal intelligence to compensate for pragmatic difficulties. On the other hand, the correlation between verbal intelligence and the computation of scalar implicatures was not found in the Asperger group and in the control

[2] This distinction in the autistic spectrum still existed in the previous version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR: American Psychiatric Association, 2000) but has been removed in the new DSM-5 (American Psychiatric Association, 2013).

group. According to these authors, the verbal-intelligence differences between the two ASD subgroups might be a limitation of this study. Furthermore, authors wrote that 'with regard to the dominant cognitive theories on autism' they 'think that it is more plausible to assume that deriving scalar implicatures is related to the theory of mind account' (p. 615), but ToM abilities had not been experimentally tested in their paradigm.

Similarly, Chevallier et al. (2010) tested 22 ASD adolescents (6 HFA and 16 Aspergers) and 22 controls on the pragmatic interpretation of underinformative connectives. The experimental paradigm was a Truth-Value Judgment Task in which participants had to decide whether they agreed or disagreed with sentences like "there is a sun or a train" when both a sun and a train were present in a picture. Like in Pijnacker et al. (2009), the rate of pragmatic interpretations in the critical condition did not diverge across the two groups and verbal intelligence correlated with pragmatic answers only in the ASD group. Researchers concluded that some semantic and pragmatic abilities are preserved in people in the HFA and Asperger spectrum.

More recently, Su and Su (2015) extended those results testing 28 Mandarin-speaking children with ASD with an age that ranged from 4 to 15 years, divided in two age-groups (mean age of the younger group: 6.6; mean age of the older group: 11.7), 15 with Asperger syndrome, 12 with autistic disorder, 1 diagnosed with pervasive developmental disorder). Children had been assessed on a Truth-Value Judgment Task on scalar implicatures associated with the scalar items 'some' and 'or'. In the underinformative conditions, children had to judge sentences like 'Some children found sea snails' in a scenario in which *all* children found sea snails, or sentences like 'Every child got a sea star or a shell' in a scenario in which *all* children had *both* a sea star *and* a shell. Authors divided participants in two groups based on mean age, one with younger participants ($M = 6.6$) and one with older ($M = 11.7$) participants. When looking at pragmatic responses for underinformative items, older

ASD participants answered similarly to their typically developing (TD) peers in the *some*-condition (ASD = 93%; TD = 96%) but not in the *every…or*-condition, in which they tended to reject the underinformative sentences significantly less than their peers (ASD = 59%; TD = 71%). The younger ASD participants answered as poorly as their TD peers in the *every…or*-condition (ASD = 36%; TD = 38%) and similarly to their peers in the *some*-condition (ASD = 64%; TD = 79%). Authors concluded that, in general, ASD participants did not seem to differ significantly from TD participants of the same age. Su & Su's study is of particular interest because they assessed children in the critical age for the computation of scalar implicatures, provided that they also included children as young as 4 in their group. On the other hand, the study has some limits: ToM had not been tested and the age range of the participants is too wide to perform comparisons across age groups. Moreover, they included children with Asperger, ASD and pervasive developmental disorder in the clinical group, and for this reason it is difficult to generalize this result. Indeed, the authors recognized that a limit of their study is the fact that they lacked the gold standards for ASD diagnosis.

A recent study by Hochstein et al. (2017) criticized the use of the Truth-Value Judgment Task since it is generally agreed that an underinformative statement is neither 'true' nor 'false'; rather, participants' answers can be either 'logic' (i.e. accepting the statement) or 'pragmatic' (i.e. rejecting it). They suggested that in all previous tasks it is not clear whether participants answered regarding the truthfulness or the felicity of the sentences they had to judge. For this reason, these authors decided to test adolescents with ASD (N = 18) with the use of interactive tasks on the role of epistemic reasoning in scalar implicatures and ignorance implicatures that require inferring speaker's ignorance. In these tasks, two puppets made a comment about something that happened, and the child had to detect which of the two puppets talked. Crucially, in the task for ignorance implicatures, one of the two characters was blindfolded, thus commented on the scenario without really knowing what happened. In

the task for scalar implicatures, instead, one of the two characters was introduced as smart and the other as silly. To further manipulate the epistemic state of the speaker, a third task was introduced, in which a puppet uttered a statement about the content of a box, sometimes after he peeped inside the box and sometimes after he didn't. They found that ASD participants inferred scalar implicatures both for fully knowledgeable and for partially knowledgeable speakers, suggesting that they computed implicatures without considering the interlocutors' epistemic states: even if they inferred the ignorance of the speakers they were not able to spontaneously work with such awareness when processing scalar implicatures.

Summing up, previous researches showed that adolescents and adults with ASD might have spared abilities to compute scalar implicatures, even if these results are not entirely clear due to the heterogeneity of the groups with respect to age and diagnosis, and the tasks used. All reported literature also shows that computing scalar implicatures has a strong connection with verbal abilities, with the caveat that all the tasks employed in those studies relied a lot on verbal abilities. As for the interpretation of these results, on the other hand, these have been explained in quite different ways. On the one hand Chevallier et al., 2010 proposed that HFA have spared pragmatic abilities; on the other, Pijnacker et al., 2009 proposed that scalar implicatures are related with easier, first-order ToM abilities that are typically acquired by adults with ASD and not with the most complex second-order ToM abilities. Finally, other authors suggested that ASD participants might compute scalar implicatures using different processing strategies that are not related with epistemic reasoning about others' mental states (Hochstein et al., 2017).

2. The present study

In our study we assessed the derivation of conversational inferences in ASD by controlling for different factors. First of all, we tested children at an interesting stage for

implicature computation, as attested from the acquisition literature on typically developing children. Second, we controlled for linguistic, cognitive and ToM abilities by assessing children with a battery of standardized tests on such skills. Finally, we aimed at extending previous results by comparing two kinds of implicatures with the same task: one that is linked to the use of a scalar term and thus requires the lexicalization of a semantic scale (scalar implicatures), and one that arises by contextual features of the context (ad-hoc implicatures). We decided to use a Picture Selection Task for both scalar and ad-hoc implicatures, provided that this task relies less on metalinguistic abilities, and no verbal answer is required.

Our main aim was threefold: first, we aimed at assessing ASD children's pragmatic performance on different types of implicatures, namely, not only scalar implicatures but also ad-hoc implicatures to extend previous findings. Second, we aimed at investigating how this performance is related to general intelligence, linguistic skills, and ToM skills, so as to shed light on which linguistic and cognitive factors might play a role in the process of computing implicatures in ASD children. Particularly, ToM is taken into consideration; we think that it is important to consider it, since our participants are in an age in which ToM skills develop, as well as the ability to compute implicatures, as we know from previous research on TD children. Third, we aimed at contributing to the theoretical debate on the mechanisms underlying conversational inferences; despite the fact that the main focus of our study is testing pragmatic abilities in on ASD children, our data might also contribute for the general theoretical debate on implicature processing.

2.1. Methods

2.1.1. Participants

Twenty-six TD children and 26 ASD children between 4 and 9 years participated to this study. One child in the ASD group had been previously removed after a re-diagnosis of

Social (Pragmatic) Communication Disorder. The two groups were matched for age and were tested with standardized tests for linguistic abilities (lexicon and syntax), non-verbal IQ and ToM (all tasks will be detailed in the next section). Statistical comparisons between the two groups revealed no difference in age, IQ, and lexicon abilities. The ASD group syntactic and ToM abilities were significantly lower than the control group (Table 1). ASD participants had been recruited and diagnosed at the Hospital 'Azienda Provinciale per i Servizi Sanitari' (APSS) in Trento, Italy; in this structure, child psychiatrists used standard clinical criteria as a diagnostic tool for their evaluation (i.e., ADOS: Autism Diagnostic Observation Schedule). ADOS scores for each individual child are reported in details in the Appendix A. Trained specialists at the APSS regularly follow all the children. TD children had been recruited in kindergartens and primary schools from two provinces in the northeast of Italy: Trento and Vicenza. According to Ethical Committee's guidelines, parents were informed about the experiments and they gave written consent to their children being part of our study. This study is part of a more extensive research on pragmatic deficits in ASD population that has been approved by University of Trento's Ethical Committee.

Table 1. Participant's age, IQ, lexicon abilities, syntax abilities and Theory of Mind (ToM) skills.

	TD children $N = 26$		ASD children $N = 26$		*t* test
	Mean (*SD*)	Range	Mean (*SD*)	Range	
Age	84.88 (21.70)	50-111	87.08 (24.28)	45-123	$t = .343$; $p = 0.73$
IQ	27.23 (6.71)	14-36	24.50 (6.56)	11-36	$t = -1.48$; $p = 0.14$
Lexicon	28.65 (11.26)	10-41	25.42 (10.10)	9-41	$t = -1.09$; $p = 0.28$

Syntax	32.81 (6.22)	22-40	28.81 (5.96)	19-39	$t = -2.37$; $p = 0.02$
ToM	3.50 (.648)	2-4	2.00 (1.33)	0-4	$t = -5.18$; $p < 0.001$
ADOS			12.96 (4.19)	8-22	

2.1.2. Materials and Procedure

We tested children with two similar pragmatic tasks, one for scalar implicatures and one for ad-hoc implicatures. In addition, participants have been tested for their non verbal-IQ by means of the Raven Coloured Progressive Matrices (Italian standardization by Belacchi, Scalisi, Cannoni & Cornoldi, 2008); for their lexical and morphosyntactic abilities by means of two tasks taken from the Batteria di Valutazione del Linguaggio 4-12: a test for receptive vocabulary, that includes 18 items for pre-schoolers and 42 items for primary school kids (Marini, Marotta, Bulgheroni & Fabbro, 2015); a test of grammatical competence that includes 40 items varying in level of morphosyntactic difficulty. Moreover, we assessed first order ToM skills with a battery of tests adapted from the first four tasks of Wellman & Liu (2004): *Diverse Desires*, i.e. the ability to judge that people might have different desires; *Diverse Beliefs*, i.e. the ability to judge that people might have different beliefs; *Knowledge Access*, i.e. the ability to judge that people might have different knowledge; and *Content False Belief*; i.e. the ability to understand that people can have a false belief.

Children have been tested alone in a quiet room; we decided to split the experiment session in two or three days, depending on the individual child's level of attention, to avoid possible tiredness and distraction.

2.2.1 Picture Selection Task for scalar implicatures

Moving from Surian & Job (1987) and Stiller, Goodman & Frank (2015), we created a *Picture Selection Task* to assess children's abilities in computing pragmatic inferences. This

task was the same used by Foppolo et al. (2018) with TD children in a different study. We considered this task as more ASD oriented compared to the classical *Truth Value Judgement Task* because children were visually provided with all the alternatives and they did not need to answer verbally to the experimenter. Our task required participant to pinpoint the requested target (among four pictures) after listening to a sentential prompt. Prior to the test phase, we introduced the main character (a boy called Daniele) to children, explaining that he wanted to play a guessing game with them: following his clues, children had to find the object that Daniele asked for on the screen. All Daniele's sentences had been previously recorded to control for prosody.

In the scalar-implicature task, children were then presented with a warm-up trial and in this phase the experimenter could correct them in case of inaccurate answers. After this trial, children began the test phase that consisted of two sentences with the quantifier *all*, four sentences with the quantifier *some* and one control trial (with no quantifier). We decided to use a pseudo-randomized order in which the first *all*-item had always been presented before the first *some*-item; in this way, we were confident that children were provided with the most informative element in the scale before listening to the less informative one (Foppolo et al., 2012). In *Figure 1* there is an example of a *some*-scenario. The items unfolded as follows. First of all, Daniele appeared on the screen saying, for example, "Guess which my cake is, I give you a clue"; then children were presented with a scenario in which there were four cakes and listened to Daniele's clue: "On my birthday cake, some of the candles are burning". By exploiting this linguistic clue, children had to spot Daniele's birthday cake among a cake with no candles (distractor), a cake with no burning candles (distractor), a cake with some (3 out of 5) burning candles (pragmatic target) and a cake with all burning candles (underinformative-target). If children interpreted the quantifier *some* pragmatically as *some but not all,* then we expected them to select the cake with *only some* of the candles burning; alternatively, if

children interpreted *some* logically as *some and possibly all* we expected them to select the cake with all the candles burning. If children selected one of the two distractors, then they showed a poor understanding of the sentence or the task. In *Figure 2* there is an example of an *all*-scenario. In this *all*-scenario, Daniele's sentence was "Guess which my favorite playground is, I give you a clue: at my favorite playground, all the flowers are red": the target picture was the one with 5 red flowers out of 5; the other pictures displayed a playground without flowers (distractor), a garden with 3 red and 2 blue flowers and a garden with 3 blue and 2 red flowers (subset competitors). The expected reasoning that the child should follow to select the target cake in the *some* example was the following: if Daniele's cake was the one with all the candles lit, then he should have clearly used the most informative quantifier *all*. Since Daniele used the less informative quantifier *some* instead, the intended target had to be the cake with *not all* the candles burning.

What is interesting about this particular experimental design is that participants are provided (both linguistically and visually) with the more informative scalar alternative *all* and, in this way, access to the relevant alternative and its contextual relevance should increase. Moreover, the epistemic state of the speaker can be easily inferred: in the way the game is posed, it is clear that (i) Daniele has full knowledge of the facts and (ii) the target is uniquely identifiable.

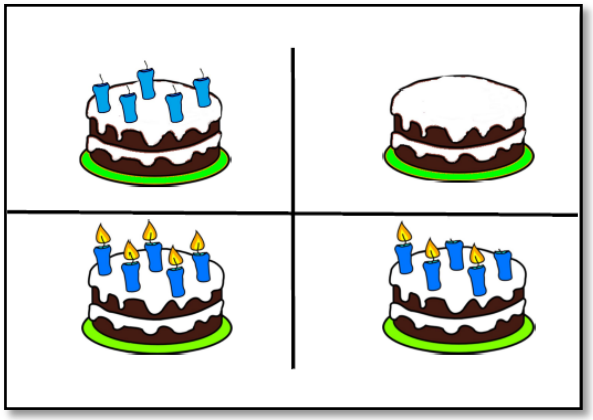


Figure 1. Example of a *some* underinformative item in the Picture Selection Task for scalar implicatures: *Guess which my cake is, I give you a clue: on my birthday cake, some of the candles are burning*.

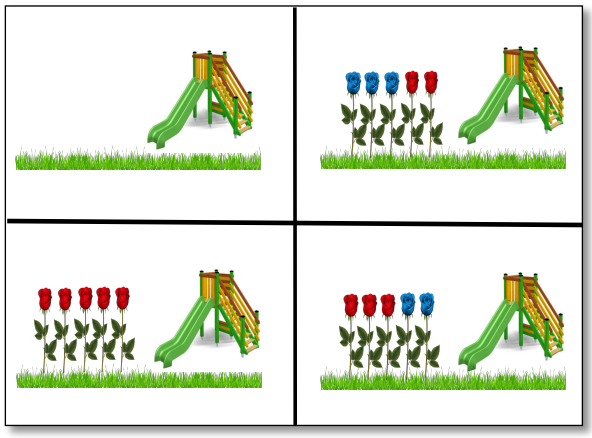


Figure 2. Example of an *all* true item in the Picture Selection Task for scalar implicatures: *Guess which my favorite playground is, I give you a clue: at my favorite playground, all the flowers are red.*

2.2.2 Picture Selection Task for ad-hoc implicatures

The Picture Selection Task for ad-hoc implicatures was modeled after Stiller et al. (2015). The procedure was the same of the task for scalar implicatures, since the goal was to compare the two kinds of implicatures with the same experimental design. For this reason, the main character remained the same, Daniele, and even in this task children had to play the spotting game finding a target, among four pictures, after listening to a prompt. The experiment consisted of two control trials and four test trials for ad-hoc implicatures. In *Figure 3* there is an example of a test item: first, Daniele appeared saying, for example, "Guess which my bed is, I give you a clue"; as in the task for scalar implicatures, four different beds appeared on the screen and Daniele added: "On my bed there is a teddy bear". As before, children had to select the target among two distractors (a bed with a penguin and

an empty bed), the pragmatic target (the bed with only the teddy bear) and the underinformative target (the bed with both a teddy bear and a penguin). The expected reasoning that the child should follow to select the target bed was the following: if Daniele's bed was the one with both the teddy bear and the penguin, then he should have clearly said this. Since Daniele mentioned only the teddy bear, the correct target had to be the bed with only the teddy bear.

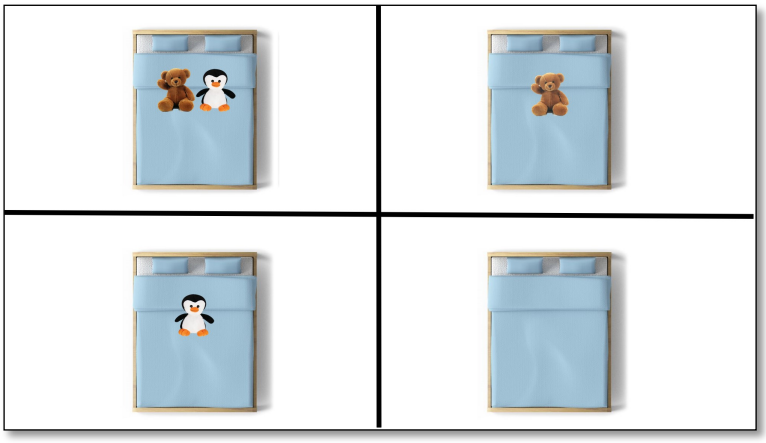


Figure 3. Example of an underinformative item in the Picture Selection Task for ad-hoc implicatures: *Guess which my bed is, I give you a clue: on my bed there is a teddy bear*.

3. Results

The responses were coded as "correct" if the child selected the pragmatic target in the test conditions, and the correct target in controls. Children's accuracy on controls was above 95% in both groups (95.4 % in the ASD group and 97.7% in the TD group), while the rate of implicature derivation was lower than controls in both groups, in the ASD group especially (65.4 % in the ASD group and 84.6% in the TD group).

We run a logit mixed effects model to predict accuracy (as a binomial variable) considering subjects and items as random factors, and Group (ASD vs. TD), Task (scalar implicatures vs. ad-hoc implicatures), Condition (implicature vs. controls) and Age in months

as fixed effects. For a better fit of the model, Age was centered prior to analysis. All mixed effects models were fit in R using the lmerTest package. A significant effect of Condition (Est = -3.734, *SE* = .956, *z* = -3.900, *p* < .001), Group (Est = 2.001, *SE* = .564, *z* = 3.549, *p* < .001) and Age (Est = .065, *SE* = .013, *z* = 5.122, *p* < .001) emerged: accuracy of children on implicature items is significantly lower than on control items, performance is worse for the ASD group compared to the TD group and it is worse for the younger kids. No significant interactions were observed, so these were removed from the model.

Considering children's performance with scalar and ad-hoc implicatures separately, we observe that: children's accuracy on ad-hoc implicatures is higher than scalar implicatures for both groups, the difference being greater in the TD group (70.2% vs. 60.6% in the ASD group; 91.3% vs. 77.9% in the TD group); TD children performed better than ASD in both kind of implicatures (Figure 4).

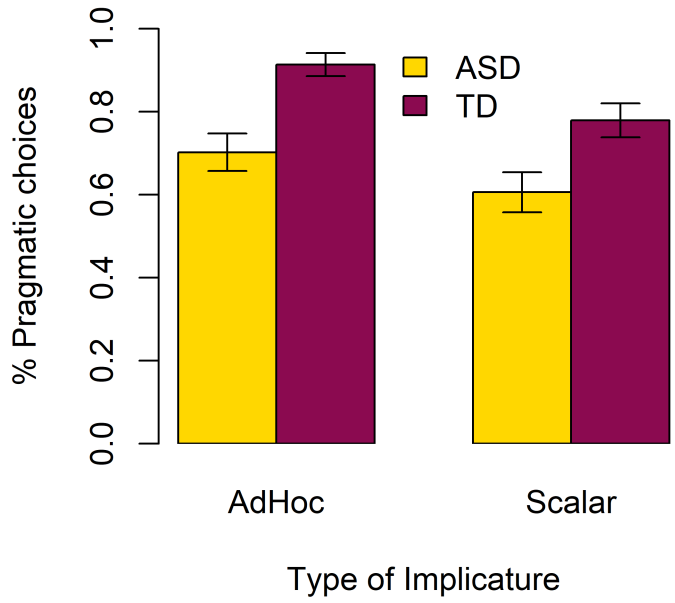


Figure 4. Mean pragmatic choices made for Ad-Hoc- and Scalar-implicature items in the two groups of ASD and TD children (bars represent Standard Error of the mean).

We further inspect children's distribution in each group and counted the number of times children in each group selected the pragmatic target in each implicature task (Table 2). This number ranged from 0 to 4 for each task provided that each participant was tested with 4 scalar and 4 ad hoc critical items The distribution of pragmatic choices (i.e. the number of times children selected the pragmatic target, that ranged from 0 to 4 times, Table 2)was compared between groups and resulted to be significant for both types of implicatures (scalar implicatures: $W = 312$, $p < .001$; ad hoc implicatures: $W = 416$, $p < .001$).

Table 2. Distribution of pragmatic choices (Scalar vs. Ad Hoc) and Group (ASD vs. TD)

	Scalar Task					Ad Hoc Task				
Group	0	1	2	3	4	0	1	2	3	4
TD	3	1	1	6	15	0	0	3	3	20
ASD	4	6	0	7	9	2	4	3	5	12

Despite the fact that the ASD kids perform worse than the TD kids in the implicature condition, it is noteworthy that 16 (62%) children in the scalar implicature task and 17 (65%) in the ad-hoc task selected the pragmatic target most of the times (at least 3 out of 4 times).

By splitting children on the basis of their schooling, we further observe that the gap between ASD and TD children increases with age: while TD kids in primary school (N=17) are at ceiling with all conditions, the ASD kids in primary school (N = 18) keep showing a difficulty with the computation of both pragmatic inferences. Moreover, while TD children in

the kindergarten (N = 9) perform better with ad-hoc implicatures than with scalar implicatures (paired Wilcoxon-test: $W = 864$, $p =. 004$), ASD younger children (N = 8) perform poorly with both implicatures, independently of type ($W = 544$, $p = .614$).

Given that controls were (almost) at ceiling in both groups, independently from Age or Type of implicature, we performed a separate analysis focusing on implicature items only. As before, we run a logit mixed effects model to predict accuracy (as a binomial variable) considering subjects and items as random factors, and Group (ASD vs. TD), Condition (scalar implicatures vs. ad-hoc implicatures), and Age in months as fixed effects, as well as their interactions. The model that best fits the data reveals a significant effect of Group (Est = 2.900, $SE = .772$, $z = 3.759$, $p < .001$) and Age (Est = .066, $SE = .013$, $z = 4.912$, $p < .001$), showing that accuracy on implicatures is higher in the TD group than in the ASD group, and it increases with age. A marginally significant interaction between Type of implicature and Group was also found (Est = -1.270, $SE = .695$, $z = -1.827$, $p = .07$).

To understand the observed interaction, we conducted further analyses by means of a series of Wilcoxon non parametric tests: accuracy on ad-hoc implicatures was found to be higher than accuracy on scalar implicatures in the TD group only ($T = 178.5$, $p = .002$), but not in the ASD group, for which the difference between ad-hoc implicatures and scalar implicatures is only marginally significant ($T = 425.5$, $p = .097$). We also found a significant difference between ASD and TD children in the rate of derivation of ad-hoc implicatures ($W = 4264$, $p < .001$) and scalar implicatures ($W = 4472$, $p = .007$), but not in the rate of accuracy in controls ($W = 8320$, $p = .521$), that remains high for both groups, as we said previously. Accuracy on control is particularly relevant when assessing children's competence, because it reveals that participants are understanding and attending to the task.

To further understand the factors underlying ASD performance with implicatures, we considered the results obtained by each child in the standardized tests for linguistic and

cognitive abilities, ToM tasks, as well as their ADOS score. These are detailed for each individual participant in the appendix (Appendix A-B). The distribution of each group in each task is also plotted in Appendix C (see also Table 1 for mean scores and pairwise comparisons between groups).

Considering ToM tasks, ASD children showed difficulties compared to TD children, especially with the more complex tasks. Table 3 summarizes the mean accuracy (and standard deviation) for each sub-task of 1st order ToM for each group.

Table 3. Participant's mean accuracy (and standard deviation) for each sub-task of 1st order ToM for each group.

	TD children *N* = 26	ASD children *N* = 26	Fisher's Exact Test
	Mean (SD)	Mean (SD)	
ToM Diverse Desires	.92 (.27)	.81 (.40)	$p = .419$
ToM Diverse Beliefs	.96 (.20)	.54 (.51)	$p = .001$
ToM Knowledge Access	.96 (.20)	.42 (.50)	$p < 0.001$
ToM Content False Belief	.65 (.48)	.23 (.43)	$p = .005$

To test the contribution of these factor on implicature computation, we ran a mixed model analysis on ASD children's accuracy with implicature, testing the different scores obtained in the linguistic, cognitive, and ToM tasks and their ADOS score as predictors, with random intercepts for subjects and items. Following a backward elimination procedure, non

significant factors and interactions were removed from the final model. The model that best fits the data reveals a significant effect of Raven (Est = .170, *SE* = .061, *z* = 2.805, *p* = .005), a significant interaction of type of implicature and ToM (Est = .984, *SE* = .369, *z* = 2.667, *p* = .007) and a marginally significant effect of Lexicon (Est = -3.731, *SE* = 2.151, *z* = -1.735, *p* = .08). Children with higher IQ computed more implicatures. Furthermore, ToM seems to play a role just for scalar implicatures and not for ad-hoc implicatures: the higher the ToM skills, the more scalar implicatures were computed, while the correlation between ToM scores and pragmatic answers in the ad-hoc task seems weaker (Figure 5).

A parallel analysis was conducted on TD children. In this case, the model reveals a significant effect of type of implicature (Est = -2.56, *SE* = 1.256, *z* = -2.043, *p* = .04) and Syntax (Est = .389, *SE* = .091, *z* = 4.273, *p* < .001), but no other significant effects nor interactions.

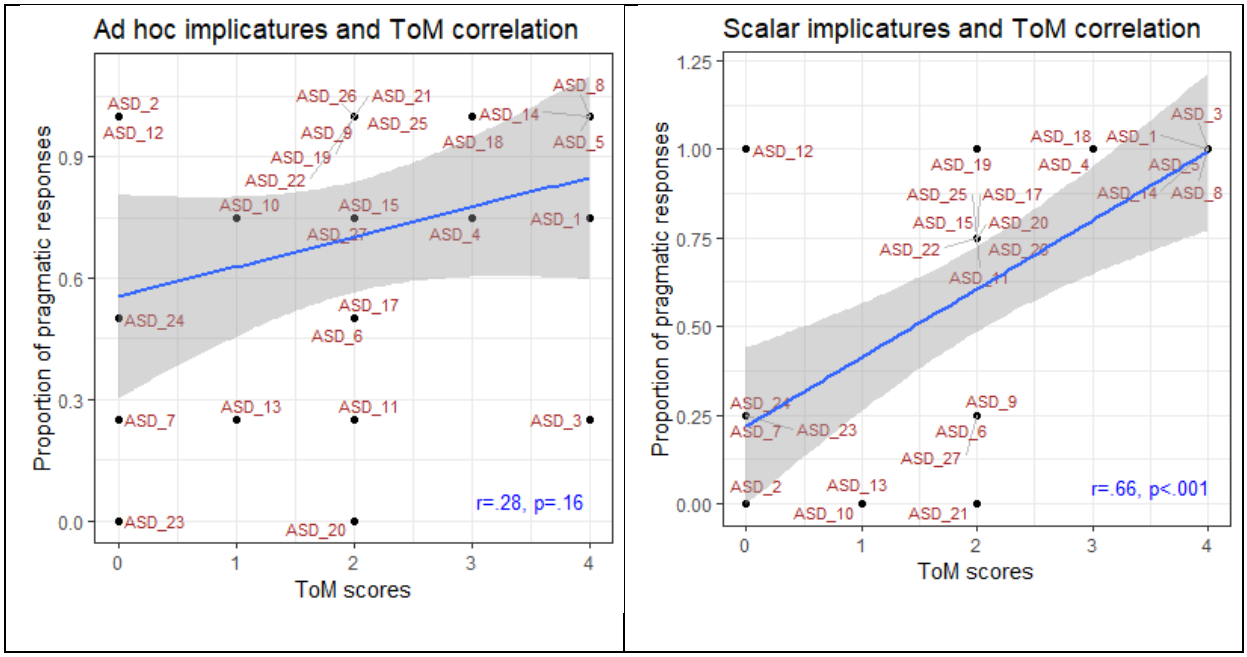


Figure 5. Correlation between ToM scores (from 0 to 4) and proportion of pragmatic answers in the ad-hoc task (left panel) and the scalar task (right panel) for all 26 individuals in the ASD group.

4. Discussion

In our study, we tested ASD children's ability to derive two types of conversational implicatures, *scalar implicatures* - which are based on linguistically encoded scales- and *ad-hoc implicatures*, which are based on contextually retrieved scales. Both kinds of implicatures arise from the cooperative exchange among speakers, that are prompted to be maximally informative and cooperative in their communications, and recipients of a message, that ascribe their interlocutors this intent and draw inferences on the basis of such assumption. In our study, we tested the pragmatic abilities of ASD children aged between 4 and 10 years and compared them with a group of TD peers matched by age. We know from the acquisition literature on scalar implicatures that preschool TD children may find difficulties in the derivation of these inferences, and that they develop this ability during school years. Nonetheless, most previous studies on scalar-implicatures computation in ASD tested adolescents or adult individuals with ASD (Chevallier et al., 2010; Hochstein et al., 2017; Pijnacker et al., 2009). The high performance obtained in those groups might simply hide a developmental delay in the acquisition of this ability. Moreover, none of the groups of ASD tested in previous studies were homogeneous and always included a subgroup of individuals with Asperger syndrome: it is not clear how this inclusion contributed to the results obtained in those studies.

From the results obtained in our study, we can draw two main conclusions. First, ASD children show some pragmatic competence: in our study, about 65% of the ASD children tested could compute implicatures and their performance increases with age in both types of implicatures, in line with previous conclusions that found a surprising ability in ASD adults in deriving scalar implicatures. However, by comparing the group of ASD children and TD children, we observed that ASD kids have more difficulties than their TD peers in deriving

pragmatic inferences of both of the kinds tested. Such difficulties are not due to general problems with the task, given that they do not differ from their peers in the control conditions. Also, the gap between ASD and TD children persist during primary school: while TD kids from the age of 6 perform at ceiling in all conditions, even the older ASD kids in our group keep showing a difficulty with the computation of pragmatic inferences, though the evidence suggests a developmental trend and an above chance performance.

Our results on scalar implicatures are somewhat at odd with findings on children with ASD reported by Su & Su (2015). Both the age of the tested population and the type of task used in the two studies may account for the inconsistency in the results. With respect of the participants' age, Su and Su tested participants of a wider and older range of age (4 to 15 years) than the present study (4 to 9 years) and this might have reduced the chance to look at the developmental stage in which scalar implicatures led to computational difficulties also in TD children. Both looking at our developmental data and at other studies on ASD participants of older age (Chevallier et al., 2010; Hochstein et al., 2017; Pijnacker et al., 2009), a developmental growth in the ability to compute scalar implicatures is quite visible, possibly due or linked to increased linguistic and cognitive skills. Results reported by Whyte and Nelson (2015) go in this direction. They considered pragmatic development in ASD and, by mean of a cross-sectional trajectory analysis, they showed that ASD children display slower development of inferential abilities.

With respect to the task, on the other hand, Su and Su (2015) assessed their children through a *Truth Value Judgement Task* on underinformative sentences that are not 'false' *per se* but are simply 'not felicitous'. Furthermore, such task is not designed to consider the context and the role of the speaker's mind and answers might be given even without considering it. Conversely, our *Picture Selection Task* has been designed to enhance the contextual salience of alternative descriptions without enhancing little task demands. Indeed,

in a recent study, Schaeken,Van Haeren and Bambini (2018) showed that ASD children tested on implicatures with a binary judgment tasks (i.e., agree/disagree) can derive them as much as TD controls. However, when tested with a ternary option (i.e., totally agree/agree a bit/totally disagree) ASD children showed a dichotomized pattern of answers, by opting to either 'fully agree' or 'fully disagree' with underinformative statements, differently from TD children, who opted for the middle option. The authors suggested that ASD children presented difficulties with implicatures, but such difficulties might have been evened out with the use of a standard binary task. They pointed out the importance of the task when assessing pragmatic abilities, especially with impaired population. Similarly, when Pastor-Cerezuela, Tordera Yllescas, González-Sala, Montagut-Asunción and Fernández-Andrés, (2018) tested children with a (different) ternary option task, they found difficulties in the comprehension of implicatures.

The second goal of the present study was to investigate the links between performance on the implicatures task and measures of intelligence, language development and social cognition. Previous works on ASD children showed a correlation between linguistic abilities and the rate of derivation of scalar implicatures. However, none of the previous studies on ASD children analyzed correlations between scalar-implicatures computation and other cognitive factors, such as ToM. After examining the influence of those skills in the computation of implicatures in the ASD group, evidence shows that – surprisingly – ToM only predicts the performance for scalar implicatures and that general intelligence (IQ) is a good predictor of the performance for both scalar implicatures and ad-hoc implicatures. Considering other studies, the role of IQ is not entirely surprising (see also Schaeken et al., 2018); what is puzzling, instead, is the role of ToM in implicature computation, although this is in line with recent findings by Foppolo et al. (2018) on TD children. Following Neo-Gricean pragmatic accounts, we should have expected a role of ToM in both kinds of

implicatures, whereas following lexicalist/grammatical accounts we should have expected a modulation of scalar implicatures by linguistic abilities. Indeed, by modelling accuracy on different predictors, it seems that ad-hoc implicatures and scalar implicatures rely on different mechanisms, while data on accuracy shows that their development seems to follow a common path in ASD children, that is, interestingly, clearly different from that of TD children.

A tentative explanation of this picture might be offered by considering the role of IQ in the ASD group on the one hand, and the developmental trajectory of ToM on the other. This relation between pragmatic ability and general intelligence also goes in the direction of studies that explored the idea that people in the autistic spectrum might compensate for their deficits using specific strategies. Indeed, also for ToM skills it has been demonstrated that ASD population might solve tasks that require this ability through compensatory learning (Senju, Southgate, White, & Frith, 2009; Schneider, Slaughter, Bayliss, & Dux, 2013). Similarly, it has been shown that in pragmatic tasks ASD population might use different processing strategies compared to typical population (Pexman, Rostad, McMorris, Climie, Stowkowy, & Glenwright, 2011). Moreover, a functional MRI study on irony processing in ASD population (Wang, Lee, Sigman, & Dapretto, 2006) showed a different pattern of brain activation (in the right inferior frontal gyrus and in the bilateral temporal regions) in ASD compared to TD: authors concluded that the greater activation may be related with more effortful processing. Moving from these results, we might speculate that ASD participants learn with age and with increased IQ skills how to deal with pragmatic tasks, even in the absence of a full mastery of ToM skills. Building on this conjecture, we can combine the evidence that scalar implicatures are more difficult for TD children with the significant role of ToM in scalar-implicature computation in ASD children: a full mastery of ToM is presumably neither necessary nor sufficient to derive scalar implicatures; for this kind of implicature, an additional step should be achieved, either by exploiting linguistic competence, as shown by

the TD group, or compensated by other mechanisms, as shown by the role of IQ in the ASD group.

Our third goal was to contribute to the understanding of the different mechanism behind ad-hoc implicatures and scalar implicatures. As we outlined in the introduction, ad-hoc implicatures and scalar implicatures might rely on different mechanisms for their derivation. In particular, while the alternative for ad-hoc scales are purely retrieved on a contextual basis, the alternatives for scalar implicatures need to be retrieved from the lexicon. According to some theories of scalar implicatures, this special kind of inferences might be grammaticalized, and a delay in their development might be linked to a delay in the lexicalization of the scale and/or in some difficulty in scale retrieval. As we said, the theoretical debate is still open with respect to the nature and differences of pragmatic inferences. Although our main aim in this paper is not to evaluate competing pragmatic theories, we think that the data from ASD children might inform the theoretical debate as to the nature of different implicatures.

With respect to this point, previous works demonstrated that TD children are more competent with ad-hoc implicatures than with scalar implicatures (Foppolo et al., 2018; Horowitz et al., 2017; Stiller et al., 2015) and we replicated the same finding in our group of TD children. In addition to this, for the first time we compared ad-hoc implicatures and scalar implicatures in ASD children and, interestingly, we observed a different pattern: specifically, for ASD children, ad-hoc implicatures are found to be as difficult as scalar implicatures. These results might be evidence that, despite their differences, some underlying general mechanism is shared in the derivation of these inferences, one that is not spared in ASD population. The best candidate for this role is indeed ToM, which has been found to be one parameter that significantly differentiated the two populations tested, one which is specifically impaired in ASD children and that we found was significantly associated with the performance on scalar implicatures.

In conclusion, it seems that ASD children can compute implicatures even if they display some difficulties compared to TD peers. Such difficulties slowly decrease with age, probably due to the acquisition of more advanced linguistic and cognitive skills, as well as to compensatory learning. Future longitudinal studies might shed light on factors affecting this developmental path that seems, however, different and slower compared to the one of TD children. Future studies should also test an ASD population with lower IQ and verbal skills to better understand the role of the various factors involved in pragmatic processing of people in the autistic spectrum. This will be helpful both to provide a better understanding of the mechanisms that rule cooperative exchanges and to highlight effective strategies of intervention to improve communication.

Acknowledgments

This work has been supported by grants from the Fondazione ONLUS Marica De Vincenzi. We are grateful to the support from the clinicians of the Azienda Provinciale per i Servizi Sanitari Provincia Autonoma di Trento, and particularly to Dott. Stefano Calzolari. We also thank the children and families who took part in our research and Dott.ssa Giulia Guglielmetti for data collection.

Appendix A. Participant's ADOS Scores – ADOS versions and modules are specified. ASD_16 is missing because of a re-diagnosis of Social (Pragmatic) Communication Disorder.

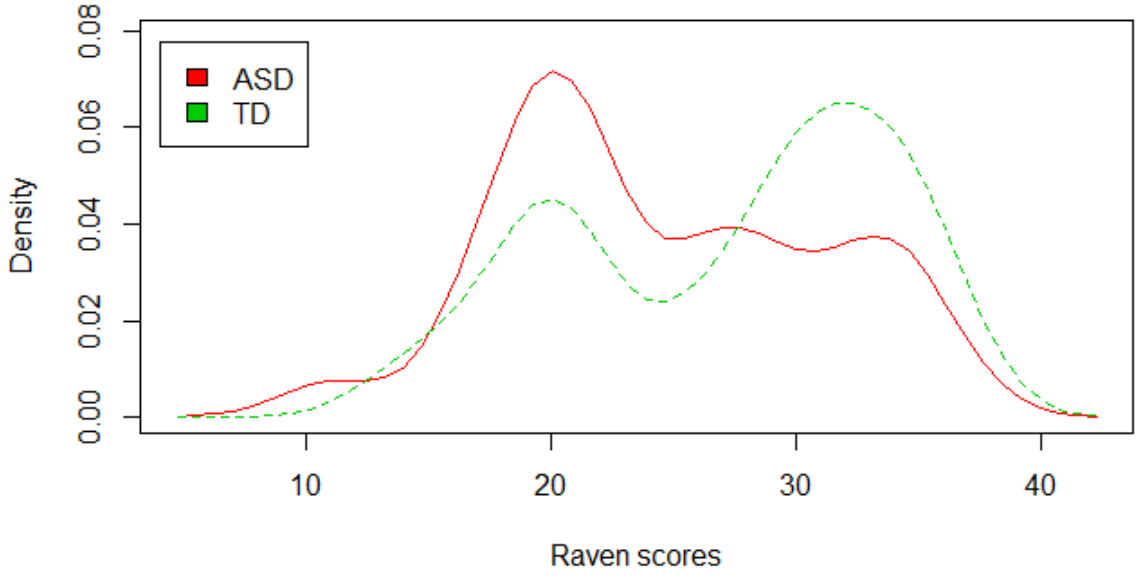
ID	ADOS edition	Module	Score
ASD_1	2	3	14
ASD_2	2	1	18
ASD_3	2	1	22
ASD_4	2	2	10
ASD_5	2	3	10
ASD_6	1	2	15
ASD_7	2	1	14
ASD_8	1	2	19
ASD_9	2	3	9
ASD_10	1	1	17
ASD_11	2	1	16
ASD_12	1	1	14
ASD_13	2	1	12
ASD_14	2	2	8
ASD_15	2	1	16
ASD_17	2	3	8
ASD_18	2	2	10
ASD_19	1	2	11
ASD_20	2	2	8
ASD_21	2	3	13
ASD_22	1	2	13
ASD-23	2	1	9
ASD-24	2	1	11
ASD-25	2	2	9
ASD-26	2	2	9
ASD-27	2	1	22

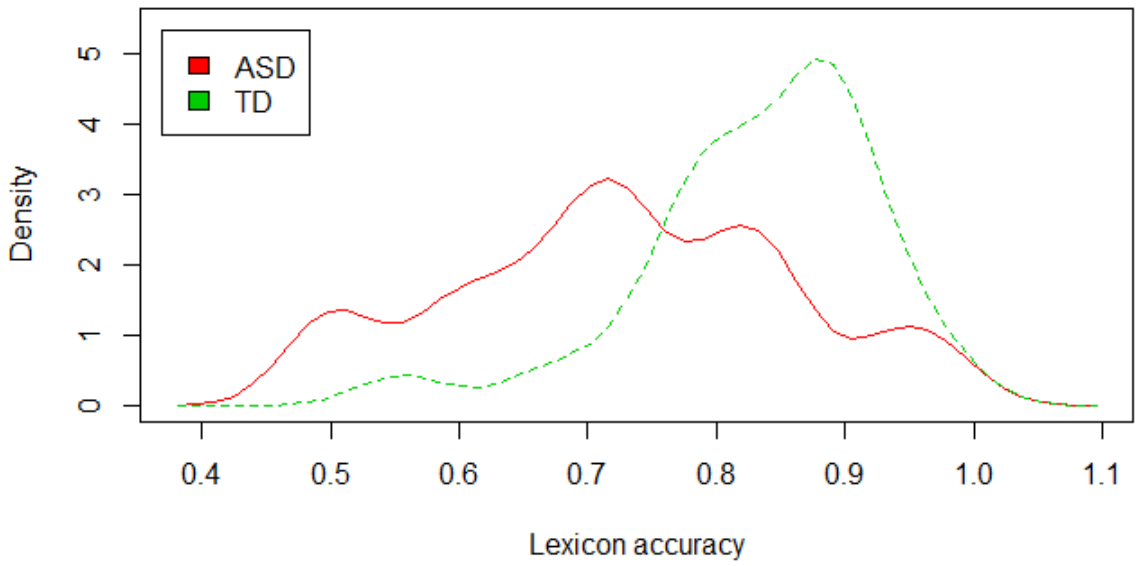
Appendix B. Participant's scores on different standardized tests and subtests. ASD_16 is missing because of a re-diagnosis of Social (Pragmatic) Communication Disorder.

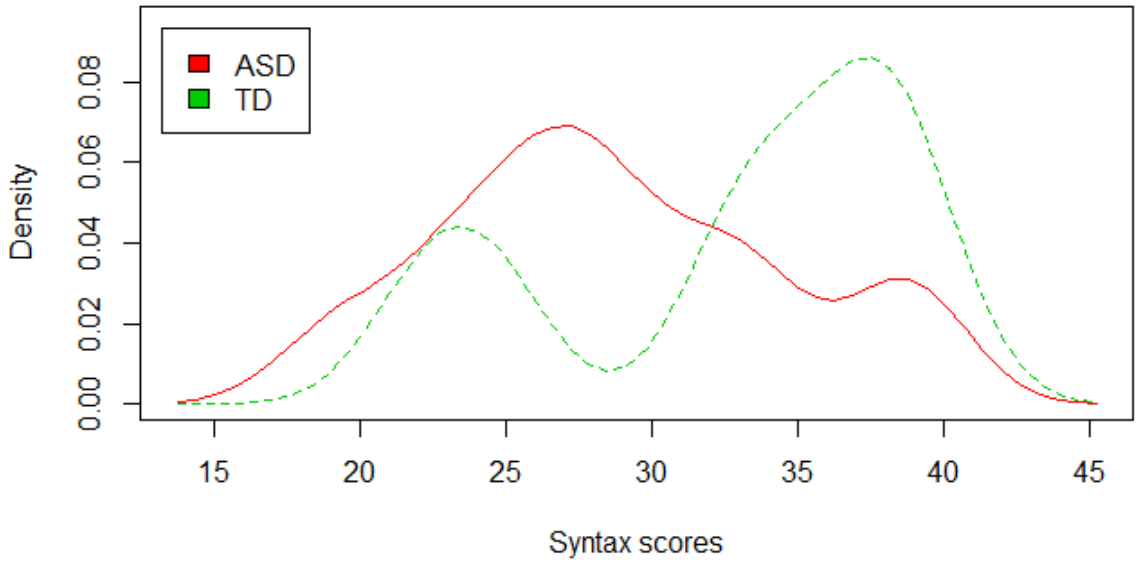
		Language tests (accuracy)		1st order ToM tasks (1=correct response)			
Participant	Raven	Lexicon	Syntax	Diverse Desire	Diverse Belief	Knowledge Access	Content False Belief
ASD_1	27	71%	83%	1	1	1	1
ASD_2	23	67%	60%	0	0	0	0
ASD_3	17	83%	68%	1	1	1	1
ASD_4	26	62%	80%	1	1	1	0
ASD_5	30	83%	80%	1	1	1	1
ASD_6	22	50%	70%	1	0	1	0
ASD_7	20	50%	58%	0	0	0	0
ASD_8	36	93%	98%	1	1	1	1
ASD_9	19	74%	65%	1	1	0	0
ASD_10	20	71%	73%	1	0	0	0
ASD_11	21	83%	73%	1	0	1	0
ASD_12	34	81%	75%	0	0	0	0
ASD_13	11	72%	60%	1	0	0	0
ASD_14	34	81%	98%	1	1	1	1
ASD_15	20	71%	68%	1	1	0	0
ASD_17	21	60%	53%	1	0	0	1
ASD_18	31	74%	85%	1	1	1	0
ASD_19	26	57%	70%	1	1	0	0
ASD_20	33	95%	95%	1	0	1	0
ASD_21	34	98%	98%	1	1	0	0
ASD_22	28	62%	65%	1	0	1	0
ASD_23	17	78%	48%	0	0	0	0
ASD_24	21	67%	58%	0	0	0	0
ASD_25	28	86%	48%	1	1	0	0

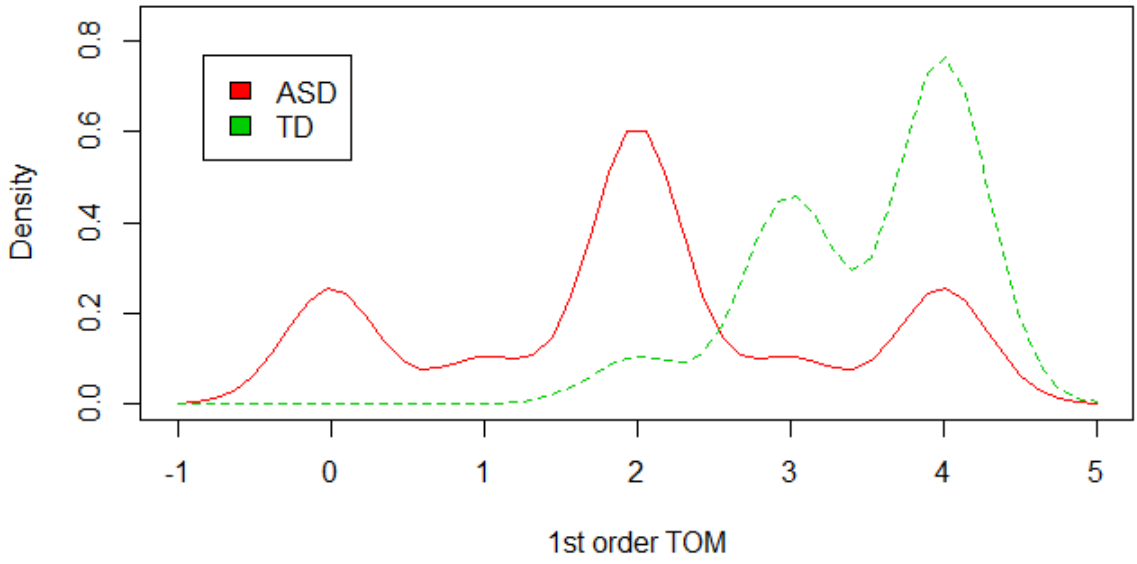
ASD_26	20	69%	85%	1	1	0	0
ASD_27	18	50%	65%	1	1	0	0
TD_1	36	93%	95%	1	1	1	0
TD_2	20	89%	55%	0	1	1	0
TD_3	21	78%	55%	1	1	1	1
TD_4	35	90%	98%	0	1	1	1
TD_5	31	88%	95%	1	1	1	1
TD_6	34	81%	93%	1	0	1	0
TD_7	19	78%	60%	1	1	1	0
TD_8	36	81%	95%	1	1	1	0
TD_9	34	95%	98%	1	1	1	1
TD_10	34	88%	98%	1	1	1	1
TD_11	19	56%	63%	1	1	0	1
TD_12	21	81%	85%	1	1	1	0
TD_13	28	88%	90%	1	1	1	0
TD_14	26	90%	83%	1	1	1	1
TD_15	31	78%	83%	1	1	1	1
TD_16	26	76%	80%	1	1	1	0
TD_17	33	83%	88%	1	1	1	1
TD_18	30	88%	90%	1	1	1	1
TD_19	31	88%	95%	1	1	1	1
TD_20	30	98%	100%	1	1	1	1
TD_21	32	88%	93%	1	1	1	1
TD_22	14	67%	63%	1	1	1	1
TD_23	16	72%	60%	1	1	1	0
TD_24	29	81%	83%	1	1	1	1
TD_25	19	83%	55%	1	1	1	1
TD_26	22	89%	85%	1	1	1	1

Appendix C. Plots of the distribution of each group in each task.






ASD
TD
Density
0.00
0.02
0.04
0.06
0.08
15
20
25
30
35
40
45
Syntax scores


ASD
TD
Density
0.0
0.2
0.4
0.6
0.8
-1
0
1
2
3
4
5
1st order TOM

PART 3 –

ADULT COMPETENCE

WITH SCALAR IMPLICATURES

Chapter 3.

Scalar Implicature Computation and the Role of the Autistic Quotient

This chapter is based on the following original article:

Mazzaggio, G., & Surian, L. (2017). The propensity to compute scalar implicatures is linked to autistic traits.

Manuscript accepted in Acta Linguistica Academica.

Abstract

We investigated whether there is an association between autistic traits in the broader phenotype and the ability to compute scalar implicatures. Previous studies found that the frequency of autistic traits is higher in students of science than of humanities; therefore, the two experiments reported here recorded the degree of rejection of underinformative scalar items in students enrolled either in a science or in a humanities curriculum. Also, we assessed their autistic traits using the Autism-Spectrum Quotient questionnaire. We found that students enrolled in science curricula provided fewer rejections compared to students enrolled in humanities curricula. Moreover, rejections were associated negatively with autistic traits and positively with performance on Theory-of-Mind tasks. These findings suggest that autism cognitive phenotype is negatively associated with a propensity to spontaneously derivate scalar implicatures.

1. Introduction

Grice (1975) made an important distinction between saying and conveying, which corresponds to a distinction between utterance meaning (i.e., the semantic meaning) and speaker meaning. Specifically, the speaker meaning is the meaning that the speaker intends to convey to the addressee. The distinction is reflected in the fact that speaker meaning can differ from utterance meaning, which is the conventional meaning of the words and syntax. An example is given in (1).

(1) A: Do you want to come out for a drink?

B: I have a job interview tomorrow.

The literal meaning of B's response (the utterance meaning) does not provide a direct answer to A's question. However, the answer is provided by the speaker meaning, it seems that B is unable to accept A's invitation. A's recognition of this meaning depends upon the specific (particular) contextual features. In a neutral context, the utterance 'I have a job interview tomorrow' does not convey that the speaker is unable to go out for a drink. In this example, such information is implicated and this is why we speak about 'implicatures'.

Grice distinguished particularized conversational implicatures, like in (1), from generalized conversational implicatures, which are, differently from the former, context-independent. An example of generalized conversational implicatures are 'scalar implicatures', which refer to sets of lexical items that constitute a scale in which the items are ordered with respect to their informativeness. The speaker's use of a weaker term in a scale implicates the negation of all the stronger terms (Horn, 1972). Considering the quantifiers scale <some, all> in (2) B's use of *some* implicates that he did not buy *all* the books for the exam, even if the semantic meaning of the utterance in (2B) is that B bought *some and possibily all* the books.

(2) A: Did you buy the books for the English exam?

B: I bought *some*.

According to Grice (1975), addressees derive an implicature as a result of assuming that speakers adhere to a *Cooperative Principle*, which, in turn, requires them to obey a set of implicit rules or maxims. Specifically, speakers are expected to make their contribution as informative as required and not more or less so (Quantity maxims). They are to believe that their contribution is true and they should not say that for which they lack adequate evidence (Quality maxims). They are to make their contribution relevant (Maxim of Relation), and they are to be perspicuous; in particular, they are to avoid obscure and ambiguous expressions, and they are to be brief and orderly (Manner maxims). Thus, in the example in (2), if B bought all the books for the exam, then B's utterance violates the first Quantity maxim (i.e. make your contribution as informative as required) because the quantifier *some* is less informative than *all.* The reason for which the existential quantifier *some* is less informative than the universal quantifier *all* is because the latter can be true in less circumstances.

Indeed, as expressed in (3a), a sentence like ‘All As are Bs’ can be true if, and only if (iff), all the elements present in the set of As are also present in the set of Bs. On the other hand, as expressed in (3b), a sentence like ‘Some As are Bs’ is true iff the interception of As and Bs is not empty (Katsos et al., 2016). This means that for (3a) to be true, (3b) must be true as well, but (3b) can be true also if (3a) is not. In other words, the quantifier “some” can logically be interpreted as “some and possibly all” but, since there is a more informative quantifier (i.e., *all*) available in the scale, “some” is interpreted pragmatically as “some but not all”. This kind of reasoning can be applied for several other sets of terms than can be ordered in a scale, such as logical operators (<or, and>), modals (<might, must>), adjectives

(<warm, hot>), verbs (<start, finish>), adverbs (<sometimes, often, always>).

(3) (a)‘ All As are Bs’ is true iff A∩B=A

(b) ‘Some As are Bs’ is true iff A ∩ B ≠ ⌀

Two main theoretical views on the computation of scalar implicatures have been proposed (for a review of all the models see Chemla & Singh, 2014a). An ongoing debate concerns whether scalar implicatures are computed through linguistic (e.g. semantic or syntactic) or pragmatic mechanisms that require mental state reasoning. For example, Chierchia's (2013) *Extended Standard Gricean Theory* assumes that a scalar implicature is derived from an utterance such as (4a) by means of an exhaustivity operator (*O*), which is a covert counterpart of *only* as shown in (4b). The result of the operator is shown in (4c). The ‘exhaustification process’ operates on the set of lexical scalar alternatives when they are relevant to the conversational goals. If deriving the implicature is not advantageous in a particular context, the exhaustification process will not occur.

(4) a. Some linguists are smart.

b. Only some linguists are smart.

c. *O* [some linguists are smart].

As we have previously seen, other researchers attribute the computation of scalar-implicatures to the ability to recognize the speaker's communicative intentions. Based on Grice’s account, if we consider the dialogue in (2), the interlocutor A, after listening the sentence “I bought *some*” uttered by B, should take into consideration the alternative sentence

"I bought *all*" that B could have uttered instead. Because A knows (and knows that B knows too) that *some* is less informative than *all* and that B used *some* instead of *all*, A assumes that if B bought *all* the books, then B would have used the quantifier *all* as required by the Quantity maxim. This ability to 'read the mind' of another person is known in the literature as Theory of Mind (Premack & Woodruff, 1978) or mentalizing ability (Frith, Morton & Leslie, 1991). Chemla and Singh (2014) suggested that both the Gricean view and the grammatical theory (i.e. Extended Standard Gricean Theory), theorize the involvement, at a certain point, of a pragmatic decision in scalar-implicature computation. However, there is no consensus on the role of Theory of Mind in pragmatic inferences. While some authors support the idea of a central role of mentalizing abilities (Nieuwland et al. 2010; Surian et al. 1996), other authors do not (Andrés-Roqueta & Katsos, 2017). Studies on population with less mature Theory-of-Mind skills might be of interest for this specific theoretic aspect of the debate.

People with Autism Spectrum Disorder (ASD) present impairment in Theory-of-Mind development (Baron-Cohen, Leslie & Frith, 1985) and this lack of or non-mature Theory of Mind has often been linked to pragmatic deficits (Baron-Cohen, 1988). ASD population is well known for presenting pragmatic difficulties, for example with obeying Gricean Maxims (Capps, Kehres & Sigman, 1998; Surian, Baron-Cohen & Van der Lely, 1996), with comprehending jokes (Baron-Cohen, 1997a; Reddy, Williams & Vaughan, 2002), irony and metaphors (Happé, 1993; MacKay & Shaw, 2004). A recent study reported that adolescents with ASD can compute scalar implicatures, but they seem do so without reasoning about the interlocutor's epistemic states (Hochstein, Bale & Barner, 2017).

The aim of the present study was to further investigate scalar-implicature computation and the role of epistemic reasoning by testing typically developing adult students enrolled in scientific and humanistic curricula (Experiment 1) and by assessing the presence of autistic traits (Experiment 2). Many previous works by Baron-Cohen and colleagues found that

autistic traits, such as weak Theory-of-Mind skills, are associated both to gender, with males scoring higher than females, and to vocational choices, with peoples involved in some scientific or technical profession scoring higher than people involved in social professions (Baron-Cohen, 1997; Baron-Cohen, Wheelwright, Scott, Bolton & Goodyer, 1997; Baron-Cohen, Bolton, Wheelwright, Scahill, Short, Mead & Smit, 1998; Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001). If scalar-implicature computation is linked to mental state reasoning, then one should expect less pragmatic answers and more logical interpretations of underinformative sentences (i.e., sentences including *some*) in students of scientific curricula and in males, than in students of humanities curricula and in females.

Baron-Cohen (1997) proposed that people with high-functioning autism have weak 'folk psychology' abilities, such as inferring mental states from people's behavior, but well developed 'folk physics' abilities, such as inferring the physical causes of natural events. These folk physics skills are superior compared to typical population (Baron-Cohen, Leslie & Frith, 1986; Sigman, Ungerer, Mundy & Sherman, 1987; Baron-Cohen, 1989; Leekam & Perner, 1991; Leslie & Thaiss, 1992; Jolliffe & Baron-Cohen, 1997).

Coherently with such proposal, Baron-Cohen et al. (1997) found that fathers and grandfathers of children with autism are more likely to be employed in fields such engineering and informatics than fathers and grandfathers of children without autism. In line with this, Baron-Cohen et al. (1998) showed that students of physics, engineering and mathematics have more biological relatives with autism compared to students of literature. Baron-Cohen et al. (2001) also found that male students and scientists score higher on autistic traits compared to females and to students of humanities and social sciences.

Thus, Experiment 2 included the 'Autism-spectrum Quotient', which is a standardized tool to measure the extent of autistic traits in the typical adult population with an average IQ (Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001). We were particularly

interested in the relation between scalar-implicature computation and autistic traits measured by the Autism-spectrum Quotient, such as social skills, attention switching skills, attention to detail skills, communication skills, imagination skills and, as we proposed, Theory-of-Mind skills.

A strong difference in autistic traits, when comparing two groups, is of course found when one compares a group of people with autism with a control group of people without such disorder. In our case in which autistic traits in the broader phenotype are assessed, if we assume that the computation of scalar implicatures involves some reasoning about mental states and this type of reasoning is negatively associated with the presence of autistic traits, then we should predict that the amount of pragmatic responses obtained in a scalar-implicature task would be a function of autistic traits. Moreover, since these traits are also associated with different vocational choices, we should also predict that pragmatic responses in the scalar-implicature task would vary as a function of the course attended by students.

2. Experiment 1

2.1. Method

2.1.1. Participants

Participants were 428 students (277 females, mean age 23.2 years, *SD* = 4.6) who were attending a university in Italy. Students were recruited online and received no compensation in exchange for participation. Following Baron-Cohen et al. (2011), participants were divided into a 'science group' (*N* = 176, 99 males, mean age 22.8, *SD* = 3.76) and a 'humanities group' (*N* = 252, 199 females, mean age 23.5, *SD* = 5.12). The science group consisted of students enrolled in mathematics (*N* = 74), chemistry (*N* = 4), medicine (*N* = 21), engineering (*N* = 21), physics (*N* = 19), informatics (*N* = 19), neuroscience (*N* = 8) and natural sciences (*N* = 10). The humanities and social sciences group consisted of students enrolled in psychology

(*N* = 201), sociology (*N* = 5), law (*N* = 6), literature and philosophy (*N* = 6), foreign languages and literatures (*N* = 8), economy (*N* = 15), archeology and architecture (*N* = 3), history (*N* = 5), and communication sciences and social services (*N* = 3). Two participants were eliminated because they were less than 18 years old.

2.1.2. Materials and procedure

Participants completed an online *Sentence Evaluation Task*, that has been used in a number of previous studies of scalar implicatures (see Noveck, 2001; Guasti, Chierchia, Crain, Foppolo, Gualmini & Meroni, 2005; Pouscoulous, Noveck, Politzer & Bastide); sentences were adapted from Bott and Noveck (2004). On each trial, a sentence with the basic form, All/Some X are Y, was presented and participants indicated whether they agreed by using the mouse to click on *Agree* or *Disagree* displayed below the sentence. There were 32 sentences presented, half began with *all* and half began with *some*. Of the sentences with *all*, 8 were universally true (All-true) and 8 were false or absurd statements (All-false). Of the sentences with *some*, 8 were true (Some-true) and 8 were underinformative (Some-underinformative). In Table 1 some examples are presented. The complete list of sentences translated from Italian is given in Appendix 1.

Table 1. Typologies, examples and expected answers of the sentences used in Experiments 1 and 2.

Type	Example Sentence	Expected Answer
All-True	All snakes are reptiles	Agree
All-False	All animals are carnivorous	Disagree
Some-True	Some dogs are Labrador	Agree
Some-Underinformative	Some children are humans	Agree (Logic);

Disagree (Pragmatic)

2.2. Results and discussion

Agree was a correct response for the 16 true sentences (8 All-true and 8 Some-true), and *Disagree* was a correct response for the 8 All-false sentences. Both the science group and the humanities group were highly accurate in responding to all three sentence-types (mean correct responses range from 7.85 to 7.97).

Agree responses to the underinformative sentences were considered logical responses whereas *Disagree* responses were considered pragmatic responses. Consistent with our prediction, the number of logical responses in the science group ($M = 2.91$, $SD = 3.58$) was significantly higher than in the humanities group ($M = 2.10$, $SD = 3.15$; Kruskal-Wallis H test, $\chi^2 (1) = 5.29$, $p = .02$, $d = .20$), as is visible in *Figure 1*.[3] Moreover, we found an effect of gender: the mean number of logical responses in males was 2.99 ($SD = 3.57$) and it was significantly higher than in females 2.14 ($SD = 3.19$; Kruskal-Wallis H test, $\chi^2 (1) = 9.13$, $p = .003$, $d = .28$). However, the gender effect on the mean number of logical responses only approached significance in the science group ($M_{MALE} = 3.35$, $SD = 3.72$; $M_{FEMALE} = 2.37$, $SD = 3.33$; Kruskal-Wallis H test, $\chi^2 (1) = 3.73$, $p = .05$, $d = .25$), and it was not significant in the humanities group ($M_{MALE} = 2.32$, $SD = 3.21$; $M_{FEMALE} = 2.05$, $SD = 3.14$; Kruskal-Wallis H test, $\chi^2 (1) = 1.96$, $p = .16$).

[3] As an anonymous reviewer noted, the scientific group did not answer 'logically' above chance in the underinformative utterances. Actually, we never expected that typical adults might answer 'logically' *tout court*; indeed, in the literature such behavior is characteristic of children before the age of 5 or 6 (for a short review, Chemla & Singh, 2014b: 390-391). Our prediction was that, overall, students of scientific disciplines would interpret underinformative statements in a logical way more often than students of humanities.

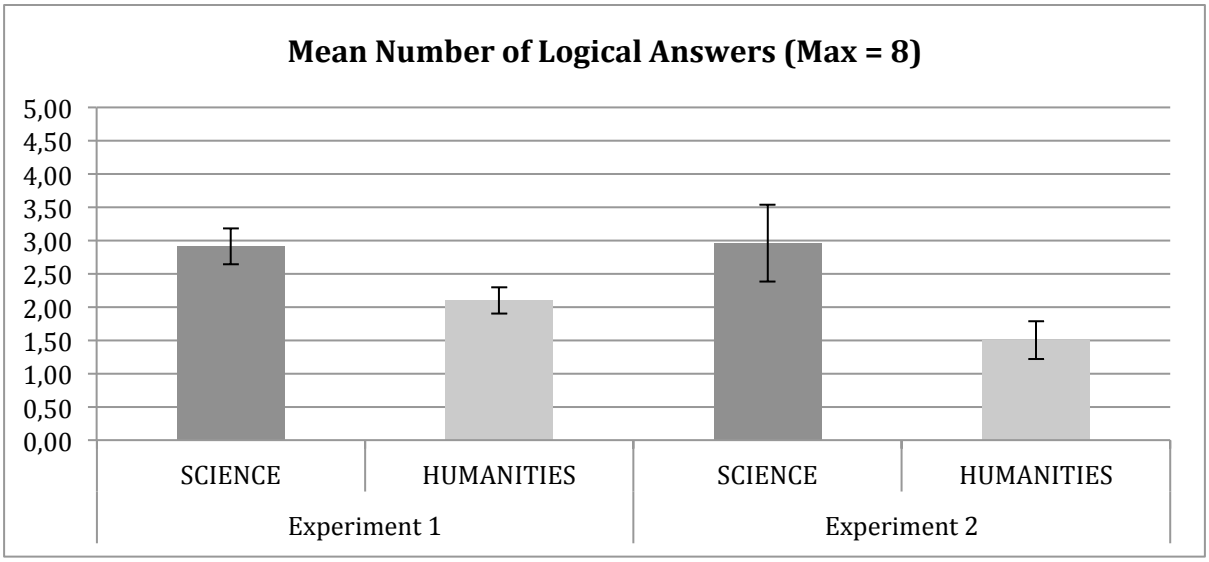


Figure 1. The Science group and Humanities group's mean number of logical responses to underinformative statements with *Some* in Experiments 1 and 2. Error bars are standard errors of the mean.

3. Experiment 2

Experiment 2 was aimed at consolidating the results of Experiment 1 by adding a direct measure of autistic traits. Thus, Experiment 2 was identical to Experiment 1 except that participants also completed the Italian version of Baron-Cohen et al's (2001) Autism-spectrum Quotient questionnaire (Ruta, Mazzone, Mazzone, Wheelwright & Baron-Cohen, 2012).

3.1. Method

3.1.1. Participants

Participants were 198 Italian university students (89 males, mean age 23.8 years, *SD* = 3.08), who were recruited online and were not compensated for their participation. As in Experiment 1, Baron-Cohen et al.'s (2011) criteria were used to divide participants into a science group (*N* = 90, 38 females, mean age 23.7 years, *SD* = 2.27) and a humanities group

(*N* = 108, 37 males, mean age 23.8 years, *SD* = 3.63). The science group consisted of students enrolled in mathematics (*N* = 38), physics (*N* = 4), engineering (*N* = 27), medical and medical biotechonologies schools (*N* = 3) and informatics (*N* = 18). The humanities group consisted of students enrolled in literature and philosophy (*N* = 22), foreign languages and literatures (*N* = 35), communication sciences (*N* = 6), history (*N* = 2), sociology (*N* = 2), art (*N* = 12), law and economy (*N* = 7), pedagogy and social sciences (*N* = 22).

3.1.2. Materials and procedure

Experiment 2 used the materials and procedure for the online *Sentence Evaluation Task* as in Experiment 1. In addition, participants completed Baron-Cohen et al's (2001) Autism-spectrum Quotient before the scalar-implicature task. The measure consists of 50 questions with 10 questions assessing each of five skills/traits: social (e.g., "I prefer to do things with others rather than on my own."), attention switching (e.g., "It does not upset me if my daily routine is disturbed."), attention to detail (e.g., "I usually notice car number plates or similar strings of information."), communication (e.g., "Other people frequently tell me that what I've said is impolite, even though I think it is polite.") and imagination (e.g., "If I try to imagine something, I find it very easy to create a picture in my mind.").

All questions were presented with four answer choices: definitely agree, slightly agree, slightly disagree or definitely disagree. Participants had to click on one of the four answers and responses were scored as 1 or 0. Half of the questions are created to elicit an "agree" or "slightly agree" answer, thus with 1 point assigned if one of those two answers are given; half of the questions are created to elicit a "disagree" or "slightly disagree" answer, thus with 1 point assigned if one of those two answers are given. The higher the obtained score, the higher the presence of autistic traits; according to Baron-Cohen et al's (2001), a score of 32+ might be a cutoff to reveal the presence of clinically significant levels of autistic traits.

However, obtaining a score of 32+ does not mean the individual actually has an autistic disorder. The maximum score is 50, and Baron-Cohen, et al. (2001) reported an average score of 16 for a group typical adults, an average score of 36 for a group with Asperger Syndrome or high-functioning autism (HFA), and an average score of 25 for a group of winners of the UK Mathematics Olympiad.

3.2. Results and discussion

Both the science and humanities groups were highly accurate in responding to the All-true, Some-true, and All-false statements (mean correct responses ranged from 7.68 to 7.93).

Like in Experiment 1, the mean number of logical answers in the science group was significantly higher than in the humanities group (*Figure 1;* $M_{SC.} = 2.96$, $SD = 3.61$ vs. $M_{HUM} = 1.55$, $SD = 2.67$; Kruskal-Wallis H test, $\chi^2(1) = 4.33$, $p = .04$, $d = .26$). The difference between the males and females in the number of logical responses was not significant, $M_{MALE} = 2.63$, $SD = 3.46$; $M_{FEMALE} = 1.83$, $SD = 2.95$, Kruskal-Wallis H test, $\chi^2(1) = 2.14$, $p = .14$).

When looking at the AQ, we replicated Baron-Cohen's (2001) data. *Figure 2* represents the means of the scores obtained in the Autism-spectrum Quotient by males and females in the two groups. The science group had a higher average Autism-spectrum Quotient than the humanities group (22.84 vs. 19.33, respectively, Kruskal-Wallis H test, $\chi^2(1) = 18.31$, $p < .001$, $d = .62$). If we consider a score of 32+ as a cutoff for clinically significant levels of autistic traits, in our sample no students in the humanities group equaled that high score, but 6 individuals in the science group did, as it is visible in *Figure 3*. In addition, males had a higher average Autism-spectrum Quotient than females (22.21 vs. 19.88, respectively, Kruskal-Wallis H test, $\chi^2(1) = 6.81$, $p = .009$, $d = .35$). Correlation between the participants' Autism-spectrum Quotient score and their number of logical responses is marginally

significant (Pearson $r = 0.136$, $N = 198$, $p = .06$). Such correlation, however, is also significant when considering students in the humanities group separately (Pearson $r = .202$, $N = 108$, $p = .04$), but it is not significant when analyzing students of the scientific group separately (Pearson $r = -.021$, $N = 90$, $p = .85$).

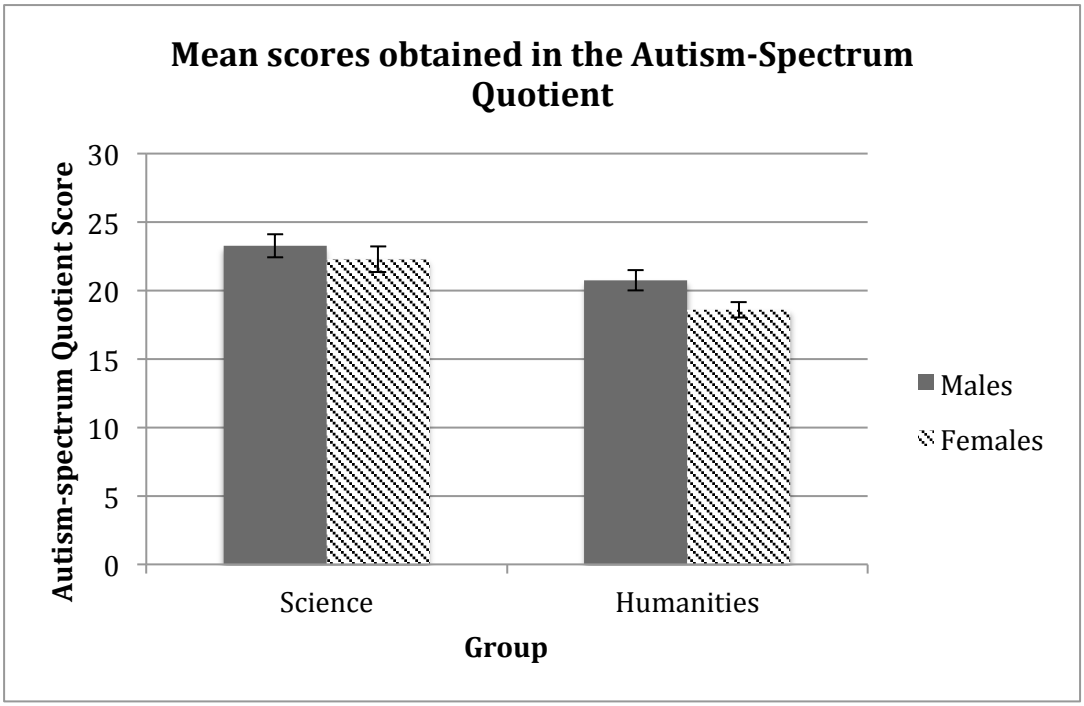


Figure 2. The mean number of scores obtained in the Autism-spectrum Quotient in Experiments 2, divided by science / humanities and males /females. Error bars are standard errors of the mean.

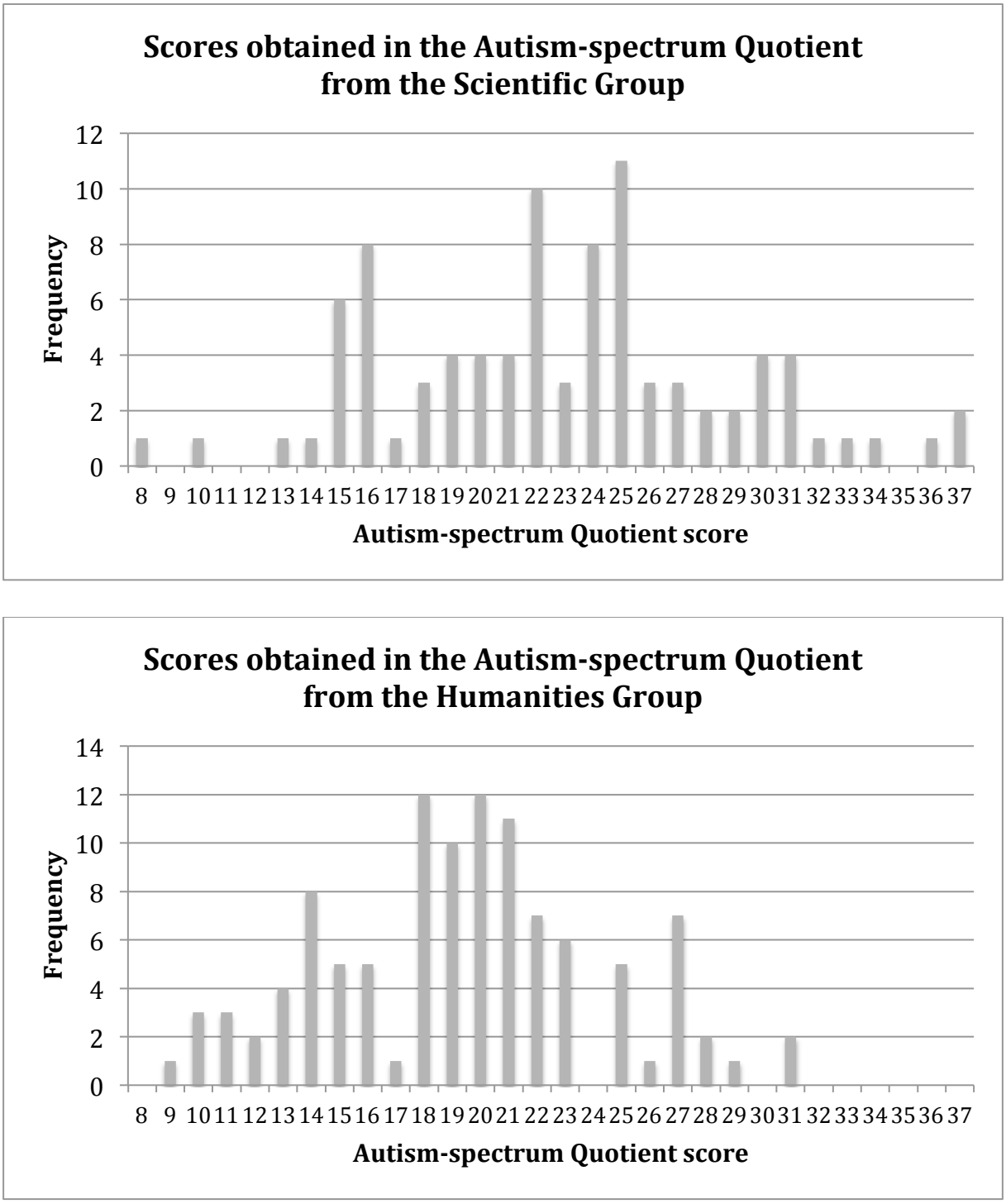


Figure 3. Frequency of obtained scores in the Autism-spectrum Quotient by the Scientific group and the Humanities group.

Thirteen items in the Autism-spectrum Quotient appeared to tap into Theory-of-Mind skills, for example, 'When I'm reading a story, I find it difficult to work out the characters'

intentions', ‘When I was young, I used to enjoy playing games involving pretending with other children’, ‘I find it difficult to imagine what it would be like to be someone else’. These items were distributed across the categories of ‘social skills’, ‘communication skills’ and ‘imagination skills’. The complete list is given in Appendix 2. Our results show a positive relationship between participants’ score for these thirteen Theory-of-Mind items (a high score implies low Theory-of-Mind skills) and the number of logical answers for underinformative items (r = .165, p = .02). Again, such correlation is not shown when considering only students of the science group (Pearson r = -.069, N = 90, p = .52), but remains significant when considering only students in the humanities group (Pearson r = .322, N = 108, p = .001). In contrast, there was no relationship between the score for the other 47 items and number of logical responses (r = .086, p = .226).

4. General discussion

In Experiments 1 and 2 we found that science students gave less pragmatic answers and were more likely to respond logically to scalar-implicature statements than humanities students. These findings are consistent with the observation of higher incidence of autistic traits among science majors compared to humanities majors (Baron-Cohen, 1997; Baron-Cohen et al., 1997; Baron-Cohen et al., 1998; Baron-Cohen et al., 2001). Experiment 2 further showed a positive relationship between typical adults' scores on the Autism-spectrum Quotient, which measures autistic traits, and the number of logical (as opposed to pragmatic) responses to scalar-implicature statements. Finally, there was a trend for lower scores in questions most relevant to the assessment of Theory of Mind to be associated with a higher number of logical responses.

Our results suggest that autistic traits are associated with weaker tendencies to draw pragmatic implicatures. In that respect, our findings are consistent with results of a recent

ERP study on autism and pragmatic processing that examined electrophysiological responses to evaluate individual differences in the pragmatic processing of underinformative statements (Nieuwland, Ditman & Kuperberg, 2010). They found individual variations in N400 responses based on participants' pragmatic skills, assessed with the Autism-spectrum Quotient. That is, participants with higher Autism-spectrum Quotient scores (and, particularly, with higher autistic traits in the Communication subscale) were less sensitive to pragmatic violations in underinformative sentences and actually showed no pragmatic N400 effect.

However, our results are surprising if one considers the outcomes of previous studies showing intact performance of teens and adults with ASD in scalar-implicature computation tasks (Chevallier, Wilson, Happé & Noveck, 2010; Pijnacker, Hagoort, Buitelaar, Teunisse & Geurts, 2009. Similar findings were also reported on children with ASD by Su and Su (2015). Pijnacker et al. (2009), tested 56 Dutch adults, 28 with ASD (11 participants with high-functioning autism (HFA) and 17 with Asperger syndrome) and 28 matched controls.[4] All participants completed a Truth-Value Judgment Task on two different types of scalar terms: <some, all> and <or, and>. Pijnacker et al. expected more logical answers from the ASD group, due to their Theory-of-Mind deficits. However, they did not find differences in the pragmatic answers of the two groups. There was also a positive correlation between the HFA group's pragmatic answers and their verbal intelligence. On the other hand, according to authors, the verbal-intelligence differences between the two ASD subgroups might be a limitation of this study (HFA $_{\text{verbal-int.}}$ = 109.8, Aspergers $_{\text{verbal-int.}}$ = 122.4). Furthermore, authors wrote that it is plausible that Theory of Mind is involved in scalar implicatures computation; however, the study did not assess Theory-of-Mind abilities.

---

[4]The distinction between Asperger syndrome and high functioning autism made by *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000) has been now abandoned in the last, fifth edition of the DSM (American Psychiatric Association, 2013).

Chevallier et al. (2010) tested 22 ASD adolescents (6 HFA and 16 Aspergers) and 22 controls on the pragmatic interpretation of underinformative connectives; also in their study they did not assess Theory-of-Mind skills. Like Pijnacker et al's (2009) findings, the rate of pragmatic interpretations was the same for the ASD and control groups, and there was a positive correlation between pragmatic responses and verbal intelligence, but only in the ASD group. Finally, Su and Su (2015) tested 28 Mandarin-speaking children (age range 4-15 yo) with ASD (12 children with autistic disorder, 15 children with Asperger syndrome and 1 child with pervasive developmental disorder) and 28 controls on the computation of underinformative sentences with the logical words “some” and “every…or”. ASD children didn’t show any delay in the derivation of scalar implicatures compared to the matched controls.

One possibility to explain the inconsistency of our results with previous findings in literature on autism is that students of scientific disciplines are trained to focus on the literal meaning when performing academic tasks or logic puzzles and this shaped their pragmatic reasoning in our task. This explanation is supported by the fact that scores in the AQ, and particularly of the items we considered related to Theory-of-Mind skills, correlated with the number of pragmatic answers only in the humanities group and not in the scientific group. If, at first glance, this result seemed in contrast with our first hypothesis, we should nonetheless consider that the main goal of our study was precisely to demonstrate that higher autistic traits might lead to more logical answers. If the idea of testing scientific students for such purpose may have been relevant in the light of previous studies (Baron-Cohen, 1997; Baron-Cohen et al., 1997; Baron-Cohen et al., 1998; Baron-Cohen et al., 2001), we now have to consider that the fact that students of scientific disciplines are trained to answer logically in academic tests, might have leveled the differences, in the implicature task, between participants with higher AQ scores and participants with lower AQ scores. On the other hand, our results remain

interesting in terms of correlation between autistic traits and pragmatic answers, with less pragmatic answers related to higher autistic traits.

The inconsistency between current and previous results might also be explained by considering task differences between our study and the others, since it has been demonstrated that, when assessing the understanding or computation of scalar implicatures, sentence-evaluation tasks are preferable to truth-value judgment tasks. Indeed, if we consider the sentence "Some dogs are animals" we cannot judge a logical answer (i.e. true) as incorrect, as it is judged in the truth-value judgment task; we should be simply interested in a felicity evaluation, which is whether the participant agrees or disagrees with the content of the statement.

Moreover, people in the autistic spectrum are aware of their pragmatic difficulties and might use compensatory strategies to deal with their deficit, as suggested by Hochstein et al. (2017). This may account in part for the divergence between the findings of previous studies on scalar implicatures in ASD patients and the findings of the present study. Finally, by testing typically developing students with a high Autism-spectrum Quotient, we could assess a high number of participants, whereas previous studies on populations with autism were carried out using smaller samples and it is not yet confirmed whether the same results would also be found using larger samples. Further research is, however, needed to fully explain the origins of the inconsistency in available evidence and clarify the roles of Theory-of-Mind abilities in scalar-implicature computation both in children and adults.

Acknowledgments

This work has been supported by grants from the Fondazione ONLUS Marica De Vincenzi. We are grateful to Prof. Kathleen Eberhard and Prof. Remo Job for reading the proofs and helping with improving the paper.

Appendix 1 - Statements presented in the Sentence Evaluation Task

All true

All snakes are reptiles

All cats are animals

All men are humans

All birds are animals

All cobras are snakes

All dogs are animals

All horses are mammals

All sunflowers are flowers

All false

All animals are carnivorous

All cats are dogs

All stones are singers

All flowers are professors

All pens are animals

All children are grandmothers

All televisions are cars

All books are drinks

Some true

Some dogs are Labrador

Some children are blonde

Some flowers are red

Some cats are Persian

Some houses are rented

Some mobiles are iPhones

Some dresses are blue

Some lakes are big

Some underinformative

Some children are humans

Some salmons are fish

Some horses are animals

Some tulips are flowers

Some dogs are animals

Some women are humans

Some giraffes are animals

Some roses are flowers

Appendix 2 - Items from the Autism-spectrum Quotient that are related to Theory of Mind.

1. I find it easy to "read between the lines" when someone is talking to me.
2. I know how to tell if someone listening to me is getting bored.
3. I find it easy to work out what someone is thinking or feeling just by looking at their face. (Social skills)
4. I am good at social chit-chat.
5. When I was young, I used to enjoy playing games involving pretending with other children.
6. I find it very easy to play games with children that involve pretending.
7. Other people frequently tell me that what I've said is impolite, even though I think it is polite.
8. When I'm reading a story, I find it difficult to work out the characters' intentions.
9. I don' t particularly enjoy reading fiction.
10. I frequently find that I don't know how to keep a conversation going.
11. When I talk on the phone, I'm not sure when it's my turn to speak.
12. I am often the last to understand the point of a joke.
13. I find it difficult to work out people's intentions.

Chapter 4.

Scalar-Implicature Computation in Second-Language Oral Processing

This chapter is based on the following original article:

Mazzaggio, G., & Surian, L. (2018). Scalar implicatures are not generated by default: Evidence from second-language oral processing.

Manuscript submitted in Journal of Pragmatics.

Abstract

We investigated the effect of a second language (L2) on scalar-implicatures processing. To ensure that L2 processing was more effortful than first-language (L1) processing, 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 (N = 479), participants were asked to perform a *Sentence Evaluation Task* either in their L1 (Italian) or in their L2 (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. We found more rejections when participants listened to L1 rather than L2 utterances. These results provide support for the non-default models of scalar implicatures.

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 quality and quantity 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* $\subseteq$ *some* (*all* logically entails *some*). The logical interpretation of *some and possibly all* is the lower-bound interpretation in (2a) whereas the pragmatic interpretation of *some but not all* is the upper-bound interpretation in (2b). The latter arises from listeners assuming that a speaker chose the most informative quantifier from the scale. Furthermore, under certain circumstances, the pragmatic interpretation can be cancelled without logical contradiction, as

in ‘Some students got an A, indeed all of them got an A’.

(2) a. ||some|| = λPλQ ∃ x[P(x)∧Q(x)]

b. ||some|| = λPλQ ∃ x[P(x)∧Q(x)] ∧ ¬∀x [P(x) → Q(x)]

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 Horn (1972) and by Levinson (2000), the pragmatic interpretation is automatic and the ‘default’ meaning: “default means that relatively weak terms prompt the inference automatically […] Also, a scalar inference can be cancelled. If this happens, it occurs subsequent to the production of the scalar term.” (Bott & Noveck, 2004, p. 439). 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.

In contrast, the non-default models, such as those proposed by Carston (1998) and by Sperber and Wilson (1986, 1995) claim that 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.

Many studies addressing the scalar-implicature debate (among them, Bott & Noveck, 2004; De Neys & Schaeken, 2007; Guasti, Chierchia, Crain, Foppolo, Gualmini, & Meroni, 2005; Marty & Chemla, 2013; Noveck, 2001; Papafragou & Musolino, 2003; Pouscoulous, Noveck, Politzer, & Bastide, 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. Several studies have investigated these effects using populations for which linguistic competence is deemed not fully developed, i.e. children and L2 speakers.

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 & Rumain, 1981; Chierchia, Crain, Guasti, Gualmini, & Meroni, 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 domain-related general cognitive limitations (Reinhart, 1999), from limitations in their lexical knowledge, preventing the access to relevant lexical scales (Lexical Account: Barner, Brooks, & Bale, 2011; Chierchia et al., 2001; Foppolo, 2007), and from difficulties in understanding the difference between 'appropriate' and 'true' (Miller, Schmitt, Chang, & Munn, 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004).

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 increased cognitive effort and a longer processing time (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Degen & Tanenhaus, 2011; Dieussaert,Verkerk, Gillard, & Schaeken, 2011; Huang & Snedeker, 2009a; Noveck & Posada, 2003; Politzer-Ahles & Gwilliams, 2015; Tomlinson, Bailey, & Bott, 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 less 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; an increase of 16% in pragmatic answers was found when more time was available. When participants had less cognitive resources available, less 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 contextual support led to no differences in terms of reaction times (Grodner, Klein, Carbary, & Tanenhaus, 2010).

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: on the one hand, 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); on the other hand, balanced bilinguals might show cognitive strengthening compared to monolinguals (Bialystok, Craik, Green, & Gollan, 2009). While most of available data from both children and adults show that pragmatic implicatures are costly to make, evidence on bilinguals is more mixed. Some recent studies report no differences between L1 and L2 processing in pragmatic answers for scalar implicatures (Antoniou & Katsos, 2017; Antoniou, Veenstra, Kissine, & Katsos, 2018; Dupuy, Stateva, Andreetta, Cheylus, Déprez, Henst, Jayez, Stepanov, & Reboul, 2018; Syrett, Austin, Sanchez, Germak, Lingwall, Perez-Cortes, Arias-Amaya, & Baker, 2016; Syrett, Lingwall, Perez-Cortes, Austin, Sánchez, Baker, Germak, & Arias-Amaya, 2017).

On the contrary, other studies found an increase in pragmatic answers by testing scalar implicatures in both early bilingual children (Siegal, Matsuo, Pond, & Otsu, 2007) and bilingual adults (Slabakova, 2010; Snape & Hosoi, 2018). Puzzlingly, explanations for such results were attributed to both increased cognitive skills (i.e., bilinguals give more pragmatic answers because to compute implicatures is costly and they have more cognitive resources), in line with the non-default models, and to decreased processing resources (i.e., 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.

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 scalar-implicatures improves, reaching the native-speakers' competence after 4 ½ years of living in the L2 foreign country. On the other hand, Korean has two words that could be translations of English 'some': *etten* and *ilbu*. Slabakova used *etten* because – according to native Korean speakers – it is closer to the English *some*. Thus, the increase in pragmatic responses may have been brought about by the asymmetry between the participants' native language and their L2 with respect to the quantifier. Given this state of affairs, it would be useful to investigate whether the Korean-English bilinguals' results can be replicated with different L2 learners, as well as with speakers with lower fluency in L2.

Our study employed both an L2 with a quantifier-system similar to L1 and an L2 with a different system. Italian language (L1) has two different existential quantifiers used with countable nouns: *qualche* that must be used followed by nouns in the singular form and *alcuni* (masculine form) or *alcune* (feminine form) that must be followed by plural nouns. In our experiment we tested the form *alcuni/e*, the most commonly used form in Italian experiments on scalar implicatures (e.g., Guasti et al., 2005). As L2 we tested either English (Experiments 1 and 2), a language with only one form corresponding to the Italian *qualche* and *alcuni*, i.e. *some* or Spanish (Experiment 2), a language that, like Italian, has two different terms, *unos* and *algunos* (for a detailed explanation of differences, see Gutiérrez-Rexach, 2001).

The present study aimed at providing a more stringent test of the competing models (i.e., default and non-default) by testing 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). In addition, differently from other studies on bilinguals, our procedure imposed a time limit for interpreting sentences, thereby adding to the resource demands of the task. 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. Therefore, if the pragmatic interpretations are the non-default interpretations of underinformative utterances, they should be more frequent in the L1 than in the L2 group.

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 divided into a low proficiency group (*N* = 8) at the A1-A2 level, an intermediate-low group (*N* = 24) at the B1 level, an intermediate-high group (*N* = 17) at the B2 level and a high group (*N* = 6) at the C1 level. 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. 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 number of pragmatic interpretations of the underinformative sentences. A one-way between subjects ANOVA was conducted to compare the effect of an L2 on the processing of scalar implicatures; there was a significant effect of group (L1 vs. L2) on the number of pragmatic answers ($F(1, 84) = 4.06$, p = .05). In the L2 group, proficiency levels are not predictors of the number of pragmatic interpretations (low proficiency – medium proficiency: $U = 133.5$, $Z = -.83$, $p = .40$; low proficiency– high proficiency: $U = 16$, $Z = -1.04$, $p = .30$; medium proficiency – high proficiency: $U = 76$, $Z = -1.82$, $p = .07$).

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 = 6.5%; L2 = 16%). Figure 2 shows the percentage of each group's participants who were classified as consistent pragmatic or logical responders. Pragmatic responders were more frequent in the L1 (64.5 %) than in L2 group (41%), whereas logical responders were more frequent in the L2 group that in the L1 group (43% and 29%, respectively) and this difference approached significance: $\chi^2 (1) = 2.93$, $p = .09$, $d = .40$.

In true sentences with *some* or *all* "Yes" responses were correct, whereas in false sentences with *all* "No" responses were correct. As expected, the L1 group's accuracy was 100% for true sentences with *some* and false sentences with *all* and was 99.6% for true sentences with *all*. However, the L2 group's accuracy was lower for all three types of sentences ($p \leq .001$). Specifically, their average percentage of correct responses for true

sentences with *some* and with *all* was 89.3% and 81.3 %, respectively, and their average percentage of correct responses for false sentences with *all* was 88.2%.

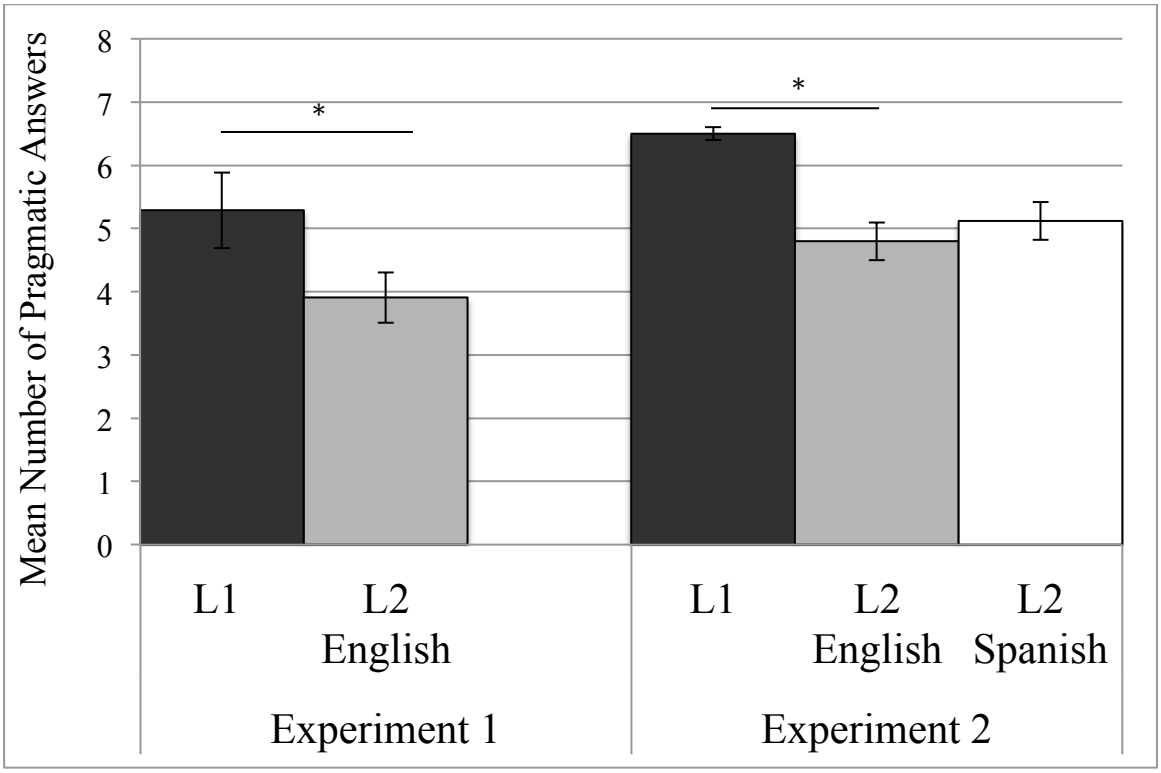


Fig 1. Mean number of pragmatic interpretations of the underinformative sentences for the L1 group and the L2 groups in Experiments 1 (left panel) and 2 (right panel). Error bars are standard error of the mean.

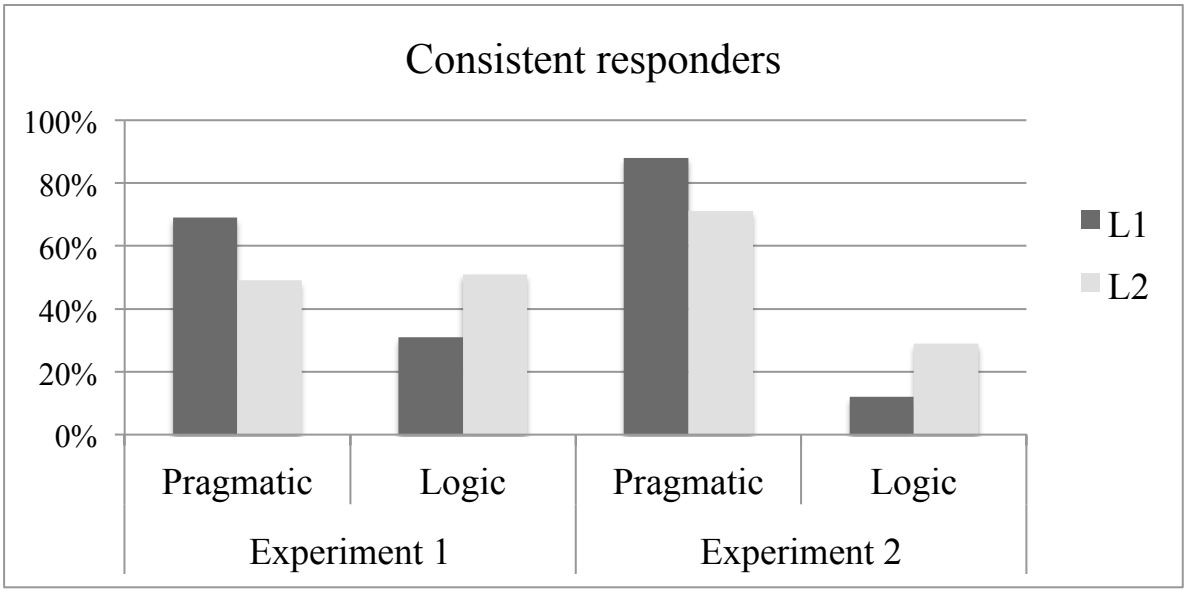


Fig 2. Percentage of participants that responded consistently – pragmatically or logically– in the L1 and L2 groups in Experiments 1 and 2.

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 – a more similar language to Italian on the quantifiers system. 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 consisted of a low-proficiency group (N = 8) with an A1-A2 level, a low-intermediate group (N = 29) with a B1 level, a high-intermediate group (N =12) with a B2 level, a low-advanced group (N = 9) with a C1 level and a high-advanced group (N = 3) with a C2 level. The Italian-Spanish bilinguals consisted of a low-intermediate group (N = 31) with a B1 level, a high-intermediate group (N = 50) with a B2 level and a low-advanced group (N = 5) with a C1 level. 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).

The L1 group and the two L2 groups' mean numbers of pragmatic interpretations of the underinformative sentences are given in Figure 1. A one-way between subjects ANOVA was

conducted to compare the effect of a L2 on the processing of scalar implicatures; there was a significant effect of group (L1 vs. $L2_{Eng}$ vs. $L2_{Spa}$) on the number of pragmatic answers ($F(2, 390) = 16.69$, $p < .001$). The mean number of pragmatic interpretation was 6.5 for the Italian L1 group, 4.8 for the English L2 group, and 5.1 for the Spanish L2 group. Post-hoc comparisons using the Tukey HSD test revealed that the number of pragmatic answers in the L1 group was significantly higher than in the L2 English group ($p < .001$) and L2 Spanish group ($p < .001$). There was no significant difference in the pragmatic answers given by the two L2 groups ($p = .74$). We divided L2 participants in three groups based on proficiency (i.e., A1-A2 = low proficiency, B1-B2 = medium proficiency, C1-C2 = high proficiency) and we compared each group's mean number of pragmatic answers for the two different L2s: there were no differences between the two L2 groups (low $proficiency_{Eng}$ – medium $proficiency_{Spa}$: $U = 279.5$, $Z = -.65$, $p = .51$; medium $proficiency_{Eng}$ – medium $proficiency_{Spa}$: $U = 1520.5$, $Z = -.77$, $p = .44$; high $proficiency_{Eng}$ – high $proficiency_{Spa}$: $U = 22.5$, $Z = -.80$, $p = .42$). Moreover, the differences between Experiment 1 and Experiment 2 in the L2 groups' mean numbers of pragmatic interpretations is not significant ($\chi^2 (1) = 2.53$, $p = .11$, $d = .23$). In the L2 groups, proficiency levels did not predict the number of pragmatic interpretations (low $proficiency_{Eng}$ – medium $proficiency_{Eng}$: $U = 159$, $Z = -.14$, $p = .89$; low $proficiency_{Eng}$ – high $proficiency_{Eng}$: $U = 40$, $Z = -.63$, $p = .53$; medium $proficiency_{Eng}$ – high $proficiency_{Eng}$: $U = 197.5$, $Z = -1.04$, $p = .30$; medium $proficiency_{Spa}$ – high $proficiency_{Spa}$: $U = 199$, $Z = -.07$, $p = .95$).

Individual participants were classified as consistent pragmatic, consistent logical or non-consistent responders. There were 27 (11%) non-consistent respondent in L1 and 37 (25,2%) in L2 (English and Spanish combined). The pragmatic responders were more frequent in the L1 group than in L2 group (78% vs. 53%, respectively), whereas logical responders were more frequent in the L2 group that in the L1 group (21.8% vs. 11%,

respectively), χ^2 (1) = 12.9, p = .0004, d = .40.

The L1 group exhibited a high level of accuracy in their responses to the other three types of sentences, with the average percent correct being 99.1% for the true sentences with *some*, 97.5% for the true sentences with *all*, and 98.6% for the false sentences with *all*. The L2 groups' accuracy was lower ($p < .001$). For the true sentences with *some*, the English L2 group's average percent correct was 93.6%, and the Spanish L2 group's average percent correct was 96.1%. The average percent correct for the true sentences with *all* was 87.1% for the English L2 group and 91.4% for the Spanish L2 group. The average percent correct for the false sentences with *all* was 95.1% for the English L2 group and 96.2% the Spanish L2 group.

4. General Discussion

In two experiments, 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, when participants were tested in their L2, they were less likely to derive a pragmatic interpretation of underinformative sentences than when they were tested in their L1. On the assumption that L2 oral processing, under time constraints, is more resource demanding than L1 processing, the decrease in pragmatic interpretations of underinformative sentences can be taken as evidence that deriving such pragmatic interpretations is not automatic.

The current pattern of results is consistent with the non-default models' view of scalar implicatures and runs against the alternative default models. According to the non-default models, 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").

Some empirical evidence relevant to the evaluation of this computational analysis can be found in Marty and Chemla (2013). They compared the number of pragmatic answers in underinformative sentences like 'Some snakes are reptiles' and in sentences with *only*, like 'Only some snakes are reptiles'. In the case with *only* the pragmatic interpretation is semantically imposed and should not require the extra pragmatic-enrichment step from the semantic interpretation to the pragmatic one. Results showed less pragmatic answers in the condition with 'some' compared to the one with 'only some'. According to the authors, these results are in line with the idea that the pragmatic interpretation is costly to make and that this cognitive cost might be related to the *decision* of deriving the implicature more than to the derivation *per se*.

Whilst our results point to a disadvantage of participants while they were tested in their L2 rather than L1, another recent study on bilinguals found no differences between bilinguals tested in L1 or in L2 (Dupuy et al., 2018). Considering the rate of pragmatic answers by bilinguals tested in the previous studies and in the current one, we see the following patterns of results: Slabakova found more pragmatic responses in L2 than in L1 (in Korean bilinguals); we found the opposite (in Italian bilinguals) and Dupuy et al's found similar rates in L1 and in L2 (in French bilinguals).

Dupuy and colleagues tested scalar-implicature processing in French adults learning English or Spanish as their L2 in two experiments. Participants performed a Truth-Value Judgment Task on control items (true all, false all and felicitous some) and on target items in which 'some' was used in an infelicitous way. Participants were tested in a within-subject design (they saw both the L1 and the L2 sentences) or in a between-subject design and in both cases the rates of pragmatic answers in the L1 and L2 conditions were similar.

There are two main reasons that may account for the discrepancy between our results and Dupuy et al.'s. First, our task was time constrained whereas their was not. Second, we assessed processing of oral rather than written stimuli. Thus, processing oral information using L2 under time constraints may have exacerbated the difficulty of the task, requiring more resources for performing the yes/no evaluation.

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. Furthermore, 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.

In order to explain the higher rates of pragmatic answers in L2 in Slabakova's work, another reason that we might take into consideration concerns the nature of participants: Slabakova's bilinguals were *immersed* in an L2 environment and they daily used their L2 more often than their L1. Thus, it is likely than the L2 processing in her bilinguals was as automatic as L1 processing. 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 scalar 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, Tretter, & Kinny, 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, Lavin, Thompson, & Ungerleider, 2010; Bialystok & Senman, 2004; Bialystok & Shapero, 2005; Kushalnagar, Hannay, & Hernandez, 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, Surian, Matsuo, Geraci, Iozzi, Okumura, & Itakura, 2010). Current results are consistent with findings showing that bilingual adults give more logical answers when tested in their L2 than in L1 (Costa, Foucart, Hayakawa, Aparici, Apesteguia, Heafner, & Keysar 2014, pp. 4-5). In other words, we suggest that the metacognitive advantages only show up in the case of high proficiency in L2.

Now, we compare our results 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 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 oral stimuli and the presence of time constraints) might help to make sense of the inconsistency among the available results. We suggest that highly proficient immersed bilinguals, because of their enhanced cognitive resources, show an increase in pragmatic responses to underinformative sentences, whereas bilinguals tested in their L1 environment without time constraints show no difference in pragmatic responses with respect to L1, because they have enough time to derive

the scalar implicatures. Finally, bilinguals tested in their L1 environment under time constraints show a decrease in pragmatic responses to underiformative sentences, because of limited resources that can be allocated to the computation of non-automatic responses.[5]

The present results do not allow deciding between competing non-default models on the computation of scalar implicatures, such as the Relevance Theory account, discussed in the introduction, and the Lexical account (Barner et al., 2011; Foppolo, Guasti, & Chierchia, 2012). This latter 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, e.g., Surian & Job, 1987). In this framework, the problems with scales - such as quantifiers – are a consequence of limitations in representing lexical items as members of a psychological 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 (e.g. Relevance Theory and Lexical accounts); a main focus should be on oral processing of scalar implicatures with participants

[5] We are greatful to Prof. Remo Job for helping us improving the paper with the proposal of this account to interpret our data.

immersed in a L2 environment, with or without time constraints.

In conclusion, our experiment brings new data to the study of scalar implicature computation in bilinguals by comparing L1 and L2 oral processing with time constraints. In both experiments, we found that pragmatic responses were more frequent in L1 than in L2 and this provides further support for non-default models of scalar implicature computation.

Acknowledgments

This work has been supported by grants from the Fondazione ONLUS Marica De Vincenzi. We are grateful to Prof. Kathleen Eberhard and Prof. Remo Job for reading the proofs and helping with improving the paper. We also thank Università degli Studi di Verona with the Linguistic Center (CLA) and all Professors for letting us recruiting and assessing their students.

Appendix 1 - Statements presented in the Sentence Evaluation Task

All true		
English (L2)	Spanish (L2)	Italian (L1)
All snakes are reptiles	Todas las serpientes son reptiles	Tutti i serpenti sono rettili
All cats are animals	Todos los gatos son animales	Tutti i gatti sono animali
All men are humans	Todos los hombres son personas	Tutti gli uomini sono persone
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
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)
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	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
Some salmons are fish	Algunos salmones son peces	Alcuni salmoni 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

Chapter 5.

The cost of scalar implicature: inference or infelicity?

A Reaction-Time and Eye-Tracking Study

This chapter is based on the following original article:

Mazzaggio, G., Reboul, A., Caretta, C., Darblade, M., Van der Henst, J., Cheylus, A., & Stateva, P. (2018). The cost of scalar implicature: inference or infelicity?

Manuscript in preparation.

Abstract

Whether there is a cost in deriving scalar implicatures (e.g., to interpret *some* as *some but not all*, instead of the logical interpretation *some and possibly all*) is still under a heated debate, with two principal accounts: the *neo-Gricean accounts* do not predict a cost for the pragmatic interpretation, while the *post-Gricean accounts* do. Since Bott & Noveck (2004), the debate has turned to the experimental field with a majority of works that seem to support the post-Gricean view. The present study addressed the topic of the cost of scalar-implicatures computation through three experiments. Experiment 1 ($N = 57$) had the goal of replicating Bott and Noveck's *Sentence Evaluation Task* in Experiment 3; we replicated their results, finding longer reaction times when participants interpreted pragmatically (i.e., disagree) underinformative sentences like 'Some elephants are mammals'. In Experiment 2 ($N = 58$), we obtained similar results, with the use of pictures and sentences containing pseudo-words (e.g., *Some blicks are mammals*), in order to have identical sentences in felicitous and infelicitous *some* conditions. With this experiment we excluded the possibility that a greater cost is due to greater difficulty in moving up in conceptual hierarchies, compared to moving down. Experiment 3 ($N = 54$) was a Sentence Evaluation task with a context; we used recorded self-paced reading, RTs, and eye-tracking techniques to look into the process of computation more in depth. Results showed that when there is support from the context, the cost of computation disappears. Overall, our results seem to support a post-gricean view of scalar-implicature computation. We propose possible future directions of these studies.

1. Introduction

Grice (1975; 1989) introduced the notion of an implicature from examples such as the following:

(1) The pianist played some Mozart sonatas.

Grice noted that *some* can either be interpreted semantically (i.e., logically), in which case it means *some and possibly all*, or pragmatically, in which case it means *some but not all*. This provides (1) with two different interpretations:

(2) The pianist played *some and possibly all* Mozart sonatas. [semantic interpretation]

(3) The pianist played *some but not all* Mozart sonatas. [pragmatic interpretation]

Implicatures like (1) (interpreted as in (3)) are called *scalar implicatures*. One of the most central debates in semantics and pragmatics for the past thirty to forty years has concerned how one goes from (1) to (3). There are two main families of theories, those that see the process as essentially linguistic (with two versions: lexical (see Levinson 2000) and grammatical (see Chierchia 2013)) and those that see it as essentially pragmatic and context-based (see Sperber & Wilson, 1995; Carston, 2002). The first type of theories is called *neo-Gricean*, while the second is called *post-Gricean*. Beyond the theoretical debate, since the beginning of the twenty-first century and spurted by pioneering work by Noveck (2001) and Bott and Noveck (2004), the debate has shifted to the experimental field and centered on the notion of *cost*. Very roughly, the neo-Gricean accounts predict that the semantic interpretation should be costlier than the pragmatic interpretation, while the post-Gricean accounts predict the reverse.

The present study tries to address the problem of the cost of scalar implicatures. The cost of scalar implicatures was first evidenced in a paper by Bott and Noveck (2004) where they carried out four experiments to test the cognitive cost of deriving the implicature. In all four experiments, they used categorical sentences, targeting *all* (true, false, absurd) and *some* (infelicitous, felicitous, absurd). The main measures were made on the infelicitous *some* condition, of which an example is "Some elephants are mammals". Participants were given the choice between two responses, *true* and *false*. The *true* response corresponds to the semantic interpretation of the sentence (*Some and maybe all elephants are mammals*), while the *false* response corresponds to the pragmatic interpretation of the sentence (*Some but not all elephants are mammals*). The two measures were the rate of pragmatic versus semantic responses and the reaction times (RTs) between the presentation of the sentence and the response.

We would like to outline that one of the main problem when using a sentence evaluation task regarding scalar implicatures is that, short of giving participants instructions on how to interpret *some,* there is no "correct" answer in the infelicitous *some* condition. The pragmatic and the semantic answers are both correct. This is why experimentalists have ignored the felicitous *some* condition. There, while the majority of participants evaluate the sentence positively, there is no way of knowing whether the participants choose the pragmatic interpretation (*Some but not all mammals are elephants*) or the semantic interpretation (*Some and possibly all mammals are elephants*). Both are indeed true. It is only in the infelicitous *some* condition that, while both *true* and *false* answers are correct, the pragmatic interpretation clearly corresponds to *false* while the semantic interpretation clearly corresponds to *true*.

In the first experiment, participants completed two different experimental sessions. In one of them, they were instructed to interpret *some* as *some and possibly all* (i.e., to go for the

semantic interpretation), and in the other to interpret *some* as *some but not all* (i.e., to go for the pragmatic interpretation). The results were interpreted relative to the instructions, i.e, relative to whether the participants had followed the instruction for the given experimental session: correct responses for the pragmatic-interpretation experimental session were *false*, while correct responses for the semantic-interpretation experimental session were *true*. The authors compared the rate of correct answers in the pragmatic condition (60%) and in the semantic condition (90%), and the difference was significant. There was also a difference in RTs: to answer correctly in the pragmatic condition took significantly longer than to answer correctly in the semantic condition. Additionally, in both conditions, the rate of correct answers and the RTs for the control conditions (felicitous *some*, absurd *some*, true *all*, false *all,* absurd *all*) were similar to the correct answers to infelicitous *some* in the semantic condition and significantly different from those in the pragmatic condition. This first result seems to support the notion that deriving scalar implicatures is a costly cognitive process as, even when instructed to do so, participants have more difficulty sticking to the instructions and take longer than when instructed to interpret *some* semantically. In the second experiment, Bott and Noveck addressed whether the results of Experiment 1 could be explained by a response bias toward *true* and were able to show that this was not the case.

Experiment 3 is the experiment we are especially interested in here. Participants were presented with the same sentences as before and again had to say whether the sentences were *true* or *false*, but they received no instruction as to how to interpret *some*. As before, the measures were rates of pragmatic vs. semantic answers and RTs. The comparison was made between *false* (pragmatic) and *true* (semantic) answers in the infelicitous *some* condition for participants who gave both types of responses (9 participants who responded exclusively pragmatically or semantically were excluded). Pragmatic answers in the infelicitous *some* condition took significantly longer than the semantic answers; additionally, they also took

significantly longer than all control conditions excepting the true *all* condition. By contrast, RTs for the semantic answer were not significantly different from those for the control conditions. Finally, the rate of pragmatic answers in the infelicitous *some* condition was 60%.

The fourth Experiment was intended to support the notion that deriving scalar implicatures is costly by making participants give their responses under time constraints. Participants were divided into two groups, one of which was instructed (and trained) to answer in short time lag (900 ms; Short condition) and the other of which was instructed and trained to answer in a long time lag (3s; Long condition). The comparison was between the rates of semantic vs. pragmatic answers in the Short and in the Long conditions. Participants in the Short condition gave 72% of semantic answers, significantly more than they did in the Long condition (57%). Again, this result, showing that time constraints directly impact the rate of semantic vs. pragmatic answers, argues for a specific cost in deriving scalar implicatures.

Bott and Noveck's paper was extremely influential in the debate between neo- and post-Gricean accounts of scalar implicatures and was interpreted as supporting the post-Gricean accounts. It was nevertheless the object of intense controversies, which are still active today. Part of the controversies lies in alternative ways to test cost, which have given partially contradictory results. To begin with those that argue for cost, De Neys and Schaeken (2007), using an interference paradigm where the same task as in Bott and Noveck's third Experiment was performed under a visual memory load, were able to show that in the load condition, the rate of pragmatic answers was significantly lower than in the control condition. De Neys and Schaeken also compared RTs for pragmatic answers between the load and the control conditions and found that they were significantly longer in the load than in the control condition. There was no such impact on the RTs for semantic answers. Both of these results reinforce the notion that deriving scalar implicatures is cognitively more costly than giving a

semantic interpretation. Another alternative test for cost used the visual world paradigm. Participants had to choose, among four candidate pictures (one corresponding to the pragmatic interpretation, one to the semantic interpretation, two distractors), which best corresponded to the sentence. The measures were both the rates of pragmatic interpretations and RTs. Here, the results are mixed, with some studies showing no significant differences between the pragmatic and the semantic interpretations (Grodner, Klein, Carbary, & Tanenhaus, 2010; Breheny, Ferguson, & Katsos, 2013a; 2013b; Foppolo & Marelli, 2017) and some showing significantly longer RTs for pragmatic interpretations (Huang & Snedecker, 2009a). But the main questions have come from queries as to how to interpret Bott and Noveck's results. Indeed alternative interpretations have been proposed.

First of all, based not on Bott and Noveck (2004), but on Noveck (2001), Guasti, Chierchia, Crain, Foppolo, Gualmini and Meroni (2005) have argued that some positive answers in his third experiment might in fact correspond to pragmatic rather than to semantic interpretations. In that experiment, which tested both children and adults, Noveck used semi-categorical sentences, e.g., *Some elephants have trunks.* As Guasti et al. noted, some participants might have constructed a complement set of trunkless elephants, thus verifying the sentence with a pragmatic interpretation of *some*. While it is of course much less clear whether participants in Bott and Noveck's experiments, where categorical sentences were used, could have built complement sets of non-mammal elephants, it cannot be excluded, and if this is the case, some apparently semantic responses in the infelicitous *some* conditions of their four experiments might turn out to correspond to pragmatic felicitous interpretations.

Additionally, while Bott and Noveck interpreted their results in terms of a higher cognitive cost induced specifically for the derivation of the implicature, other interpretations are possible. For instance, participants who give pragmatic answers might first try and fail to build complement sets of the kind described above in the infelicitous *some* condition, leading

to longer RTs. It might also be the case that the delay in the production of pragmatic sentences is due to the difficulty of understanding upper-bound sentences, leading to a trade-off between speed and accuracy. The idea is that the pragmatic interpretation would be accessed quickly but that ascertaining the truth-value of the sentence would be a time-consuming procedure.

These hypotheses have been investigated in two studies, the first one (Bott, Bailey, & Grodner, 2012) using a very similar sentence verification task with categorical sentences to that used in Bott and Noveck (2004). Bott et al. used a speed accuracy trade-off procedure, allowing to separate speed and accuracy (truth value assessment). In their three experiments, participants had to respond when they heard an auditory cue at different time lags after the presentation of the sentence. In the first experiment, participants were separated into two groups following the same procedure as in Bott and Noveck's first two experiments, thus allowing to distinguish correct and incorrect answers to the infelicitous *some* condition. This allowed Bott et al. to compute the intercept (the earliest point at which the rate of correct answers in the infelicitous *some* condition departed from chance). The results showed that the intercept occurs earlier in the semantic than in the pragmatic condition, suggesting that the longer RTs for pragmatic interpretation are not a speed-accuracy trade-off. In Experiment 2, Bott et al. compared infelicitous *some* with an explicit formulation of the pragmatic interpretation (i.e., *only some elephants are mammals*). Both types of sentences are equally complex in terms of meaning, but it is only the *some* sentences in which the pragmatic interpretation is optional. The same procedure as in Experiment 1 was used. Again, the intercept for *some* in the pragmatic condition was delayed relative to the intercept for false *only some*. In the third experiment, Bott et al. compared *some* in the semantic condition to its explicit formulation, i.e. *at least some*. While the results for this third experiment are less clear (suggesting that it might be the interpretation of *some* that is costly), a comparison

between the results of Experiment 3 and those of Experiment 2 shows that the difference in Experiment 2 is significantly larger than the difference in Experiment 3. This supports the conclusion that the pragmatic interpretation of *some* is more costly than its semantic interpretation.

Another paper, using a different paradigm, based on the interference task used by De Neys and Schaeken (2007), addresses again the accuracy problem, i.e., the possibility that the longer RTs for the pragmatic interpretation might be due to checking the truth-value of the pragmatic interpretation, rather than accessing the pragmatic interpretation *per se*. Marty and Chemla (2013) compared the rate of *false* responses for infelicitous *some* and to false *only some* categorical sentences in two conditions of visual memory interference: high load and low load. They found that the rate at which participants produced the *false* answer for *only some* was not significantly different in the two conditions. By contrast, it was significantly different for the infelicitous *some* sentences. Marty and Chemla, following the alternative-based account proposed by, e.g., Chierchia (2013), propose a two-step account of the derivation of scalar implicatures: the first step is a decision procedure on whether or not to derive the implicature; the second step corresponds to the generation of a set of alternatives and to the exclusion of those alternatives. While the interpretation of *some* and the interpretation of *only some* both involve the second step, it is only the interpretation of *some* that also involves the first step. On this basis, Marty and Chemla interpret their results as showing that the cost of deriving the implicature lies only in the first, decisional, step, and not in the second, interpretative step.

This, however, suggests that the problem might lie in the infelicity of the crucial experimental conditions in all these studies using sentence evaluation tasks, rather than in anything else. This, by the way, seems to agree with the results of studies using contexts that encourage the derivation of the implicature by making relevant the pragmatic interpretation in

infelicitous *some* sentences (see Dupuy et al. 2016). Such a view is also encouraged by Bott and Noveck's (2004) Experiment 3. As explained above, authors generally focus on the infelicitous *some* conditions and ignore felicitous *some* conditions, that are considered as control conditions, on a par with true and false *all*. However, this is a debatable decision: while it is true that it is only in the infelicitous *some* condition that the pragmatic answers can be distinguished from the semantic answers (with the proviso indicated above that some positive answers might correspond to pragmatic felicitous interpretations), it might still be interesting to look at RTs in the felicitous *some* condition. So, let us return to the results of Bott and Noveck's third experiment, and more specifically to the RTs for the different conditions:

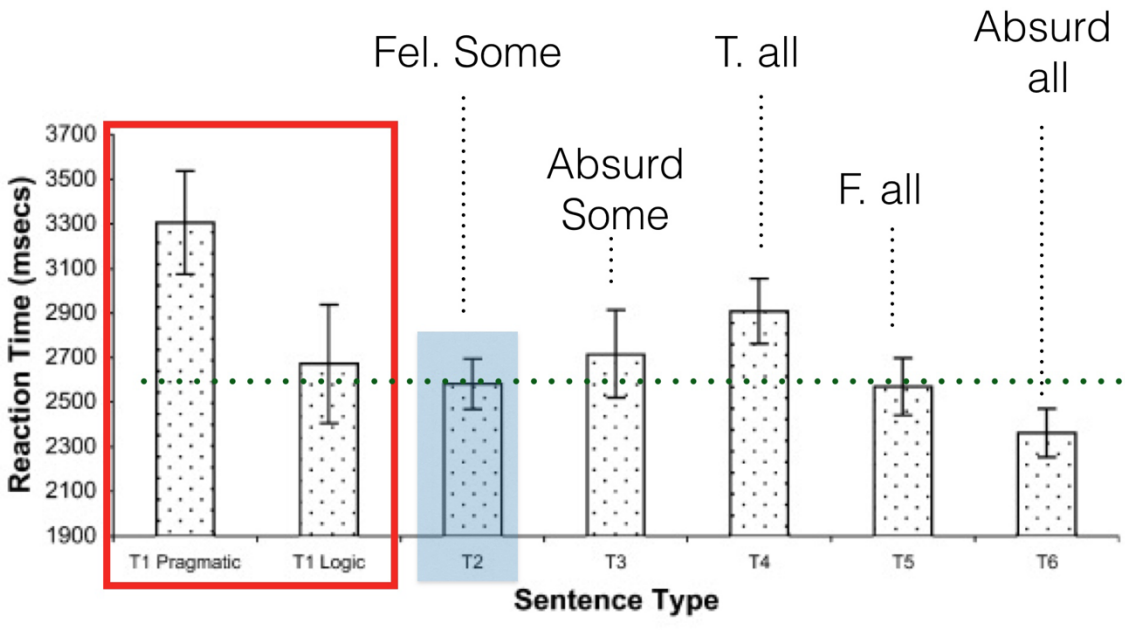


Figure 1: RTs in Bott and Noveck's Experiment 3

(Answers to the infelicitous *some* condition are indicated in the red box)

As can be seen, responses to sentences of type T2, corresponding to felicitous *some*, are not only significantly faster than the pragmatic interpretations for infelicitous *some*, they are as fast (and even slightly faster) than the semantic interpretations in that condition. Remember that felicitous *some* sentences are sentences such as *Some mammals are elephants*. This makes it slightly difficult to suppose that participants would choose **not** to interpret *some* as *some but not all*. While it is possible that some participants interpret felicitous *some* sentences semantically, it is unlikely that all participants do.

But, on the reasonable assumption that some participants at least interpret *some* pragmatically in the felicitous *some* condition, this leaves us with something of a mystery: why are responses to felicitous *some* so much faster than pragmatic answers in the infelicitous *some* condition? Here, there are two answers, given what we saw above. The first one, following Marty and Chemla, is that the decision to draw the implicature is much easier to take in the felicitous *some* condition. A second possibility is that it is indeed, as suggested indirectly by Guasti et al. (2005), assessing the truth-value of the pragmatic interpretation in the infelicitous *some* condition that is more costly. After all, constructing a complement set (which is what verifying the pragmatic interpretation comes to) in the felicitous *some* condition is very easy while it is difficult, if not impossible, in the infelicitous condition. It should be noted that these two explanations are perfectly compatible. Indeed, regarding the decision step postulated by Marty and Chemla, it seems again easier to decide to draw the implicature in the felicitous than in the infelicitous condition. But then one would have to conclude that it is the very infelicity of the condition that explains the longer RTs, rather than the decision process, the derivation of the implicature, or its verification. If this is the case, it might be a good idea to abandon sentence evaluation task for experiments on scalar implicatures. An additional indication that this might be a good idea is that in most of the sentence evaluation task studies (with the intriguing exception of that by De Neys &

Schaeken), the rate of pragmatic vs. semantic answers in the infelicitous *some* condition is not significantly different from chance, hovering between 40 and 60%.

2. The present study

In the present study, we will present three experiments that aim at clarifying the question of the cost of implicatures. The first experiment is basically a replication of Bott and Noveck's (2004) Experiment 3. To our knowledge it has never been replicated before. We aim at replicating it, thus showing that there is a cost to the pragmatic answer in the infelicitous *some* condition, while no such cost arises in the felicitous condition. We made a few minor changes that are detailed in the *Material and Procedure* sections in *Section 3* below.

The second experiment aims at avoiding two problems with Bott and Noveck's experimental material. First, as noted by Bott et al. (2012), the infelicitous *some* sentences all had subjects combining *some* with a noun for a basic category (e.g., *some elephants*). Equally the felicitous *some* sentences all had subjects combining *some* with a noun for a super-ordinate category (e.g., *some mammals*). This would facilitate the adoption of a systematic semantic or pragmatic strategy in the infelicitous condition by allowing participants to identify the condition from the start of the sentence. To avoid this problem, Bott et al. added another type of sentence, using subjects referring to basic categories with predicates referring to subordinate categories, e.g., *Some elephants are African*. However, there is an additional worry with Bott and Noveck's experimental material that this addition does not solve. It might be the case that it is easier to go down in a conceptual hierarchy (from superordinate categories such as *mammals* to basic categories such as *elephants*) rather than to go up in a conceptual hierarchy (from basic to super-ordinate category). This worry receives some tentative support from the fact that in Bott and Noveck's third experiment, the only control condition for which RTs were not significantly shorter than for the pragmatic answers in the

infelicitous *some* condition was the true *all* condition, which corresponds to sentences such as *All elephants are mammals*, going in the same upward direction. Experiment 2 is designed to eschew these difficulties, using pseudo-words referring to artificial categories and allowing us to use identical sentences in felicitous and infelicitous *some* conditions.

Experiment 3 aims at looking more acutely into the process through which participants give pragmatic answers in the infelicitous *some* condition. It uses a sentence-evaluation task, but the sentences are not categorical sentences. Participants are presented with a picture and a sentence, where the picture verifies or falsifies the sentence. In the infelicitous *some* condition, the picture verifies the semantic interpretation and falsifies the pragmatic interpretation. Participants are asked to say whether they think that the sentence is a good description of the picture by pressing one of two keys (corresponding respectively to *agree* and *disagree*). Experiment 3 recorded self-paced reading, RTs, and used eye tracking.

3. Experiment 1

3.1. Participants

Fifty-seven participants (34 females, mean age = 23.16, *sd* = 3.23) took part to our experiment[6]. All participants were native French speakers.

3.2. Material

While this experiment is intended as a replication of Bott and Noveck's (2004) third experiment, and while we mainly used the same materials of the original paper, i.e. we used categorical sentences of the type in Table 1 and asked participants to provide Agree/Disagree

[6] This study was carried out in accordance with the recommendations of the Comité de Protection des Personnes Sud Est II, who gave it its agreement (IRB number: 11263). All participants gave written informed consent in accordance with the Declaration of Helsinki.

judgments, there are nevertheless a few important modifications that we describe below. As in the original experiment, participants could be presented with sentences that were blatantly true (All-True, Some-True and Only-Some true conditions), blatantly false (All-False condition) or underinformative (Some-Underinformative condition). We decided not to assess participants in the Some-False and in the absurd condition because we believed that those categories were not useful for our purpose. Furthermore, we added the Only-Some True condition that we considered interesting in light of Marty and Chemla (2013).

Sentences were of the form "Only some/Some/All Xs are Ys", where Xs and Ys are superordinate/subordinate elements of a natural category: for example, considering the "mammal" category (superordinate) we had 9 subordinate exemplars, like the elephants' one. All categories and members of each category are described in *Appendix 1*. We also decided not to use the "Shellfish" category, originally present in Bott and Noveck, and to previously test all the exemplars of the other categories to check whether they were easy recognizable as part of the corresponding superordinate category. We changed some of the items that had not been easily recognized with others more common. Moreover, in the original paper, Bott and Noveck asked whether the sentence was True/False but since the underinformative sentence is neither "true" nor "false" *per se*, it is preferable to use a Sentence Evaluation Task than a Truth Value Judgement Task. Thus, participants had to choose between Agree and Disagree.

Condition	Example Sentence	Expected Answer
All-True	All elephants are mammals	Agree
All-False	All mammals are elephants	Disagree
Some-True	Some mammals are elephants	Agree

Only Some-True	Only some mammals are elephants	Agree
Some-Underinformative	Some elephants are mammals	Agree (Logic); Disagree (Pragmatic)

Table 1. Typologies, examples and expected answers of the sentences used in Experiments 1.

3.3. Procedure

As in Bott and Noveck, participants were presented with 9 examples of 5 kinds of sentences so each participant saw 45 experimental items. The stimuli were randomly generated in the same way as the original paper, i.e. from a base of 5 categories and 9 exemplars from each of these categories and with each category-exemplar used just once during the experiment. 16 training sentences (4 for each category: All-true, All-false, Some-True and Some-Underinformative) and 5 dummy sentences were presented at the beginning of the experimental session to train participants.

Participants sat in front of a computer in a quiet room. At the beginning instructions and the training part were presented: each participant then could decide when to start the experimental session by pressing the space bar on the keyboard. Before the presentation of each sentence participants saw a fixation point on the screen. Words then appeared consecutively onto the screen with a gap of 240 ms between each word; then the sentence remained on the screen until participants gave their response using the computer keyboard. They didn't receive any feedback. Response times were measured from the beginning of the sentence.

3.4. Results

As in Bott and Noveck, we analyze results of choice proportions and RTs. Table 2 shows the mean of correct answers for the control items and the mean of logical answers in the underinformative condition. Percentages of correct answers for the All true, All false, Some true and Only some true conditions are all above 91%. In the underinformative *some* condition, participants gave more pragmatic answers (66.7%) than logical answers (33.3%) (Wilcoxon signed rank: $V = 402$, $p = 0.002$).

Condition	Example Sentence	Mean	*sd*
All true	All elephants are mammals.	91.6	11.6
All false	All mammals are elephants.	96.1	6.5
Some true	Some mammals are elephants.	94.9	9.6
Only some true	Only some mammals are elephants.	92	11.1
Underinformative some	Some elephants are mammals.	33.3	34.8

Table 2. Mean and standard deviation (*sd*) of correct answers; for the underinformative sentences we considered logical correctness (i.e., agree).

To look for the difference in RTs between the logical answer and the pragmatic answer, we analyzed data of participants that had both logical and pragmatic answers in the underinformative condition (N = 32 subjects). Figure 2 shows overall results for RTs in all conditions. As in Bott and Noveck, we divided participant's answers to underinformative sentences into logical answer or pragmatic answer, in order to assess whether one kind of answer is faster than the other. Particularly, in the original paper the logical answers were faster than the pragmatic ones. A log-linear mixed model analysis on 54 participants (we

considered just correct responses for which RTs were > 0.5 s and < 10 s) with items and participants as random variables and the combination of condition and answer as fixed factor confirmed the significant difference in terms of RTs between the faster logical answer and the pragmatic answer ($z = 2.74$, $p = .006$). There is no difference in terms of RTs between the logical answer in the underinformative condition and the Only some condition ($z = -1.41$, $p = .16$). There are also no differences in terms of RTs between the logical answer in the underinformative condition and the Some true condition ($z = -.86$, $p = .39$) and the All true condition ($z = 1.77$, $p = .08$). The All false condition is slightly faster than the logical answer in the underinformative condition ($z = 2.02$, $p = .04$). The pragmatic answer in the underinformative condition takes longer compared to the Some true ($z = 2.70$, $p = .007$), Only some true ($z = 1.98$, $p = .05$), All true ($z = 6.04$, $p < .001$) and All false ($z = 6.42$, $p < .001$) conditions.

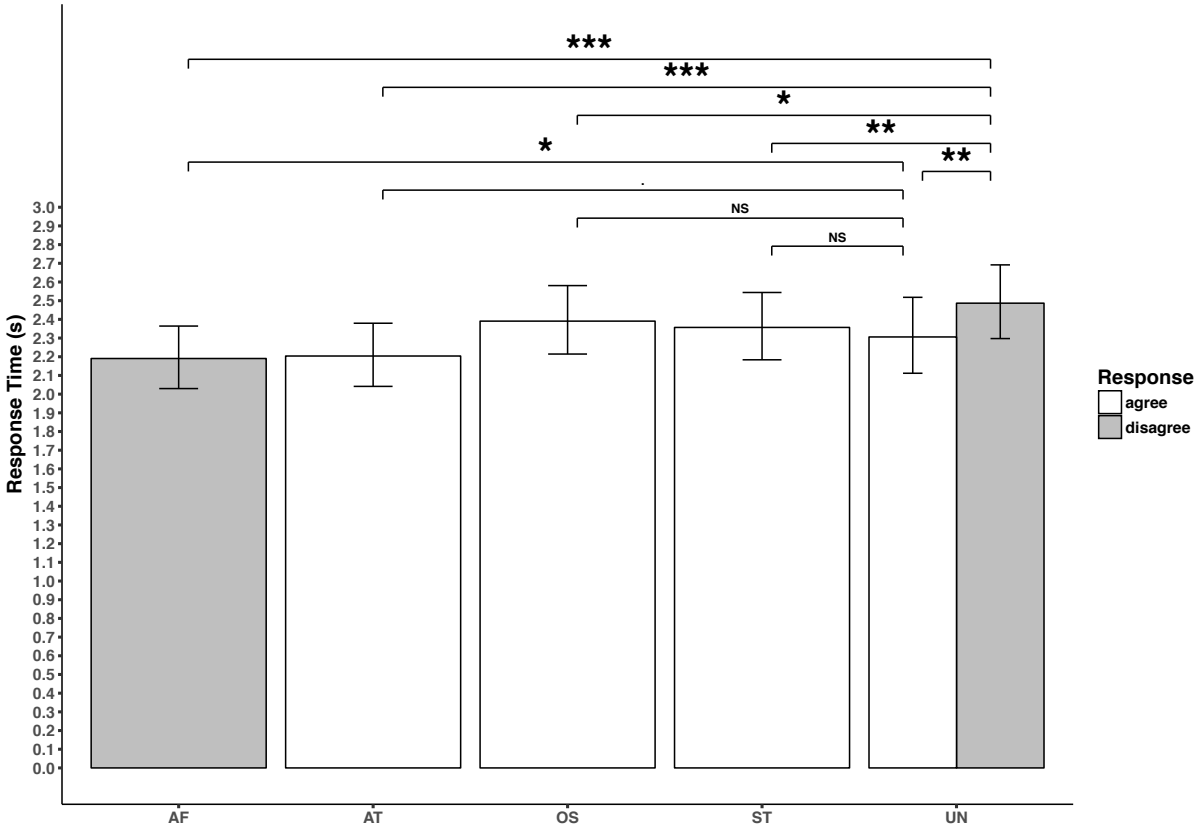


Figure 2. Response time as a function of condition and answer. Values were estimated with a log-linear mixed model analysis on 54 participants, correct responses for which RTs were > 0.5 s and < 10 s, items and participants as random variables and the combination of condition and answer as fixed factor. Conditions were abbreviated as follows: AF: all false; AT: all true; OS: only some; ST: some true; UN: underinformative. UN agree is compatible with a logical interpretation while UN disagree is compatible with a pragmatic interpretation. Error bars stand for 95% confidence intervals.

3.5. Discussion

This study was intended as a replication of Bott and Noveck's (2004) third experiment, with a few modifications (see above, section *Materials*). Results replicate those of the original work. Considering the critical underinformative condition (e.g., *Some elephants are mammals*), participants gave significantly more pragmatic (disagree) than logical answers (agree). Moreover, to answer pragmatically required significantly more time (longer RTs) than to answer logically, and more time than to answer other control conditions (Some true, Only some true, All false, All true). By contrast, when participants answered logically we recorded similar RTs compared to the control conditions (Some true, Only some true, All true). In conclusion, with this first experiment we confirmed that to derive a scalar implicature is costly, as Bott and Noveck did.

4. Experiment 2

4.1. Participants

Fifty-eight participants (22 males, mean age = 23.14, *sd* = 3.20) took part to our experiment.[7] All participants were native French speakers.

4.2. Material

It was decided to compare two groups who would see the same categorical sentences, but relative to two different sets of images. This allowed the introduction of artificial categories and of using the same sentences in the *some* conditions, one of which would be true in the pragmatic interpretation (in, e.g., Group 1), while the other would be false in the pragmatic interpretation (in, e.g., Group 2). Images introduced the artificial categories by presenting 10 silhouettes of animals and/or plants (six belonging to a given basic category, four to another basic category), which would or not fall under the same superordinate category. All silhouettes were tested to check that they were recognizable. The image was accompanied by a sentence of the form "Look! Here are Xs" (where Xs was a pseudo-word in French, supposed to correspond to the composite artificial category represented on the picture). This was followed by the test sentence.

Here is an example of an image that would verify the pragmatic interpretation for the sentence "Some Xs are mammals":

[7] This study was carried out in accordance with the recommendations of the Comité de Protection des Personnes Sud Est II, who gave it its agreement (IRB number: 11263). All participants gave written informed consent in accordance with the Declaration of Helsinki.

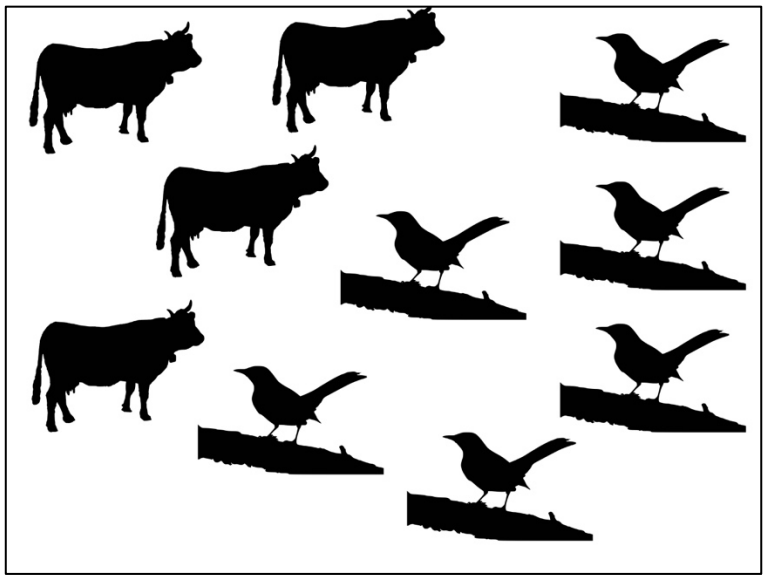


Figure 3: Picture verifying the pragmatic interpretation

If the picture above was presented to Group 1, Group 2 would see the picture below that falsifies the pragmatic interpretation:

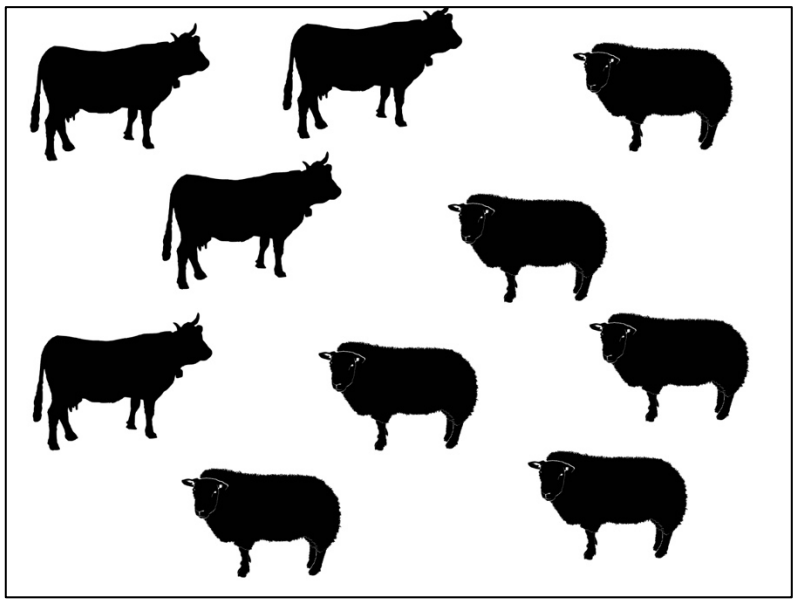


Figure 4: Image falsifying the pragmatic interpretation

Each group saw one version of the picture, which was presented five different times, coupled with the following types of sentences:

Some blicks are mammals. (test/control)

Only some blicks are mammals. (test/control)

Some mammals are blicks. (filler)

All blicks are mammals. (control)

All mammals are blicks. (filler)

There were eight different sets of picture-sentences. Each participant saw 40 such picture-sentence items and each group saw an equal number of true/false items (respectively of infelicitous/felicitous *some* items).

4.3. Procedure

Participants sat in front of a computer in a quiet room. At the beginning instructions and the training part were presented. Participants saw 8 training sentences with both some and all in which the used categories (furniture, vehicles) and pseudo-words were different from the experimental session's ones. Each participant then could decide when to start the experimental session by pressing the space bar on the keyboard. The task was a sentence evaluation task. At the beginning the image appeared on the screen for 1 second and then a sentence introducing the new category was presented (e.g., *Look, here are Xs*) for 1 second. After that, the target sentence appeared progressively (word by word) on the screen, with every new word appearing after 240ms and remaining on the screen. After the last word appeared participants read the question that stayed until they gave their response using the computer keyboard. They didn't receive any feedback. Response times were measured from the appearance of the first word of the target sentence.

4.4. Results

We decided to exclude from the analysis 16 participants: 1 because the participant asked a question during the trial invalidating RTs data, 11 that made three errors or more in the All

true and All False control conditions and 4 more to restore an equal number of 21 participants in each group. As in Experiment 1, we analyzed results of choice proportions and RTs. In Table 3 we present the mean of correct answers for the control items and the mean of logical answers in the underinformative condition. It seems that some difficulties had been experienced with the Only some false condition (70.8%). In the underinformative condition participants gave more logical answers (62.5%) than pragmatic answers (37.5%) (Wilcoxon signed rank $V = 358.5$, $p = .025$).

Condition	Example Sentence	Mean	*sd*
All true	All blicks are mammals.	91.1	13.3
All false	All blicks are mammals.	92.3	12.9
All reverse false	All mammals are blicks.	56.8	30.9
Some true	Some blicks are mammals.	92.9	15.9
Some reverse true	Some mammals are blicks.	79.2	26.3
Only some true	Only some blicks are mammals.	90.5	16.5
Only some false	Only some blicks are mammals.	70.8	31.7
Underinformative some	Some blicks are mammals.	62.5	35

Table 3. Mean and standard deviation (*sd*) of correct answers; for the underinformative sentences we considered logical correctness (i.e., agree).

Further analyses focused on the difference in RTs between answers. To look at differences in RTs between the logical and the pragmatic answers in underinfomative sentences, we ran a log-linear mixed model analysis on 42 participants (we considered just correct responses for

which RTs were > 0.5 and < 10) with items and participants as random variables and the combination of condition and answer as fixed factor (Figure 5). The analysis revealed a significant difference in terms of RTs between the logical answer and the pragmatic answer (z = 3.61, p = .002), with longer RTs for participants that answered pragmatically. Indeed, RTs for pragmatic answers are also significantly longer than RTs for all the other conditions, that are the All true (z = 4.49, p < .001), All false (z = 5.80, p < .001), All reverse false (z = 4.07, p < .001), Some true (z = 4.82, p < .001), Only some true (z = 3.27, p = .004) and Some reverse true (z = 4.01, p < .001) conditions, except for the Only some false condition (z = 1.84, p =.13). Differently, RTs for logical answers in the underinformative condition do not differ from RTs for the other All true (z = .89, p =.46), All reverse false (z = .29, p =.79), Only some true (z = -.51, p = .68), Only some false (z = -2.01, p =.09), Some true (z = -1.27, p =.30) and Some reverse true (z = -.05, p =.95) conditions, except for the All false condition which is faster (z = 2.41, p = .04).

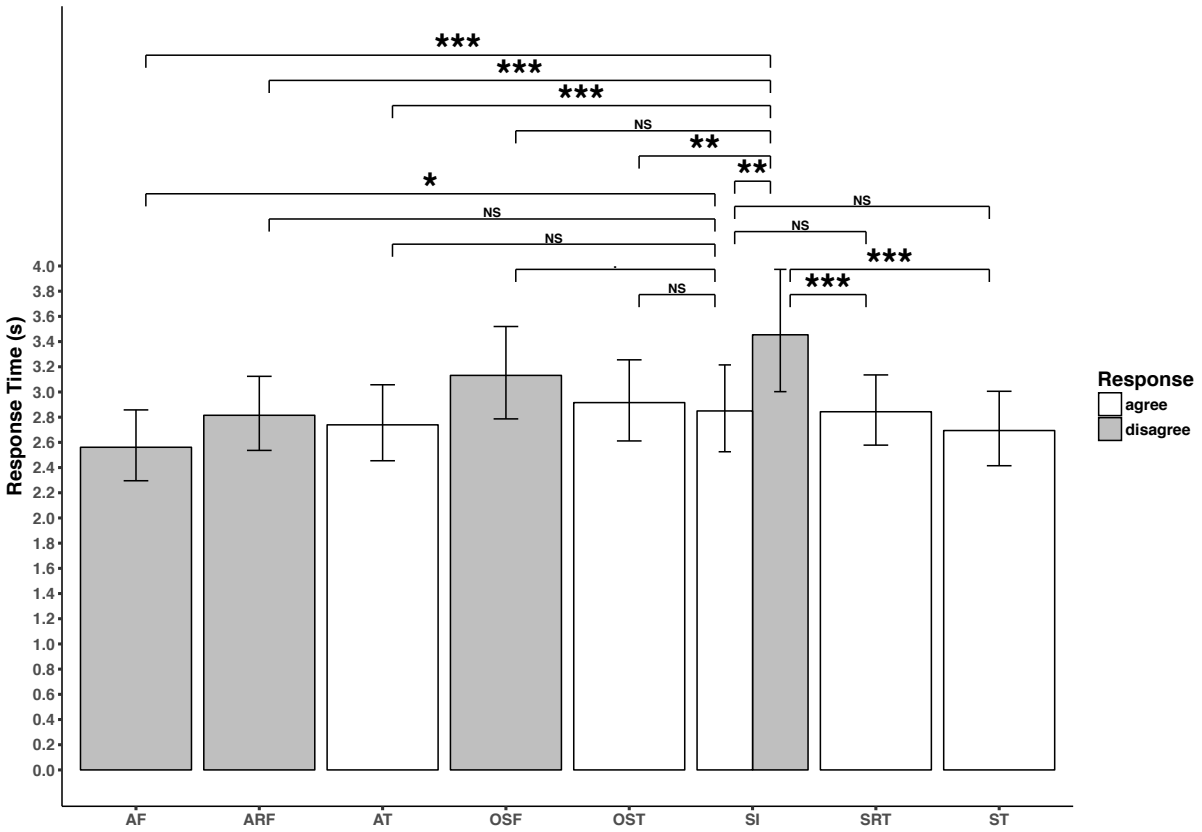


Figure 5. Response time as a function of condition and answer. Values were estimated with a log-linear mixed model analysis on 42 participants, with correct responses for which RTs were > 0.5 s and < 10 s, with items and participants as random variables and the combination of condition and answer as fixed factor. Conditions were abbreviated as follows: AF: all false; ARF: all reverse false; AT: all true; OSF: only some false; OST: only some true; SI: some infelicitous; SRT: some reverse true; ST: some true. SI agree is compatible with a logical interpretation while SI disagree is compatible with a pragmatic interpretation. Error bars stand for 95% confidence intervals.

4.5. Discussion

Considering the longer RTs when participants answer sentences like “Some elephants are mammals” compared to “Some mammals are elephants”, with this second experiment we had one main goal: to establish whether the cost associated with the computation of scalar

implicatures may be due to a greater difficulty in moving up in conceptual hierarchies (e.g., elephants > mammals), than in moving down (mammals > elephants). Thanks to the use of pseudo-words we created a Sentence Evaluation Task in which the Some-sentences were the same in the true and underinformative conditions (e.g. *Some blicks are mammals*). Again, results show a cost when computing pragmatic interpretations in underinformative sentences, compared to both semantic interpretation in underinformative sentences and some-true sentences; thus, we excluded the possibility that such cost is due to discrepancy in conceptual hierarchy's movement. Interestingly, and differently from Bott & Noveck's and from the first experiment's results, we found more logical answers than pragmatic answers in the underinformative sentences. This might be related with the difficulty of our task, a cognitively effortful one that probably led to a decrease of pragmatic answers.

5. Experiment 3

5.1. Introduction

In the first two experiments, we have found results similar to those found by Bott and Noveck (2004) in their third experiment. In other words, the RT is significantly longer for the pragmatic answer (*agree*) than for the semantic answer (*disagree*) in the infelicitous *some* condition. This argues for the notion that there is indeed a specific cost for the pragmatic interpretation in the infelicitous *some* condition. However, it has nothing to say about why this should be the case. Thus far, hypotheses relative to the cost of the pragmatic interpretation of *some* in the infelicitous condition can be divided depending on whether they take into account the infelicity itself or ignore it.

The only hypothesis on the side that ignores the infelicity is the post-Gricean interpretation proposed by Bott and Noveck (2004), and Noveck and Sperber (2007),

according to which the cost is due to the inferential process necessary to draw the interpretation. All other hypotheses factor in the infelicity itself as the major reason for the cost. According to Guasti et al. (2005), the cost of the pragmatic answer does not lie in accessing the pragmatic interpretation, but in verifying it: this entails constructing a complement set, which, in the infelicitous condition (but not in the felicitous one) is difficult, if not impossible. The next suggestion comes from Marty and Chemla (2013). They suggest that, in the infelicitous condition, it is not the interpretative process of deriving the pragmatic interpretation that is costly, but the decision to choose the pragmatic interpretation. While in the felicitous *some* condition, the decision is easy (as the pragmatic interpretation is obviously verified), it is much more difficult to make the decision in the infelicitous *some* condition (where it is not). Finally, and this is more or less a variant of Marty and Chemla's position, one could argue that interlocutors operate according to a Principle of Charity (see Davidson 1974;1984), which enjoins them to choose an interpretation that maximizes speaker's rationality[8]. Clearly, such a Principle of Charity would be difficult to reconcile with the pragmatic interpretation in the infelicitous *some* condition, while it is compatible with the logical interpretation.

We decided to test these hypotheses, using a sentence evaluation task, but with a different paradigm relative to that used in Bott and Noveck's (2004) third experiment. We did not use categorical sentences, but sentences referring to one of two characters and to a set of objects (see below), which had to be verified relative to an image. And, in keeping with Marty and Chemla (2013), we had two *only some* conditions (one true and one false). The measures were self-paced reading, RTs and eye-tracking, differentiating on the screen between three areas of interest (AOIs), one corresponding to the sentence, one corresponding to the vertical half of the image where the protagonist (referred to in the subject of the sentence) appears and

[8] There is an obvious relationship between Davidson's Principle of Charity and Grice's Principle of Cooperation, but it is beyond the ambition of the present paper to discuss it.

one corresponding to the vertical half of the image where the antagonist (not mentioned in the sentence) appeared.

The predictions are as follows:

1. On the hypothesis that the cost is due to the inferential process incurred in deriving the pragmatic interpretation (Bott & Noveck 2004, Noveck & Sperber 2007), one would predict either a longer reading time for *some* or for the complement noun in the infelicitous condition if the reader gives the pragmatic interpretation than if she gives the logical interpretation, providing that the quantifier is interpreted locally. If it is interpreted globally, the prediction is that the RT will be longer for the pragmatic (*disagree*) than for the logical (*agree*) answer. Regarding the eye-tracking, one would expect participants giving the pragmatic interpretation to concentrate on the text AOI.
2. On the hypothesis that the cost of the pragmatic answer in the infelicitous condition is to the verification of the pragmatic interpretation (Guasti et al. 2005), one would not expect a longer reading time for the quantifier and/or the complement noun, as the cost occurs *after* the pragmatic interpretation is accessed (and it is accessed locally in that hypothesis). On the other hand, one would expect a longer RT for the pragmatic interpretation as the verification process occurs after the interpretation itself. Finally, given that the verification process amounts to the verification of the existence of a complement set, one would expect participants giving the pragmatic answer to look more at the antagonist AOI than those that give the semantic answer.
3. Under the hypothesis that the cost of the pragmatic answer in the infelicitous condition is due to a decision to draw or not to draw the pragmatic interpretation (Marty & Chemla 2013), one would expect the reading time for the quantifier and/or for the complement noun to be longer for participants giving the pragmatic answer than for participants giving the logical answer. On the other hand, one would not expect the

RT to be longer for the pragmatic answer than for the logical answer in the infelicitous *some* condition, as well as for the false *only some* condition. Regarding the eye tracking, there are no specific predictions.

4. On the hypothesis that the cost is due to a conflict between the pragmatic interpretation and the Principle of Charity in the infelicitous *some* condition, one would expect a longer reading time for the quantifier and/or the complement noun for the pragmatic answer than for the logical answer, and a longer RT for the pragmatic answer than for the logical answer (because the logical interpretation is compatible with the Principle of Charity) as well as for answers to false *only some*. Regarding the eye-tracking, one would again predict that the participants that give the pragmatic answer will look at the antagonist AOI more than those that give the logical answer, to check whether, indeed, the pragmatic answer is not verified.

5.2. Participants

Fifty-four participants (21 males, mean age = 23.25, *sd* = 3.26) took part to our experiment.[9] All participants were native French speakers.

5.3. Material

Experiment 3 was again a sentence evaluation task, but it did not use categorical sentences. Rather, participants were asked to evaluate the sentence relative to an image which presented two easily distinguishable characters (e.g., a boy and a girl) and a set of six objects which were distributed among the characters or in the exclusive possession of one of them. In the

[9] This study was carried out in accordance with the recommendations of the Comité de Protection des Personnes Sud Est II, who gave it its agreement (IRB number: 11263). All participants gave written informed consent in accordance with the Declaration of Helsinki.

felicitous *some* condition, the character mentioned in the sentence (e.g., *The girl has some cars*) had two of the objects mentioned in the sentence, while the other character had four:

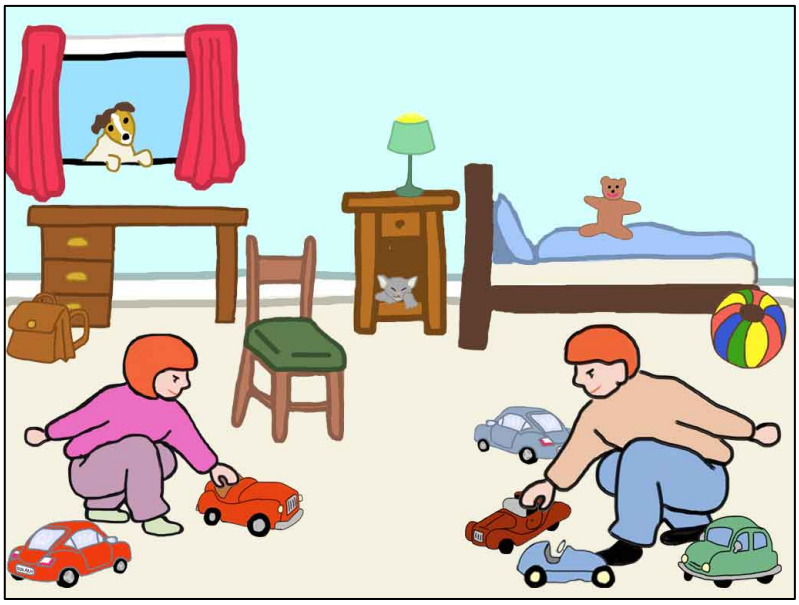


Figure 6: Felicitous *some*

In the infelicitous *some* condition, the character mentioned in the sentence has all of the objects as shown below:

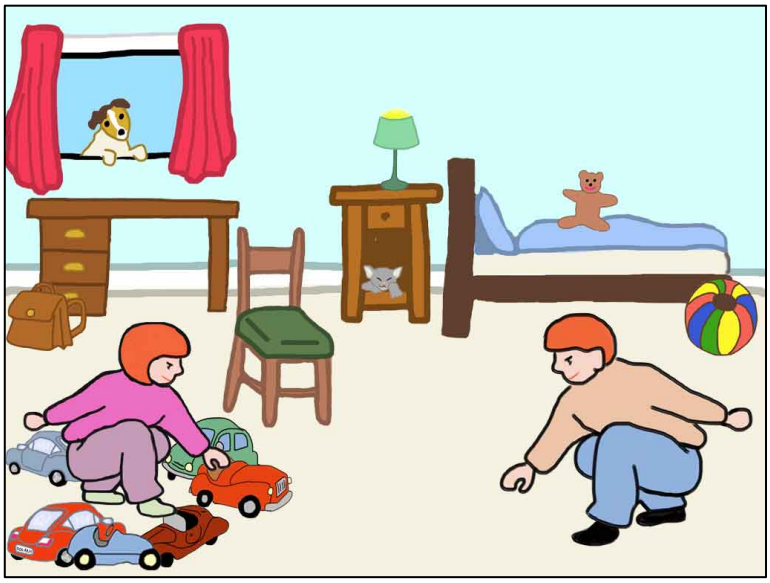


Figure 7: infelicitous *some*

There was one control condition, *all* (four true and four false items) and two test conditions, *only some* (four true and four false items), and *some* (four felicitous and four infelicitous items). There were also three filler conditions: *exactly one* (two true and two false items), *exactly two* (two true and two false items), and *exactly three* (two true and two false items).

As in Experiment 2, participants were divided into two groups who saw the same sentences, but different pictures. Each group saw exactly the same number of items and the same number of true/false, felicitous/ infelicitous items. When one saw the felicitous (or true) picture, the other saw the corresponding (infelicitous or false) picture with the same characters, background and set of objects, but with a different distribution of the objects. And vice versa.

5.4. Procedure

Participants were first presented with an image with two characters for a duration of one second. Then the objects came into the picture with different distributions in the different conditions (see above). The resulting picture was observed for one second before the first group of words appeared. The first group of words corresponded to the subject NP (e.g., *The girl*) and participants had to push on the space key to see the next word (e.g., *has*), etc., for, successively, the quantifier (e.g., *some*) and the complement word (e.g., *cars*). On the appearance of each new word or group of words, the previous one disappeared. This was to avoid the participants pressing on the space key to access the whole sentence. Thus, reading was self-paced. When the participants had read the whole sentence, they were asked to indicate whether they agreed or disagreed by pressing one of two keys. Both their answers and their RTs were recorded. The whole process was done while the participants were facing an eyetracker, allowing to record their gaze direction. Figure 8 is an example of trial composition and of how we defined RTs.

Keyboard interaction RTs and data selection

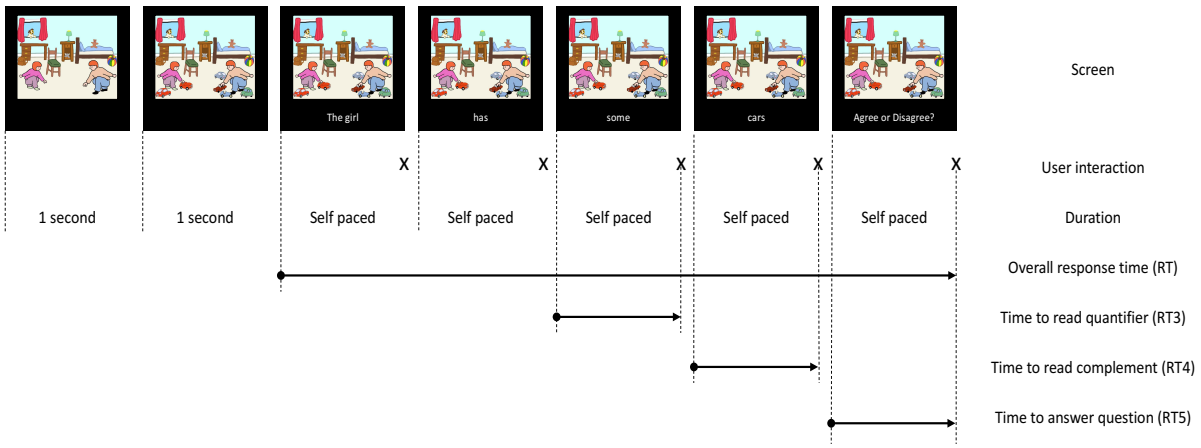


Figure 8: trial composition and RTs definitions

Several RTs related to keyboard interactions were computed. The overall response time (RT) includes both the reading time and the time to answer the question. It is measured between when the first group of words appears on the screen, and when the subject presses a key on the keyboard to answer the question. Trials with RT > 10s or RT < 0.5s or incorrect answers were discarded. The time to read quantifier (RT3) corresponds to the third group of words, which can be “all”, “some”, or “only some” and is measured between when the word appears on the screen and when the subject presses the spacebar on the keyboard. The time to read complement corresponds to the fourth group of words. The time to answer the question is measured between when the question appears on the screen, and when the subject presses one of the two answer keys on the keyboard.

5.5. Eyetracking data processing

The eyetracking data was collected at a rate of 60Hz with a Tobii® X120 eyetracker. Prior to the experiment, the subjects underwent a 5-point calibration phase with the Tobii Studio Software that allowed matching gaze direction to pixels on the screen. Eyetracking data was

collected with Presentation® software. It was the average of right and left eye gaze direction, expressed in pixels. The calibration was checked every 6 trials.

5.6. Computation of fixation data

Missing data for durations less than 400ms were interpolated with a nearest neighbor algorithm. Fixations were isolated from saccades by computing a moving average of order 10, and considering as saccades the time samples for which the distance in pixels between two successive averages was greater than 35 pixels. When multiple saccades were detected in segments of 6 consecutive time samples, only the largest saccade was kept. Segments of more than 6 consecutive time samples between saccades where considered as fixations.

5.7. Areas of Interest

Three Areas of interest (AOI) were defined for fixation positions, as in Figure 9. The image displayed on the screen was separated in two rectangular AOIs with a width 400 pixels and a height 600 pixels, one being the part of the picture where the protagonist of the sentence appears ("Protagonist AOI") and the other side displayed the other character ("Antagonist AOI"). A "Text AOI" was also defined around the words displayed on the screen. It was 1200 pixels wide and 200 pixels high.

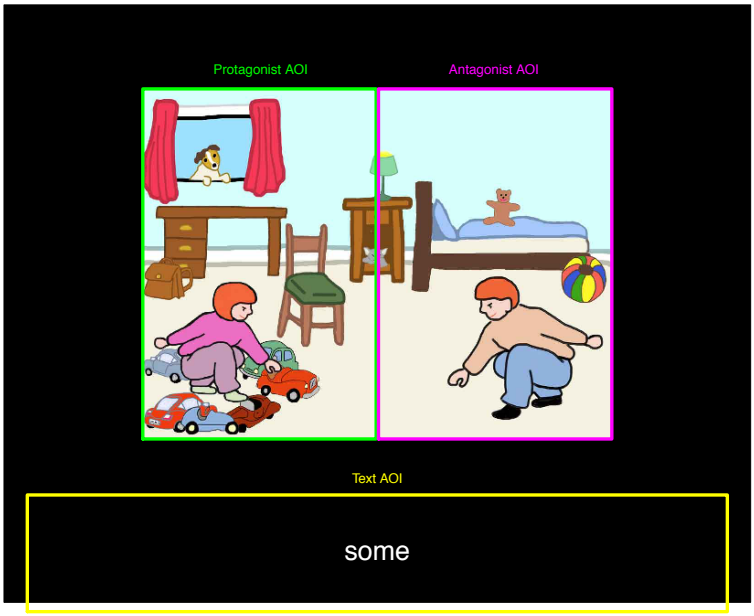


Figure 9: Areas of interest (AOI) definitions

5.8. Fixation data statistics

For each AOI, we computed a fixation count that corresponds to the number of fixations on this area between when the quantifier words appear and the end of trial. We also computed a fixation duration, which is the sum of the duration of each of these individual fixations.

5.9. Results

First of all, we analyzed results of choice proportions. In Table 5 we present the mean of correct answers, considering the mean of logical correctness in the underinformative condition. Participants did the experiment with ease and the mean of correct answers is high (> 95%). Considering the underinfomative condition (Some semantic), participants gave much more pragmatic answers (72.2%) than logical answers (27.8%) (Wilcoxon signed rank: $V = 275$, $p < 0.001$).

Condition	Example Sentence	Mean	*sd*
All true	The boy has all the apples.	98.6	5.8
All false	The boy has all the apples.	100	0
Only some true	The girl has only some cats.	97.7	7.3
Only some false	The girl has only some cats.	95.4	14.6
Some pragmatic	The girl has some cars.	98.6	5.8
Some semantic	The girl has some cars.	27.8	37.2

Table 5. Mean and standard deviation (*sd*) of correct answers; for the underinformative sentences we considered logical correctness (i.e., agree).

We decided then to focus on the time participants spent to read the quantifier in the conditions in which *some* was used felicitously or infelicitously (Figure 10). We run a log-linear mixed model analysis with items and participants as random variables and the combination of condition and answer as fixed factor. The analysis showed no significant differences in terms of RT between the conditions (felicitous - pragmatic infelicitous: $z = .89$, $p = .71$; felicitous - logical infelicitous $z = .2$, $p = .84$; pragmatic infelicitous - logical infelicitous: $z = .72$, $p = .72$).

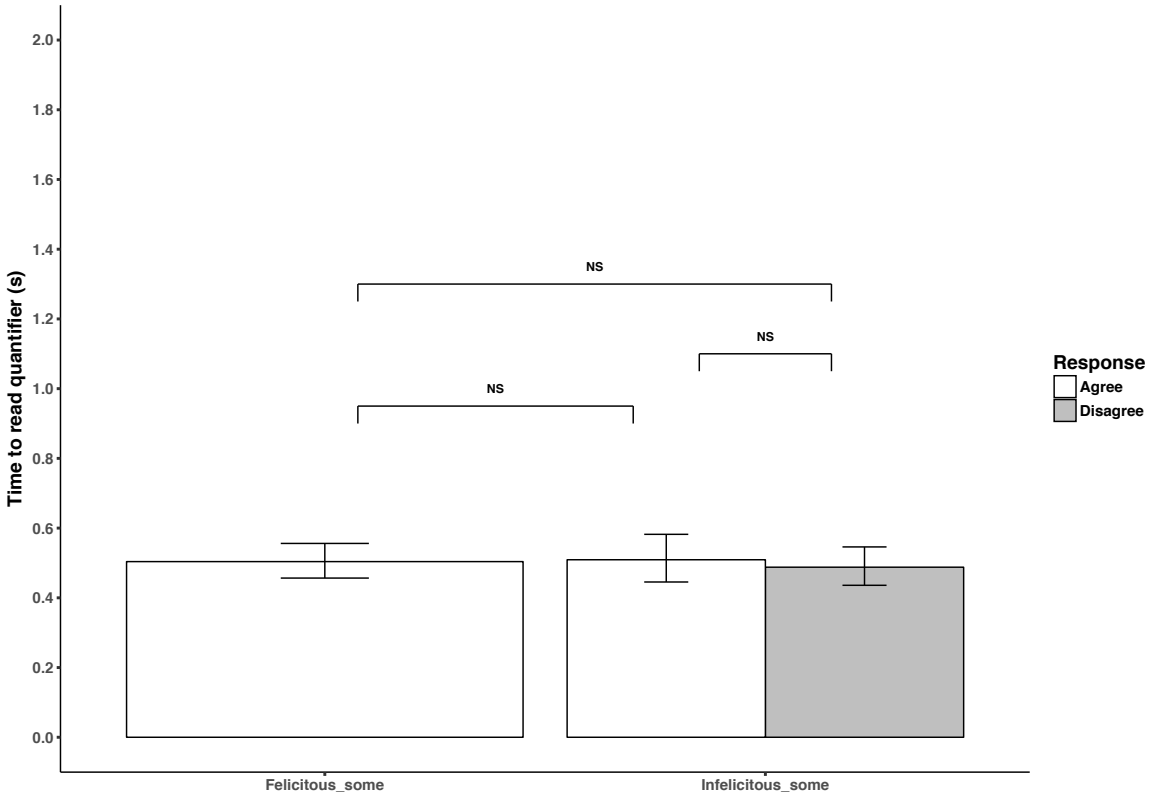


Figure 10: Time to read quantifier

Time to read quantifier as a function of condition and answer in the felicitous and infelicitous some conditions. Values were estimated with a log-linear mixed model analysis on 54 participants and 8 items, with correct responses for which overall RTs were > 0.5 s and < 10 s.

A log-linear mixed model analysis with items and participants as random variables and the combination of condition and answer as fixed factor had also been run to analyze the time participants need to read the complement as a function of condition and answer in the felicitous and infelicitous some conditions (Figure 11). No significant differences were found in terms of RT between the conditions (felicitous - pragmatic infelicitous: $z = 1.27$, $p = .31$; felicitous - logical infelicitous $z = .99$, $p = .32$; pragmatic infelicitous - logical infelicitous: $z = 1.66$, $p = .29$).

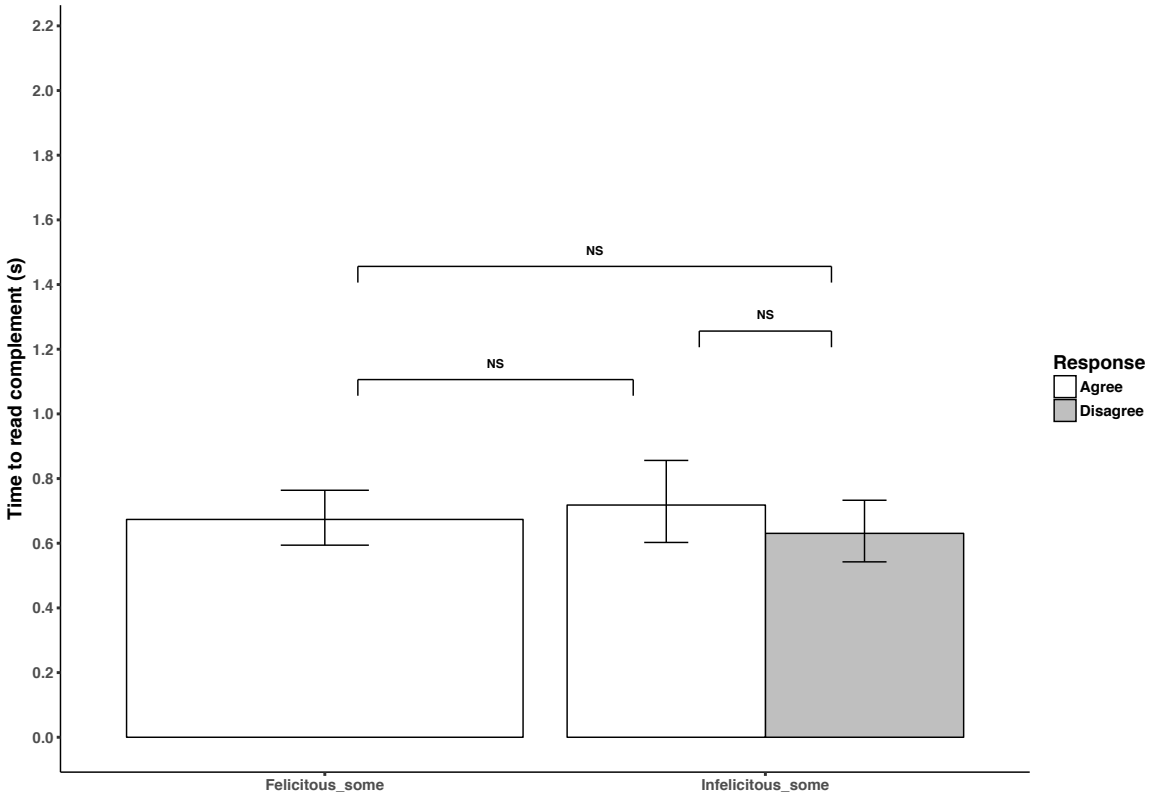


Figure 11: Time to read complement

Time to read complement as a function of condition and answer in the felicitous and infelicitous some conditions. Values were estimated with a log-linear mixed model analysis on 54 participants and 8 items, with correct responses for which overall RTs were > 0.5 s and < 10 s.

Finally, we analyzed the time participants took to answer the question as a function of condition and answer in the control and test conditions. We run a log-linear mixed model analysis with items and participants as random variables and the combination of condition and answer as fixed factor (Figure 12). No significant differences were found in terms of RT between the infelicitous some conditions (pragmatic infelicitous - logical infelicitous: z = 1.22, p = .31). The time to answer the question was higher in the infelicitous some condition

compared to all other conditions (see Table 6). No significant differences were found in terms of RT between the only some conditions (z = .23, p = .86)

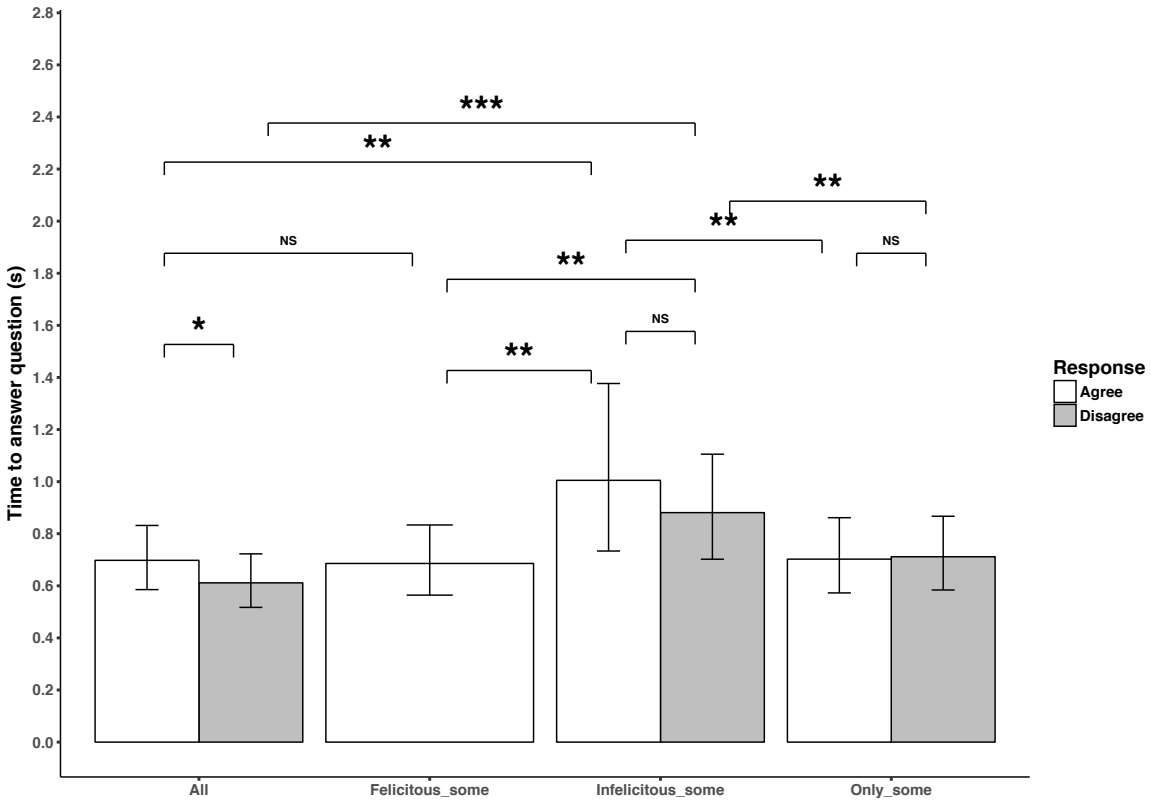


Figure 12: Time to answer the question

Time to answer the question as a function of condition and answer in the control and test conditions. Values were estimated with a log-linear mixed model analysis on 54 participants and 24 items, with correct responses for which overall RTs were > 0.5 s and < 10 s.

Table 6 – Comparisons on the time participants took to answer the questions in the different conditions.

Contrasts	*z value*	*Pr(>|z|)*
All Disagree - All Agree	-2.53	.022
Felicitous_some Agree - All Agree	-.29	.86

Contrasts	*z value*	*Pr(>|z|)*
Infelicitous_some Agree - All Agree	3.43	.003
Infelicitous_some Disagree - All Agree	3.21	.004
Only_some Agree - All Agree	.11	.91
Only_some Disagree - All Agree	.32	.86
Felicitous_some Agree - All Disagree	1.88	.09
Infelicitous_some Agree - All Disagree	4.61	< .001
Infelicitous_some Disagree - All Disagree	4.87	< .001
Only_some Agree - All Disagree	2.22	.043
Only_some Disagree - All Disagree	2.36	.032
Infelicitous_some Agree - Felicitous_some Agree	3.80	.001
Infelicitous_some Disagree - Felicitous_some Agree	3.52	.002
Only_some Agree - Felicitous_some Agree	.41	.84
Only_some Disagree - Felicitous_some Agree	.57	.74
Infelicitous_some Disagree - Infelicitous_some Agree	-1.22	.31
Only_some Agree - Infelicitous_some Agree	-3.29	.004
Only_some Disagree - Infelicitous_some Agree	-3.20	.004
Only_some Agree - Infelicitous_some Disagree	-2.95	.007
Only_some Disagree - Infelicitous_some Disagree	-3.10	.005
Only_some Disagree - Only_some Agree	.23	.86

We analyzed also fixation counts (i.e., the number of fixations on a specific area between when the quantifier words appear and the end of trial): specifically, we analyzed the proportion of items for which the subject looked more often to the antagonist rather than on the protagonist, as a function of condition and answer. A logistic mixed model analysis had

been run, with correct responses and looks on the image, with items and participants as random variables and the combination of condition and answer as fixed factor. As we can see in Figure 13, we found no significant differences between the felicitous some condition and the infelicitous some condition (infelicitous some agree – felicitous some: z = -.68, p = .50; infelicitous some disagree – felicitous some: z = .98, p = .49; infelicitous some disagree – infelicitous some agree: z = 1.26, p = .49;).

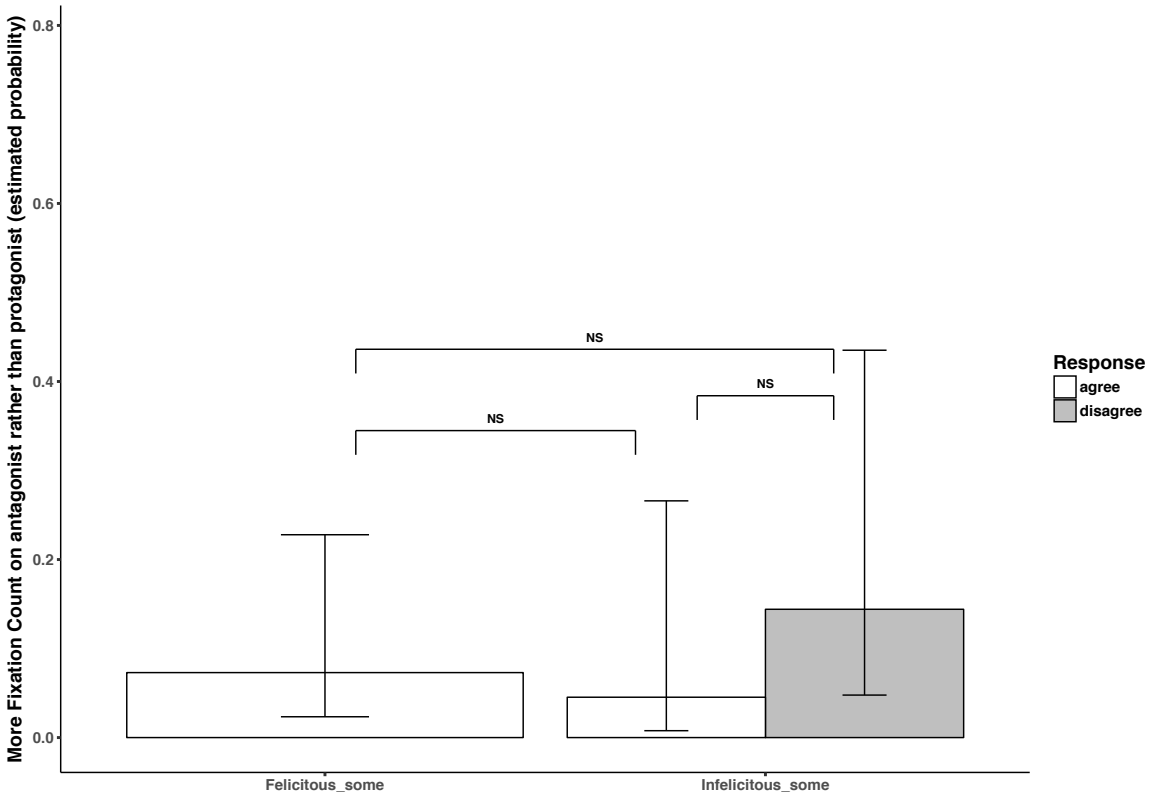


Figure 13. Proportion of items for which the subject looked more often to the antagonist rather than on the protagonist, as a function of condition and answer. Values were estimated with a logistic mixed model analysis on 51 participants, with correct responses and looks on the image, with items and participants as random variables and the combination of condition and answer as fixed factor. Error bars stand for 95% confidence intervals.

To conclude we analyzed fixation durations (i.e., the sum of the duration of each of the individual fixations). Specifically, we looked at proportion of items for which subjects looked longer to the antagonist rather than on the protagonist, as a function of condition and answer. We ran a logistic mixed model analysis with correct responses and looks on the image, with items and participants as random variables and the combination of condition and answer as fixed factor. As we can see in Figure 14, again no significant differences had been found between the felicitous some condition and the infelicitous some condition (infelicitous some agree – felicitous some: z = -.79, p = .43; infelicitous some disagree – felicitous some: z = 1.02, p = .43; infelicitous some disagree – infelicitous some agree: z = 1.40, p = .43;).

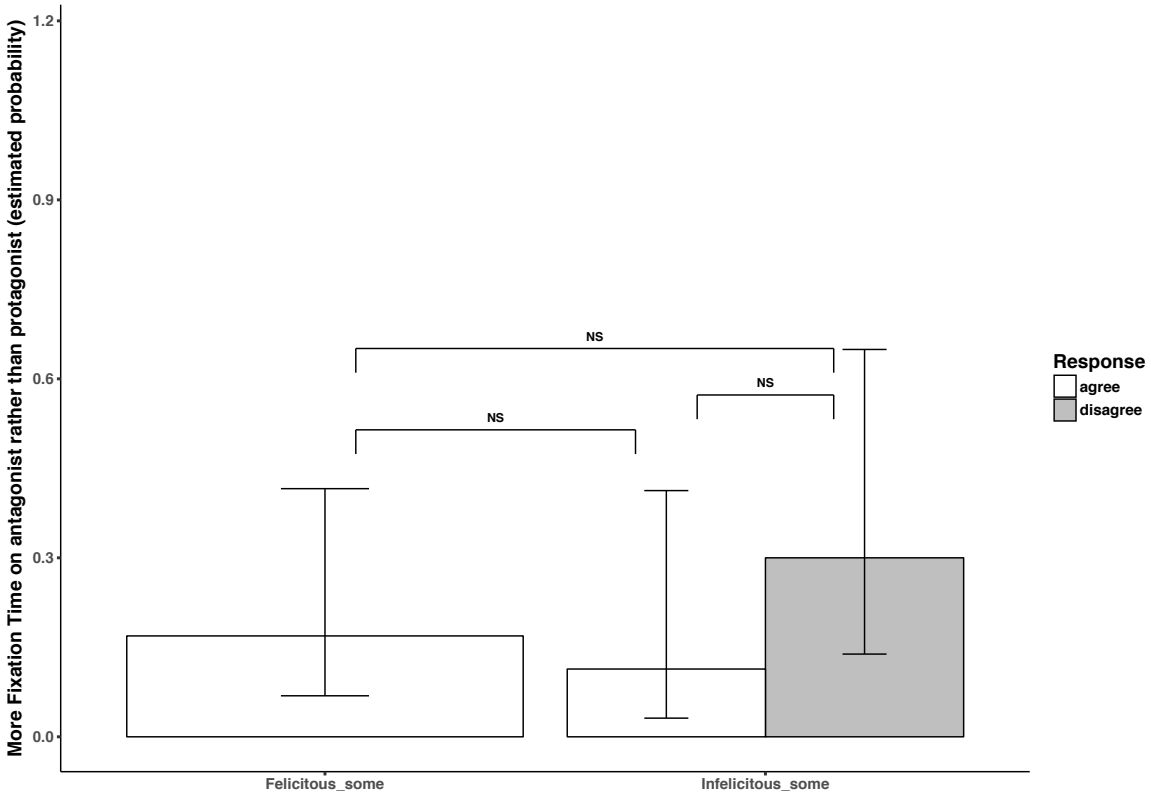


Figure 14. Proportion of items for which the subject looked longer to the antagonist rather than on the protagonist, as a function of condition and answer. Values were estimated with a logistic mixed model analysis on 51 participants, with correct responses and looks on the

image, with items and participants as random variables and the combination of condition and answer as fixed factor. Error bars stand for 95% confidence intervals.

5.10. Discussion

After confirming that there is a cost in deriving scalar-implicatures (Experiment 1 and 2) and that this cost is unrelated to differences in moving upward or downward the conceptual hierarchy (Experiment 2), Experiment 3 had the goal to address such cost in scalar implicature computation with self-paced reading, RTs and eye-tracking measurements. We had specific predictions regarding the outcomes and they are based on four main hypotheses.

According to the first approach, the cost of deriving the pragmatic computation lies in the inferential process necessary to draw the interpretation (Bott & Noveck 2004, Noveck & Sperber 2007); under such a view, we might predict a) longer reading times for the quantifier and/or for the complement noun for participants that answered pragmatically in the underinformative condition (if the quantifier is interpreted locally) or b) longer RTs to answer pragmatically in the underinformative condition (if the quantifier is interpreted globally) and c) more fixation counts/longer fixation times at the text AOI for pragmatic interpretations in the underinformative condition.

According to the second approach, the cost of deriving the pragmatic computation lies in the verification of the pragmatic interpretation (Guasti et al., 2005); under such a view, we might predict a) no longer reading times for the quantifier and/or for the complement noun for participants that answered pragmatically in the underinformative condition, b) longer RTs to answer pragmatically in the underinformative condition and c) more fixation counts/longer fixation times at the antagonist AOI for pragmatic interpretations in the underinformative condition.

According to the third approach, the cost of deriving the pragmatic computation lies in the decision to draw the implicature (Marty & Chemla 2013); under such a view, we might predict a) longer reading times for the quantifier and/or for the complement noun for participants that answered pragmatically in the underinformative condition and b) no longer RTs to answer pragmatically in the underinformative condition. There are no specific predictions regarding the eye-tracker results.

Finally, according to the fourth approach, the cost of deriving the pragmatic computation lies in the conflict between adhering to the Principle of Charity and to the pragmatic interpretation; under such a view, we might predict a) longer reading times for the quantifier and/or for the complement noun for participants that answered pragmatically in the underinformative condition, b) longer RTs to answer pragmatically in the underinformative condition and c) more fixation counts/longer fixation times at the antagonist AOI for pragmatic interpretations in the underinformative condition.

Looking at reading time results, we did not find any differences either between infelicitous and felicitous conditions, or between pragmatic and logical answers in the infelicitous condition. These results seem to go against localist approaches of the quantifier's interpretation. Indeed, if the quantifier *some* is interpreted locally in the sentence (instead of globally), we might have expected longer reading times in the infelicitous condition compared to the felicitous conditions. However, we should also consider alternative explanations for such results, for example participants might adopt a strategy in the self-paced reading in which they press keys systematically until the end of the sentence. Thus, our arguments should not be considered as decisive because our goal was not precisely to test the localist interpretation.

Moving to results based on RTs, it took longer to answer the question in the infelicitous some condition compared to all other conditions, but – surprisingly – we did not

replicate results of Experiment 2 and 3, since there is no RTs difference between the logical answer and the pragmatic answer. What differs in this third experiment, compared to Experiments 1 and 2, is the context.

Finally, data based on the eye-tracker again showed no significant pattern of results. Independently of condition, we saw more and longer fixations at the protagonist AOI, in line with the idea that the protagonist is the most salient in the context because it is the one that is mentioned in the sentence. Looking at graphs, we can see that there is a trend for longer fixation times and more fixation counts on the antagonist AOI in participants that interpreted the underinformative sentence pragmatically instead of semantically. One possibility is that this trend is not significant due to the great variance and/or issues of statistical power or due to an experimental artifact.

In conclusion, our study is the first of several steps that aim to better address the debate on the cost of deriving scalar implicatures with the use of eye-tracking techniques. We found longer RTs when answering at underinformative sentences compared to control sentences but no longer reading times for the quantifier or the complement noun. Overall, these results seem to go against localist approaches of the quantifier's interpretation, but we need to conduct other studies to better understand where the cost lies.

6. General Discussion

In this paper we mainly addressed the debate regarding the cost of deriving scalar implicatures, using different experimental techniques (i.e., RTs, reading times and eye-tracking). There are two competing accounts on the computation of scalar inferences. On the one hand, there are the neo-Gricean approaches (Levinson, 2000; Chierchia, 2013) according to which the computational process is linguistically driven, with the semantic interpretation that is more cognitive demanding than the pragmatic interpretation. On the other hand, there

are the post-Gricean approaches (Sperber & Wilson, 1995; Carston, 2002) according to which the computational process is pragmatic and context-based, with the pragmatic interpretation that requires cognitive effort to be computed.

Since Bott and Noveck (2004), many studies tried to experimentally assess whether or not there is a cost when we compute an inference and if a cost exists.

Findings seem to support the post-Gricean approaches because when considering RTs, pragmatic answers require more time than logical answers and cognitive loads have an impact for both the rate and RTs of implicature computations (Bott, Bailey, & Grodner, 2012; Bott & Noveck, 2004; De Neys & Schaeken, 2007; Huang & Snedeker, 2009a; Noveck & Posada, 2003; Tomlinson, Bailey, & Bott, 2013). Considering a sentence that contains an underinformative element in a scale, the context in which it is uttered seems to play a role (Breheny, Katsos, & Williams, 2006; Degen & Tanenhaus, 2015). Regarding the nature of the cost, there is further debate: some theories claim that the cost is simply related to the inferential process that is necessary to pragmatically enrich the literal meaning of the sentence (Bott & Noveck, 2004), other theories found the inner cause of the cost in different processes, for example that of verification of the pragmatic interpretation (Guasti et al., 2005), or in the decisional process of drawing the implicature (Marty & Chemla, 2013) or in the resolution of the conflict between the Principle of Charity and the pragmatic interpretation. Our study aimed to better understand the predictions of each approach and to find supportive data.

With Experiment 1, we replicated the third experiment of Bott & Noveck (2004) in which, through a Sentence Evaluation Task with categorical sentences, they demonstrated that more time is needed to answer underinformative sentences pragmatically than logically. Similarly, we found that when participants had to evaluate sentences such as “Some elephants are mammals”, they answer pragmatically (i.e., false) more often than logically (i.e., true) but that RTs are longer when they draw the implicature. Thus, we confirmed that there is a

cognitive effort in deriving the implicature and it is visible in longer RTs (compared to both the logical interpretation of underinformative sentences and of control sentences).

The second experiment might be considered as a follow-up of Bott & Noveck's (2004) study. Indeed, our goal was to exclude the possibility that the cost of the implicature derivation is related to an experimental artifact, specifically that it is easier for participants to move down in the conceptual hierarchy (eg. *Some mammals are elephants*) than to move up (e.g., *Some elephants are mammals*). Through a similar Sentence Evaluation Task, we compared two groups of people who saw the same categorical sentences: for each sentence, there were two different sets of images, one validating and one not validating the sentence, with one group seeing the validating one, and one group seeing the other one. In this way, the same *some*-sentence was true in the pragmatic interpretation of one group, while it was false in the pragmatic interpretation of the other group. This was possible thanks to the creation of artificial categories named with a pseudo-word. The task demands were high (participants that failed in more than 2 control sentences were eliminated from the analysis) and this led to a higher number of logical answers. However, RTs remained longer when participants gave pragmatic answers compared to logical answers, confirming again the cost of drawing an implicature.

Once confirmed that there is actually a cost, our third experiment aimed to look more in depth at the source of the cost, considering not only RTs but also reading times and eye-movements. We created a variety of images with two characters (and three distinct areas of interest in the screen) and objects with different distributions in the different conditions. Then participants, through self-paced reading, read a sentence that could be either true or false (in the control conditions) or pragmatically true or pragmatically false (in the underinformative condition). They were asked to indicate whether they agreed or disagreed with the presented sentence, based on the picture. We recorded their answers, their RTs and their gaze direction.

Results have been quite surprising, since we did not find statistically significant data overall. There were no differences in reading times either between infelicitous and felicitous conditions, or between pragmatic and logical answers in the infelicitous condition. Participants took longer to answer questions in the infelicitous some condition compared to all other conditions, but no differences have been detected between the logical answer and the pragmatic answer, contrary to Experiments 1 and 2. It might be the case that having a visual context (offered by images) helped. It should be worth creating a new within-subject experiment in which participants are tested both with a classic Sentence Evaluation Task without context (like the one in Experiment 1) and then with a Picture Selection Task with context. If our predictions are right, we expect to find the majority of participants who answered logically in the without-context experiment, to move to pragmatic answers in the with-context experiment, and this is consistent with the idea that they are not more willing to answer logically regardless but they do so simply because of an experimental artifact.

Finally, eye-tracker data showed more fixations at the protagonist for a longer time in the infelicitous some condition but no significant results have been found about fixations at the antagonist; we expected to find more/longer fixations at the antagonist when participants answered pragmatically but, even though a trend in this direction exists, it is not statistically relevant. The future direction of the study is to create a new eye-tracking experiment in which we manipulate how the sentence is presented: by presenting the image after the entire sentence has been read, we can better focus on the participants' fixations. Indeed, in the present experiment we analyzed fixations between the quantifier words' appearance and the end of trial but participants could look at the entire picture before being presented with the sentence. This might be a limit of the study and a possible explanation for the not-significant pattern of results.

In conclusion, further studies must be done to adjudicate between the different theories regarding the cost; our data, however, seem to support a post-gricean view of scalar-implicature computation, a view according to which there is a cost in deriving scalar implicatures. Nonetheless, this cost seems to disappear under particular contexts.

Appendix 1 – Categories and exemplars used in Experiment 1

Bird	Fish	Insect	Mammal	Reptile
Canaries	Anchovies	Ants	Bears	Alligators
Crows	Barracudas	Beetles	Cats	Chameleons
Eagles	Carps	Butterflies	Cows	Cobras
Owls	Cods	Caterpillars	Dogs	Crocodiles
Parrots	Piranhas	Cockroaches	Elephants	Grass snakes
Peacocks	Salmons	Flies	Horses	Iguanas
Pigeons	Sardines	Mosquitos	Monkeys	Lizards
Sparrows	Trouts	Spiders	Pigs	Tortoises
Vultures	Tunas	Wasps	Sheeps	Vipers

GENERAL DISCUSSION AND PERSPECTIVES

General Discussion and Perspectives

1. Main Findings

The present dissertation mainly addresses the debate around the computation of specific type of implicatures, namely generalized (in this thesis scalar implicatures: Chapters 1, 2, 3, 4 and 5) and particularized implicatures (or ad-hoc implicatures: Chapters 1 and 2). As argued in the General Introduction part, those two types of implicatures have differences based on how alternatives are set. The former are inferences that are lexically-determined; by contrast, the latter are context-determined. Unlike other implicatures, such as the Conventional ones, they are both cancellable. As we have seen herein this thesis, in the Pragmatics' experimental turn different theories and aspects related to the computation of implicatures are considered based on the population under investigation, particularly children or adults. For this reason, the General Discussion part will be divided in sub-topics.

1.1. The Scalar/Ad-hoc Implicatures Debate

As previously suggested, there are divergent views on implicatures: on the one side there are views according to which the distinction between scalar and ad-hoc implicatures exists (Chierchia, 2004; Grice, 1975; Levinson, 2000), on the other side there are views according to which there is a unique mechanism of inferences (Carston, 1998; Degen & Tanenhaus, 2015; Geurts, 2010; Sperber & Wilson, 1986; 1995).

In Chapter 1 (Experiment 2), with a Picture Selection Task that reduced task demands, we demonstrated that younger children can successfully compute ad-hoc implicatures but they have difficulties with scalar implicatures. Similar results have been found also in Horowitz, Schneider and Frank (2017); however, since we did not only test preschoolers but also

schoolers, we could also discuss the developmental trajectory of the two types of implicatures. In our study, preschoolers easily computed ad-hoc but not scalar implicatures; on the latter, they reach an optimal performance at school age.

In Chapter 2, with the same Picture Selection Task of Chapter 1, we demonstrated that, contrary to expectations, children in the autistic spectrum were also computing more ad-hoc than scalar implicatures, even if they don't reach the level of their typically developing peers for both type of implicatures. Typically developing children behave like children in Chapter 1, replicating the same pattern of results (preschoolers compute more ad-hoc than scalar implicatures, while at school-age their performance is at ceiling for both types of implicatures).

The overall picture seems to lead us to consider scalar and ad-hoc implicatures as two distinct phenomena.

1.2 The Debate around the Acqusition of Scalar Implicatures in Typically Developing Children

In Chapter 1 we didn't just focus our attention on the scalar-ad-hoc implicatures debate, but also on the one concerning the acquisition of scalar implicatures in typically developing children (Experiment 1). As we have seen, acquisition studies found that children before the age of seven (or five) might present difficulties in computing scalar implicatures, i.e. they often accept underinformative sentences (for a review, Chemla & Singh, 2014b: 2.3). Different theories tried to explain children's behaviour. On the one side, there are 'pragmatic accounts' that interpret children's failure with scalar implicatures by considering their pragmatic system as not fully developed. Specifically, children have been considered as unable to recognize that scalar terms constitute relevant alternatives in certain contexts (Skordos & Papafragou, 2016) or as simply more tolerant of pragmatic violations compared

to adults (Katsos & Bishop, 2011). On the other side, there are 'lexicalist accounts' according to which children's failure can be explained in terms of difficulties in retrieving the lexical scale (Barner, Brooks, & Bale, 2011; Foppolo, Guasti & Chierchia, 2012). There is also a 'processing account' based on a Relevance-Theoretic framework (Pouscoulous, Noveck, Politzer, & Bastide, 2007): children difficulties arise from the labour related to the implicature processing and thus, task complexities might lead to greater cognitive efforts and less computations.

In our first experiment (Chapter 1) we decided to run a within-subject design testing the same children with both a classic Truth Value Judgment Task and with a new Picture Selection Task, in which we provided children with visual and linguistic cues to derive scalar alternatives. The two tasks led to similar results (children had a performance around 55% of pragmatic answers) and thus visual alternatives didn't help children with scalar implicature derivation. Our results are difficult to explain in terms of pragmatic accounts, since:

i) If contextual relevance matters we should have expected a worse performance in the Truth Value Judgment Task compared to the Picture Selection Task, where the 'all' alternative to 'some' is made salient visually and verbally,

ii) Children can be more tolerant but this can account for a Truth Value Judgment Task, not for a Picture Selection Task where they have to select between a pragmatic and a logical answer. Again, we should have expected a worse performance in the Truth Value Judgment Task.

Our results are difficult to explain also in terms of a processing account, since the Picture Selection Task has the relevant alternatives visually presented and it is less complex than a Truth Value Judgment Task in which a metalinguistic judgment must be provided, so it should have elicited much more pragmatic answers. Lexical accounts can explain our results,

since children's difficulty is considered to be related with the lexical scales (the scale is not yet lexicalized or there are difficulties with the retrieval of alternatives).

Results of Experiment 2 (Chapter 1) have also been analyzed considering the 'lexical accounts': the observed difference in the acquisition of scalar and ad-hoc implicatures can be explained considering that scalar implicatures require the lexicalization of relevant scales, a step that is not involved in ad-hoc implicatures (typically context-dependent).

1.3 The Default-Non-Default Debate of Scalar-Implicatures Computation

As presented in the Introduction, one of the liveliest debates on scalar implicature is related to the processing time that is needed to integrate semantic and pragmatic information to successfully infer the speaker's meaning, with two main theoretical proposals (Chemla & Singh, 2014a,b, for a review). According to what we can define 'Strong Default approach', we compute implicatures without additional processing costs and context plays no role in the computation; the cancelation of the implicature involves a cognitive cost. The main proponent of this view is Levinson (2000). According to 'Non-Default approaches', with recent new elaboration into a 'Literal-First/Two Stage models' (Huang & Snedeker, 2009a), the literal interpretation is accessed first and implicatures are computed, as a second step, with additional time and cognitive effort. Recently, a third account known as 'Constraint-Based approach' (Degen & Tanenhaus, 2015) has been developed. In this new processing theory of implicatures computation, concepts like 'naturalness', 'availability of alternatives' and 'context' are paramount in order to evaluate whether an implicature will be interpreted with an extra cost or not. Indeed, Degen and Tanenhaus asserted:

> "When we embed the issue of when, and how quickly, upper- and lower-bound some are interpreted within a Constraint-Based approach with expectations and adaptation, then questions about time course no longer focus on distinguishing between the claim

> that scalar implicatures are computed by default and the claim that they are only computed after an initial stage of semantic processing. When there is more probabilistic support from multiple cues, listeners will compute scalar inferences more quickly and more robustly. Conversely, when there is less support, listeners will take longer to arrive at the inference, and the inference will be weaker" (2015, p. 672).

It can be of interest to consider whether our results can be analyzed accordingly to their proposals, even if the essays proposed in this thesis had not the scope of assessing these theories (clear-cut predictions based on Degen and Tanenhaus' proposal are not easyly conceivable).

In Chapter 4, we decided to assess the cost of deriving implicatures without contextual support but within a population that might present a cognitive disadvantage (i.e., cognitive load) for our purposes. We assessed non-balanced bilinguals of two different second languages (i.e., English and Spanish) with a Sentence Evaluation Task, in which they had to judge whether they agree or disagree with underinformative categorical sentences of the type 'Some Xs are Ys', when actually all the Xs are Ys. Furthermore, we assessed oral processing under time constraints for the answer. In two between-subject experiments, we found that bilingual participants tested in their second language computed significantly less pragmatic answers compared to bilinguals tested in their first language. Our results have been evaluated considering both the cognitive load of interpreting a sentence in a second language – when participants are non-balanced and non-immersed bilinguals – and the cognitive load of oral processing under time constraints. Our proposal is that both factors contributed to load the cognitive resources available to participants tested in their second language, causing a decrease of pragmatic answers. Although offering new data against the Default approach, this experiment does not permit to adjudicate between different Non-Default models.

In Chapter 5, we again focused on the processing-cost debate with three experiments (and different experimental techniques, such as RTs, reading times and eye-tracking), two

with sentences interpreted without context (Experiments 1 and 2) and one with contextual (visual) support (Experiment 3). Since Bott and Noveck (2004) was the first experimental study testing the cost of scalar implicatures, we decided to replicate their third experiment in our first one. This is, to our knowledge, the first time that a study aims to systematically replicate their results and this can be a good starting point, due to the replication crisis in Psychology. Using the same Sentence Evaluation Task (with little modifications), we replicated their results that showed how answering pragmatically requires more time than answering logically to underinformative sentences like "Some elephants are mammals".

Our second experiment in Chapter 5 moves from Bott and Noveck (2004) and had the goal of excluding the possibility that the proved cost is related to an experimental artifact (i.e., that it is easier to move down in the conceptual hierarchy than to move up). Again we used a Sentence Evaluation Task with the use of pseudo-words. The first interesting result was that we obtained a high number of logical answers, due to our task's high demands. Moreover, we again found loger RTs when participants gave pragmatic answers compared to logical answers, excluding the possibility that this is due to an experimental artifact. Once more, our results seem to counter the idea that pragmatic inferences are accessed by default.

The third experiment of Chapter 5 is a Sentence Evaluation Task but we introduced contextual supports. Being an eye-tracking experiment, we did not consider just RTs but also reading times and eye-movements. Surprisingly, we did not find differences in RTs, reading times or gaze directions (related to Areas of Interests in the screen) between pragmatic and logical answers in the infelicitous condition. Results are thus different from the ones described before. It might be the case that having a visual context (offered by images) helped, suggesting that it might be more of interest to analyze the cost related with the context, more than the cost *per se*. This idea would be supportive of Constraint-Based approaches.

We are planning two new experiments to better address the core and limits of our study. We will run both a within-subject experiment in which participants will be tested with a Sentence Evaluation Task both with and without context. In line with Context-based approaches (e.g., Constraint-Based approach), we expect to find an increase of pragmatic answers in the with-context experiment. Furthermore, we are creating a new eye-tracking experiment in which we will present the image after the entire sentence has been read, contrary to what we did in Experiment 3 (Chapter 5).

We can conclude saying that our studies' results do not support 'Strong defaultism' views but future studies in the field should identify "the cues that listeners use in computing scalar implicatures", in order "to provide a unified account of why in many situations this computation appears to be delayed and to come at a cost, whereas in other situations, it is more rapid, and less resource-intensive" (Degen & Tanenhaus, 2015, p. 704). Indeed, when we speak about 'context' it is not always easy to establish which aspects of the context we should refer for. Foppolo and Marelli wrote:

> "From the point of view of Constraint-Based Models, the role of context in the generation of scalar inferencing is founded on three basic considerations about what listeners do during sentence processing: they rapidly integrate multiple sources of information; they generate expectations of multiple types (phonological, syntactic, semantics, etc.) about 'what is coming next' in the sentence and they can adapt their expectations to different interlocutors and situations. In general, the notion of context might involve probabilistic constraints about the type of sentence uttered (including phonetic–acoustic properties and syntactic–semantic expectations); speakers' meaning and intentions; considerations about the situation and the goals of the conversations; integration of different sources of informa- tion, like acoustic and visual cues (Degen & Tanenhaus 2015). Theoretically, one first very broad distinction is between pragmatic and default approaches: according to strong defaultism, as we said, context bears no role in the process, as SIs are recursively factored in the computation whenever a scalar term is encountered, independently of contextual features. According to pragmatic approaches, instead, context might intervene in modulating SI computation, for example, by modulating the saliency of the relevant

> alternatives by means of different 'linguistic' factors like word order, questions under discussion and focus (Geurts 2010). More specifically, 'context' might also refer to the syntactic traits of the sentence in which the scalar term appears, and to the semantic properties of the structure in which the scalar term is embedded that are known to affect SI computation [cf. Chierchia (2013) for a discussion]." (2017, p. 662).

All those contextual aspects have to be considered in future studies.

1.4 The Role of Theory of Mind in the Computation

Whether or not Theory of Mind has a role in pragmatic computation, and particularly in implicature computation, is still under debate. Results in the literature are mixed, some arguing for a role of ToM (Nieuwland, Ditman, & Kuperberg, 2010; Surian, Baron-Cohen, & Van der Lely, 1996) and others for none (Andrés-Roqueta, & Katsos, 2017; Antoniou, Cummins, & Katsos, 2016). Andrés-Roqueta and Katsos (2017), based on conflicting results in the literature on ToM skills and pragmatic phenomena, proposed a distinction between a "linguistic-pragmatics" where "structural language and competence with pragmatic norms are enough to perform successfully in the task" and a "social-pragmatics" where "in addition to structural language and pragmatics, the child needs competence with ToM, and specifically the ability to represent other people's intentions, desires and beliefs". The difference, however, has nothing to do "with pragmatic phenomena *per se*, but with the communicative situation" (Andrés-Roqueta & Katsos, 2017, pp. 2-3). In their proposal, sensitivity to informativeness is included in "linguistic-pragmatics": to compute an implicature, people mainly rely on linguistic knowledge (e.g., semantics of quantifiers) and the role of ToM is minimal.

Moving to our results, in Chapter 1 and 2 in both typically developing children and children with Autism Spectrum Disorders we found a role of ToM in scalar implicature computations. However, roles of morphosyntax skills have also been found, while no roles of

ToM skills have been found for ad-hoc implicatures. This lack of correlation between ToM and ad-hoc implicatures – together with a positive correlation with scalars – may look surprising at first sight. We analyzed those results considering the fact that both scalar implicatures and ToM skills are at maturational stage in children of the age we tested, while ad-hoc implicatures are computed also at a younger age. It might be that those children that are in a stage in which ToM skills are more developed are also able to compute scalar implicatures. Vice versa, children that have less mature ToM skills, have also problems with such derivation. As a consequence, it might be that ToM is not a prerequisite for implicature computation; otherwise it should be the same for ad-hoc implicatures. Another aspect to take into account is that ASD children did not reach the same level of TD children with both types of implicatures: this might be related with ToM skills as well as linguistic skills (our participants where not matched for syntactic abilities). Since individuals in the autistic spectrum pass first- and second-order ToM with a significant delay (Happé, 1994), it might be that implicatures are also succesfully computed with a delay. This would explain also the inconsistency between our results and results with adults in the spectrum.

In Chapter 3, we tried to address the debate on ToM and implicatures computation assessing the inference-ability in people in in the broader phenotype but with a risk of presenting higher autistic traits. As previously demonstrated in a series of works discussed in detail in Chapter 3, students of scientific disciplines were good candidates for having higher autistic traits compared to students of humanities. We found that students enrolled in scientific curricula were less likely to compute scalar implicatures than students enrolled in humanities curricula and that AQ measures (and particularly ToM measures) correlate with pragmatic answers. If these results seem to be in contrast with Andrés-Roqueta and Katsos's (2017) proposal, we must also consider that in our experiment the correlation between autistic traits and pragmatic answers was present only in the humanities group and not in the scientific

group. This fact leaves open the possibility that students of scientific disciplines are simply trained to focus on the task's logical aspects. Still, a correlation between ToM scores and pragmatic answers has been found in the humanities group, though there is a limitation in our study: the ToM items used to assess the correlation were not originally described as ToM-related by the authors of the AQ test (indeed, they never delimited a specific category for ToM). We selected the items that, in our opinion, targeted ToM skills. In the future it would be useful to assess whether those items are recognized as ToM-related also from other experts in the fields or from other people instructed on the definition of ToM.[10]

To sum up, even if some role of ToM has been detected in our experiments, our results cannot be conclusive for the debate. It is worth considering that when researchers try to correlate performances in ToM tasks with other pragmatic skills, it is not clear whether they employed ToM tasks target the specific abilities required when considering the interlocutor's perspective in pragmatic inferences (Andrés-Roqueta & Katsos, 2017, p. 2). Longitudinal studies might be of interest for the debate here described but they should consider ToM skills more related to the particular pragmatic aspect under investigation. However, another possible interesting consideration takes again the context into account. In Degen and Tanenhaus's view (2015), one of the contextual cues that have to be considered is related to the fact that the listener generates expectations on the speakers'meaning and intentions. We can suggest that in a Constraint-Based framework, ToM skills' role might be more or less involved based on how much the speakers' intentions are clear in the specific context: if, for example, the speaker is supposed to be ignorant about something, implicatures might not arise. Even if we have presented some studies related to this view (e.g., Hochstein, Bale, & Barner, 2018), future developments might consider this possibility more in depth with different experimental techniques and populations.

---

[10] We would like to thank Prof. Ira Noveck and participants of his brown bag at CNRS Institute for Cognitive Sciences-Marc Jeannerod (Lyon) for suggesting this solution.

2. Conclusion

This dissertation collected a series of studies that had the goal of disentangling – without any claim of exhaustiveness – some open issues on the topic of implicatures computation. We used a variety of experimental techniques (Sentence Evaluation Tasks, Picture Selection Tasks, RTs, reading times, gaze detection) and different tested population (typically developing children, children with Autism Developmental Disorders, adults in the broader phenotype with higher autistic traits, bilinguals and typical adults). We detected a distinction between the computation of generalized and particularized implicatures (using the same task to assess both), with the latter being computed more easily than the former. We proposed that a lexical account might explain the difference, considering that generalized implicatures, differently from particularized implicatures, require the lexicalization of relevant scales. We also showed that children with ASD compute less implicatures (both generalized and particularized) compared to their typically developing peers and that this might be related to ToM and morphosyntactic difficulties. We also proposed that general abilities might help ASD children in compensating with their difficulties.

All in all, our data seem to oppose a 'Strong defaultism' view of implicature computations, demonstrating that a certain role of the speaker's epistemic reasoning is needed (Chapters 1, 2 and 3) and that less scalar implicatures are computed under a cognitive load and in tasks without contextual support (Chapters 4 and 5). However, with particular contextual support, no differences have been detected between logical and pragmatic answers in terms of RTs and reading times (Chapter 5). Those latter results seem to go more in the direction of Constraint-Based approaches, even if the difference that we found between generalized and particularized implicatures seems not find support in this theory. Grammatical-approaches, instead, support both the distinction between generalized and particolarized implicatures and consider the role of the speaker's epistemic state. Moreover, a

certain role of the ‘context’ is expected too since, as previously suggested, with context we might also refer to the syntactic traits of the sentence in which the scalar term appears, and to the semantic properties of the structure in which the scalar term is embedded that affects the computation of scalar implicatures (Foppolo & Marelli, 2017, p. 662).

To conclude, it is not always easy to neatly reconcile theoretical and empirical perspectives. As Chemla & Singh (2014b, p. 395) suggested, “the relations between theory and experiment in the domain of scalar implicature are in their infancy”. For this reason a joint approach is the only one that we must consider to move forward in the comprehension of the diverse range of issues addressed in this thesis. It is hoped that the new field of Experimental Pragmatics might take advantage of the always more sophisticated experimental techniques (e.g., eye-traking techinques, EEG, ERP etc) and will apply them to the more theoretical field of Pragmatics. We will “benefit from joint experimental and theoretical scrutiny” (Chemla & Singh, 2014b, p. 395). Future directions in this field should focus more on defining contextual features and reconcile results of different population (children vs. adults, typical vs. non-typical) and of different techniques to create a unique proposal.

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.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: Author.

Andreou, G., & Karapetsas, A. (2004). Verbal abilities in low and highly proficient bilinguals. *Journal of Psycholinguistic Research*, *33*(5), 357-364.

Andrés-Roqueta, C., & Katsos, N. (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. *Frontiers in Psychology, 8*, 996.

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, *99*, 78-95.

Antoniou, K., & Katsos, N. (2017). The effect of childhood multilingualism and bilectalism 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.

Austin, J. L. (1962). How to do things with words: The William James lectures. *Cambridge, MA*.

Baltaxe, C. A. (1977). Pragmatic deficits in the language of autistic adolescents. *Journal of Pediatric Psychology*, *2*(4), 176-180.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*(1), 40-62.

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.

Baron-Cohen, S. (1988). Social and pragmatic deficits in autism: Cognitive or affective?. *Journal of Autism and Developmental Disorders*, *18*(3), 379-402.

Baron-Cohen, S. (1989). Are autistic children "behaviorists"? An examination of their mental-physical and appearance-reality distinctions. *Journal of Autism and Developmental Disorders*, *19*(4), 579-600.

Baron-Cohen, S. (1997). Are children with autism superior at folk physics?. *New Directions for Child and Adolescent Development*, *1997*(75), 45-54.

Baron-Cohen, S. (1997). Hey! It was just a joke! Understanding propositions and propositional attitudes by normally developing children and children with autism. *Israel Journal of Psychiatry and Related Sciences*, *34*, 174-178.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology*, *4*(2), 113-125.

Baron-Cohen, S., Bolton, P., Wheelwright, S., Scahill, V., Short, L., Mead, G., & Smith, A. (1998). Autism occurs more often in families of physicists, engineers, and mathematicians. *Autism 2*(3). 296-301.

Baron-Cohen, S., Wheelwright, S., Stott, C., Bolton, P., & Goodyer, I. (1997). Is there a link

between engineering and autism?. *Autism*, *1*(1), 101-109.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of Autism and Developmental disorders*, *31*(1), 5-17.

Belacchi, C., Scalisi, T. G., Cannoni, E., & Cornoldi, C. (2008). *CPM coloured progressive matrices: standardizzazione italiana: manuale*. Firenze, Italy: Giunti OS.

Bezuidenhout, A. L., & Morris, R. K. (2004). Implicature, relevance and default pragmatic inference. In *Experimental pragmatics* (pp. 257-282). Palgrave Macmillan, London.

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.

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

Bott, L., & Noveck, I. A. (2003). The time course of scalar implicature. In *Abstract for the International Workshop "Where Semantics meets Pragmatics"*, pp. 11-13.

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. (2013a) Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes, 28,* 443-467.

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

Capps, L., Kehres, J., & Sigman, M. (1998). Conversational abilities among children with autism and children with developmental delays. *Autism*, *2*(4), 325-344.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In *Relevance theory: applications and implications.* R. Carston and S. Uchida. Amsterdam, Benjamins, 179-236.

Carston, R. (2002) *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell Publishing.

Chemla, E., Cummins, C., & Singh, R. (2013). Reading Hurford disjunctions: processing evidence for embedded implicatures. Poster presented at XPRAG 2013.

Chemla, E., & Singh, R. (2014a). Remarks on the experimental turn in the study of scalar implicature, Part I. *Language and Linguistics Compass*, *8*(9), 373-386.

Chemla, E., & Singh, R. (2014b). Remarks on the experimental turn in the study of scalar implicature, Part II. *Language and Linguistics Compass*, *8*(9), 387-399.

Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*(9), 1104-1117.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry*, *37*(4), 535-590.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. OUP: Oxford.

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 from the Annual Boston University Conference on Language Development*, *25*, 157–168.

Chierchia, G., Fox, D., & Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Handbook of semantics*, eds. by P. Portner, C. Maienborn, and K. Heusinger, New York, NY: Mouton de Gruyter. Consulted online the 26th of July 2018 at https://pdfs.semanticscholar.org/26d8/b73c795cc50061c55e6fb68df236ac68b0c0.pdf?_ga=2.142847285.1223715563.1532601199-1340705153.1532601199, pp. 1-43

Chierchia, G., Guasti, M. T., Gualmini, A., Meroni, L., Crain, S., & Foppolo, F. (2004). Semantic and pragmatic competence in children's and adults' comprehension of or. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 283–300). New York, NY: Palgrave Macmillan.

Chin, H. Y., & Bernard-Opitz, V. (2000). Teaching conversational skills to children with autism: Effect on the development of a theory of mind. *Journal of Autism and Developmental Disorders*, *30*(6), 569-583.

Clifford, S. M., & Dissanayake, C. (2008). The early development of joint attention in infants with autistic disorder using home video observations and parental interview. *Journal of Autism and Developmental Disorders*, *38*(5), 791-805.

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. *Aspects of Bilingualism*, 127-138.

Davidson, D. (1984). *Inquiries into Truth and Interpretation: philosophical Essays*. Oxford: Clarendon Press.

De Villiers, J. (2007). The interface of language and theory of mind. *Lingua*, *117*(11), 1858-1878.

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. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667-710.

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〉. 〈hal-01803048〉

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. (2012). *The logic of pragmatics. An experimental investigations with children and adults*. LAP Lampert Academic Publicher.

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.

Foppolo, F., & Marelli, M. (2017). No delay for some inferences. *Journal of Semantics*, *34*(4), 659-681.

Foppolo F., Mazzaggio G., Panzeri F., & Surian L. (2018). What's behind some (but not all) implicatures. *Frontiers in Psychology - Conference Abstract: XPRAG.it 2018 - Second Experimental Pragmatics in Italy Conference.*

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.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In *Presupposition and implicature in compositional semantics* (pp. 71-120). London: Palgrave Macmillan.

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder: Autism. *Trends in Neurosciences*, *14*(10), 433-438.

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

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: Academic Press.

Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Syntax and semantics 9: Pragmatics*, pp. 113–128. New York: 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, T.M., 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.

Happé, F. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, *48*(2), 101-119.

Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154.

Hochstein, L., Bale, A., & Barner, D. (2018). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Language Learning and Development*, *14*(3), 224-240.

Hoffmann, Maria (1987). *Negatio Contrarii: A Study of Latin Litotes.* Assen: Van Gorcum.

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

Horn, L. R. (1984). Toward a new taxonomy for scalar inference. In D. Schiffrin (Ed.), *GURT*. Washington, DC: Georgetown University Press.

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2017). The Trouble With Quantifiers: Exploring Children's Deficits in Scalar Implicature. *Child Development*. doi: 10.1111/cdev.13014. [Epub ahead of print]

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.

Jackson, S., & Jacobs, S., (1982).Ambiguity and implicature in children's discourse comprehension. *Journal of Child Language*, *9*, 209–216.

Jolliffe, T., & Baron-Cohen, S. (1997). Are people with autism and Asperger syndrome faster than normal on the Embedded Figures Test?. *Journal of Child Psychology and Psychiatry*, *38*(5), 527-534.

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.

Laurence R, H., & Gregory L, W. (2006). *The handbook of pragmatics.* Oxford: Blackwell.

Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit?. *Cognition*, *40*(3), 203-218.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, *43*(3), 225-251.

Levinson, Stephen C. 1983. Pragmatics. Cambridge: Cambridge University Press.

Levinson, S. C. (1995). Three levels of meaning. In *Grammar and meaning: Essays in honour of Sir John Lyons* (pp. 90-115). Cambridge University Press.

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

Loukusa, S., Leinonen, E., Kuusikko, S., Jussila, K., Mattila, M. L., Ryder, N., Ebeling H., & Moilanen, I. (2007). Use of context in pragmatic language comprehension by children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, *37*(6), 1049-1059.

MacKay, G., & Shaw, A. (2004). A comparative study of figurative language in children with autistic spectrum disorders. *Child Language Teaching and Therapy*, *20*(1), 13-32.

Marini, A., Marotta, L., Bulgheroni, S., & Fabbro, F. (2015). *Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni*. Firenze: Giunti OS.

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

Mazzaggio G., Panzeri F., Giustolisi B., & Surian L. (2018). The atypical pattern of irony comprehension in children with high-functioning autism. *Frontiers in. Psychology Conference Abstract:* XPRAG.it 2018 - Second Experimental Pragmatics in Italy Conference.

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.

Mill, J. S. (1867). *An Examination of Sir William Hamilton's Philosophy* (3d edn.). London: Longman.

Miller, K., Schmitt, C., Chang, H. H., & Munn, A. (2005). Young children understand some implicatures. In *Proceedings of the 29th Annual Boston University Conference on Language Development*, 389-400.

Naigles, L. R., Cheng, M., Rattanasone, N. X., Tek, S., Khetrapal, N., Fein, D., & Demuth, K. (2016). "You're telling me!" The prevalence and predictors of pronoun reversals in children with autism spectrum disorders and typical development. *Research in Autism Spectrum Disorders*, *27*, 11-20.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*(3), 324-346.

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

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

Noveck, I. A., & Sperber, D. (Eds.). (2004). *Experimental pragmatics*. Basingstoke: Palgrave Macmillan.

Ozonoff, S., & Miller, J. N. (1996). An exploration of right-hemisphere contributions to the pragmatic impairments of autism. *Brain and Language*, *52*(3), 411-434.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantic-pragmatics interface. *Cognition*, *80*, 253–282.

Papafragou, A., & Skordos, D. (2016). Scalar implicature. *The Oxford Handbook of Developmental Linguistics*, 611-632.

Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, *12*, 71–82.

Paris, S. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology*, *16*, 278-291.

Pastor-Cerezuela, G., Tordera Yllescas, J. C., González-Sala, F., Montagut-Asunción, M., & Fernández-Andrés, M. I. (2018). Comprehension of Generalized Conversational

Implicatures by Children With and Without Autism Spectrum Disorder. *Frontiers in Psychology*, *9*, 272.

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.

Pexman, P. M., Rostad, K. R., McMorris, C. A., Climie, E. A., Stowkowy, J., & Glenwright, M. R. (2011). Processing of ironic language in children with high-functioning autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *41*(8), 1097-1112.

Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*(4), 607-618.

Politzer-Ahles, S., & Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography. *Language, Cognition and Neuroscience*, *30*(7), 853-866

Potts, C. (2005). *The logic of conventional implicatures* (No. 7). Oxford University Press on Demand.

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.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences*, *1*(4), 515-526.

Reboul, A. (2005) Pragmatics, point of view and theory of mind. à paraître dans *Intellectica*.

Reddy, V., Williams, E., & Vaughan, A. (2002). Sharing humour and laughter in autism and Down's syndrome. *British Journal of Psychology*, *93*(2), 219-242.

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.

Rundblad, G., & Annaz, D. (2010). The atypical development of metaphor and metonymy comprehension in children with autism. *Autism*, *14*(1), 29-46.

Ruta, L., Mazzone, D., Mazzone, L., Wheelwright, S., & Baron-Cohen, S. (2012). The Autism-Spectrum Quotient—Italian version: A cross-cultural confirmation of the broader autism phenotype. *Journal of Autism and Developmental Disorders*, *42*(4), 625-633.

Sampath, K. K. (2005). Effect of bilingualism on intelligence. In *Proceedings of the 4th International Symposium on Bilingualism*, 2048-2056.

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*, *129*(2), 410-417.

Schaeken, W., Van Haeren, M., & Bambini, V. (2018). The understanding of scalar implicatures in Children with Autism Spectrum Disorder: Dichotomized responses to violations of informativeness. *Frontiers in Psychology*, *9*, 1266.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883-885.

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).

Sigman, M., Ungerer, J., Mundy, P., & Tracy, S. (1987). Cognition in autistic children. In D. Cohen and D. Donnellan (eds.). *Handbook of Autism and Pervasive Developmental Disorders*. New York: Wiley. 103-120.

Slabakova, R. (2010). Scalar implicatures in second language acquisition. *Lingua*, *120*(10), 2444-2462.

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*, 153, 6-18.

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.

Sperber, D., & Wilson, D. (1986;1995). Relevance: Communication and cognition (2nd ed.) Cambridge, MA: Harvard University Press.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176-190.

Storto, G., & Tanenhaus, M. (2005). Are scalar implicatures computed online? In *Proceedings of Sinn und Bedeutung 9*. Ed. by C. B. Emar Maier and J. Huitink, 431–445. Nijmegen: Nijmegen Centre for Semantics.

Su, Y. E., & Su, L. Y. (2015). Interpretation of logical words in Mandarin-speaking children with autism spectrum disorders: Uncovering knowledge of semantics and pragmatics. *Journal of Autism and Developmental Disorders*, *45*(7), 1938-1950.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*(3), 395.

Surian, L., Baron-Cohen, S. & Van der Lely, H. (1996). Are children with autism deaf to Gricean maxims? *Cognitive Neuropsychiatry*, *1*(1), 55-72.

Surian, L., & Job, R. (1987). Children's use of conversational rules in a referential communication task. *Journal of Psycholinguistic Research* 16, 369–82.

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).

Tieu, L., Romoli, J., Zhou, P., & Crain, S. (2015). Children's knowledge of free choice inferences and scalar implicatures. *Journal of Semantics*, *33*(2), 269-298.

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.

Wang, A. T., Lee, S. S., Sigman, M., & Dapretto, M. (2006). Neural basis of irony comprehension in children with autism: the role of prosody and context. *Brain*, *129*(4), 932-943.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523-541.

Whyte, E. M., & Nelson, K. E. (2015). Trajectories of pragmatic and nonliteral language development in children with autism spectrum disorders. *Journal of Communication Disorders*, *54*, 2-14.

Wilson, D., & Sperber, D. (2002). Relevance theory. In *Handbook of Pragmatics*. Blackwell.

Acknowledgments

"I have not always chosen the safest path.
I've made my mistakes, plenty of them.
I sometimes jump too soon and fail to appreciate the consequences.
But I've learned something important along the way:
I've learned to heed the call of my heart.
I've learned that the safest path is not always the best path
and I've learned that the voice of fear is not always to be trusted".
Steve Goodier

This is the end of a long and complicated path: there are so many people I want to thank that I find it hard to begin. It is difficult to overstate my gratitude for my PhD tutor Prof. Luca Surian: he met me three years ago when I was a naïve dreamer and worked a lot to make me a more pragmatic (no pun intended) researcher. He taught the unsophisticated linguist in me how to reason scientifically and how to write scientific papers. There is still a lot to learn but I would not have reached this point without his patience: I will always be thankful. I must also thank my co-advisor Prof. Remo Job for giving me the opportunity to have the scholarship and to work with him in these three years. I am proud to have had my work supported by the Fondazione ONLUS Marica De Vincenzi, joining this large group of researchers.

I want also to thank another person who believed in me and gave me the strength not to give up three years ago: Prof. Maria Teresa Guasti, who introduced me to the vibrant research environment of Miano-Bicocca. There, I met my "unofficial" tutors that worked as well with me in the past three years and to whom I'm very indebted: Prof.ssa Francesca Foppolo and Prof.ssa Francesca Panzeri. Francesca & Francesca (as I always referred to them) are scholars with amazing personalities have been a panacea during my research journey: I have learnt a lot from them and I could always count on their help. I will miss our meetings in P03 office at Milano-Bicocca.

Along that journey, I spent 6 months abroad, at the University of Notre Dame (Indiana) and at the CNRS Institut des Sciences Cognitives “Marc Jeannerod”, Lyon, where I was lucky enough to meet two mentors who proved very influential: Prof. Kathleen Eberhard and Prof. Anne Reboul. Those figures of strong women and researchers have been a steady source of inspiration. There are no words to express how much I’m glad to have had the opportunity to meet them. Their consistent encouragements had been a beacon in the night.

There are many other people that I met at Università degli Studi di Trento and that cheered me up during my journey. First of all, I won’t forget Gisella, Federica & Carlotta with whom I laughed, I cried, I screamed, I panicked, I traveled, I talked: we shared unforgettable moments. I will also remember Salvatore and Edoardo: I could always count on them for a smile even in the darkest day.

Outside the academic environment, I want to thank other people that supported me all the way. First of all, I must thank my husband Michelangelo. He has been my first supporter and without him I would have not reach this “finish line”. He always complained about my little self-esteem and he tried to push me every single day. However, Michelangelo didn’t help me just with words but he also “worked” with me, giving advice, listening my talks, reading my papers, traveling with me, and so on. He is my soul mate and now a great expert in scalar implicatures and Pragmatics.

I warmly thank my family for being next to me not just in this path but also in my entire life. I learned a lot from every single member of my huge and lovely family: my mum Giusy taught me to be an independent woman, my dad Corrado taught me to be patient in pursuing my goals, my sister Teresa taught me not be take myself too seriously. Now a new member is

teaching me unconditional love, my nephew Alessandro. My family is my strength, including those that did not live under my same roof but that I love nonetheless: my grandparents Lina and Aldo, my uncles and cousins that I cannot name because we are too many! During those years we lost some of them in the terrestrial form but they are inside my heart.

I also thank my friends of a lifetime for being there even after my long stays abroad: Lisa, Sara, Paola, Melissa, Giulia, Ilaria, Bob, Mary have always shared my happy moments. We are growing up: life sometime creates distance in space that our memories will always bridge.

I want to conclude thanking a very special person in my life, to which I dedicate my dissertation, my grandmother Teresa. Your passing away earlier this year has been something that changed me in depth and is still hard to accept. When my heart is full of sadness, I think about the gift that God gave me in having you in my life for twenty-eight years. It is from you that I learned something that you cannot read on a book, even with a PhD: how good a person can be. You are my life and this thesis is for you. I know that you are looking at me. Thank you.

Ringraziamenti

"Non sempre ho scelto il cammino più sicuro.
Ho commesso i miei errori, molti errori.
Spesso mi lancio troppo presto e non analizzo le conseguenze.
Lungo la strada, però, ho imparato qualcosa di importante:
Ho imparato ad ascoltare la chiamata del mio cuore.
Ho imparato che il percorso più sicuro non è sempre il percorso migliore
e ho imparato che non bisogna sempre fidarsi della voce della paura".
Steve Goodier

Questa è la fine di un percorso lungo e complicato: ci sono così tante persone che voglio ringrazire che quasi non so da dove iniziare. Non è facile esprimere a parole la mia gratitudine per il mio tutor accademico, il Prof. Luca Surian: mi ha conosciuta tre anni fa quando ero un'ingenua sognatrice e ha lavorato molto per rendermi una ricercatrice più pragmatica (in tutti i sensi). Lui ha insegnato alla linguista "grezza" che era in me come ragionare e lavorare in modo scientifico. C'è ancora molto da imparare ma non avrei raggiunto questo traguardo e questa consapevolezza senza la sua pazienza: gli sarò sempre grata. Devo anche ringraziare il mio co-tutor, il Prof. Remo Job per la fiducia datami nell'assegnarmi la borsa di studio e per avermi dato la possibilità di lavorare con lui in questi anni. Sono orgogliosa di aver lavorato grazie al supporto della Fondazione ONLUS Marica De Vincenzi, inserendomi in questa grande famiglia di ricercatori che ha lavorato negli anni in suo nome.

Non posso poi non ringraziare un'altra persona che ha creduto in me e che mi ha sempre dato la forza di non mollare, cosnigliandomi per il meglio: la Prof.ssa Maria Teresa Guasti, la quale mi ha anche introdotto al fervente gruppo di ricerca dell'Università Milano-Bicocca. In questo gruppo ho avuto la fortuna di incontrare le mie tutor "ufficiose", le quali hanno lavorato con me in tutti questi anni e cui sono davvero grata: la Prof.ssa Francesca Foppolo e la Prof.ssa Francesca Panzeri. Francesca & Francesca (come amo riferirmi a loro) sono ricercatrici brillanti e con una personalità dirompente, per me sono state un'ancora di salvezza

nel mio percorso accademico: ho potuto sempre contare sul loro aiuto e ho imparato un sacco di cose. Mi mancheranno i nostri meeting nell'ufficio P03 in Bicocca.
Nel mio cammino, ho trascorso sei mesi all'estero, alla University of Notre Dame (Indiana) e al CNRS Institut des Sciences Cognitives "Marc Jeannerod" a Lione dove ho avuto l'onore di incontrare due mentori che si sono rivelate davvero importanti per me: la Prof.ssa Kathleen Eberhard e la Prof.ssa Anne Reboul. Queste due figure di donne e ricercatrici dal carattere forte ma al contempo premuroso sono state fonte di ispirazione. Non ho parole per esprimere quanto sono grata di aver avuto la possibilità di incontrarle. Le loro parole di incoraggiamento sono state un faro nella notte.
Ci sono molte altre persone che ho incontrato all' Università degli Studi di Trento e che mi hanno rallegrato in questi anni. Prima di tutto, non scorderò Gisella, Federica e Carlotta con le quali ho riso, pianto, 'sclerato', viaggiato e parlato: abbiamo condiviso momenti indimenticabili. Ricorderò anche Salvatore ed Edoardo: ho sempre potuto contare su di loro per un sorriso anche nelle giornate più difficili.

Al di fuori dell'ambiente lavorativo, devo ringrazire molte altre persone che mi hanno supportato e sopportato lungo la via. Prima di tutto, devo ringraziare mio marito Michelangelo. Lui è stato il mio primo sostenitore e senza di lui non avrei mai raggiunto questo traguardo. Michelangelo si è sempre lamentato della mia poca autostima e ha provato a spingermi ogni giorno a fare meglio e a credere in me stessa. Per di più, non mi è stato vicino solo moralmente ma ha anche lavorato al mio fianco, dandomi consigli, ascoltando infinite volte le mie presentazioni, leggendo i miei articoli, viaggiando insieme a me e molto altro. Come dice il libro che abbiamo comprato a gennaio 2016 ad Indianapolis, Michelangelo è la mia anima gemella, il mio 'Once in a lifetime' … e ora un grande esperto in implicature scalari e Pragamtica Sperimentale!

Per continuare, ringrazio con eterno amore la mia famiglia per essermi stata accanto non solo in questo percorso ma per tutta la vita. Ho imparato da ognuno di voi qualcosa che mi è stato utile nella mia avventura di dottoranda: mamma Giusy mi ha insegnato ad essere una donna forte ed indipendente, papà Corrado mi ha insegnato ad essere paziente e perseverante nella strada verso un obiettivo, mia sorella Teresa mi ha insegnato a non prendermi troppo sul serio. Ora, un nuovo membro, mi sta insegnando l'amore incondizionato: mio nipote Alessandro. La mia famiglia è la mia forza, inclusi coloro che non hanno vissuto sotto il mio stesso tetto ma che amo allo stesso modo: i miei nonni Lina e Aldo, i miei zii e cugini che non posso nominare uno ad uno perché sono veramente troppi! In questi anni abbiamo perso alcuni dei nostri amati nella loro forma terrestre ma li conservo tutti nel mio cuore.

Ci tengo anche a ringraziare i miei amici di una vita per esserci nonostante i miei numerosi periodi di lontananza: Lisa, Sara, Paola, Melissa, Giulia, Ilaria, Bob, Mary hanno sempre condiviso i miei momenti felici. Stiamo crescendo: la vita a volte crea distanza fisica ma i nostri ricordi sono un ponte tra di noi.

Voglio concludere ringraziando una persona molto speciale nella mia vita, la persona a cui dedico la mia tesi: mia nonna Teresa. Il tuo lasciarci quest'anno è stato qualcosa che mi ha segnato nel profondo e che ancora fatico ad accettare. Quando il mio cuore è colmo di tristezza, penso al regalo che Dio mi ha fatto donandomi ventotto anni di vita insieme a te. È grazie a te se ho imparato qualcosa che non puoi leggere in un libro, nemmeno con un Dottorato di Ricerca: quanto una persona possa essere buona nell'animo. Tu sei la mia vita e questo lavoro è dedicato a te. So che mi stai guardando. Grazie.