# Pre-proceedings of

# Trends in Experimental Pragmatics

## Workshop at Center for General Linguistics

## Berlin, Germany; January 18–20, 2016

Workshop organizers: Uli Sauerland and Petra Schumacher

Pre-proceedings eds. Fabienne Salfner and Uli Sauerland

funded by XPRAG.de & ZAS

XPRAG.de (Priority Program 1727)
Zentrum für Allgemeine Sprachwissenschaft
Schützenstr. 18
10117 Berlin
http://www.xprag.de

## Foreword

We are very happy to host the workshop 'Trends in Experimental Pragmatics' as part of the activities of the priority program SPP 1727 "XPRAG.de – New Pragmatic Theories based on Experimental Evidence" of the German Research Foundation (DFG) at the Centre for General Linguistics (ZAS), Berlin. XPRAG.de continues the tradition – once started by the EURO-XPRAG network – of international experts collaborating on various facets of experimental pragmatics. Even though it is a German program, having Ira Noveck, considered as one of the founders of experimental pragmatics, and Jesse Snedeker from Harvard University as mercator fellows underlines that collaboration with the international community is essential to XPRAG.de.

The researchers in XPRAG.de seek to advance pragmatic theory by simultaneously formulating formally explicit models of the cognitive mechanisms underlying pragmatics and testing these models using experimental methods. XPRAG.de, coordinated by Petra Schumacher and Uli Sauerland, is funded from 2014 until 2020 with about 1.8 Million Euro per year. During two consecutive periods of three years each, researchers at various universities and research institutions in Germany are collaborating in individual scientific projects. In the current period, 16 projects in seven different cities (Berlin, Cologne, Tübingen, Potsdam, Göttingen, Konstanz, and Bielefeld) are funded.

This workshop is intended to play a key role for the further direction of the research field of experimental pragmatics. We have received an outstanding number of 36 high-quality submissions for which we thank all authors. Unfortunately, we only had nine slots for a talk presentation, therefore we decided to add a poster session. In addition, the workshop program features four keynote talks by Jesse Snedeker, Richard Breheny, Bart Geurts and Ira Noveck. We are proud that we could win these experts to debate together with Uli Sauerland and Petra Schuhmacher about the "future of experimental pragmatics" in a panel discussion. The thoughts and ideas they put up for discussion are included in these pre-proceedings as well. We hereby also use the opportunity to thank the four invited speakers for helping with reviewing the submitted abstracts.

The purpose of this pre-proceedings is to present the talks and poster presentations in an informal way. The pre-proceedings will be available only in electronic form and for a limited time. Papers collected here should only be cited until a more formally published version of the research becomes available. The copyright to the individual papers remains with the authors. We hope that the pre-proceedings contribute to make this workshop a successful forum for the exchange of ideas for both presenters and participants. We thank all the authors for preparing their contributions, and look forward to the presentations.

Financial support for the workshop comes primarily from the DFG grants SA 925/11-1 and SA 925/12-1 within the SPP 1727. Additional financial support comes from the German Federal Ministry for Research, BMBF grant 01UG1411.

Berlin, January 2016,
*Fabienne Salfner and Uli Sauerland*

# TABLE OF CONTENTS

# The time course of verbal irony comprehension and context integration

Rachel M. Adler
*University of Maryland*

Jared M. Novick
*University of Maryland*

Yi Ting Huang
*University of Maryland*

**Abstract** The present study tests two hypotheses of how listeners use contextual cues to interpret irony in real-time. According to the Early Context Account, listeners use contextual information to rapidly generate an expectation for irony and constrain the literal interpretation. In contrast, the Late Context Account posits that listeners initially access the literal interpretation and use it to guide the identification of relevant contextual cues. Subjects heard two speakers, one literal and one ironic, describing visually depicted referents while their eye-movements were measured. Fixation patterns showed that listeners were slower to reach ironic interpretations compared to literal interpretations. Critically though, even highly conventional, contextually-supported ironic utterances led to processing delays compared to literal utterances. These findings support the Late Context Account: listeners initially access the literal analysis and use this to guide context cue retrieval.

**Keywords:** Irony, context, eye-tracking, visual world, nonliteral language, comprehension

## 1 Introduction

The relationship between semantics and pragmatics during real-time comprehension has long been a topic of interest in psycholinguistics. One debate concerns the degree to which semantic analysis must occur prior to drawing a pragmatic inference. Irony serves as a useful test case, as it offers a situation in which semantic and pragmatic interpretations come into direct conflict. For example, if a speaker says, *"What a fabulous chef Fred is,"* we might conclude that he cooks well (literal interpretation). However, if we had just witnessed Fred making a mess, we would instead infer that he is a terrible chef (ironic interpretation). It is clear that contextual cues play a vital role in

inferring irony; for example, comprehenders may exploit knowledge of speaker tendencies or the fact that irony is used more often to criticize than to compliment (Gibbs 2000). However, it remains unknown how context is incorporated with literal meaning during real-time comprehension. One possibility is that salient contextual cues may immediately inform comprehension and constrain the literal meaning of utterances, particularly when ironic interpretations are highly conventional (Early Context Account). For example, Fred's mess and the conventional use of irony as criticism might generate an expectation for irony and thus rapidly guide the ironic interpretation. Alternatively, listeners might initially analyze the literal meaning of utterances and use it as the basis for recruiting relevant contextual cues (Late Context Account). For example, we may initially interpret "fabulous" literally, but then use that semantic analysis to identify relevant contextual cues, such as the mess Fred made.

Prior studies testing these accounts have often used either reading times to index processing difficulty (e.g., Giora et al. 2007, Schwoebel et al. 2000) or ERPs to assess information access (Regel et al. 2010, Spotorno et al. 2013). For example, Giora et al. (2007) presented subjects with dialogues between two friends. In each dialogue, one speaker produced an ironic utterance and then later, produced another utterance (target sentence) that could be interpreted ironically or literally, based on the dialogue context. Target sentence reading times were slower following ironic-biasing dialogues compared to literal-biasing ones, suggesting that context effects may be delayed during comprehension. Critically, however, it is possible that the contextual cues provided were not strong enough for the subjects to use. Subjects did not receive any explicit information about the speakers, and they encountered new speakers with each dialogue.

Indeed, recent studies that manipulate speaker characteristics provide support for an early effect of context. Regel et al. (2010) presented subjects with brief discourses that biased a final, critical utterance (e.g., *"That game was fantastic"*) toward either an ironic or literal interpretation. Afterwards, one speaker uttered 70% of the ironic-biased utterances, while the other speaker uttered 30%. Utterances that were consistent with the speakers' communicative style (e.g., highly ironic speaker producing irony) elicited a P200 after the onset of the sentence final word (e.g., *"fantastic"*). This suggests that knowledge of likely interpretations is available early during comprehension. Importantly, however, these findings do not show whether comprehenders can actually use this information so early. The subjects' only task was to read the presented sentences, and ERP deflections by themselves do not provide information about how comprehenders might reach the final ironic interpretation (e.g., via the literal analysis). The P200 may simply reflect the detection of an incongruity

between the utterance and its speaker, not necessarily the integration of contextual information with linguistic input necessary to comprehend irony.

Thus, prior research leaves open critical questions about how rapidly comprehenders make use of contextual cues to generate ironic interpretations and the degree to which the identification of relevant cues is guided by the semantic analysis. To address this, we manipulated context by comparing two cases of irony. The conventional use is *ironic criticism* (Gibbs 2000): this refers to an individual that did something poorly using a positive statement (e.g., saying "What a fabulous chef Fred is" after he makes a mess). The less conventional type is an *ironic compliment*: using a negative statement to ironically describe a successful individual (e.g., saying "What a terrible chef Sally is" after she makes a beautiful cake). If contextual cues are accessed early, then ironic criticisms should be processed as quickly as literal uses of positive adjectives. However, if context effects are delayed, ironic criticisms should take more time to comprehend.

In the current study, subjects watched two characters, Fred and Sally, perform actions on a computer screen. Next, they heard a speaker describe one of the characters using a positive or negative adjective ("What a fabulous/terrible chef s/he is"). One speaker was always literal and one speaker was always ironic. While Targets were unambiguously identified by pronoun gender, adjective valence could provide an earlier cue. With respect to semantic analysis, studies show that negative words (e.g., "terrible") are interpreted more slowly than positive ones (e.g., "fabulous"; Kuchinke et al. 2005, Schacht & Sommer 2009). Critically, when the literal speaker speaks, subjects should look to Sally shortly after *"fabulous"* and Fred after *"terrible."* However, looks to Fred should be delayed compared to looks to Sally, given evidence that negative words are interpreted more slowly than positive words. Importantly, when the ironic speaker speaks, fixations to the target referent will reveal the extent to which context guides early interpretation.

## 2 Method

### 2.1 Subjects

Thirty-five undergraduates from the University of Maryland participated for course credit. One subject's data were not analyzed due to equipment malfunction, and two more subjects were excluded due to low task accuracy. Data from the remaining 32 subjects were analyzed (27 female, mean age = 19.3).

### 2.2 Procedures & materials

Subjects were seated in front of a computer and their eye movements were recorded using an Eyelink 1000 desktop mount eye-tracker. Trials unfolded in two parts. First, subjects heard vignettes describing visually depicted events featuring two different-gender characters (Figure 1). For example, Sally baked a beautiful cake while Fred made a mess.

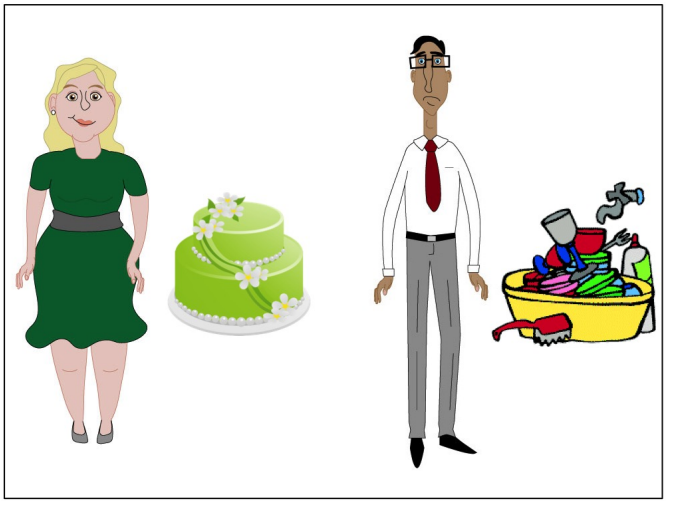


*Figure 1.* Example display.

Next, subjects heard either the ironic or literal speaker describe Fred or Sally using a positive or negative adjective ("What a *fabulous/terrible* chef s/he is"). Subjects were told at the start of the study that one speaker would always be literal and one speaker would always be ironic. For half the subjects, the ironic speaker was male and the literal speaker was female; for the other half, the genders were switched. The speakers who pre-recorded the literal and ironic statements were instructed to use an enthusiastic tone of voice, which was felicitous with an ironic interpretation, but did not preclude a literal one. Therefore, subjects heard the same recordings, regardless of whether the ironic speaker was male or female. The subjects' task was to click on the character that the speaker described.

Table 1
Design Crosses Adjective Valence, Interpretation Type, and Speaker Gender

| | Positive (*"fabulous chef"*) | Negative (*"terrible chef"*) |
|---|---|---|
| Literal interpretation | (a) Positive Target | (b) Negative Target |
| Ironic interpretation | (c) Negative Target | (d) Positive Target |

*Note.* (c) is an ironic criticism and (d) is an ironic compliment. For half the subjects (n = 16), the male speaker was ironic and the female speaker was literal; for the other half, the genders were reversed.

The experiment employed a 2 x 2 x 2 design, with adjective valence (positive, negative) and interpretation type (literal, ironic) as within-subject factors, and speaker gender (ironic male and literal female; ironic female and literal male) as a between-subjects factor (Table 1). Twenty critical items were rotated through the eight conditions across eight lists. Subjects were randomly assigned to lists, with an equal number per list. An additional 24 filler trials were constructed in which both characters were either successful or unsuccessful at a given action. Therefore, for filler items, subjects could not identify the Target until the pronoun. The presentation of critical and filler trials was randomized across critical trials.

According to both the Early and Late Context Accounts, looks to the Target character should be slowest in the ironic compliment condition ("What a terrible chef she is") due to the negative adjective and the infrequent use of ironic compliments. However, the accounts make different predictions about ironic criticisms ("What a fabulous chef he is"). The Early Context Account predicts Target looks should be just as fast as when positive adjectives are used literally ("What a fabulous chef she is" about Sally), because contextual cues should facilitate rapid irony comprehension. However, the Late Context Account predicts that ironic criticisms should be slower than both literal conditions, but faster than the ironic compliment condition (due to its positive adjective).

## 3 Results

Eye movements were divided into three time regions of interest: pre-adjective ("What a"), adjective-noun ("fabulous chef"), and pronoun ("he is"). For each region, looks prior to 200ms were removed to account for the time it takes to launch a saccade. In addition, all incorrect trials (i.e., where the subject did not click on the Target) were removed from analysis. We coded the two characters as Target (who the speaker described) and Distractor (the other character on the screen). Our primary dependent measure examined the proportion of looks to the Target, calculated as Target looks divided by Target plus Distractor

looks. These values were analyzed in R (version 3.2.2; Team 2015) using a linear mixed effects model with subjects as random intercepts and adjective valence (positive, negative), interpretation type (literal, ironic), and speaker gender as fixed effects. There were no main effects or interactions involving speaker gender ($p$s $>.10$) in any regions of interest, so it will not be discussed further.

During the pre-adjective region, the proportion of looks to the Target character in all conditions was at chance ($M = 0.50$, $SE = 0.02$). There were no significant main effects of adjective or interpretation type, and there was no interaction between adjective and interpretation type (all $p$s $>.63$). Similarly, when referents were fully disambiguated during the pronoun region (he vs. she), there were no significant main effects of adjective, interpretation type, or interaction between the two (all $p$s $>.09$).

During the critical adjective-noun region, the proportion of looks to the Target was higher when the adjective was positive versus negative (63% vs. 55%). This was confirmed by a significant main effect of adjective ($F(1, 593.55) = 8.32$, $p < .01$). Thus, consistent with prior work on semantic analysis (Kuchinke et al. 2005), access to positive adjectives was faster than negative adjectives. However, adjective valence did not affect Target looks uniformly. When the adjective was negative, there was a larger difference between the proportion of looks to the Target in the literal (61%) and ironic conditions (49%) than there was when the adjective was positive (literal = 66%, ironic = 60%). Thus, relative to literal conditions, there was a greater delay for irony when it was used to compliment than when it was used to criticize. This was reflected in a significant interaction between adjective and interpretation type ($F(1, 593.81) = 11.75$, $p < .001$).

To examine the fine-grained time course of how these effects unfolded, we divided Target looks into 50-ms intervals and compared them to chance (50%), which represents no character preference (Figure 2). In the literal conditions, Target looks became and remained greater than chance within 150ms after the positive adjective, and 500ms after the negative adjective, $t(31) = 2.41$, $p < .05$; $t(31) = 2.01$, $p < .05$. Again, this is consistent with prior findings showing that processing the literal meaning of positive adjectives is faster than negative ones.

Next, we compared the timing of ironic interpretations to literal analysis. In the less conventional, ironic compliment condition, looks to the Target were significantly greater than chance at 1150ms, $t(31) = 2.13$, $p < .05$, constituting a 650-ms delay compared to the literal use of a negative adjective (700ms). Thus, as predicted by both the Early and Late Context Accounts, processing the ironic meaning of negative adjectives takes longer than processing the literal

meaning. When the adjective was positive and used ironically (ironic criticism), Target looks did not exceed chance until 700ms post-adjective onset, $t(31) = 1.7$, $p < .05$, representing a 550-ms delay compared to when the adjective was used literally (150ms). This supports the Late Context Account, which predicts slower interpretations when positive adjectives are used ironically compared to literally.

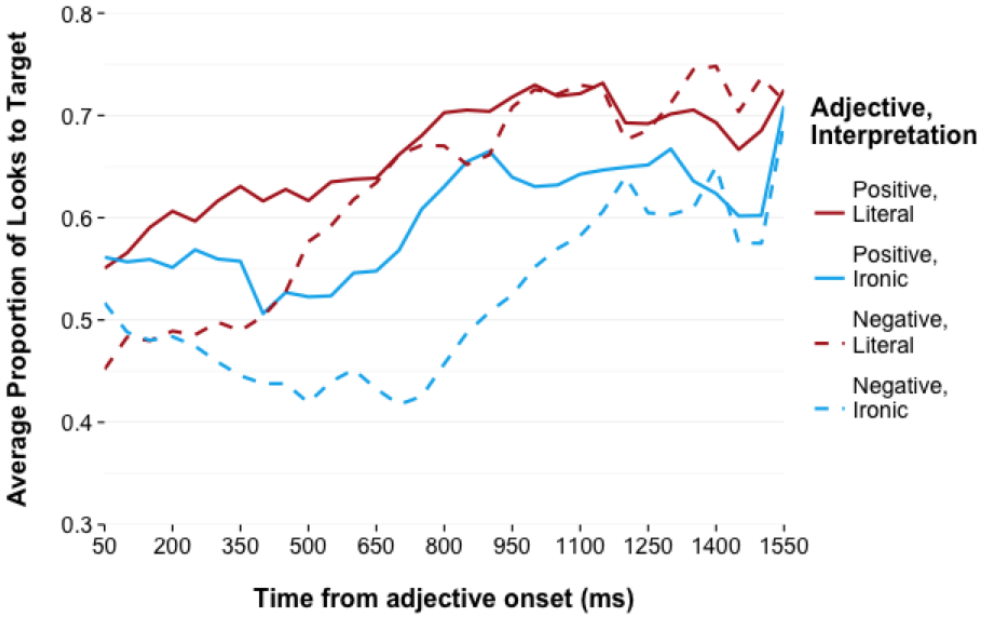



*Figure 2.* Average proportion of looks to Target character in 50-ms intervals post-adjective onset by adjective valence and interpretation type.

## 4 Discussion

In this study, we tested two accounts of how comprehenders use contextual cues to interpret ironic utterances in real-time. We manipulated context by comparing a conventional case of irony (ironic criticism; e.g., saying "What a fabulous chef he is" after Fred makes a mess) with a less conventional case (ironic compliment; e.g., saying "what a terrible chef she is" after Sally makes a beautiful cake). Consistent with both the Early Context Account and Late Context Account, less conventional ironies that used negative adjectives were processed more slowly than both literal conditions and the more conventional ironic condition. Critically, using positive adjectives ironically (ironic criticisms)

led to slower processing than using positive adjectives literally. This finding is consistent only with a Late Context Account, since the speed with which the ironic utterance is processed is mediated by the corresponding literal analysis. Indeed, while the ironic compliment was accessed more slowly than the ironic criticism, the magnitude of their respective delays as compared to the literal use of the same adjective was approximately equal (650ms for the ironic compliment, 550ms for the ironic criticism).

One potential limitation of the present study is the kind of context we used. We aimed to provide very strong contextual cues, but the resulting implementation (i.e., someone who always speaks ironically) may not have been sufficiently realistic. To address this issue, future work should vary the context probabilistically (e.g., a speaker that is ironic 75% of the time), rather than categorically, and require subjects to learn these probabilities implicitly. The current categorical manipulation may have led subjects to use strategies, such as always doing the opposite of what the ironic speaker says, rather than actually computing the pragmatic inference generated by irony. We are conducting a follow-up study to test this in which the ironic speaker is replaced with a speaker who always says "the opposite" of what s/he actually means. If the subjects in the present study were just doing the opposite of what the ironic speaker said, then we'd expect to replicate the present findings in this follow-up. However, if comprehending ironic utterances requires pragmatic processes beyond simply computing the opposite, then we would predict smaller time delays relative to the literal conditions for the opposite conditions versus the ironic conditions.

In sum, the current work demonstrates that listeners use the literal meaning of ironic utterances to guide the retrieval of relevant contextual cues. This suggests that certain pragmatic inferences, at least those used for nonliteral language comprehension, rely on initial lower-level semantic processes.

## References

Gibbs, Raymond. 2000. Irony in talk among friends. *Metaphor and Symbol* 15(1-2). 5–27. http://dx.doi.org/10.1080/10926488.2000.9678862.

Giora, Rachel, Ofer Fein, Dafna Laadan, Joe Wolfson, Michal Zeituny, Ran Kidron, Ronie Kaufman & Ronit Shaham. 2007. Expecting irony: Context versus salience-based effects. *Metaphor and Symbol* 22(2). 119–146. http://dx.doi.org/10.1080/10926480701235346.

Kuchinke, Lars, Arthur M. Jacobs, Claudia Grubich, Melissa L.-H. Vo, Markus Conrad & Manfred Herrmann. 2005. Incidental effects of emotional valence

in single word processing: An fMRI study. *NeuroImage* 28. 1022–1032. http://dx.doi.org/doi:10.1016/j.neuroimage.2005.06.050.

Regel, Stefanie, Seana Coulson & Thomas C. Gunter. 2010. The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research* 1311. 121–135. http://dx.doi.org/10.1016/j.brainres.2009.10.077.

Schacht, Annekathrin & Werner Sommer. 2009. Time course and task dependence of emotion effects in word processing. *Cognitive, Affective, and Behavioral Neuroscience* 9(1). 28–43. http://dx.doi.org/10.3758/CABN.9.1.28.

Schwoebel, John, Shelly Dews, Ellen Winner & Kavitha Srinivas. 2000. Obligatory processing of the literal meaning of ironic utterances: Further evidence. *Metaphor and Symbol* 15(1-2). 47–61. http://dx.doi.org/10.1080/10926488.2000.9678864.

Spotorno, Nicola, Anne Cheylus, Jean-Baptiste van der Henst & Ira A. Noveck. 2013. What's behind a P600? Integration operations during irony processing. *PLoS ONE* 8(6). 1–10. http://dx.doi.org/10.1080/10926480701235346.

Team, R Core. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rachel M. Adler
Department of Hearing and Speech Sciences
University of Maryland College Park
0100 Lefrak Hall
College Park, MD 20742
radler1@umd.edu

Jared M. Novick
Department of Hearing and Speech Sciences
University of Maryland College Park
0100 Lefrak Hall
College Park, MD 20742
jnovick1@umd.edu

Yi Ting Huang
Department of Hearing and Speech Sciences
University of Maryland College Park
0100 Lefrak Hall
College Park, MD 20742
ythuang1@umd.edu

# The cognitive foundations of pragmatic development

Kyriakos Antoniou and Napoleon Katsos

## 1. Introduction

Philosopher Paul Grice (1975) suggested that a good deal of what is communicated in everyday conversation relies on the appreciation of certain conversational expectations or maxims. According to his account, speakers design their utterances with respect to the cooperative principle and maxims, while listeners expect speakers to adhere to these conversational principles and calculate the speaker's meaning on the basis of this expectation. These maxims enjoin communicators to be no less and no more informative than is required for the purpose of the talk exchange (maxim of quantity I and II), tell the truth and avoid statements for which they have not adequate evidence (maxim of quality), be relevant (maxim of relation), and be brief, orderly, and avoid ambiguity and obscurity (maxim of manner) (Grice 1975, pp. 44-45).

In many communicative situations consideration of the cooperative principle and maxims prompts interlocutors to draw inferences about a speaker's intentions through which they attribute to speakers an implicit meaning that goes beyond what they literally said. These inferences are what Grice (1975, p. 45) called *conversational implicatures*. For example, consider the following mini-discourse:

(1) *A: Did all of his students fail the exam?*
*B: Some of his students failed the exam.*
(2) *Not all of his students failed the exam.*

B's utterance in (1) implies (2) even though this has not been stated explicitly. This implied proposition is an inference known as a *scalar implicature* (henceforth, *SI*).

From a Gricean perspective, the complete derivation of the implicature in (2) is accomplished through a reasoning process about the speaker's intentions that involves taking into account a rich array of information: (a) what the speaker explicitly said, (b) the linguistic and non-linguistic context, (c) the assumption that B is cooperative, (d) sensitivity to the maxim of quantity I –that is, sensitivity to the fact that there exists a more informative proposition using the term *all* that could have been used but wasn't, (e) the assumption that the speaker is knowledgeable of the situation and that s/he would assert the more informative proposition with *all* if s/he knew it to be true, and (g) the assumption that all the above information is available to both interlocutors and that both interlocutors assume this to be the case. In a similar vein, a broad class of implicatures can be inferred by exploiting the other maxims.

Grice (1975, p. 56) also introduced a distinction between *particularized* and *generalized conversational implicatures*. For instance, given a different question *Is he a promising teacher?*, B's answer in (1) would still imply (2) but will also imply the proposition in (3) or a related one. The implied proposition in (2) represents a case of a generalized conversational implicature (henceforth, *GCI*) and (3) is a case of a particularised conversational implicature (henceforth, *PCI*).

(3) *It is not certain that he is a promising teacher.*

GCIs are associated with a specific form of words and seem to be stable across contexts. PCIs, on the other hand, seem to be more context-dependent and are not associated with specific linguistic items. The example above illustrates the difference between the two types of implicature: while the GCI in (2) arises in both contexts, the PCI in (3) emerges only in the second as a result of the specific question asked.

For Grice (1975) and other so-called contextual theories of implicature (e.g., Geurts, 2010; Sperber & Wilson, 1986/1995) the generalized-particularized distinction is not

 In: Fabienne Salfner and Uli Sauerland (eds.), *Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 10–17.
All papers in this collection are unedited prepublication DRAFTs!

necessarily of particular theoretical importance. Other theories of implicature (Chierchia, 2004; Levinson, 2000), however, suggest that GCIs are distinguished from PCIs in that the former arise by default mechanisms, which are distinct from the processes involved in PCIs.

Implicature understanding is a skill that is routinely employed in everyday conversation and represents an important and, perhaps, the most sophisticated aspect of children's pragmatic competence. However, developmental studies have shown that young children often face difficulties with various facets of this pragmatic ability and that they do not reach adult-like levels of performance until early school age or even later in childhood (with the timing at which they achieve adult-like performance depending on the type of implicature) (see e.g., Waggoner & Palermo 1989; Winner 1997; Winner et al. 1988). In this experiment we aimed to examine the cognitive factors that underpin this aspect of pragmatic development in children. Specifically, we wanted to investigate whether children's implicature understanding skills are affected by executive functions, a set of interrelated cognitive processes that include working memory (the ability to simultaneously maintain and manipulate information in memory), inhibition (the ability to suppress irrelevant information), and switching (the ability to flexibly switch between, tasks, rules, or representations) (Miyake et al. 2000). We were also interested in examining the psychological validity of the generalized-particularized distinction by testing children's performance in various implicature types (based on different Gricean maxims).

**1.1 The relation between implicature understanding and executive control**

Several theoretical and empirical factors suggest that EC might be positively associated to implicature understanding. To start with, according to Grice's (1975) framework, the generation of implicatures requires an inferential process that takes into account various linguistic and contextual data (see previous section). This requirement to coordinate different pieces of information while interpreting language potentially poses demands on listeners' EC resources (see e.g., Breheny et al. 2013, Grodner & Sedivy 2005, Tomlinson et al. 2013, for some evidence that, in deriving implicatures, interlocutors indeed consider some of the information proposed by Grice).

Similarly, according to Relevance theory (Sperber & Wilson 1986/1995), implicature interpretation requires extra cognitive effort (as compared to literal meaning). Sperber and Wilson (1986) do not explicitly characterise the specific cognitive-psychological nature of this additional cognitive effort. However, several researchers have interpreted it in terms of employing extra cognitive resources (such as working memory) (e.g., De Neys & Schaeken 2007; see also Huang and Snedeker 2009b and Siegal & Surian 2007, who explicitly suggest a link between executive functions and the ability to generate implicatures in children); which may also manifest as extra processing time (e.g., Bott & Noveck 2004; Breheny et al. 2006).

There is also experimental evidence that directly or indirectly points to a link between implicature understanding and EC. First, several experimental investigations with adults have documented that the time course of SIs is associated with an additional processing cost relative to conventional meaning (see e.g., Bott & Noveck 2004, Breheny et al. 2006, Huang & Snedeker 2009a, Tomlinson et al. 2013; but see e.g., Grodner et al. 2010). This, in turn, suggests that computing SIs and, perhaps, implicatures in general is a non-automatic controlled process, which possibly relies on EC resources.

Moreover, empirical data in studies with adults suggests that comprehending SIs specifically involves working memory (henceforth, *WM*) resources. De Neys and Schaeken (2007) reported that burdening adults' WM resources with the requirement to remember a complex dot pattern before evaluating various types of statements, significantly decreased their rates of SI responses to under-informative sentences but did not affect their accuracy when judging sentences where no implicature was involved (see also Marty & Chemla 2013).

### 1.2 The present experiment

In the following study a group of young children (five to twelve years of age) were administered a novel pragmatics test on the ability to understand several types of implicatures. Children were also given a battery of tasks measuring all aspects of EC.

Various other variables were also tested to ensure that potential relations between EC and implicature comprehension were not due to third factors. Specifically, children's age and language proficiency were measured because these are also factors that have been suggested in the literature to affect children's implicature understanding skills (see e.g., Norbury 2005, Rundblad & Annaz 2010 for effects of language ability and Guasti et al. 2005, Noveck 2001, Waggoner & Palermo 1989, and Winner et al. 1988 for effects of age). Finally, socioeconomic status and non-verbal fluid intelligence were also evaluated as control background variables.

## 2. Method

### 2.1 Participants

Participants were 140 children (73 girls and 67 boys, aged 4;2–12;2, mean age 7;6, SD 1;6 years). They had variable language backgrounds and included monolingual children (speakers of Standard Modern Greek), bi-dialectal children (speakers of Cypriot Greek and Standard Modern Greek), and multilingual children (speakers of Cypriot Greek, Standard Modern Greek, and one or two additional languages).

### 2.2 Materials and procedure

All children were tested in two sessions taking approximately 50-60 minutes each. For EC, they were administered the following tests (measures taken from each test are given in parentheses). The Simon task (Simon effect) (Simon 1969) and an online version of the Stop-signal task (Stop Signal Reaction Time) (Ellefson et al. 2011-2014) for inhibition, an online version of the Color-Shape task (switch cost) (Ellefson et al. 2011-2014) for switching, and online versions of the forward and backward Corsi blocks task (Ellefson et al. 2011-2014), and the Backward Digit Span task (number of correctly recalled trials) (Wechsler 1949) for WM. The standardised Greek version of the Word Finding Vocabulary Test (henceforth WFVT) (Vogindroukas et al. 2009) and the Wechsler Matrix Reasoning test (Wechsler 1999) were given to assess expressive vocabulary and non-verbal fluid intelligence, respectively. Information about the participants' age and socioeconomic status was obtained through a *Socioeconomic status and language background questionnaire*. Three indicators of socioeconomic status were extracted from the questionnaire: the family's wealth as measured by the *Family Affluence Scale* (henceforth, *FAS*) and the parents' levels of education.

In the following sections we describe the implicatures test. A detailed description of the rest of the tasks and material can be found in Antoniou et al. in press.

#### *Implicatures test*

This was a novel task testing children's comprehension of four types of implicature: quantity I (scalar), relevance, manner, and quality (metaphors) implicatures. The implicatures included in the test can be further categorised in terms of the generalized-particularized distinction. Specifically, for the researchers who subscribe to this distinction, the manner and quantity I (scalar) implicatures can be considered cases of the generalized type while the relevance and quality (metaphor) implicatures are exemplars of the particularised type.

There were 15 implicature items (3 per implicature sub-test), 48 filler items, and one practice item. All multilingual and bi-dialectal children took the test in CG, and monolinguals in SMG. Moreover, there were two task versions of the pragmatics test and each child was tested in one of the two versions. Children's performance in the 48 filler items of the

implicatures test was considered a measure of their language comprehension skills in the language of testing (*language comprehension score*). A total comprehension score was calculated for each child by transforming the child's scores in the filler items of each implicature sub-test into z scores and then averaging the three z scores.

*Relevance implicature items*

Children were instructed that they would hear stories about a young male character named George and his mother, and that at the end of each story they had to point to a picture that showed how the story ended. All items were based on a previous study with children conducted by Bernicot et al. (2007). Each item was composed of two slides. In the first slide the target story was heard. Target stories in the three critical items were of the following format: George asked his mother a question and his mother replied with an utterance that implied either a negative or a positive answer. In the second slide the experimenter asked *What happened at the end of the story?* and introduced two pictures as possible endings.

*Quality implicatures (metaphors)*

The sub-test was designed based on a previous study by Waggoner and Palermo (1989). Children were told that they would hear stories about George and his father and that at the end of each story they should point to a picture that showed how George's father felt. They heard three stories ending in metaphors describing either the emotion of sadness or anger (e.g., *George's father was a thundering cannon*). Again, each critical trial was composed of two slides. In the first slide the target story was heard. In the second slide the experimenter asked *How did George's father feel at the end of the story?* and presented two pictures -a picture of a sad man and a picture of an angry man. All metaphors were embedded in contexts that introduced the two emotions but did not give away which of the two emotions was expressed by the metaphorical sentence. Novel and apt metaphors were used.

*Manner implicatures*

The general design of this sub-test was a sentence-to-picture-matching task. Participants were informed that they would hear George describing a picture from a book and that they had to point to a picture that matched George's description. Critical items were causatives for which a lexicalised and an opposed periphrastic alternative are available (e.g. *Opened the door* as opposed to *Made the door open*). Lexicalised causatives are associated with a normal, more stereotypical causation while their periphrastic alternatives are associated with a non-normal, non-stereotypical causation (Levinson 2000). Again, each item was composed of two slides. In the first slide the target sentence was heard. The second slide featured two pictures as possible matches to the description. In the critical items the two pictures contrasted an unmarked, stereotypical way of causation with a marked, non-stereotypical way of causation.

*Scalar implicatures act-out task*

This sub-test was a PowerPoint version of the action-based task used by Pouscoulous et al. (2007). Participants were presented with slides depicting five boxes and a selection of animals. There were three scenarios. In the 5/5 scenario all boxes contained the same animal, in the 2/5 scenario two of the five boxes contained the same animal, and in the 0/5 scenario none of the boxes had any animals. For each scenario, children heard statements constructed with the quantifiers *all*, *some*, and *none* and three types of animals. This resulted in a total of 27 test items. Children were told that they would hear George describing the display and that they had to make the display match the description using the mouse. Critical items were statements with the quantifier *some* (e.g. *There are turtles in some of the boxes*) in the 5/5 scenario.

*Scalar implicatures binary judgment task*

In each trial of this test the participant saw a depiction of five cards face down. An auditory stimulus was then played, *There are <X> on <Q> of the cards*, where X was the item type (rings, hearts, or stars) and Q the quantifier (*all*, *some*, or *none*). When the auditory stimulus ended, the cards were immediately 'turned over' to reveal the items. Participants were instructed to press a green key if the utterance were true and a red key if it were false, responding as quickly and accurately as possible. There were three critical under-informative cases using the quantifier *some*. The rest of the items comprised an equal number of true and informative, and semantically false utterances with the quantifiers *some*, *all,* and *none*.

## 3. Results

### 3.1 Preliminary analyses

***Principal component analyses***

A principal component analysis (henceforth, *PCA*) was conducted on the implicature measures extracted from the various parts of the implicatures test (number of accurate responses in each implicature sub-test). The PCA on the four implicature indicators revealed only one factor with an eigenvalue above Kaiser's criterion of 1. This factor explained 36.8% of the variance.

Similarly, Antoniou et al. (in press) reported that a PCA on the six EC dependent measures revealed two components. Participants' scores in the forward and backward conditions of the Corsi Blocks task, and in the Backward Digit Span Task clustered on the first component, which they interpreted as representing the Working Memory aspect of EC. The switch cost, Simon effect, and SSRT measures loaded on the second component, which they interpreted as representing the Inhibition aspect of EC.

***Composite scores***

A composite score was computed for overall performance in the implicatures test (*overall implicature score*). This score was calculated by averaging participants' scores in each of the sub-tests of the implicature task except for the relevance part in which ceiling performance was observed (95% accuracy). The following composite scores were also used: *general language ability*, *SES*, *WM*, and *Inhibition*. The general language ability score was computed by transforming into z scores and then averaging the WFVT and the language comprehension score. SES was calculated by combining (as above) maternal level of education, paternal level of education, and FAS score into a single score indicating socioeconomic status. Finally, separate WM and Inhibition scores were computed from the individual measures that loaded on each of the two EC factors (as for the other composite scores).

### 3.2 Main analyses

***The relation between implicature performance and EC***

In order to explore the relation between implicature performance and EC, a regression analysis was conducted on children's overall implicature scores. The following variables significantly correlated with the overall implicature score (based on initial bivariate correlations) and were included as predictors in the regression model: Age, General language ability, WM, IQ, and Task Version. Overall, this regression model was significant ($F(5, 127)=15.465$, $p<.05$) accounting for 38% of the variance in the dependent measure. When looking at the coefficients, only Age, General language ability, and Task Version significantly predicted overall implicature performance, the association with the first two measures being positive and the association with Version being negative ($t(127)=3.400$, $p<0.05$, $t(127)=2.745$, $p<.05$, $t(124)=-3.793$, $p<.05$). Results of this linear regression analysis are presented in Table 1.

**Table 1: Results of regression analysis on overall implicature performance (n=133).**

| | B (SE)[a] | B |
|---|---|---|
| Constant | **1.629 (0.306)** | |
| Age | **0.135**[a] (0.040)** | **0.333** |
| WM | **0.083 (0.073)** | **0.110** |
| IQ | **0.002 (0.008)** | **0.026** |
| Version | **-0.334** (0.88)** | **-0.27** |
| General language ability | **0.196** (0.071)** | **0.227** |

*Note 1:* $R^2$= .38. F-Test (5, 127)=15.465 (p<.05). * p<0.05, **p<.01

*Note 2:* Age=age in years, WM=working memory composite score, IQ=scores in the WASI matrix reasoning test, General language ability=general language ability composite score in the language of testing, Version=task version of implicatures test.

## 4. Discussion

The main goal of the present study was to explore potential relations between children's pragmatic language understanding and cognitive skills such as executive control. We also looked for evidence in children's performance with various types of implicature, that would justify the theoretical distinction between generalized and particularized implicatures. Two findings emerged from this study and these are briefly discussed in the following sections.

### 4.1 The components of pragmatic ability

A Principal Component Analysis on children's scores from four implicature sub-tests (on metaphor, manner, and scalar implicature comprehension) revealed only a single factor of implicature performance. This was not a trivial result given that the scores for each implicature type were extracted from different sub-tests that varied methodologically (e.g., in terms of instructions, method of response, verbal demands). In terms of the generalized-particularized implicature debate, this finding provides support to pragmatic theories (e.g. Geurts, 2010; Grice, 1975; Sperber & Wilson, 1996) that treat all types of pragmatically inferred meanings as the outcome of a single pragmatic interpretation process that involves uncovering the speaker's intentions behind an utterance. On the contrary, this result is not easily accommodated by theories that postulate a categorical distinction in the mechanisms involved in deriving PCIs and GCIs (Chierchia, 2004; Levinson, 2000).

### 4.2 The cognitive foundations of implicature understanding in children

Another purpose of this study was to investigate whether implicature understanding is a skill that depends on children's EC abilities. Initial bivariate correlations provided some partial support for the view that working memory possibly plays a positive role (WM was positively associated with the overall implicature score). Nevertheless, the WM effect disappeared in the regression analyses when controlling for the influence of other variables that also correlated with pragmatic performance. Instead, language proficiency in the language of testing and age proved to be strong positive predictors of pragmatic language understanding and remained statistically significant even after partialing out the effect of other third factors. Thus, the hypothesis that implicature understanding in children is an ability that draws on EC resources was not empirically verified in the current study. The findings of our experiment suggest that implicature understanding is a pragmatic-communicative skill that largely depends on

children's language abilities.

**References**

Antoniou, Kyriakos, Kleanthes K. Grohmann, Maria Kambanaros & Napoleon Katsos. in press. The effect of childhood bilectalism and multilingualism on executive control. *Cognition*.

Bernicot, Josie, Virginie Laval & Stéphanie Chaminaud. 2007. Nonliteral language forms in children: In what order are they acquired in pragmatics and metapragmatics? *Journal of Pragmatics* 39(12). 2115–2132.

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

Breheny, Richard, Heather J. Ferguson & Napoleon Katsos. 2013. Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition* 126(3). 423-440.

Breheny, Richard, Napoleon Katsos & John Williams. 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.

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

Corsi, Philip M. 1973. *Human memory and the medial temporal region of the brain.* Montreal: McGill University dissertation.

Currie, Candace E., Rob A. Elton, Joanna Todd & Stephen Platt. 1997. Indicators of socioeconomic status for adolescents: The WHO Health Behaviour in School-aged Children Survey. *Health education research* 12(3). 385-397.

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

Ellefson, Michelle R., Zewelanji Serpell & Teresa Parr. July 2011-June 2014. *Exploring the malleability of executive control*. (R305A110932). Grant awarded to the University of Cambridge by the Institute for Educational Sciences, United States Department of Education.

Geurts, Bart. 2010. *Quantity Implicatures*. Cambridge: Cambridge University Press.

Grice, Paul H. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and semantics,* vol. 3, 225-242. New York: Seminar Press.

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

Grodner, Daniel J. & Julie Sedivy. 2005. The effect of speaker-specific information on pragmatic inferences. In Neal J. Pearlmutter & Edward Gibson (eds.), *The processing and acquisition of reference*, 239-272. Cambridge, MA: MIT Press.

Guasti, Teresa Maria, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini & Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes* 20(5). 667-696.

Huang, Yi Ting & Jesse Snedeker. 2009a. On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology* 58(3). 376-415.

Huang, Yi Ting & Jesse Snedeker. 2009b. Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental psychology* 45(6). 1723-1739.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Marty, Paul P. & Emmanuel Chemla. 2013. Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology* 4, 403. 1-12

Miyake, Akira, Naomi P. Friedman, Michael J. Emerson, Alexander H. Witzki, Amy Howerter & Tor D. Wager. 2000. The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology* 41(1). 49-100.

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

Norbury, Courtenay Frazier. 2005. The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology* 23(3). 383-399.

Pouscoulous, Nausicaa, Ira A. Noveck, Guy Politzer, and Anne Bastide. 2007. A developmental investigation of processing costs in implicature production. *Language Acquisition* 14(4). 347-376.

Renfrew, Catherine E. 1995. *Word Finding Vocabulary Test*. 4th Edition. Oxon: Winslow.

Rundblad, Gabriella & Dagmara Annaz. 2010. Development of metaphor and metonymy comprehension: Receptive vocabulary and conceptual knowledge. *British Journal of Developmental Psychology* 28(3). 547-563.

Siegal, Michael & Luca Surian. 2007. Conversational understanding in young children. In Erika Hoff & Marilyn Shatz (eds.), *Blackwell handbook of language development*, 304-323. Oxford: Blackwell.

Simon, Richard J. 1969. Reactions towards the source of stimulation. *Journal of Experimental Psychology* 81(1). 174–176.

Sperber, Dan & Wilson, Deirdre. 1986/1995. *Relevance: Communication and cognition*. Oxford: Blackwell.

Tomlinson, John M., Todd M. Bailey & Lewis Bott. 2013. Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of memory and language* 69(1). 18-35.

Vogindroukas, Ioannis, Athanasios Protopapas & Georgios Sideridis. 2009. Test of Expressive Vocabulary [in Greek]. Chania: Glafki.

Waggoner, John E. & David S. Palermo. 1989. Betty is a bouncing bubble: Children's comprehension of emotion-descriptive metaphors. *Developmental psychology* 25(1). 152-163.

Wechsler, David. 1949. *Wechsler Intelligence Scale for Children*. San Antonio, TX, US: Psychological Corporation.

Wechsler, David. 1999. *Wechsler Abbreviated Scale of Intelligence (WASI)*. San Antonio: Pearson Assessment, Inc.

Winner, Ellen. 1997. *The point of words: Children's understanding of metaphor and irony*. Cambridge, MA: Harvard University Press.

Winner, Ellen, Jonathan Levy, Joan Kaplan & Elizabeth Rosenblatt. 1988. Children's understanding of nonliteral language. *Journal of Aesthetic Education*, 22(1). 51-63.

Kyriakos Antoniou, Department of Theoretical and Applied Linguistics, University of Cambridge and Département de Langues et Lettres/LaDisco, Université libre de Bruxelles, ka353@cam.ac.uk; Napoleon Katsos, Department of Theoretical and Applied Linguistics, University of Cambridge, nk248@cam.ac.uk.

# Trends in (Experimental) Pragmatics

Richard Breheny
*University College London*

## A. No more X- or T-

One hopes that in the not-too-distant future, 'Experimental Pragmatics' will disappear from the academic landscape; and with it, 'Theoretical Pragmatics'. Any division like this is clearly nonsensical (as Jesse Snedeker once pointed out to me when I invoked it in a discussion). Pragmatics as we understand it is a part of a broader 'Cognitive Science' or 'Cognitive Psychology' enterprise. We have insights, theories, hunches, make generalisations, etc.. We try to build these into systematic statements about the nature of utterance interpretation in context. Then we need to test and revise. I think that, to an extent, the trend toward the end of X- vs. T-Prag is already beginning. One sees increasingly more systematic methods of collecting data (surveys, laboratory experiments, etc.) appearing in presentations at key international conferences like SALT and S&B; also in key journals, like J. of Semantics, Semantics & Pragmatics and elsewhere. Increasingly, on the back of the increasing interest in Rational / Probabilistic / Bayesian models (see below), research on phenomena traditionally studied in pragmatics is appearing beyond the small-scale Linguistics journals in such places as Science, Cognitive Psychology, Cognitive Science, Cognition and so forth.

## B. The latest dance craze ...

Like elsewhere in cognitive psychology / science, a lot of interest is being generated by Rational / Probabilistic / Bayesian (RPB) approaches to pragmatic phenomena. In many respects, Pragmatics seems an ideal target for this approach. To date, the bulk of pragmatics (and semantics) research has consisted of more-or-less formally explicit insights or generalisations about language phenomena. RPB approaches seem to offer a useful tool to develop these insights or generalisations into a computational-level model (in Marr's sense) that can be compared to data collected via surveys or in the lab. There has been a debate about the value of RPB approaches more broadly in the CogSci literature (Chater et al., 2006; Kemp & Tenenbaum, 2008; Jones & Love, 2011; Marcus & Davis, 2013). Here I will briefly discuss two recent papers (Potts et al., 2015; Degen et al., 2015) highlighting some of the strengths as well as the points of criticism that have been outlined elsewhere and seem relevant


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 18–21.

to Pragmatics. First I should say that these are great papers that report neat studies and provide us with good data. Second, both demonstrate the flexibility of RPB modelling, which is seen as a potential source of strength (Jones & Love, 2011; Chater et al., 2011) and as a source of weakness due to the temptation to overfit in various ways (Jones & Love, 2011; Marcus & Davis, 2013). A display of perhaps both these sides of RPB can be found in Degen et al., (2015). The question there is why (1) still gives rise to the quantity implicature (Geurts, 2010).

(1) I threw ten marbles in the pond. Some of them sank.

Degen et al. establish by survey that it does give rise to the implicature and that this contradicts the prediction of the standard RSA model of Frank & Goodman (2012), based as it is on the very high prior probability that marbles sink. The solution offered is to change the RSA model to allow for variable priors. Degen et al.'s account of what goes on in these cases is that listeners accommodate 'wonky' background assumption about the world when the utterance seems to dictate. This paper highlights the positive and negative side of the framework: On the one hand, it shows the flexibility of RPB modelling with regards to accommodating any new kind of insight or data. On the other, the final account offered here leaves open completely the question raised by Geurts' question in the first place, why should a defeasible pragmatic inference not be blocked by world knowledge in this case? Where does the explanation for this lie? Potts et al. (2015) demonstrate the strengths of RPB modelling, providing a comparison in performance between a literal-only model, a standard Gricean model and two that allow for local enrichment - one completely unconstrained and one constrained according to neo-Gricean proposals. From a positive perspective, the paper opens the way for further research on how local enrichment processes are constrained. That being said, the results of this comparison raise more questions than the RPB framework itself seem able to settle. That a constrained local-enrichment model seems to do better (on certain items) than an unconstrained one calls for an explanation. In fact, given the constraints used it seems to call for an answer to the very old question, where do scales come from? My bet is that this question will be answered at a different level of explanation (see below). To see why I think that, imagine if were we to conduct a similar study using figurative/metaphorical items.[1] It does not even make sense to consider neo-Gricean constraints on local metaphoric operations. Thus any such constraints would be

1 Note that, like scalars, metaphors are pragmatic effects that are subject to embedding, as the pair in (ia,b) - adapted from a poem by Carl Sandburg - illustrate:

(i) a. The fog comes in on little cat feet.
b. The fog rarely comes in on little cat feet.

specific to local scalar implicatures. Constraints on (embedded) metaphors would presumably arise from a different set of mechanisms that are intimately linked to the structure of the lexicon, how activation spreads through long-term memory, how features become salient for a metaphoric interpretation, and so forth (see Wilson & Carston, 2007). This brings me to the third comment, which echoes the main message of Jones & Love (2011). To consider optimal design only at the level of behaviour (as RPB approaches do) seems to get things the wrong way around. Assuming that optimal behaviour results, in part even, from evolutionary pressure, then those pressures affect behaviour via the level of mechanism (or bias, heuristic etc.). The bulk of research in experimental psychology investigates the mind at the level of mechanism. 'Enlightened Bayesianism' (ibid) needs to engage with this research, providing a framework in which theoretical claims can be compared. In pragmatics, as elsewhere in psychology, theory needs to framed in terms of the mechanisms that underpin behaviour.

### C. The near future (for me)

As mentioned, I see pragmatics as part of cognitive psychology. I see an important contribution to pragmatic theory coming from experimental psychological research. To me it seems necessary and somewhat overdue. The existence of key phenomena in pragmatics, to do with scalar implicature, common knowledge, non-literal language among many others, are hard to square with the assumptions about interlocutors that Grice and his contemporaries made. For example, as many researchers since Kroch (1972) have noted, scalars are inexplicable using Grice, without some further assumptions. I am sceptical that there can be a linguistic/structural theory of alternatives (Breheny et al., under review). I think the right explanation for scalars will make reference to a set of extra-linguistic mechanisms, possibly those related to the confirmation bias that exists elsewhere in cognition. In my presentation I will discuss common ground and the presence of something like, 'common knowledge' in discourse in terms of processing biases that are present in infancy, and possibly innate (Csibra, 2010).

### References

Breheny, R., Klinedinst, N., Romoli, J., & Sudo, Y. (under review). Does the structural approach to alternatives give us just enough alternatives to solve the symmetry problem? Ms., University College London and Ulster University.

Chater, N. Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M. & Tenenbaum, J. B.. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. Behavioural and Brain Sciences

Chater, N., Tenenbaum, J. B. & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. Trends in Cognitive Sciences 10(7):287-91.

Csibra, G. (2010). Recognizing communicative intentions in infancy. Mind & Language, 25, 141-168.

Degen, J., M. H. Tessler, & N. D. Goodman. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336, 998.

Geurts, B. (2010). Quantity Implicatures. Cambridge: Cambridge Univ Press

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioural and Brain Sciences, 34(4), 169-188.

Kemp, C. & Tenenbaum, J. B. (2008) The discovery of structural form. Proceedings of the National Academy of Sciences USA 105:10687-92.

Kroch, Anthony. 1972. Lexical and inferred meanings for some time adverbs. In Quarterly progress report of the research laboratory of electronics 104, MIT.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? Psychological Science, 24, 2351-2360.

Potts, Christopher; Daniel Lassiter; Roger Levy; and Michael C. Frank. (2015). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. To appear in Journal of Semantics.

Wilson, Deirdre & Robyn Carston. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In Noel Burton-Roberts (ed.), Pragmatics, 230-259. Basingstoke and New York: Palgrave Macmillan

# Is children's referential informativity associated with their visual or linguistic abilities?

Catherine Davies[1], Helene Kreysa[2]

**Abstract**

4-year-old and 7-year-old children took part in a referential communication task. Their referring expressions were measured for informativity, and their eye movements were analysed to investigate whether fixations to a contrast object predict referential informativity. Performance on a battery of standardised tests was also measured. In line with previous work, we found a developmental trajectory towards greater informativity as children mature. The eye tracking data suggest that even though 4-year-olds engage in comparison activity to a similar extent as 7-year-olds and adults, their scanning behaviour is not linked to their ensuing referential informativity. Like adults, older children appear to make greater use of information gleaned from their visual scanning, supported by their more advanced linguistic skills. Results support a processing-based (cf. pragmatic-based) account of referential informativity.

## 1. Introduction

In learning to communicate effectively children must learn to refer to objects unambiguously by using at least minimally informative referring expressions (e.g., *the small apple* to refer to the smaller of a pair of apples). In doing so, they must consider the referential context, for example its visual, social, and functional aspects, such that their addressee can identify their intended referent. The ability to produce informative expressions develops throughout early childhood, with children first passing through a phase of habitual underinformativity in which they produce expressions such as *the apple* in a two-apple context, before they master the ability to produce felicitous referring expressions (hereafter REs) at around 7 years of age (Davies & Katsos, 2010; Sonnenschein, 1982; Matthews, Lieven & Tomasello, 2007).

The developmental trajectory of referential communication has been investigated by a substantial collection of studies (for reviews see Dickson, 1982; Graf & Davies, 2013). This body of work has put forward several explanations for early underinformativity, for example, difficulties in understanding that a RE must describe differences between target and distracter items (Whitehurst, 1976; Whitehurst & Sonnenschein, 1981); performance-related demands (Matthews et al., 2007); lack of perspective-taking (Nadig & Sedivy, 2002, i.a.), as linked to executive function skills (Nilsen & Graham, 2009).

The existing literature on the development of reference has focused on children's concurrent cognitive and linguistic capabilities but has not yet comprehensively addressed the question of how *visual scanning behaviour* might affect referential informativity (appealed for by Deutsch & Pechmann, 1982: 178; investigated in adults by Brown-Schmidt & Tanenhaus, 2006, and recently in 4-5 year-olds by Nilsson, Catto & Rabagliati, 2014). Pechmann (1989: 98) suggested that incomplete visual scanning may be a reason for failures in informativity, but did not provide developmental data to support this. The current study speaks to this gap by examining the relationship between children's eye movements and the form of their REs. It combines experimental methods from language production and those using eye movements as an index of cognitive processes, and reveals differences in the rate at which children between 4 and 7 years of age integrate information from the visual scene into their referential choices. In line with

[1] Corresponding author. Dept. of Linguistics & Phonetics, University of Leeds, UK. c.n.davies@leeds.ac.uk
[2] Dept. of Psychology, Friedrich Schiller University, Jena, Germany. helene.kreysa@uni-jena.de



 In: Fabienne Salfner and Uli Sauerland (eds.), *Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 22–30.

previous work, it also measures children's cognitive and linguistic profiles. We ask three main research questions:

1. What is the developmental trajectory in informativity when children refer to objects in simple and more complex visual scenes? *H1: 4-year-old children will largely be underinformative in this simple referential task, especially in complex displays, whereas 7-year-olds will provide more informative expressions, though not to the same extent as the adult comparison group.*

2. Do children who tend to provide underinformative referring expressions have a common linguistic / cognitive profile? *H2: Children who tend to provide underinformative referring expressions have a common linguistic / cognitive profile.*

3. *a)* What is the pattern of fixations before informative vs. underinformative referring expressions as a function of age? *b)* What is the pattern of fixations during informative vs. underinformative referring expressions as a function of age? *H3a. For all age groups, the contrast object will be fixated more frequently before informative referring expressions than before underinformative referring expressions. H3b. For all age groups, the contrast object will be fixated more frequently during informative referring expressions than during underinformative referring expressions.*

## 2. Method

*Design*. The experiment had a mixed design. For measuring the form of REs from participants' **production data** (Section 3.1), the experiment had a 2 x 2 x 2 design (age group x contrast x display complexity). Age group was between-participants (4-year-olds; 7-year-olds). Contrast (present or absent = two referents vs. one referent from the same noun category) and display complexity (four or eight objects) were within-participants. The dependent variable was utterance type: underinformative, informative, or overinformative[3]. For measuring the relationship between **eye movements and informativity** (Section 3.3), the contrast variable was dropped from the analysis, that is, only contrast-present items were included since this analysis focused on looks to the contrast object (which was of course absent in the contrast-absent condition). Utterance type was included as an independent variable. The dependent variable was the presence of fixations to the contrast object during two time windows (pre- and during-utterance).

*Participants.* Table 1 contains participant profile information. All were monolingual native speakers of British English. 24 adults were also recruited for a separate study with a similar methodology (see Davies & Kreysa, in prep.), and acted as the comparison group herein.

*Materials and Procedure:*

**Referential communication task.** The stimuli consisted of 44 displays of everyday objects, grouped into semantically related sets, e.g., animals, food, household objects, clothes. 16 displays were critical items, 24 were fillers and four formed the practice block. Of the critical items, half of the displays contained four objects and half contained eight objects, constituting simple and complex displays respectively (see Figure 1 for example displays). Half of the critical displays contained a no-contrast display with only one referent of each noun category (e.g., a ball, a doll, a teddy and a car) and half contained a contrast display featuring two referents of the same noun category (e.g., a large apple, a small apple, a sausage and a sandwich), one of which was the target thus requiring modification for disambiguation. Target objects differed from their

[3] Because overinformativity was rare in the data (2% of all utterances by the younger children and 8% by the older children), statistical comparisons are restricted to rates of optimal informativity vs. underinformativity.

contrast mates by size (large vs. small). These 16 critical items all appeared in four pseudorandomised lists, counterbalanced for target attribute and for block order, meaning that half the participants saw for example, the small apple as the target while the other half of the participants saw the large apple as the target. No target object appeared more than once throughout the experiment, and the position of the target and the contrast object was rotated around each slot within the four- and eight-object displays. Stimuli were presented and eye movements were recorded using Tobii Studio software. The sequencing of each trial was as follows: a fixation cross was displayed for one second, a preview of the displays (target not highlighted) was displayed for three seconds for four-object displays and for four seconds for eight-object displays. A fixation cross in the form of a red star then appeared on screen within the preview for one second, then the fixation star disappeared and the target was highlighted with a red frame around the object for five seconds, during which time the participant produced their utterance of the form 'click on the X'.

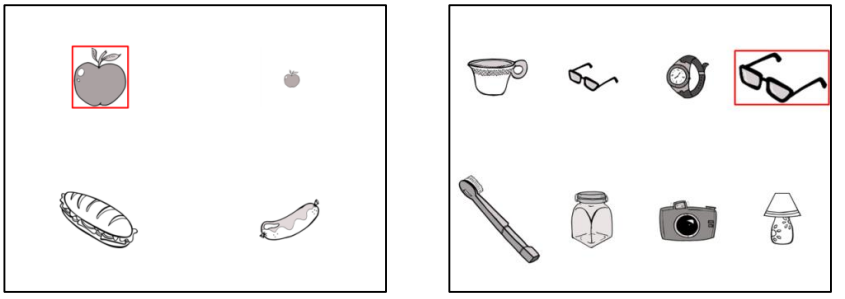


Figure 1. Visual stimuli. Left hand panel shows a four-object item; right hand panel shows an eight-object item. Both panels are two-referent displays, target highlighted.

Participants were seated in front of a Tobii X120 remote desk-mounted eye tracker and monitor, with the experimenter seated at a laptop nearby. The two monitors were not mutually visible. A five-point calibration was performed, then participants were instructed as follows: *We're going to play a game. Your job is to help me find some pictures. You'll see some pictures on the screen. I can see them too, but they're not in the same place on my screen. Look at the pictures on your screen. A red box will appear around one of them for you. You should tell me to click on that picture, like "click on the number 7". You'll also see a red star - you should always try to look at the red star when it appears. We'll practice a few times first and then we'll play the game.* We emphasised that their role was to tell the experimenter to click on the highlighted item. During the experiment, the experimenter clicked a mouse to signal that they had found the referent roughly one second after the offset of the participant's utterance, regardless of the form of RE used. No other feedback was given.

**Standardised tests.** Three tests of linguistic and cognitive abilities were administered to correlate participants' profiles with their informativity in the referential communication task. As an index of receptive language ability, the British Picture Vocabulary Scale (BPVS-III) was used, normed for 3 – 16 year-olds (Dunn, Styles & Sewell, 2009). For visual search efficiency, the Bug Search task from the WPPSI-IV battery was used (Wechsler, 2013). This is a processing speed subtest for ages 4;0 – 7;7 and measures participants' perceptual speed, short-term visual memory, cognitive flexibility, visual discrimination, and concentration whilst they match images within a field of five to a reference image. As a measure of perspective-taking ability within a discourse context, the Short Narrative subtest from the DELV-ST (Diagnostic Evaluation of Language Variation), recommended for use with 4 – 9 year-olds (Seymour, 2003). The whole testing session lasted approximately 30 minutes.

## 3. Results

3.1 Referential communication task: Production data

In an analysis of all production data (contrast and no contrast conditions; four and eight object displays), 4-year-olds were numerically equivocal in the informativity of their REs (42% underinformative and 52% informative) whereas 7-year-olds were more optimal in their referential choices (23% underinformative and 68% informative).

For the contrast items only, across the two levels of display complexity, 4-year-olds were largely underinformative in their referential choices (83% underinformative and 12% informative) whereas 7-year-olds were more equivocal (46% underinformative and 53% informative). Adults were largely informative at a mean rate of 79% (see Figure 2). A mixed ANOVA found a main effect of age on informativity: 4yo mean = 12% (SE = 5), 7yo mean = 53% (SE = 6), $F(1, 44) = 24.22$, $p < .001$, $\eta^2p = .36$. There was a main effect of display complexity: four-object mean = 43% (SE = 5), eight-object mean = 21% (SE = 4), $F(1, 44) = 42.24$, $p < .001$, $\eta^2p = .49$. There was also a significant interaction between age and complexity in that increased complexity compromised informativity for the 7-year-olds to a greater extent than the 4-year-olds, $F(1, 44) = 18.12$, $p < .001$, $\eta^2p = .29$. This is likely driven by floor effects in the younger group. Within-group pairwise comparisons were performed using Wilcoxon signed-rank tests with a Bonferroni correction applied, resulting in a significance level set at $p < 0.025$. For the 4-year-olds, mean rates of informativity in the four-object condition (15%, SD = 29) were significantly higher than in the eight-object condition (8%, SD = 24), $Z = -2.33$, $p = .020$, $r = -.45$). For the 7-year-olds, mean rates of informativity in the four-object condition (71%, SD = 38) were similarly significantly higher than in the eight-object condition (35%, SD = 30), $Z = -3.34$, $p = .001$, $r = -.77$). Within-complexity pairwise comparisons were performed using Mann Whitney tests with the same correction applied. For the four-object displays, mean rates of informativity by the 4-year-olds (16%, SD = 29) was significantly lower than by the 7-year-olds (71%, SD = 38), $U = 78.0$, $z = -4.26$, $p < .001$. For the eight-object displays, mean rates of informativity by the 4-year-olds (8%, SD = 24) was significantly lower than by the 7-year-olds (36%, SD = 30), $U = 97.5$, $z = -4.01$, $p < .001$.

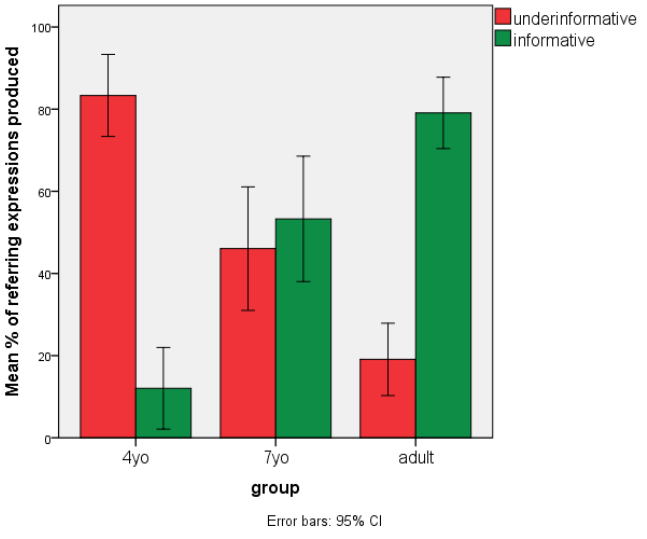



Figure 2. Mean rates of informativity of REs as a percentage of all expressions produced; contrast condition only.

As predicted by our first hypothesis, the younger children were largely underinformative when referring to objects for their addressee, whereas their older counterparts were less so, though not

as informative as the adult comparisons. Both child groups produced more underinformative expressions when displays were complex (as did the adults), though this effect was more pronounced in the 7-year-olds.

### 3.2 Correlational analyses with standardised tests

A Pearson correlation coefficient was computed to assess the relationship between informativity of REs and performance on the standardised tests (see Table 1 for scores). Amongst the 7-year-olds, rates of underinformativity (contrast condition only) were negatively correlated with performance on receptive vocabulary test ($r$ = -.41, $p$ = .08), but not on visual search or perspective-taking measures. Among the 4-year-olds and the adult comparison group, there were no significant correlations between informativity and any of the standardised measures (all $p$s > .1; all $rs$ <.3), though this may have been driven by floor and ceiling effects. These results partially support our second hypothesis that children who tend to provide underinformative referring expressions have a common linguistic / cognitive profile, in that they tentatively indicate that language ability may underpin informative referring as children mature.

| | **4 yos (n=27; 13 males)** | **7 yos (n=19; 8 males)** | **adults (n=24; 4 males)** |
|---|---|---|---|
| age (y;m) | 4;7 (0;5) | 7;9 (0;6) | 19 (1;5) |
| range | 4;0 - 5;6 | 6;9 - 8;6 | 18 - 23 |
| BPVS (raw) | 74.1 (11) | 110 (15.5) | 161.3 (4.1) |
| range | 54 - 99 | 84 - 140 | 151 - 167 |
| BPVS (standardised) | 109.3 (6.9) | 103.1 (13.1) | 111.4 (7.4) |
| range | 91 - 124 | 81 - 126 | 96 - 124 |
| DELV narrative | 3.5 (1.6) | 5.8 (1.3) | 5.8 (1.2) |
| range | 1 - 7 | 2 - 7 | 0 - 7 |
| WPPSI-IV Bug Search (raw) | 21.9 (8.7) | 42.4 (8.3) | - |
| range | 6 - 42 | 29 - 60 | - |

Table 1. Scores on background measures: mean (sd).

### 3.3 Eye movement data

Only the contrast condition was analysed in this section since the dependent variable is the number of fixations to the contrast object, which is absent in the no-contrast condition. Due to overall sampling validities of <40%, seven participants were excluded from the eye tracking analysis (three from the 7-year-old group). This left the remaining younger sample at n=23, mean age 4;8 years (SD 0;5), range 4;0 – 5;4, 12 males. The remaining older sample was n=16, mean age 7;10 (SD 0;7), range 6;9 – 8;6, 6 males. Separate analyses were run for the pre-utterance and the utterance time windows.

#### 3.3.1 Contrast fixations during the pre-utterance time window

To investigate the relationship between speaker informativity and whether the contrast object had been fixated, we looked at the proportion of trials in which speakers did or did not fixate the contrast object before producing the two major utterances types (informative and underinformative). Trials were divided into those involving at least .3 of a fixation to the contrast

object in the pre-utterance time window (individual fixations which spanned all three time windows were divided by 3) and those involving no fixations to the contrast in the same time window.

4-year-olds were overall more likely not to fixate (74%) than to fixate (26%) the contrast object in the pre-utterance time window, and they were much more likely (85%) to produce an underinformative than an informative utterance (15%)[4]. Importantly, the likelihood of producing an underinformative utterance after looking at the contrast object (32/42=76%) was comparable to the likelihood of producing an underinformative utterance without having looked at the contrast object (106/120=88%). Put another way, the likelihood of producing an informative utterance after looking at the contrast object (10/42=24%) was comparable to the likelihood of producing an informative utterance without having looked at the contrast object (14/120=12%). A chi-square analysis of the likelihood of producing informative and underinformative utterances missed significance ($\chi^2(1) = 3.64$, $p = .08$). This was the case for the four-object and eight-object items combined as well as when the two levels of display complexity were analysed separately (four-objects: $\chi^2(1) = .22$, *ns*; eight-objects: $\chi^2(1) = 5.17$, $p = .06$) indicating that fixating the contrast object plays only a minor role for informativity in this age group. Thus, 4-year-olds are overwhelmingly underinformative regardless of fixation to the contrast.

7-year-olds were also more likely not to fixate (64%) than to fixate (36%) the contrast object in the pre-utterance time window, and they equally likely (52%) to produce an informative as an underinformative utterance (48%). Like adults, they were more likely to produce an underinformative utterance without having looked at the contrast object (49/78=63%) than to produce an underinformative utterance after looking at the contrast object (10/44=23%). Put another way, the likelihood of producing an informative utterance after looking at the contrast object (34/44=77%) was almost double the likelihood of producing an informative utterance without having looked at the contrast object (29/78=37%). A chi-square analysis revealed a significant association between informativity and contrast fixation, ($\chi^2(1) = 18.11$, $p < .001$), which also held when the two levels of display complexity were analysed separately (four-objects: $\chi^2(1) = 5.62$, $p < .05$; eight-objects: $\chi^2(1) = 10.77$, $p < .005$). This pattern of results suggests that contrast fixations lead 7-year-olds to produce informative REs.

The pattern shown by the older child sample was mirrored in the adults' data, who were more likely not to fixate (61%) than to fixate (39%) the contrast object in the pre-utterance time window, and who were more likely to produce an informative (79%) than an underinformative utterance (20%). Crucially, they were more likely to produce an underinformative utterance without having looked at the contrast object (52/187=28%) than to produce an underinformative utterance after looking at the contrast object (11/121=9%). In other words, they were more likely to produce an informative utterance after looking at the contrast object (110/121=91%) than to produce an informative utterance without having looked at the contrast object (7135/187=2%), $\chi^2(1) = 15.82$, $p < .001$. The same boosting effect of contrast fixation was found at both levels of complexity, that is in four-object displays ($\chi^2(1) = 5.83$, $p < .05$) and in the eight-object displays ($\chi^2(1) = 11.09$, $p < .005$).

The results from the 7-year-olds support hypothesis 3a in that the contrast object was fixated more frequently before informative REs than before underinformative REs, across both types of display complexity (as was also the case for the adult group). In contrast, the results from the 4-

[4] These percentages vary slightly from those reported in the production results due to the exclusion of four participants from the eye tracking analysis.

year-olds did not support our hypothesis. Instead, the younger children's looking behaviour in the pre-utterance region was independent of later informativity.

3.3.2 Contrast fixations during the utterance time window

We ascertained in Section 3.3.1 that younger children are no more likely to be informative whether or not they fixate the contrast before starting to speak, and conversely, that older children's informativity is boosted by fixating the contrast pre-utterance, like adults. In this section, we investigate whether later contrast fixations are linked to children's choice of RE. In line with the analysis of contrast fixations during the pre-utterance region, we looked at the proportion of trials in which speakers did or did not fixate the contrast object while producing the two main utterances types.

4-year-olds were overall more likely not to fixate (62%) than to fixate (38%) the contrast object in the utterance time window. Importantly, they were more likely to produce an underinformative utterance whilst not looking at the contrast object (92/101=91%) than whilst looking at it 46/61=76%). Put another way, looking at the contrast object raised the likelihood of 4-year-olds producing an informative RE (15/61=25%) relative to them not looking at it (9/101=9%), $\chi^2(1) = 7.41$, $p < .05$. This pattern is driven by the significant relationship between contrast fixations and informativity for the eight-object displays ($\chi^2(1) = 21.9$, $p < .001$) rather than the more simple four-object displays ($\chi^2(1) = .06$, ns). Thus, despite pre-utterance contrast fixations not playing a significant role in informativity for 4-year-olds, they do provide a boost to informativity once a young child has started to produce their RE in complex displays.

7-year-olds were slightly more likely not to fixate (56%) than to fixate (44%) the contrast object in the utterance time window. Like adults, they were equally likely to produce an underinformative utterance whilst not looking at the contrast object (35/68=51%) as they were to produce an underinformative utterance whilst looking at the contrast object (24/54=44%). In other words, they were equally likely to produce an informative utterance whilst looking at the contrast object (30/54=56%) as they were to produce an informative utterance whilst not looking at the contrast object (33/68=49%), $\chi^2(1) = 0.60$, ns. The same pattern of results was found in the analysis split by display complexity, that is, 7-year-olds were equally as likely to be informative regardless of whether or not they were fixating the contrast object in both four-object displays ($\chi^2(1) = 0.48$, ns) and in eight-object displays ($\chi^2(1) = .32$, ns).

The pattern shown by the older child sample was mirrored in the adults' data, who were equally likely not to fixate (58%) as to fixate (42%) the contrast object in the utterance time window. Adults were equally likely to produce an underinformative utterance whilst not looking at the contrast object (35/179=20%) as they were to produce an underinformative utterance whilst looking at the contrast object (28/129=22%). In other words, they were equally likely to produce an informative utterance whilst looking at the contrast object (101/129=78%) as they were to produce an informative utterance whilst not looking at the contrast object (144/179=80%), $\chi^2(1) = 0.21$, ns. The same pattern of results was found in the analysis split by display complexity, that is, adult speakers were equally as likely to be informative regardless of whether or not they were fixating the contrast object in both four-object displays ($\chi^2(1) = 0.37$, ns) and in eight-object displays ($\chi^2(1) = 1.57$, ns).

As in the pre-utterance contrast fixation analysis in Section 3.3.1, older children and adults pattern similarly whilst younger children show a different relationship between contrast fixations

and informativity. In the utterance time window, younger children's informativity benefits from fixations to the contrast object, whereas fixating (or not fixating) the contrast object whilst speaking did not influence older children's and adults' concurrent tendency to produce informative expressions, with no effect of display complexity.

Thus hypothesis 3b (that the contrast object will be fixated more frequently during informative REs than during underinformative REs) is supported for the younger children but not for the older children and adults. This suggests that late looking can boost informativity in younger children in the same way that pre-utterance contrast fixations do for their older counterparts. In contrast, 7-year-olds' and adults' informativity does not benefit from these later contrast fixations.

## 4. Discussion and conclusions

We have replicated previous studies which found a developmental shift from underinformativity to full informativity as children mature from 4 to 7 years of age (Davies & Katsos, 2010; Matthews et al., 2007; Whitehurst & Sonnenschein, 1981, i.a.). Our correlational analyses using a range of linguistic and cognitive tests revealed a link between language ability (as indexed by a receptive vocabulary test) in the older children, suggesting that it is the modified noun structure which may be the most challenging aspect of this task for the younger children.

One area of ability which we did not measure and which may yield significant effects is executive functioning; a set of skills which has recently been found to boost children's perspective taking and associated referential informativity (Nilsen & Graham, 2009; Nilsen, Varghese, Xu & Fecica, 2015). These studies suggest that greater inhibitory control and working memory skills enable children to use their communicative partner's perspective. In future work we would like to investigate whether such skills can also boost more comprehensive visual scanning and/or the integration of information from contrast objects with referential choice. The 4-year-olds only fixated the contrast object before speaking on 26% of trials, which indicates a somewhat automatic 'see-the-target, say-the-target' strategy. This lack of attention on non-target items may be related to inhibitory control. However, the 7-year-olds and adults only fixated the contrast on 36% and 39% of trials respectively, yet produced higher frequencies of informative REs than their younger counterparts, suggesting that pre-utterance contrast fixations are not the whole story with regard to informativity. Further research is needed to investigate the relative influence of i) contrast fixations and ii) the integration of contrast information (in addition to language ability) on informativity across development.

Our eye tracking results show that the younger children's pre-utterance contrast fixations do not strongly influence their informativity unlike their older counterparts (cf. Nilsson et al. 2014), allowing us to rule out incomplete visual scanning as a reason for early underinformativity (cf. Deutsch & Pechmann, 1982; Pechmann, 1989). Language ability emerges as a stronger constraint on informativity: children must be in a state of linguistic readiness in order to produce fully informative referring expressions, though this is likely to be boosted by their visual scanning behaviour.

**References**

Brown-Schmidt, S. and Tanenhaus, M.K. 2006. Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, 54, 592-609.

Davies, C. and Katsos, N. 2010. Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua 120*(8): *Special Issue on Asymmetries in Child Language,* 1956-1972.

Deutsch, W. and Pechmann, T. 1982. Social interaction and the development of definite descriptions. *Cognition*, 11, 159-184.

Dickson, W. 1982. Two decades of referential communication research: A review and meta-analysis. In C. J. Brainerd and M. Presley (eds), *Verbal Processes in Children,* 1-33. New York: Springer Verlag.

Dunn L. M., Dunn D. M., Styles B., Sewell J. 2009. *The British Picture Vocabulary Scale*, 3rd Edn. (BPVS-III). London: GL Assessment.

Graf, E. and Davies C. 2014. The Production and Comprehension of Referring Expressions. In D. Matthews (Ed.) *Pragmatic development in first language acquisition, Trends in language acquisition research*, 161-181. Amsterdam: John Benjamins.

Matthews, D., Lieven, E., and Tomasello, M. 2007. How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*. *78*(6), 1744-1759.

Nadig, A. S. and Sedivy, J. C. 2002. Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science,* 13(4), 329-336.

Nilsen, E. S., and Graham, S. 2009. The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, 58, 220-249.

Nilsen, E.S., Varghese, A., Xu, Z. and Fecica, A. 2015. Children with stronger executive functioning and fewer ADHD traits produce more effective referential statements. *Cognitive Development,* 36, 68-82.

Nilsson, J., Catto, K. and, Rabagliati, H. 2014. Awareness and monitoring in children's referential communication. Poster presented at the *Boston University Conference on Language Development* (BUCLD 39), November 7-9.

Pechmann, T. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27, 89-110.

Seymour, H. N., Roeper, T. and De Villiers, J. G. 2003*. DELV-ST (Diagnostic Evaluation of Language Variation) Screening Test.* San Antonio TX: The Psychological Corporation.

Sonnenschein, S. 1982. The effects of redundant communication on listeners - when more is less. *Child Development,* 53(3), 717-729.

Wechsler, D. 2013. *Wechsler Preschool and Primary Scale of Intelligence*, 4th Edn. (WPPSI-IV) London: Pearson.

Whitehurst, G. J. 1976. Development of communication - changes with age and modeling. *Child Development,* 47(2), 473-482.

Whitehurst, G. J. and Sonnenschein, S. 1981. The development of informative messages in referential communication: Knowing when vs. knowing how. In W. P. Dickson (ed.), *Children's oral communication skills,*127-142. New York: Academic Press.

# Using corpus methods can begin to address how children acquire presupposition triggers

Rachel Dudley
*University of Maryland*

Meredith Rowe
*Harvard University*

Valentine Hacquard
*University of Maryland*

Jeffrey Lidz
*University of Maryland*

## 1 Introduction

*Know* is a presupposition trigger. We can see this when we compare *know* to a closely related word that shares the same asserted content: *think* (1).

(1) a. John knows that Mary is home
b. John thinks that Mary is home

Sentences like (1a) can only describe true beliefs while sentences like (1b) can describe either true or false beliefs. This is the case because while both (1a) and (1b) entail that John has a belief about Mary's location, only (1a) entails that Mary is actually home. Accordingly we tend to use sentences like (1a) in situations where we take it for granted that Mary is home and not in situations where Mary's location is up for debate (Stalnaker 1974). Furthermore, even when embedded in entailment-cancelling environments like the family-of-sentences (Chierchia & McConnell-Ginet 2000), the truth of the complement projects with *know* but not *think* (2).

(2) a. John doesn't know that Mary is home
b. John doesn't think that Mary is home

However, *know* is a weak presupposition trigger, so uses of it do not reliably exhibit the properties taken to be diagnostic of presupposition triggers (Abusch 2010, among others), such as projecting the truth of their complements, conveying information that is taken for granted, and supplying not-at-issue content:

(3) I didn't know that you have a daughter because you don't, in fact, have a daughter (Simons et al. 2014)

(4) Did you know that John won the lottery?

(5) Q: Where was Louise yesterday?
A: I know from Henry that she was in Princeton




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 31–38.

In cases like (3), the truth of *know*'s complement does not project out of negation (see Beaver 2010 for many other examples of presupposition cancellation with *know* and other cognitive factive verbs). Sentences like (4) can be felicitiously uttered discourse-initially, indicating that the information expressed in the complement need not be mutually familiar to all conversational participants and Spenader (2003) finds that *know* and similar verbs are used to convey addressee-new information over half the time in a corpus of speech between adults. Dialogues like (5), which comes from Simons (2007), demonstrate that the content of the complement can address the Question Under Discussion. If these examples which we find attested in adult speech are present with any frequency in speech to children, how can the language learning child be expected to discover that verbs like *know* are factive?

In acquiring *know*, the learner must discover: (i) that the complement is entailed, (ii) that the truth of the complement is taken for granted; (iii) that the truth of the complement projects out of entailment-cancelling contexts; and (iv) that the addressee should accommodate any new information expressed by the complement (all else being equal). And there is a consensus in the acquisition literature that children have discovered this quite early on. While full understanding of factivity does not develop until much later (Schulz 2003, among others), children seem to understand that the complement of *know* projects and should be accommodated by 4 (Dudley et al. 2015, Abbeduto & Rosenberg 1985, among many others).

The fact that children acquire this distinction between a factive and a non-factive verb so early indicates that there must be some signal available in their linguistic input. But what is it? Does input to children more reliably signal the presuppositon associated with *know* than speech between adults has been shown to or, alternatively, are there other cues available?

## 2 Discovering factivity

Perhaps children discover that *know* is factive by observing that it exhibits the properties we consider diagnostic of presupposition? If so, what kind of data would be useful in observing these properties?

In order to discover that the truth of the complement of *know* projects while other complements do not, the child needs to observe that *p* is true when *x knows p* is embedded in family-of-sentences contexts (6) while *p* is not necessarily true when *x thinks p* is embedded in the same contexts (7).

(6) a. John doesn't know that Mary is home
b. Does John know that Mary is home?
c. John might know that Mary is home
d. If John knows that Mary is home, he'll be happy

(7) a. John doesn't think that Mary is home
b. Does John think that Mary is home?
c. John might think that Mary is home
d. If John thinks that Mary is home, he'll be happy

To collect this data, we need compare occurences of sentences like those in (6) to the corresponding sentences in (7) within children's input. How often do children get exposed to the verbs in family-of-sentences contexts, and how often is the complement taken to be true in each case?

Alternatively, the child might discover that *know*'s complement conveys old information which doesn't address the Question Under Discussion by observing that *p* is old news or mutually familiar when *x knows p* is uttered, but not necessariliy when *x thinks p* is uttered. To collect this data, we need to look at properties of *x knows p* tokens within children's input to see whose beliefs are being discussed and what kind of information is conveyed by the complement.

## 3 Corpus study design

For this study, we coded all tokens of *know* (N=1234), *think* (N=1156) and related complement-taking verbs from the Gleason corpus in CHILDES (Gleason 1980, MacWhinney 2000) which is comprised of dinnertime conversations and play interactions between 24 children (mean age = 3.5 years) and their parents. Since the factive readings of *know* only arise when it takes declarative complements, we distinguished declarative complements from other kinds:

(8) Complement-types
a. John knows/thinks that Mary is home (declarative)
b. John knows/*thinks where Mary is (interrogative)
c. John knows/*thinks Mary's location (np)
d. John knows/thinks about Mary's location (pp)
e. John knows/thinks (null)

To examine children's experience with projection, we coded for family-of-sentences contexts (6): negation, questions, modals, and conditionals. And to get a proxy for whether the complement of *know* expresses information in the common ground, we coded for subject-types:

(9) a. Does he know that Mary is home? (3rd person)
b. You know that Mary is home (2nd person)
c. I (don't) know that Mary is home (1st person)

## 4 Corpus results

Our results indicate that children have relatively few opportunities to observe *x knows p* as compared to *x thinks p*: only 14% of *know* tokens take declarative complements, as compared to 85% of *think* tokens (Figure 1).

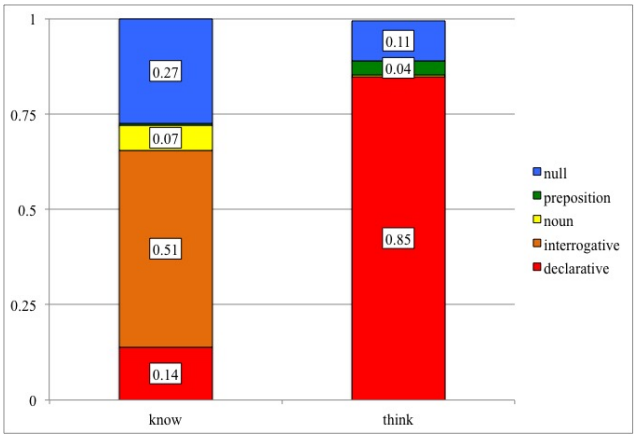



**Figure 1** Proportion of each complement type by verb

Out of those *x knows p* tokens, relatively few occur in the classic family-of-sentences contexts (Figure 2). Less than half of their experience with *x knows p* is in these potentially projective contexts (which corresponds to about 5% of all *know* tokens).

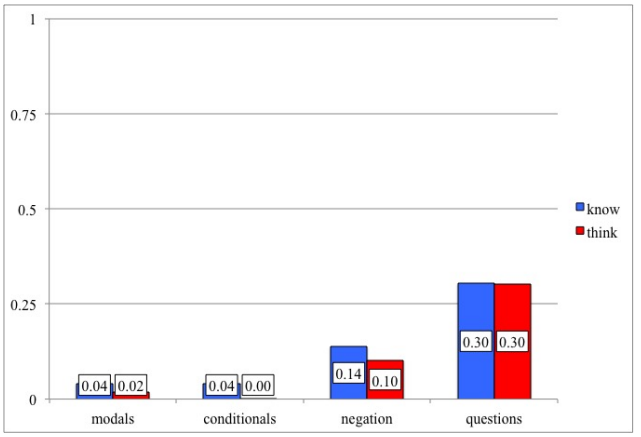



**Figure 2** Proportion of *x verbs p* tokens in family-of-sentences contexts by verb

On the other hand, children might get more opportunities to observe whether *x knows p* conveys old information instead of in projective contexts. Approximately

two-thirds of their experience with *x knows p* is in what we have identified as potentially mutually familiar contexts (Figure 3). Utterances like *Did you know that Mary is home?* were excluded because they are used to offer new information to the addressee and utterances like *I don't know that Mary is home* were excluded because they have multiple readings which aren't all factive (e.g.: I don't know if Mary is home). Future work will have to look all all of these utterances in context to determine whether they actually expressed mutually familiar information or not.

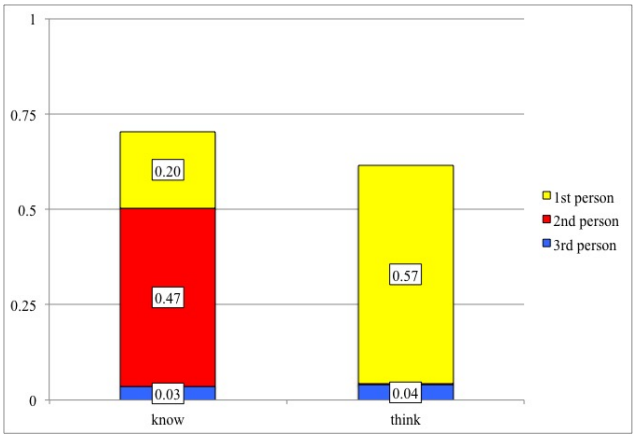



**Figure 3** Proportion of *x verbs p* tokens which possibly convey new information

This corpus study was intended as a first glimpse at what kinds of contexts the verbs are used in. Since we were evaluating thousands of tokens in this pilot study, we looked at the sentences in isolation. To get a complete picture, the next step will be to look at the sentences embedded in their discourse contexts. However, the results are indicative that children's experience does not support a learning mechanism that requires the child to observe that–overwhlemingly or more often than not–*know* exhibits properties of presupposition. Children rarely get to observe *know* in contexts where presuppositions could arise and, when they do, the distribution of *know* in those contexts is not informatively different from the distribution of *think*.

## 5 Indirect cues to factivity

While *x knows p* tokens were relatively rare, there are other kinds of contexts which we found *know* to occur in frequently and which could help the child identify *know* as a factive verb: embedding interrogatives and indirect speech acts.

*Know* takes both interrogative (10a) and non-interrogative (10b) complements, while the non-factive *think* takes only non-interrogative complements (11) and the non-factive *wonder* takes only interrogative complements (12).

(10) a. John knows where Mary is
b. John knows that Mary is home

(11) a. * John thinks where Mary is
b. John thinks that Mary is home

(12) a. John wonders where Mary is
b. * John wonders that Mary is home

These differences in the kinds of complements that the verbs embed has been linked to their factivity: only verbs which embed both kinds of complements will be factive (Hintikka 1975, Egré 2008, Ginzburg 1995, among others). And indeed in this sample, *know* and the related cognitive factive *remember* occurs with interrogative complements in addition to their declarative complements, while *think* does not (Figure 4).

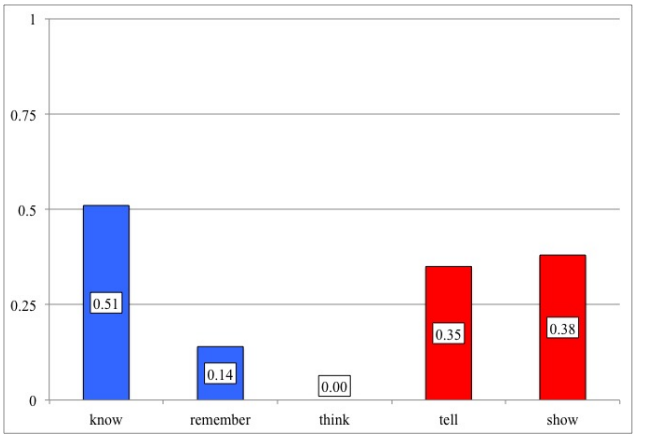



**Figure 4** Proportion of embedded interrogatives by verb

Yet we find that verbs like *tell* and *show* also occur with both interrogative and declarative complements when these verbs are not universally accepted to be factive. If we accept that these verbs have factive readings in certain contexts (e.g.: *Mary told John that she is pregnant*, see Spector & Egré 2015, Schlenker 2006), then it looks like embedding questions (in addition to other kinds of complements) is a decent cue to factivity.

Furthermore, if we zoom out to look at the whole distributions of *know* and *think* in the input, we see that they are vastly different. In two-thirds of children's experience with the verbs, they occur in different sentence frames: *Do you know where the keys are?* as compared to *I think that the keys are in the kitchen*. More often than not, children hear *know* used in second-person question with *wh*-complements

while they hear *think* in first-person assertions with declarative complements. These utterance types are used to perform different kinds of indirect speech acts which could be indicative of their different semantics. These data suggest that *know* is often used to ask and answer questions while *think* is often used to make (indirect) assertions. Since children have an understanding of indirect speech acts at this age (Spekman & Roth 1985), they might use their knowledge of indirect speech acts to uncover the verbs' (non-)factivity. More explicitly, if you often hear *Do you VERB what time it is?*, understanding the speaker to be asking for the time, you might conclude that they expect you have the (true) answer to their question and thus that verbs which occur in this context relate one to truth. This suggests that the complements that *know* and *think* embed–while superficially similar (1)–are underlyingly different: *know* embeds true propositions or facts, while *think* embeds propositions.

## 6 Conclusion

These corpus results give us a fine-grained picture of children's linguistic experience with the factive *know* (in contrast to related verbs like the non-factive *think*). This kind of data can illuminate which kinds of distributional cues to factivity are reliably available in the input and, in conjunction with further experimental work, this kind of data could lead to a greater understanding of how factive presuppositions are represented. At this stage, our results suggest that traditional contexts which are diagnostic of factivity are rare in the input, but that syntactic cues and pragmatic cues to factivity are readily available and potentially informative.

Future work will examine *know* and *think* tokens in contexts to determine (i) what the status of *p* is; (ii) what the speaker actually intends to convey with their utterance; and (iii) whether children actually use these cues in determining whether or not a verb is factive. In addition, these methods could be applied to other kinds of triggers to see if different cues are important to identifying triggers with other properties.

## References

Abbeduto, Leonard & Sheldon Rosenberg. 1985. Children's knowledge of the presuppositions of know and other cognitive verbs. *Journal of Child Language* 12(03). 621–641.

Abusch, Dorit. 2010. Presupposition triggering from alternatives. *Journal of Semantics* 27(1). 37–80.

Beaver, David. 2010. Have you noticed that your belly button lint colour is related to the colour of your clothing. *Presuppositions and Discourse: Essays Offered to Hans Kamp. Elsevier, Philadelphia, PA* 65–99.

Chierchia, Gennaro & Sally McConnell-Ginet. 2000. *Meaning and grammar: An introduction to semantics*. MIT press.

Dudley, Rachel, Naho Orita, Valentine Hacquard & Jeffrey Lidz. 2015. Three-year-olds' understanding of know and think. In *Experimental perspectives on presuppositions*, 241–262. Springer.

Egré, Paul. 2008. Question-embedding and factivity. *Grazer Philosophische Studien* 77(1). 85–125.

Ginzburg, Jonathan. 1995. Resolving questions, 1 II. *Linguistics and Philosophy* 18(6). 459–527, 567–609.

Hintikka, Jaakko. 1975. 'different constructions in terms of the basic epistemological verbs: A survey of some problems and proposals. *The Intentions of Intensionality* 1–25.

Schlenker, Philippe. 2006. Transparency: An incremental theory of presupposition projection .

Schulz, Petra. 2003. *Factivity: Its nature and acquisition*, vol. 480. Walter de Gruyter.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117(6). 1034–1056.

Simons, Mandy, E Kummerfeld, D Beaver, C Roberts & J Tonhauser. 2014. The best question: explaining the projection behavior of factives. *Discourse Process (to appear)* .

Spector, Benjamin & Paul Egré. 2015. A uniform semantics for embedded interrogatives: An answer, not necessarily the answer. *Synthese* 192(6). 1729–1784.

Spekman, Nancy J & Froma P Roth. 1985. Preschool children's comprehension and production of directive forms. *Journal of Psycholinguistic Research* 14(3). 331–349.

Spenader, Jennifer. 2003. Factive presuppositions, accommodation and information structure. *Journal of Logic, Language and Information* 12(3). 351–368.

Stalnaker, Robert. 1974. Pragmatic presuppositions. *Semantics and Philosophy, pages=197-214* .

# Cooperation and exhaustification

Giulio Dulcinati
*University College London*

Nausicaa Pouscoulous
*University College London*

**Abstract** Cooperation between interlocutors is the cardinal assumption that affords all pragmatic inferences in Grice's (1989) theoretical framework. This assumption is the background of much theoretical and experimental work in pragmatics but it has received little attention from experimental research. We used a signalling game to investigate how cooperation between interlocutors affects exhaustification inferences. The results of the experiment suggest that hearers infer less quantity implicatures from utterances of uncooperative speaker compared to utterances of a cooperative speaker. We interpret these results in support of Grice's account.

**Keywords:** Grice, Cooperation, Quantity implicatures, Exhaustification, Signalling game

## 1 Introduction

In this paper we focus on what happens to pragmatic inferences when the speaker is not fully cooperative. As Pinker et al. (2008) point out, research in pragmatics mostly studies communication in cooperative situations, which are taken as standard. However, interlocutors bring different goals to a conversation and they often have conflicting interests on some of these goals. Here we present a paradigm aimed at probing the pragmatic inferences of hearers who face an uncooperative speaker. We believe that empirical research on this topic can bring an important aspect of language use into the picture and further our understanding of the factors that affect or afford pragmatic inferences.

Grice sees conversation as an instance of the category of cooperative transactions, whose participants recognize "a common purpose or a set of purposes, or at least a mutually accepted direction" (Grice 1989: 26). His Cooperative Principle and maxims aim at capturing how a rational agent is expected to behave in such situations. Hearers derive implicatures by assuming that the speaker is striving to uphold the maxims, which is what a rational speaker should do if they are being cooperative. Therefore, the assumption that the speaker is cooperative is what ultimately affords pragmatic inferences.

Grice does concede that communication is not entirely cooperative and that it leaves room for "a high degree of reserve, hostility and chicanery" (Grice 1989: 369). Grice, and many authors after him (e.g., Attardo 1997; Sperber et al. 2010), think




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 39–45.

that although some degree of cooperation has to be in place for communication to take place at all, interlocutors may be more or less cooperative with respects to the many goals that they bring to the conversation. However, Grice does not give an account of how the hearer's expectations, and therefore their inferences, are affected by their assumptions about the cooperativity of their interlocutor.

Although Grice does not make predictions about implicatures in non-cooperative contexts, we can derive predictions from his account and we will take quantity implicatures as an example. If during a card game we ask a player what they have in their hand and they reply "I have Hearts and Spades" we can infer that they do not have Diamonds or Clubs. We can assume that a cooperative speaker fails to utter a relevant and more informative statement than the one they actually uttered (in our example "I have all suits" would be more informative than "I have Hearts and Spades") because they do not believe the alternative statement to be true. If the speaker is competent about the alternative statement (in our example, they know whether they have all suits) we can infer that the speaker is communicating that the alternative statement is false (a quantity implicature). If an uncooperative speaker fails to utter a stronger relevant statement they may have done so because they are not upholding the maxim of quantity and they are being intentionally under-informative. Because of this alternative explanation, we can derive the prediction that, other things being equal, hearers will be less likely to infer that an uncooperative speaker is communicating an implicature compared to a cooperative speaker.

We test this prediction using a signalling game where participants cooperate or compete for points (and monetary gains) with a virtual confederate. In a signalling game a player sends a message based on some private information to another player, who receives it and chooses an action. This action determines the gains for both players. Games with monetary rewards are often used in economics and psychology to study cooperation (e.g., Rand & Nowak 2013). There is also a growing literature which tries to account for pragmatic inferences by modelling context of utterance as a signalling game (Franke 2013, for an overview). Here we try to combine features from both fields and we manipulate the degree to which the players' interests overlap in a signalling game in order to study the role of cooperation for pragmatic inferences. The closest precedent to our experiment is a study conducted by Pryslopska (2013) on scalar implicatures in which she manipulated cooperativity of the speaker through a card game preceding the experimental task and she found evidence that participants inferred less implicatures from an uncooperative speaker compared to a cooperative one. Our study aims to extend her findings using a different methodology and a different kind of pragmatic inferences.

## 2 Experiment

Participants played a game with a virtual confederate who was either playing with them (cooperative) or against them (uncooperative). Participants believed that they were playing a multiplayer game called "The Grid Game" in which they were the 'guesser' and another player was the 'describer'. The actions of the other player (describer) were entirely pre-established for the purposes of the experiment. Each round of the game had a preassigned set of four grids each containing a variable number of stars. Participant believed that the describer saw only one of the four cards in the round and briefly described it. The participant (guesser) saw all the cards in the round except one that was hidden by a question mark. They listened to the description and selected the grid which they thought the other player was describing. The recorded descriptions could give rise to quantity implicatures and the choices contained a grid representing an exhaustified interpretation of the description and a grid representing a non-exhaustified interpretation. Consistently with the prediction we derived from Grice's account we predicted that participants in the uncooperative condition would be less likely to choose the 'exhaustified' grid compared to participants in the cooperative condition. In order to avoid complete distrust of the descriptions in the uncooperative condition we enforced the impossibility of lying in the game by saying that an 'inspector' had disqualified all players who gave false descriptions. A similar device is also used in game theoretic approaches pragmatics (e.g., Parikh 1992; van Rooij 2008) and it allowed us to focus on pragmatic inferences.

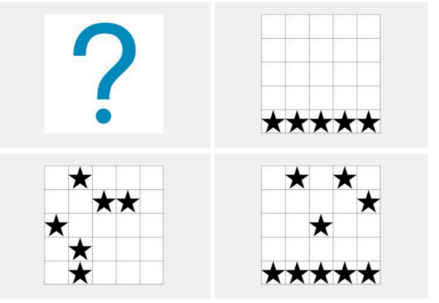


**Figure 1** *Sample choices shown to the participant in the experimental trial paired with the description "Here the squares in the bottom row have a star"*

### 2.1 Methods

**Materials** Each participant saw the same 10 experimental trials and 15 fillers. In each trial of the experiment the participant saw a page where they could play the

recorded description and see the four choices for that round in a randomized position (Fig.1). In experimental trials, the recorded description is a short positive or negative statement (e.g. Here the squares in the bottom row have/don't have a star) which can give rise to a quantity implicature concerning the unmentioned slots. In the options the participant can choose from, two of the grids fit the description: one represents a non-exhaustified interpretation of the description (bottom-right in Fig.1) and one represents an exhaustified interpretation (top-right in Fig.1).

Filler trials were similar to experimental items and they were of three types: three fillers had only one visible grid that clearly matched the description (*easy fillers*), six fillers had no visible answer that matched the description (*no-choice fillers*) and lastly six fillers had descriptions which left room for vagueness (e.g., Here two squares near the top right have a star) and two visible options that matched the description equally well (*vague fillers*).

**Conditions and scoring** Each participant was randomly assigned to either a cooperative condition or an uncooperative condition. Participants in the cooperative condition believed that the describer played the game with them as a team mate and participants in the uncooperative condition believed that the describer played against them as their opponent. Participants in both conditions gained ten points every time they selected the correct grid (i.e., the grid the describer was 'actually' describing), 0 points when they selected an incorrect grid and 3 points when they selected the question mark (unless the question mark was actually hiding the correct grid). These points were converted into money and participants could win up to £1.2 extra at the end of the experiment. Participants believed that the describer also gained a reward depending on their answers. Participants in the cooperative condition believed that the describer gained the same points (and money) as them. Conversely, participants in the uncooperative condition believed that the describer gained points from their incorrect guesses they made and no points from their correct answers. The scoring system for both conditions is summarized in Table 1.

| | *Cooperative condition* | | *Uncooperative condition* | |
|---|---|---|---|---|
| | Guesser/Perticipant | Describer | Guesser/participant | Describer |
| Correct answer | 10 points | 10 points | 10 points | 0 points |
| Incorrect answer | 0 points | 0 points | 0 points | 10 points |
| Question mark* | 3 points | 3 points | 3 points | 3 points |

*When the Question mark hid the correct grid, it counted as a correct guess.

**Table 1** *Scoring system for guesser and describer in each condition*

**Participants and procedure** Thirty-nine native English speakers (24 females, Mean Age=31.5 years) were recruited from the crowdsourcing website Prolific Academic and directed via web link to the experiment, which was hosted on the website Qualtrics. After demographic questions, participants were randomly assigned to either the cooperative (19 participants) or the uncooperative condition (20 participants) and they read the instructions for the game and the scoring system (different depending on condition). The instructions included one trial in which they saw the game from the describer's perspective and one trial in which they were given feedback on their choice and they were shown the grid that was hidden under the question mark. After the instructions participants were presented the ten experimental items in a random order (different for each participant) interspersed with the fifteen fillers (in one of three possible pseudorandom orders). Participant were rewarded £1.5 for their participation plus a reward up to £1.2 which depended on their performance in the game. The additional reward was based on their performance on fillers and not on test items.

### 2.2 Results

The frequency of participants' choices in experimental trials in each condition are summarized in Table 2. Participants choices were coded as 1 when they chose the exhaustified (pragmatic) option and 0 for any other option. Comparison of the frequency of pragmatics choices in the two conditions using Wilconxon Exact rank sum tests indicated that participants chose the pragmatic option significantly more often in the cooperative condition than in the uncooperative condition (by subjects W=118, p=.031, and by items V=45, p=.0039). In contrast, the pattern of participant's choices in each of the three types of filler trials (easy fillers, no-choice fillers and vague fillers) was not significantly different in the two conditions.

| *Condition* | *Frequency of each type of response* | | | |
|---|---|---|---|---|
| | Question mark | Exhaustified | Non-exhaustified | Wrong |
| Cooperative | 6.84% | 90% | 3.16% | 0% |
| Uncooperative | 20% | 72% | 8% | 0% |

**Table 2** *Frequency of responses in each condition*

## 3 Discussion

We found that the rate of pragmatic choices in the cooperative condition (90%) was significantly higher compared to the rate of pragmatic choices in the uncooperative condition (72%). This result supports our hypothesis that hearers are more likely

to infer implicatures from the utterances of a cooperative speaker compared to an uncooperative speaker. As we pointed out in the introduction, this prediction can be derived from Grice's account because in the uncooperative condition the speaker can be seen as being intentionally underinformative.

Our results are also consistent the behavioural findings in Pryslopska (2013). However, the eye-tracking data from her experiment suggest that participants were computing the scalar implicatures very quickly and then cancelling them. This result is consistent with Sperber et al. (2010), who propose that "whether he ends up accepting it or not, the hearer interprets the speaker as asserting a proposition that would be relevant enough to him provided that he accepted it" (Sperber et al. 2010: 368). Although we interpreted the non-pragmatic choices as instances where participants have not drawn the implicature, they might also have computed implicatures and then rejected them. Since we only have an offline measure of participants' behaviour we cannot distinguish between these two interpretations of the data. It would be interesting to adapt the paradigm used in this experiment in order to collect an online measure that could address this issue.

One interesting and unexpected aspect of our findings is the very high rate of pragmatic choices in both conditions. Although we did not have any predictions concerning the rate of pragmatic choices as this is a novel task, we found the rate surprisingly high, especially in the uncooperative condition. A first methodological concern is that showing participants both the exhaustified and non-exhaustified choices before they listened to the description may have encouraged them to think about alternatives and therefore made implicatures more accessible. This factor may be boosting the rate of implicatures in both conditions. A second concern is that since there is evidence that hearers often take what is implicated to be truth evaluable and part of what is said (Nicolle & Clark 1999; Larson et al. 2009), our participants might have thought that describers who used descriptions carrying quantity implicatures to describe 'non-exhaustified' grids would have been disqualified for lying. This might have also increased the rate of implicatures.

Putting aside our methodological concerns, we believe that the high rate of implicatures in the uncooperative condition may also be interesting from a theoretical point of view. On one hand, it is difficult for accounts which see cooperation between interlocutors as a necessary condition for implicature derivation, Grice *in primis*, to give a principled explanation to this result. They could however appeal to a number of methodological explanations. On the other hand, this high rate of implicatures is an interesting result for authors like Fox (2014) who argue that implicatures may be available even in contexts where it is not reasonable to assume that the speaker intended to communicate them. In our uncooperative condition the speaker cannot be interested in conveying information that might help their opponent give correct responses, instead they might try to deceive the hearer by implying false information.

However, it is not straightforward how accounts proposing that implicatures are available regardless of intentions can explain our main finding.

## References


Attardo, Salvatore. 1997. Locutionary and perlocutionary cooperation: The perlocutionary cooperative principle. *Journal of Pragmatics* 27(6). 753 – 779. http://dx.doi.org/http://dx.doi.org/10.1016/S0378-2166(96)00063-X. http://www.sciencedirect.com/science/article/pii/S037821669600063X.

Fox, Danny. 2014. Cancelling the maxim of quantity: Another challenge for a gricean theory of scalar implicatures. *Semantics and Pragmatics* 7(5). 1–20. http://dx.doi.org/10.3765/sp.7.5.

Franke, Michael. 2013. Game theoretic pragmatics. *Philosophy Compass* 8(3). 269–284.

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

Larson, Meredith, Ryan Doran, Yaron McNabb, Rachel Baker, Matthew Berends, Alex Djalali & Gregory Ward. 2009. Distinguishing the said from the implicated using a novel experimental paradigm. *Semantics and pragmatics: from experiment to theory* 74–93.

Nicolle, Steve & Billy Clark. 1999. Experimental pragmatics and what is said: A response to gibbs and moise. *Cognition* 69(3). 337–354.

Parikh, Prashant. 1992. A game-theoretic account of implicature. In *Proceedings of the 4th conference on theoretical aspects of reasoning about knowledge*, 85–94. Morgan Kaufmann Publishers Inc.

Pinker, Steven, Martin A Nowak & James J Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of Sciences* 105(3). 833–838.

Pryslopska, Anna. 2013. Implicatures in uncooperative contexts: Evidence from a visual world paradigm, Poster presented at XPRAG, Utrecht, Netherlands.

Rand, David G. & Martin A. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17(8). 413 – 425. http://dx.doi.org/http://dx.doi.org/10.1016/j.tics.2013.06.003. http://www.sciencedirect.com/science/article/pii/S1364661313001216.

van Rooij, Robert. 2008. Games and quantity implicatures. *Journal of Economic Methodology* 15(3). 261–274.

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi & Deirdre Wilson. 2010. Epistemic vigilance. *Mind & Language* 25(4). 359–393.

# Implicature production in children: a corpus study

Sarah F. V. Eiteljörge
*Georg-August-Universität Göttingen*

Nausicaa Pouscoulous
*University College London*

Elena Lieven
*University of Manchester*

**Abstract** Until at least age 4, children, unlike adults, interpret *some* as compatible with *all* (Noveck 2001). While there is still no consensus on why children find it difficult to draw these scalar implicatures, little is known about how their production develops. We extracted utterances containing *some* from dense corpora of five British English children aged 2;00 to 5;01 (N = 5 276) and analysed them alongside an equivalent portion of their caregivers' (N = 5 430). Utterances were coded following structural and contextual categories allowing for judgments on the probability of a scalar implicature. The findings indicate that children begin producing implicatures during their third year of life, shortly after their first *some*. Implicature production is low but matches their parents' input in frequency (and resembles that found by Degen 2015, for adults). In both children and adults, *some* appears to be multifaceted, with implicatures being infrequent, and structurally and contextually constrained.

**Keywords:** scalar implicatures, pragmatic development, production, corpora

## 1 Introduction

A lot of information conveyed in conversation is not communicated explicitly, but implicitly; it is left for the audience to infer. For instance, if a student says she "read some of the papers assigned", the listener may infer that she has not read all of them even though this was not stated. Deriving the implicit interpretation of an utterance seems challenging for young children (Noveck 2001, Papafragou & Musolino 2003). Most work on how children come to grip with implicit meaning was carried out on scalar terms such as *some*. These expressions are part of a semantic informativeness scale (e.g., <some, most, all>) and the use of a weaker term on the scale (*some* or *most*) will often be taken to imply the negation of the stronger term (*all*) giving rise to a scalar implicature.

In experimental contexts, children, unlike adults, interpret *some* as compatible with *all*, and are not found to be adult-like until seven (Guasti et al. 2005, Huang & Snedeker 2009, Noveck 2001, Papafragou & Musolino 2003). While the age at




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 46–52.

which children draw scalar implicatures has been pushed down in some paradigms, they are still not found to interpret *some* in a pragmatic way until at least 4 years of age (Katsos & Bishop 2011, Pouscoulous et al. 2007). Three main accounts of this phenomenon have been put forward. According to Katsos & Bishop (2011), young children understand the scalar implicature linked to *some*, but they are pragmatically more tolerant than adults. This leads them to accept utterances with *some* in contexts where *all* would be more appropriate even though they perceive the term as under-informative. Lexicalist accounts, on the other hand, claim that while young children know the meaning of quantifiers such as *some* and *all*, they have not yet acquired the overarching informativeness scale. This prevents them from comparing *some* to *all*, and thus, from deriving the scalar implicature (Barner, Brooks & Bale 2011). Finally, it has been argued that the processing cost of implicatures is too high for young children; while they have the ability to understand scalar implicatures, they often lack the resources to make a relatively effortful inference (Pouscoulous et al. 2007, Reinhart 2004). One of the keys to the mystery of scalar implicature development has to be production. Indeed, little is known about how the most popular scalar term, *some*, is produced by children or their caregivers; although children seem to fare better with production than comprehension (Katsos & Smith 2010). Interestingly, even in adult speech a rather small percentage of *some* instances has been found to carry scalar implicatures (Degen 2015).

In the hope to shed light on implicature competence in early childhood we conducted a corpus study looking at the perception and production of the quantifier *some* in five British English children aged two to five.

## 2 Methods

Children's utterances containing *some* were extracted with three lines of context before and after each *some* occurrence from the dense corpora of five British English speaking children aged 2;00 to 5;01 (N = 5 276). One set (Thomas) is part of the CHILDES corpus (MacWhinney 2000), while all other sets have been submitted to the CHILDES corpus and were accessed with the kind permission of the Child Study Centre, University of Manchester (Lieven, Salomo & Tomasello 2009). For each child, data were organised into age windows of three months allowing for an analysis of individual developmental trajectories. To examine inputs in the early years, we extracted the mothers' first sentences with *some* in a number equivalent to their child's production (N = 5 430). To further investigate input development, we extracted another 300 *some* utterances produced by each of the mothers after their child's birthdays (N=3 600; Total mother N = 9 030). All utterances were coded following structural and contextual categories allowing for judgments on the probability of a scalar implicature (coding scheme adapted in part from Degen 2015).

### 2.1 Structural and contextual categories

First, all the extracted *some* utterances were coded as belonging to one of the structural, mutually exclusive, categories outlined in Table 1. Utterances falling in an *Excluded* category (grey background) were not analysed further due to either errors or cases of ambiguity. Utterances falling in one of the *Included* categories (white background) were then assigned to one of four, mutually exclusive, contextual categories, which reflect their likeliness to carry an implicature based on structure and the extracted context ($\pm$ 3 utterances): *Not possible*, *Not plausible*, *Neutral*, and *Plausible*. In utterances falling in the *Not possible* category there was no established set. The speaker could therefore not refer to a subpart and intend to communicate a scalar implicature ("I need *some* help."). In *Not plausible* utterances a set could be involved, but it is unlikely that the speaker was referring to it and intended an implicature ("We need to buy *some* batteries."). In *Neutral* utterances there is a possible set and the speaker might have been referring to it and intend to communicate the implicature ("I ate *some* biscuits."). In *Plausible* utterances the context and linguistic background clearly identified a set and the speaker was making reference to it intending for the speaker to infer the scalar implicature *not all* ("The puzzle is missing *some* pieces."). As the speaker's intentions and listener's interpretation were never made explicit, these criteria were applied based on the coder's judgment of pragmatic inferences. To avoid false positives, the less implicature plausible category was chosen when in doubt.

## 3 Results

Children used *some* in its many forms with a predominance of the categories *Some mass* and *Some NP plural* (see Table 1 and Figure 1). Exclusion was highest for the *Solitary some* category. Implicature production was observable, even though *Plausible* implicatures represented a small part of the corpus contrary to *Not plausible* ones. Taken together, the findings indicate children's competence with the different uses of *some* including cases that were meant to carry an implicature and be understood as *not all*.

To establish when different types of uses and implicature production first appear, we also looked at children's individual language acquisition (see Table 2). The resulting developmental picture shows that children begin using *some* in its many forms during their third year of life. Importantly, this includes implicature production which is present shortly after the children's first usage of *some*, although low in frequency.

In the mothers' data, the categories *Some mass* and *Some NP plural* dominate. Again, exclusion was highest for the *Solitary some* category. *Plausible* implicatures

| | Category | *Some* followed by... | Example | Children % | Adults % |
|---|---|---|---|---|---|
| Included | mass | mass NP | Mummy want *some* tea. (E., 2;00) | 19.26 | 26.98 |
| | object mass | object mass NP | Get *some* fruit from there. (E., 2;11) | 0.34 | 0.41 |
| | banana | sg count NP for mass | I like *some* banana. (E., 2;00) | 5.91 | 8.42 |
| | adjective | adjectival NP | I need *some* yellow. (E., 2;00) | 0.89 | 0.29 |
| | people | pl NP for pl quantity | *Some* people love Peppa Pig. (H., 3;00) | 0.87 | 0.20 |
| | NP sg | sg count NP | *Some* little boy kissed a chair. (H., 4;01) | 2.54 | 0.98 |
| | NP pl | pl count NP | I want *some* dinosaurs. (E., 2;01) | 17.87 | 26.56 |
| | of X | partitive preposition | Mum keeps *some* of these balls. (E., 3;01) | 3.45 | 4.92 |
| | there's | sg verb + pl NP | There's *some* clothes. (E., 2;02) | 1.69 | 1.77 |
| Excluded | solitary some | no spelled-out NP | Po like *some*. (E., 2;00) | 13.93 | 7.59 |
| | more | might mean *more* | I need *some more*. (E., 2;00) | 7.92 | 5.67 |
| | more NP | *more* + NP | I get *some more* bricks (E., 2;00) | 6.27 | 6.81 |
| | scissors | pl NP for sg quantity | Need *some* scissors. (E., 2;00) | 2.29 | 2.12 |
| | several | multiple *some* NPs | I got *some* apples but I haven't got *some more* apples. (G., 3;00) | 0.68 | 2.65 |
| | conjunctives | conjunctive NPs | I've got *some* fish and chips cook. (E., 2;08) | 0.68 | 0.83 |
| | correction | *some* replaced | I've got *some* <a triangle>. (E., 2;00) | 1.59 | 0.28 |
| | incomplete | incomplete phrase | Let's play *some* +... [+ IN] (E., 2;03) | 4.09 | 1.51 |
| | coding error | transcription failure | Mummy, let's go *some* paint xxx. (E., 2;00) | 7.58 | 1.97 |
| | others | unclear utterance | I can do *some* [=? the] shopping. (E., 2;05) | 2.12 | 0.06 |

Notes. sg = singular, pl = plural, NP = noun phrase (fully compatible with DP analysis)

**Table 1** Structural categories, their description, and the results for children and parents in percentages.

represented a small proportion with most utterances being categorised as *Not plausible*. Pooling *Neutral* and *Plausible* utterances, only 14.71% of adult's and 18.63% children's uses of *some* in the corpus potentially carry an implicature (cf. Degen 2015, for similarly low rates).

Children's production and adults' child-directed speech did not differ significantly from each other in either structural or contextual categories (Mann Whitney U, $ps > 0.1$), indicating similar usage patterns of *some* for children and adults (see Figure 1).

Interestingly, parents' usage of *some* changes as a function of the child's age, with age significantly predicting an increase in *Plausible* implicatures (ANOVA: $F(1,10) = 43.283$, $p < .001$). The model accounted for 81.2% of the total variance ($R^2 = 0.812$; $Y = 11.707 * X - 14.655$). In contrast, age was a significant predictor for a decrease in *Not plausible* implicatures (ANOVA: $F(1,10) = 14.369$, $p = .004$). Here, the model accounted for 59% of the total variance ($R^2 = 0.590$; $Y = -21.310 * X + 195.983$). Thus, for each birthday of the child, the mother's number of *Plausible*

| Child | Recording | Total | Incl | Excl | 1st *some* | 1st implicature |
|---|---|---|---|---|---|---|
| Eleanor | 2;00 - 3;01 | 937 | 491 | 446 | 2;00;03 | 2;00;28 |
| Fraser | 2:00 - 3;01 | 625 | 356 | 269 | 2;00;28 | 2;03;06 |
| Thomas | 2;00 - 4;11 | 1759 | 853 | 906 | 2;00;13 | 2;09;11 |
| Gina | 3;00 - 4;07 | 962 | 493 | 469 | 3;00;01 | 3;00;04 |
| Helen | 3;00 - 5;01 | 993 | 594 | 399 | 3;00;02 | 3;00;10 |

Notes. Incl = N of utterances in Included categories, Excl = N of utterances in Excluded categories

**Table 2** Onset of *some* production and implicatures for individual children.

implicatures increased by 11.707, and the number of *Not plausible* implicatures decreased by 21.31 utterances. As many of the *Not plausible* utterances are related to food (i.e., *Some banana*), and *some* with a pragmatic interpretation is often used to contrast, it might be that the conversations evolve from focusing on more basic needs (e.g., nutrition) to complex discussions about variations in the world ("Some girls have brown hair.").

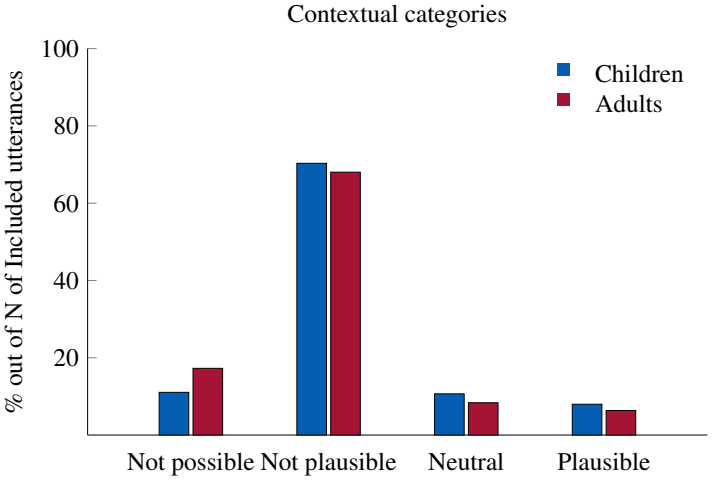



**Figure 1** Contextual categories indicating implicature plausibility. Utterances in the *Neutral* and *Plausible* categories can give rise to scalar implicatures - a potential or salient set makes reference to a subset possible, or even likely.

## 4 Discussion

The present findings indicate that children begin producing and interpreting implicatures in a pragmatic way during their third year of life, shortly after they first produce *some* (see Table 2). Thus, as soon as they acquire *some*, children produce it competently and mirror adult behaviour. Their production of *some* implicatures is low but matches their parent's input in frequency (and resembles the levels reported by Degen 2015, for adults). Interestingly, the mothers' implicature production increases as a function of the children's age, which paints a picture of growing complexity. In both children and adults *some* appears to be multifaceted, which results in implicatures being infrequent, and structurally and contextually constrained.

The low frequency of parents' implicature production corroborates the findings of Degen (2015) and seems difficult to reconcile with theories assuming that *some* commonly induces implicatures, such as syntactic accounts (e.g., Chierchia, Fox & Spector 2012) or generalized implicatures theories (Levinson 2000). Further, these results imply that children are rarely confronted with utterances meant to carry implicatures.

Children, nonetheless, produce implicatures at adult levels from the outset. These findings contrast with work showing that *some*-related implicatures are understood relatively late, and thus, call for an explanation. An account along lexicalist lines (e.g., Barner, Brooks & Bale 2011) might find it difficult to contend with such early implicature production. If toddlers have not associated *some* with its lexical scale (<many, most, all>), this should affect their ability to produce, as well as comprehend, implicatures. Therefore, a domain-general approach might be more appropriate.

Several elements may explain children's behaviour such as their pragmatic tolerance (Katsos & Bishop 2011), the relevance of the implicature in context (Papafragou & Musolino 2003, Guasti et al. 2005, Skordos & Papafragou forthcoming), and children's limited processing resources when faced with an infrequent, relatively effortful inference (Pouscoulous et al. 2007, Reinhart 2004). These factors combined with children's limited exposure to *some*-related implicatures may be sufficient to account for the discrepancy between production and comprehension. In this view, children are capable of producing and inferring *some*-related implicatures from their third year of life, and any difficulty in understanding them in experimental settings is to be attributed to factors outside their semantic and pragmatic competence.

In the past decade a lot of work has been devoted to children's comprehension of *some*. In fact, our knowledge of implicature acquisition is largely based on their understanding of this one expression. A systematic corpus analysis of how toddlers hear and produce it should therefore be an essential to any informed argument in the debate.

## References


Barner, David, Neon Brooks & Alan Clinton Bale. 2011. Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 118(1). 84–93.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In *An international handbook of natural language meaning.* Mouton de Gruyter.

Degen, Judith. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics* 8(11). 1–55.

Guasti, Maria Teresa, Gennaro Chierchia, Stephen Crain, Francesca Foppolo, Andrea Gualmini & Luisa Meroni. 2005. Why children and adults sometimes (but not always) compute implicatures. *Language and cognitive processes* 20(5). 667–696.

Huang, Yi Ting & Jesse Snedeker. 2009. Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Developmental Psychology* 45(6). 1723.

Katsos, Napoleon & Dorothy VM Bishop. 2011. Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120(1). 67–81.

Katsos, Napoleon & Nafsika Smith. 2010. Pragmatic tolerance and speaker comprehender asymmetries. In K Franich, K. M. Iserman & L. L. Keil (eds.), *Proceedings of the 34th annual boston conference on language development*, 221–232. Somerville, MA: Cascadilla Press.

Levinson, Stephen C. 2000. *Presumptive meanings: the theory of generalized conversational implicature*. MIT Press.

Lieven, Elena VM, Dorothé Salomo & Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: a usage-based analysis. *Cognitive Linguistics* 20(3). 481–507.

MacWhinney, Brian. 2000. *The CHILDES project: The database*. Psychology Press.

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

Papafragou, Anna & Julien Musolino. 2003. Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition* 86(3). 253–282.

Pouscoulous, Nausicaa, Ira A Noveck, Guy Politzer & Anne Bastide. 2007. A developmental investigation of processing costs in implicature production. *Language Acquisition* 14(4). 347–375.

Reinhart, Tanya. 2004. The processing cost of reference set computation: acquisition of stress shift and focus. *Language Acquisition* 12(2). 109–155.

Skordos, Dimitris & Anna Papafragou. Forthcoming. Children's derivation of scalar implicatures: alternatives and relevance. *Cognition*.

# Some is *not all*, sometimes

Francesca Foppolo[*], Marco Marelli[#] and Stefania Donatiello[*]
[*]University of Milano-Bicocca, [#]CiMeC, University of Trento

## 1. Introduction

In this paper, we present an eye-tracking study on the incremental derivation of the *some-but-not-all* scalar implicature associated to the scalar quantifier *some*. This question has been the matter of a vivid debate, both in linguistics and in psycholinguistics (cf. Chemla and Singh, 2014a,b; Geurts, 2010; Sauerland, 2012 for an overview). Experimentally, it was addressed in previous studies by means of eye-tracking and lead to different results: while Huang & Snedeker (2009) found evidence for a delay of *some* with respect to *all*, Grodner, Klein, Carbary & Tanenhaus (2010) argued for a rapid integration of pragmatic *some*. More recently, Breheny, Ferguson, & Katsos (2013) and Degen & Tanenhaus (2015) contributed with new data in support of the Constrain-Based approach, according to which the rate and/or the speed with which scalar implicatures are computed is tightly linked to features of the context of the utterance in which the scalar term appears. For example, Degen and Tanenhaus modulated the context by including -or not including- descriptions with number terms and demonstrated that their presence as an alternative way of addressing the scenario affected SI computation, that was quicker when the number terms were not included. Breheny, Ferguson, and Katsos provided eye-movements data in a looking while listening task in which participants heard sentences while they explored a visual scene and their eye-movements were monitored. They tested the incremental interpretation of *some* (and *all)* in sentences like "The man has poured *some* of the water with limes into the bowl on tray A and *all* of the water with oranges into the bowl on tray B" in a situation (represented in a short movie) in which the man poured (from a pitcher) some but not all the water with limes into the bowl on tray A and all of the water with oranges (from another pitcher) into the bowl on tray B. They found that convergence on the target for *some* was not delayed with respect to *all*, and occurred prior to the disambiguating point in the input, in contrast with previous findings in which a delay in the case of *some* compared to *all* was recorded.

These results are captured by a Constrain-Based approach, according to which contextual features (however these are defined) can modulate at different levels the probability and/or the availability of the enriched interpretation associated with scalar terms, affecting the speed with which this enriched meaning is derived.

Theoretically, the question about the defaultness and/or the automaticity of the derivation of SIs has been declined in three broad approaches: Levinson defines pragmatic inferences as default mode of reasoning, albeit suspendable, that are always computed as a first step in the derivation of the meaning of an utterance, independently of the context; recursive-grammatically driven approaches such as Chierchia's claim that generalized conversational implicatures (GCI) like SIs are computed "recursively and compositionally, on a par with ordinary meaning computation (and therefore are not part of a postgrammatical process)" (Chierchia, 2006:544). These approaches have been dubbed, respectively, strong and mild defaultism in Geurts (2010). In contrast, Neo-Gricean/pragmatic approaches view GCIs as genuinely post-grammatical/pragmatic processes that arise by exploiting Gricean maxims, and are added in a second step of the derivation, after compositional semantics has completed its job (a.o. Grice, 1975; Russell, 2006; Sauerland, 2004).

It is standardly assumed that focus might enhance the operation of evoking alternatives (Geurts, 2010): when an expression like *some of the Ys* is focused, the speaker draws attention to the fact that the quantity of Ys is relevant, and this immediately evokes the alternative

 In: Fabienne Salfner and Uli Sauerland (eds.), *Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 53–59.

statement *all of the Ys*, contributing in making the scalar implicature *some but not all of the Ys* available. Conversely, when the same expression comes as old information, or in topic position, the speaker draws attention to some other features of the utterance, possibly leaving the alternatives to *some* less available, or suspended, in this case. In a self-paced reading study on Greek, Breheny, Katsos and Williams (2005) showed that the SI associated to *some* was computed when the scalar trigger appeared in a focus position. In this work we contribute to this debate by employing a looking while listening task in which we tap the question of the derivation of SIs by modulating the informational status of the utterance.

## 2. Our studies

Exploiting one feature of Italian that allows for post-verbal subjects, we modulated the sentential position -and, consequently the informational status- of *some* (*alcuni*) in sentence pairs in which the quantifiers (*some* and *all*) appear in Topic or Focus position, that correspond to a pre-verbal or post-verbal position respectively in Italian (Belletti, 1989). This manipulation was done within subjects in two different experiments, that were tested on the same population in two separate sessions administered at least two weeks apart (Experiment 1 was always administered first).

### *2.1 Participants*

Twenty-nine Italian adults volunteered to participate. All participants were native speakers of Italian, had normal or corrected-to-normal vision and were naive with respect to the goals of the experiment.

### *2.2 Materials and Procedure*

Materials consisted of 72 test trials, plus 1 practice item and 3 warm-ups. Each trial consisted of a pair of Italian sentences (for a total of 144 experimental sentences), auditorily presented in sequence, and two visual scenarios, one for each sentence in the pair. The two consecutive scenarios involved the same objects and the same predicates with a different arrangement of "who is doing what". The objects represented on the screen were always sets of abstract objects like digits, shapes or letters that were rendered "animate" by the presence of an emoticon to represent a predicate (cf. Foppolo, Marelli, Meroni and Gualmini, 2015). The first sentence in each pair was always a canonical SV(O) sentence introduced by the adverbial modifier *qui* (lit. here) and served to introduce the predicate under consideration and to render the use of post-verbal subject felicitous in the second sentence of the pair. This was always a non canonical V(O)S sentence introduced by the adverbial *Adesso invece* (lit. now instead). in which a quantified subject appeared post-verbally. By modulating the quantifier (*some* vs. *all*) and the type of scenario, three different conditions were created for each experiment, that will be described separately.

### *2.3. Esperiment 1: some in Topic*

In this first experiment, Topic position is tested, resulting in three conditions in which *some* appears always in the first (SVO) sentence and *all* in the second (VOS) sentence:

| | SVO | scenario 1 | VOS | scenario 2 |
|---|---|---|---|---|
| SOME-topic (ST) | some (1a) | EARLY (Fig.1A) | all (1b) | EARLY (Fig.1B) |
| SOME-late (SL) | some (2a) | LATE (Fig.2A) | all (2b) | EARLY (Fig.2B) |
| ALL-late (AL) | some (3a) | EARLY (Fig.3A) | all (3b) | LATE (Fig.3B) |

TABLE 1. Conditions for Experiment 1, TOPIC.

An example of sentence and visual scenario for each combination is provided below:

1.
(a) Qui alcuni dei cinque stanno telefonando
*Here, some of the fives are talking-on-the-phone*
(b) Adesso invece stanno telefonando tutti i nove
*Now, instead, are talking-on-the-phone all the nines*

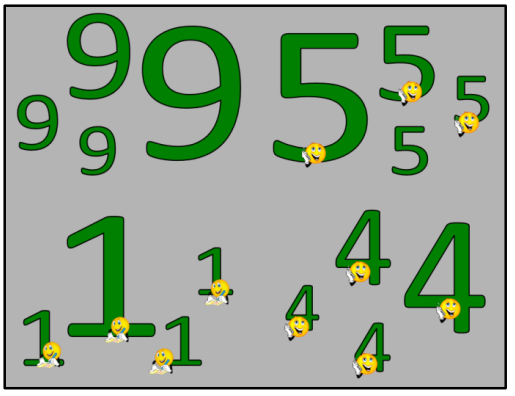


FIG. 1A

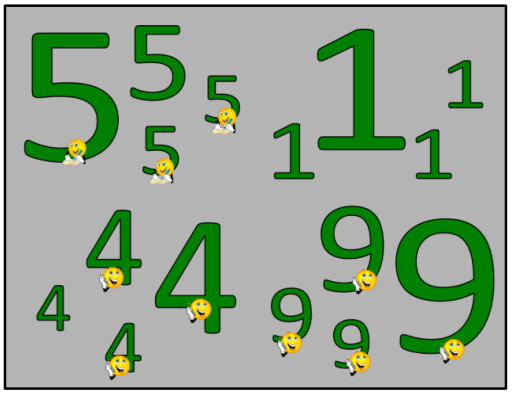


FIG. 1B

2.
(a) Qui alcune delle A stanno telefonando
*Here, some of the As are talking-on-the-phone*
(b) Adesso invece stanno telefonando tutte le U
*Now, instead, are talking-on-the-phone all the Us*

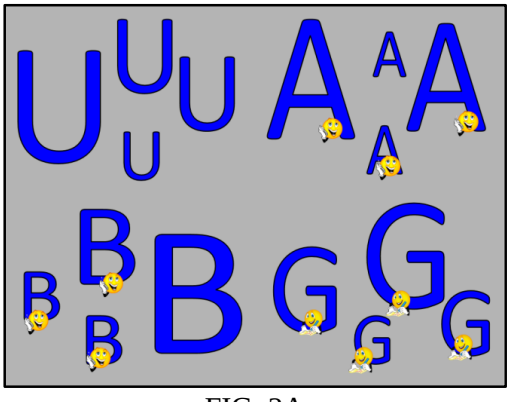


FIG. 2A

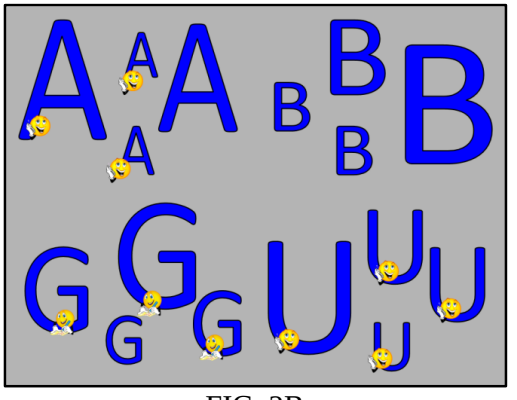


FIG. 2B

3.
(a) Qui alcuni dei pentagoni stanno telefonando
*Here, some of the pentagons are talking-on-the-phone*
(b) Adesso invece stanno telefonando tutti i cerchi
*Now, instead, are talking-on-the-phone all the circles*

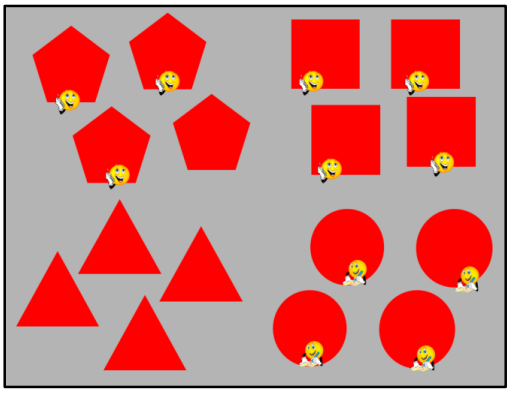


FIG. 3A

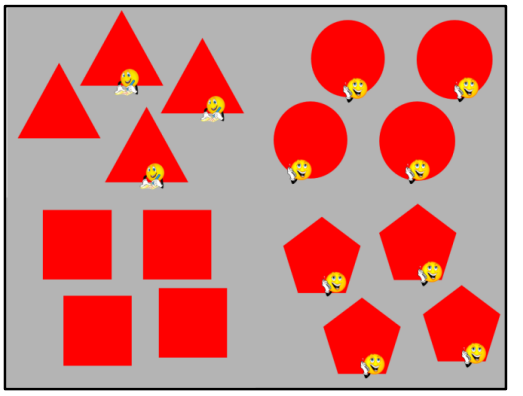


FIG. 3B

It's important to note that, when the visual scenario is labelled EARLY, the target can be spotted while hearing the quantifier. For example, in the case of sentences like (1a) "*Here, some of the fives are calling"* in front of the scenario in Fig. 1A, one can converge to the target of the fives at the quantifier region, i.e. before hearing the target noun (fives). Crucially, this happens only if the scalar implicature *some but not all* is computed incrementally at this stage. Conversely, the target can be identified only at the final noun in the LATE visual condition. For example, in the case of sentences like (2a) "*Here, some of the As are calling"* in front of the scenario in Fig. 2A, one has to wait until the final noun (As) is given, being the quantifier compatible with both As and Bs in this scenario.

### *2.4. Experiment 2: some in Focus*

Analogously to Experiment 1, three conditions were created and tested in different scenarios (cf. Table 2). Differently from the first experiment, here *some* always appears in Focus, i.e. in the V(O)S sentences like 4. Albeit employing different materials, the rationale of visual scenarios and sentences was analogous to that of Experiment 1.

| | SVO | scenario 1 | VOS | scenario 2 |
|---|---|---|---|---|
| SOME-focus (SF) | all (4a) | EARLY | some (4b) | EARLY |
| SOME-late (SL) | all | EARLY | some | LATE |
| ALL-early (AL) | all | EARLY | all | EARLY |

TABLE 2. Conditions for Experiment 2, FOCUS.

4.
(a) Qui tutte le U stanno telefonando
*Here, all the Us are talking-on-the-phone*
(b) Adesso invece stanno telefonando alcune delle A
*Now, instead, are talking-on-the-phone some of the As*

### *2.5. Results*

We recorded participants' fixations to the target in different temporal regions and found that sentential position (and, thus, the informational status) of the quantifier affected the rate/speed of the computation of the implicature (Fig. 3).

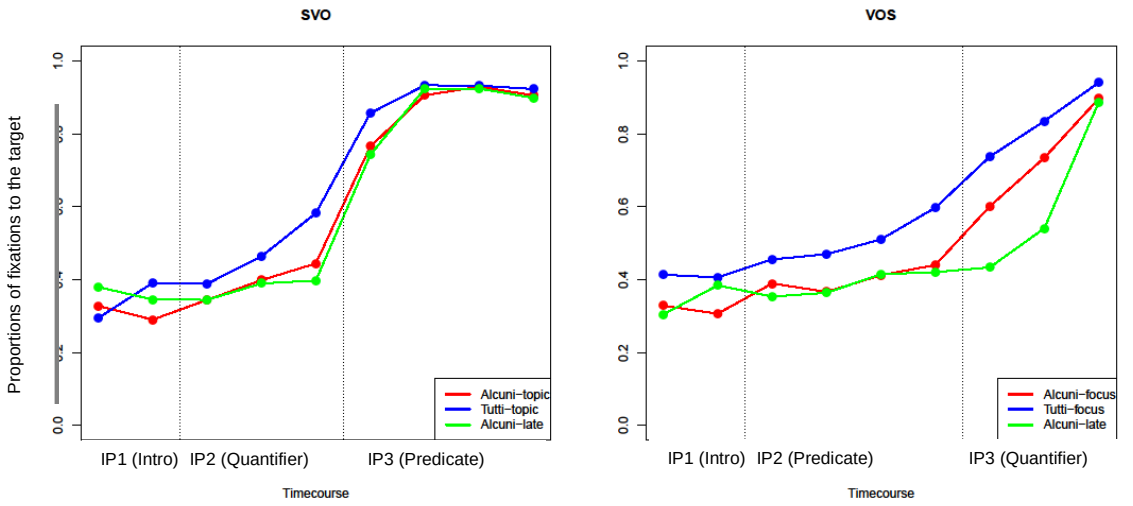



**Fig. 3** Proportions of Fixations to the Target in the 3 types of scenarios for the different quantifiers (Some-Early=RED, Some-Late=GREEN, All=BLUE) in the two experiments (Exp. 1-TOPIC on the left; Exp. 2-FOCUS on the right)

In particular, statistical analyses by means of mixed effect models (Baayen et al., 2008) showed that, in Experiment 2, when *some* appeared in Focus position, participants quickly converged on the pragmatic target (cf. red line in Fig. 3-right panel) in the EARLY-condition, and a significant difference is found in this case between EARLY (red line) and LATE (green line) conditions in the quantifier phase. Conversely, no difference is found (throughout the whole sentence) between EARLY (red line) and LATE (gree line) conditions when *some* appeared pre-verbally, i.e. in TOPIC position (Fig. 3-left panel). A summary of the statistical comparisons between Early and Late conditions related to *some* in the two experiments is provided in Table 1:

| | **IP1 (intro)** | | **IP2 (Quant/Pred)** | | **IP3 (Pred/Quant)** | |
|---|---|---|---|---|---|---|
| | Estimate | p | Estimate | p | Estimate | p |
| **TOPIC** Early vs. Late | .15 | .9194 | **-.09** | **.3417** | .06 | .629 |
| **FOCUS** Early vs. Late | .07 | .322 | -.004 | .5062 | -.56 | **<.001** |

**Table 1.** Results of the analyses on *some* in the two scenarios (Early vs. Late) carried out (separately) in the two conditions (Topic vs. Focus) tested within subjects

## 6. Conclusion

In this experiment, we show that the discourse status of weak scalar quantifiers affects their interpretation. Extending previous findings (Degen & Tanenhaus, 2015), we show that the rate and/or the speed of the scalar inference associated to *some* is also modulated by the

informational status of the quantifier: when this appears in Focus position, the pragmatic interpretation is more easily available than when it is in Topic position. This is due, presumably, to the enhanced availability of alternatives that are evoked by focused elements and/or, more generally, to the scopes of the communication which focused elements contribute to by providing new, thus crucial, information: this new information is likely to be the most informative as possible, and, crucially, more informative than given information.

We interpret our results within a Grammatical approach to SI that explicitly formalizes the idea that SIs are derived whenever the scalar alternatives are active (Chierchia, 2006; Chierchia, Fox & Spector, 2009). In Chierchia's account, each scalar item is claimed to be associated with a σ feature (mnemonic for *strong*) that can be associated with a positive or a negative sign, e.g. $some_{[\pm\,\sigma]}$: when the positive sign is associated to the feature ($some_{[+\,\sigma]}$), then the scalar alternatives to *some* are active and must lead to enrichment; when, on the contrary, the negative sign is associated to the feature ($some_{[\pm\,\sigma]}$), then the scalar alternatives are not active and thus cannot lead to pragmatic enrichment. Crucially, "scalar inferences might get suspended, i.e. not generated it certain contexts, by choosing the feature setting of the scalar item that fits the context best." (Chierchia, 2006: 547). Our results show that *some* is *not all*, but only sometimes*:* the scalar inference *some but not all* is derived incrementally only when scalar alternatives are active. We show that this happens, for example, when the scalar term appears in focus position. When it appears in topic position, and thus in a lower-bound context, the implicature is not (or not immediately) derived instead.

**References**

Belletti, Adriana (2001). Inversion as focalization. *Subject inversion in Romance and the theory of Universal Grammar*, 60-90.

Breheny, Richard, Heather J. Ferguson, & Napoleon Katsos (2013). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28* (4), 443-467.

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.

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

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

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

Chierchia, Gennaro, Danny Fox, & Benjamin Spector (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Maienborn et al. (2012), pp. 2297–2332.

Degen, Judith, & Michael K. Tanenhaus (2015). Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive science*, *39*(4), 667-710.

Foppolo, Francesca, Marco Marelli, Luisa Meroni, & Andrea Gualmini (2014). Hey Little Sister, Who's the Only One? Modulating Informativeness in the Resolution of Privative Ambiguity. *Cognitive Science*, 1-29.

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

Grice, Herbert P. (1991). *Studies in the Way of Words*. Harvard University Press.

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

Huang, Yi Ting, & Jesse Snedeker (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology*, *58* (3), 376-415.

Russell, Benjamin (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics*, *23*(4). 361–382.

Sauerland, Uli (2004). Scalar implicatures in complex sentences. *Linguistics and philosophy*, *27*(3), 367-391.

Sauerland, Uli (2012). The computation of scalar implicatures: Pragmatic, lexical or grammatical?. *Language and Linguistics Compass*, *6*(1), 36-49.

# Task types, link functions & probabilistic modeling in experimental pragmatics*

Michael Franke
*Seminar für Sprachwissenschaft*
*Eberhard Karls Universität Tübingen*

**Abstract** Recent years have seen increased interest in experimental approaches in pragmatics, but pragmatics has not been an experimental discipline from the start. As a result, a common problem is one of mapping between theory and experimental data: how do established theoretical notions carry over to precise predictions about to-be-expected data?; conversely, what exactly do particular experimental tasks measure, expressed in notions meaningful to pragmatic theory? I argue here that explicit probabilistic modeling can provide a key for tackling these fundamental issues.

**Keywords:** probabilistic pragmatics, link functions, regression models, truth-value judgement task, rating scale task

## 1 Towards theory-based statistical modeling

Experimental pragmatics is a relatively young scientific enterprise, but it builds on long traditions in especially theoretical linguistics, psycholinguistics and experimental psychology. Its developmental lineage is both virtue and vice: on the one hand, experimental pragmatics can tap into rich theoretical and methodological knowledge bases, but, on the other hand, it may unduly hamper itself by a suboptimal combination of elements from its theoretical and experimental ancestors. I argue here that experimental pragmatics can benefit from endorsing the richness of formal cognitive modeling, thereby going beyond mere hypothesis testing and out-of-the-box regression analyses, techniques which I will call "theory-free." The alternative is to spell out, in the same data-generating model, both: (i) a theoretical component (inspired by pragmatic theory), and (ii) a link function (inspired by

*Dorothea Knopp helped with designing the experiments, and processing the data. Fabian Dablander assisted in implementing the experiments. Their help is much appreciated. I am also grateful to Judith Degen and Bob van Tiel for sharing data and thoughts. This work was supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63) and the Priority Program XPrag.de (DFG Schwerpunktprogramm 1727).




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 60–67.

| | variant | | | |
|---|---|---|---|---|
| | A | B | C | D |
| task type | ordinal | ordinal | binary | binary |
| fillers *many* & *most* | present | absent | present | absent |
| no. participants in analysis | 119 | 114 | 109 | 107 |

**Table 1** Experimental variants

standard statistical modeling) that describes how the theoretical predictions map onto observable choice probabilities in a given task.

For concreteness of example, Section 2 introduces what superficially looks like inconclusive evidence obtained from two different task types: while a truth-value judgement (TVJ) task indicates no contextual interference effects, a rating-scale judgement (RSJ) task does. Discrepancies between results from different tasks are relatively common in the recent literature, often leading to deadlocks and inability to reach a consensus about important theoretical issues (e.g., the methodological debate about whether there are "embedded implicatures" (c.f. Geurts & Pouscoulous 2009, Chemla & Spector 2011, Geurts & van Tiel 2013)).

The probabilistic model of Section 3 is able to resolve this apparent tension. Putative differences between task types are accommodated by suitable link functions, so that it is possible to maintain a uniform and explicit picture of what exactly is measured in either task and how experimental manipulations relate to theoretical notions of interest.

## 2 Case study: typicality of quantifiers

van Tiel (2014) and Degen & Tanenhaus (2015) independently looked at typicality ratings for scalar *some* in sentences like "Some of the circles are black" in combination with different pictures of varying numbers of black and white circles. These studies used a rating scale task and showed that the typicality of *some* is a gradient function of the number of black circles (see left of Fig. 1). The data for this paper comes from a partial replication of these studies with two additional manipulations: (i) the task type and (ii) the presence/absence of quantifiers *many* and *most* as additional fillers in the experiment.

Participants were recruited via Mechanical Turk and assigned to one out of four variants in a between-subjects design (see Table 1). On each trial of any variant, subjects were presented with a randomly generated picture of 10 circles, some of which white, the others black. In variants *A* and *B*, subjects rated whether a sentence was a good description of a picture on a 7-point rating scale. In variants *C* and *D*,

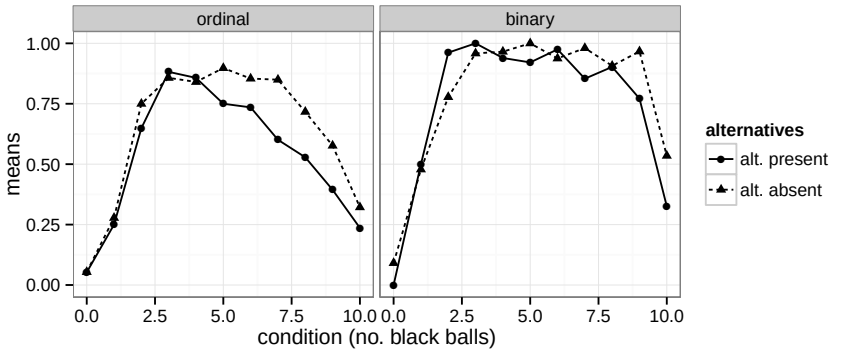



**Figure 1** Means of ordinal 7-point rating scale data (left), with the $i^{\text{th}}$ degree coded as $\frac{i-1}{6}$, and means of binary truth value judgements with *true* judgement coded as 1.

subjects judged whether a sentence was true of the given picture or not. Participants rated 13 sentences in variants *A* and *C*, which contained fillers with *many* and *most*, and 8 in variants *B* and *D*, which did not contain these fillers. Each sentence was either a random control sentence, a critical sentence with *some*, or a filler sentence with *many* or *most* (for variants *A* and *C* only). Sentences were presented in pseudo-random order with some constraints, the most important of which was that in variants *A* and *C*, exactly one filler with *many* and one with *most* preceded the first encounter of a critical *some* sentence. Mean responses from between 107 and 119 participants per variant (see Table 1) are shown in Fig 1.

Does presence or absence of fillers with *many* or *most* have an effect on the data under either task type? Visual inspection suggests that presence of alternatives *many* and *most* seems to be reflected in ordinal RSJs, but perhaps not in binary TVJs. Statistical analyses in support of this conclusion exists. Analyses that suggest otherwise do too. The issue to debate is not which statistical analysis is least careless or most adequate and certainly not which one is correct.

There are more general questions. What do TVJs and RSJs measure? Same thing or different? Is whatever is measured influenced by the presence or absence of alternatives? If so, how? Is what is measured influenced by experimental manipulations in the same way in either task? What does it even mean to measure something with a task, and, most importantly, how is whatever we measure related to a rich body of pragmatic theory? It is difficult, if not impossible, to address this by testing null-hypotheses or calculating regression models. But that does not mean that statistical modeling is at its end, of course. The key is to inject pragmatic theory.

## 3 A probabilistic model

The (simplified) structure of (Bayesian) regression modeling is this:

(1) predictor value + link function = likelihood of data.

For each data point $d$ we compute a *predictor value* $x_d$ as a (e.g., linear) function of a vector of coefficients $\vec{b}$, that represents the influence of all relevant explanatory factors, and the values for these factors associated with $d$. The predictor value $x_d$ is then mapped onto a likelihood $P(d \mid \vec{b}, \dots)$ by a *link function* suitable for the task.

Standard regression models are, in a manner of speaking, *theory-free*. The predictor value $x_d$ is retrieved from some general-purpose, mathematically convenient function of coefficients $\vec{\beta}$. In contrast, a *theoretically informed data-generating model* would mold domain-specific assumptions into a specific, tailor-made map from factors to predictor value $x_d$. Unlike in theory-free models, the latent predictor variable $x_d$ can be conceptually meaningful in theory-driven models. From the point of view of a theoretically informed model, $x_d$ is what a task measures.

Link functions can remain the same for theory-free and theory-inspired models. To ask whether two tasks could plausible measure the same thing, is to find a plausible theoretical model for $x_d$, plug it into different link functions, and see whether data from both tasks can be handled in the combined model. Let's do that.

**Pragmatic model of the predictor value.** We need a measure for the pragmatic felicity of statement "Some of the circles are black" in each of the 11 conditions (0, 1, ... 10 of the 10 circles being black). Pragmatic felicity is influenced by, among other things, the purpose of the conversation and, being a relative notion, the pragmatic felicity of other possible expressions. Inspired by game theoretic and probabilistic pragmatics, the assumption here is that pragmatic felicity is the *(scaled) expected utility* of the *some* statement, relative to that of alternative statements, to a speaker who describes a given picture for a literally interpreting listener (because such a speaker demonstrably implements Gricean language use).

Fix conditions $c \in \{0, 10\}$ for the number of black balls and messages $m \in M = \{none, one, two, three, some, many, most, all\}$, where *some* is taken to mean "at least one," *most* to mean "more than half" and *many* to mean "at least 4" (fixed *ex post* by subjects' actual judgements of *many*-sentences). Degrees of salience of alternatives to *some* are estimated from the observed data (see Franke 2014).

A literal listener interprets message $m$ as a random state in which $m$ is true. If $c$ is the actual state and the literal listener guesses $c'$, then the speaker's utility is a parameterized function of the distance between $c$ and $c'$ (Nosofsky 1986):

$$\mathrm{U}(c, c' \,;\, \pi) = \exp(-\pi\,(c - c')^2)\,. \tag{1}$$

Here, $\pi$ is a free parameter for pragmatic precision: if $\pi \to \infty$ only guessing the true state has positive utility; as $\pi$ decreases near guesses have more and more utility; for $\pi = 0$ all interpretations are equally good.

A speaker's expected utility of using message $m$ in condition $c$ is:

$$\text{EU}(m, c\ ;\ \pi) = \sum_{c'} P_{LL}(c' \mid m)\ \text{U}(c, c'\ ;\ \pi), \tag{2}$$

where $P_{LL}(c \mid m)$ is the probability that the literal listener chooses interpretation $c$ for $m$. To reflect competition between alternative expressions, as Gricean pragmatics would have us do, consider the *scaled expected utility* for *some* in each condition, relative to any set $X \subseteq M$ of entertained alternatives (this always contains *some*):

$$\text{EU}^*(c, X\ ;\ \pi) = \frac{\text{EU}(\textit{some}, c) - \min_{m \in X} \text{EU}(m, c)}{\max_{m \in X} \text{EU}(m, c) - \min_{m \in X} \text{EU}(m, c)}\,. \tag{3}$$

Speakers may not entertain a fixed set of alternatives $X$. If the probability of entertaining alternatives is given by a probability vector $\vec{s}$ of length 7, and if we assume (crudely) that probabilities of entertaining alternatives are all independent, the probability that set $X$ is entertained is $P(X \mid \vec{s}) = \prod_{m \in X} s_m \prod_{m \in M \setminus X} (1 - s_m)$. The central tendency of relative pragmatic felicity of *some* in condition $c$ is then:

$$\text{F}(c\ ;\ \vec{s}, \pi) = \sum_{X} P(X \mid \vec{s})\ \text{EU}^*(c, X\ ;\ \pi)\,. \tag{4}$$

This is a theory-driven predictor for TVJs and RSJs alike.

**Link functions.** Let's consider standard link functions from regression modeling for our task types. For binary response variables, like from TVJs, the link function is usually given by a **logistic function** whose output is fed into a binomial distribution. If the data is a number $k$ of *true* responses out of $n$ observations, then the likelihood is Binomial$(k, n, p)$ with probability $p = (1 + \exp(-\gamma(x - \theta)))^{-1}$ given by a logistic function of predictor value $x$ with threshold $\theta$ and gain $\gamma$.

For ordinal response variables, like from RSJs, $x_d$ is fed into a **thresholded probit model** and the outcome is piped into a multinomial distribution. Let $\vec{k}$ be a vector of counts with $k_d$ the number of choices of degree $d \in 1, \ldots 7$ on the 7-point rating scale, and $n$ the number of observations. Then the likelihood is Multinomial$(\vec{k}, n, \vec{p})$ where $\vec{p}$ is a probability vector of length 7, calculated as follows. Each degree $d$ is associated with an interval $I_d$, the boundaries of (some of) which are free model parameters. Intervalls for all degrees form a partition of the reals. We assume that, on each choice occasion, the predictor value $x$ is perturbed by Gaussian noise with standard deviation $\sigma$. The degree corresponding to the interval in which

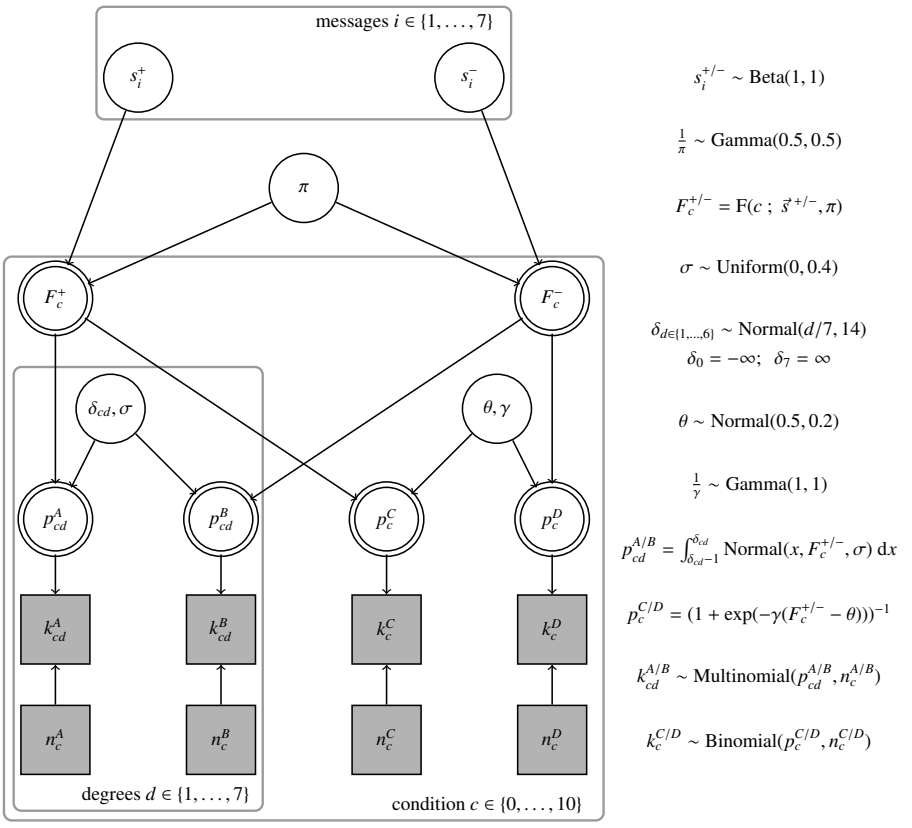



**Figure 2** Probabilistic graphical model (see Lee & Wagenmakers 2015).

the perturbed value resides is chosen. Hence, the probability $p_d$ of observing a choice of degree $d$ on the rating scale is the probability that the Gaussian perturbation of $x$ lies in $I_d$. A formalization of this idea is contained in Figure 2.

**Data-driven inference.** Figure 2 gives the full probabilistic model, using the conventions for probabilistic graphical models of Lee & Wagenmakers (2015). Arrows indicate dependencies of variables. The observed data, in shaded boxes, informs the values of the latent parameters. Latent parameters without dependencies are constrained by suitable prior distributions, as given on the right in Figure 2. The most important detail is that two vectors of salience of alternative messages are used, $\vec{s}^{\,+}$ for the case where alternatives are present, and $\vec{s}^{\,-}$ otherwise.

Estimates of the joint posterior over latent parameters, conditional on the data, were obtained by MCMC sampling using JAGS (Plummer 2003). After a burn-in of

10,000 samples, every second of another 10,000 samples entered into the analysis. Convergence was assessed by visual inspection and $\hat{R}$ values (Gelman & Rubin 1992). Posterior predictive checks confirm that the model, when using posteriors of model parameters, generates virtual data that is indistinguishable from the actual. In this weak sense, the model seems to "work" alright: it is possible to think that the same underlying value generated both TVJs and RSJs at the same time.

The most interesting inference is that of posteriors of salience of messages. Figure 3 shows estimates from the MCMC samples. Model and data suggest that *most* is made more salient by its presence, but not *many*. Empirical and theoretical consequences of this prediction remain to be explored.

## 4 Conclusions

Pragmatic notions of theoretical interest can be crafted into quantitative models of latent predictor values. Combined with standard link functions from regression modeling, we obtain theory-driven, data-generating models, with the help of which we can start to make sense of otherwise eluding pieces of data.

The particular model given in this paper is merely an example. It raises many further issues. These issues, however, are meaningful to experimental pragmatics and could not be perceived clearly and discussed stringently without any concrete model on the table. It is an empirical question whether this model, or any other is the right way to think about truth-value or rating scale judgements. To decide between competing models, statistical model comparison based on suitable data is necessary. Doing so will give a better understanding of what different tasks are measuring, how a measure is influenced by factors of relevance and what we may or may not conclude from experimental data. In sum, experimental pragmatics will benefit from explicit, theory-driven probabilistic modeling of the whole data-generating process.

## References

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

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicatures: A constraint-based approach. *Cognitive Science* 39. 667–710. http://dx.doi.org/10.1111/cogs.12171.

Franke, Michael. 2014. Typical use of quantifiers: A probabilistic speaker model. In Paul Bello, Marcello Guarini, Marjorie McShane & Brian Scassellati (eds.), *Proceedings of cogsci*, 487–492. Austin, TX: Cognitive Science Society.

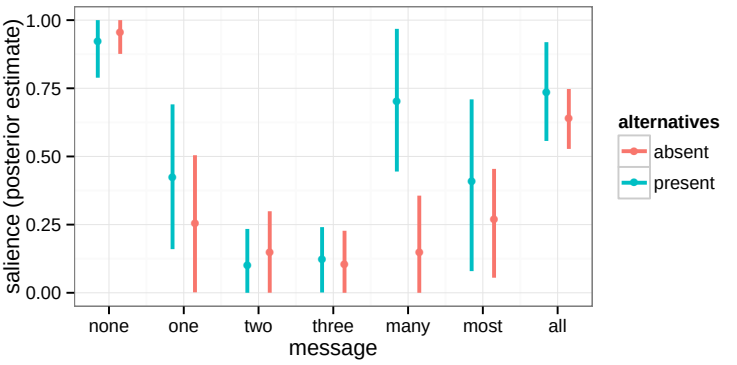



**Figure 3** Estimates of posteriors of $\vec{s}^{+}$ and $\vec{s}^{-}$. Bars are 95 % highest density intervals (i.e., an interval of values with non-negligible posterior credence levels), dots are posterior means.

Gelman, Andrew & Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 7. 457–472.

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

Geurts, Bart & Bob van Tiel. 2013. Embedded scalars. *Semantics & Pragmatics* 6(9). 1–37. http://dx.doi.org/10.3765/sp.6.9.

Lee, Michael D. & Eric-Jan Wagenmakers. 2015. *Bayesian cognitive modeling: A practical course*. Cambridge, MA: Cambridge University Press.

Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115(1). 39–57.

Plummer, Martyn. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Kurt Hornik, Friedrich Leisch & Achim Zeileis (eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*, .

van Tiel, Bob. 2014. *Quantity matters: Implicatures, typicality, and truth*: Radboud Universiteit Nijmegen dissertation.

Michael Franke
Wilhelmstraße 19
72076 Tübingen
mchfranke@gmail.com

# A wish list for experimental pragmatics

Bart Geurts
*Radboud University Nijmegen*

Several decades' worth of heated discussion have failed to establish a useful distinction between semantics and pragmatics. I think I know why: it is because there wasn't anything useful to be established in the first place. This puts me in a terminological bind. I need a word to cover everything that 'semantics' and 'pragmatics' have been taken to cover between them; and 'pragmantics' will not do. Since whatever we decide to call it, our main explanandum is communication, I will adopt 'pragmatics' as my term of choice. Experimental pragmatics is part of pragmatics thus understood.

What is experimental pragmatics? First, experimental pragmatics is not a recent invention: it was practiced well before the term was coined. Secondly, just as pragmatics is not a subfield of, e.g., linguistics or psychology, but rather an interdisciplinary field by its very nature, experimental pragmatics is not a subfield of any academic discipline, either. Thirdly, in contrast to pragmatics, experimental pragmatics is not really a field or subfield to begin with. Or at any rate, it would be a bad idea to view it thus.

What is it then? As I remember it, what happened is that, shortly after the opening of the new millennium, people working on pragmatic theories began to suspect that quantitative data might be a good thing to have, while at the same time people working in psychology were looking to pragmatic theories for inspiration, if not guidance. We were after exchange and mutual support, and got it.

It may be that, thus far, the interplay between theory and experiment hasn't gone very deep. I'm not aware of any non-trivial theoretical predictions being dramatically vindicated or refuted by experimental data, nor of experimental results revealing new vistas for theoretical analysis. For example, although there has been a lot of valuable experimental and theoretical work on scalar implicatures, the interaction between the two has been rather superficial. But no matter: the good thing is that there has been interaction in the first place.

In that spirit, I suggest that we view 'experimental pragmatics' simply as rallying cry for those of us who value multi-method research in pragmatics. This line of research has become increasingly popular over the past decade or so, and therefore it is of some interest to ask where it should be going in the coming decade or so. By way of a partial answer to that question, I present the following wish list.


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 68–70.

**[1]** In a recent paper, Michael Franke observes that 'a common problem [in experimental pragmatics] is one of mapping between theory and experimental data: how do established theoretical notions carry over to precise predictions about to-be-expected data?; conversely, what exactly do particular experimental tasks measure, expressed in notions meaningful to pragmatic theory?'

In the philosophy of science, it is a truism that the relationship between theories and data is fraught with difficulties. Elsewhere in academia, this truism is widely ignored, and thus far experimental pragmatics hasn't been an exception. Just to mention one example, experimental studies generally bypass the inconvenient truth that pragmatic theories are not about processing, so whenever claims are made to the effect that, e.g., Gricean pragmatics predicts that experimental subjects will behave in such-and-such a way, then these claims hinge on tacit assumptions that are not part of the theory proper.

So item number one on my wish list for experimental pragmatics is that we follow Michael's lead and start facing up to such issues.

**[2]** My second wish is a continuation from the first. Somewhat paradoxically, perhaps, I feel that research in experimental pragmatics has been focusing a bit too heavily on experimental studies. When I try to think what have been the main advances under the 'experimental pragmatics' banner, it is mostly experimental results that come to mind. So in a sense, I would like us to become more serious about theory. Moreover, there is a lot of research on pragmatics, or that could fruitfully interact with pragmatics, which thus far has played a minor role in experimental role in experimental pragmatics, if it has played a role at all. For example, at our conferences and workshops we haven't heard that much about ongoing developments in computational modeling, intercultural pragmatics, or the philosophy of language. As I said before, I think that the principal concern of experimental pragmatics should be to foster interdisciplinary research in pragmatics, but even if the main emphasis is on experiments, I see no reason why these disciplines should be left out.

**[3]** I expect we all agree that communication is the main explanandum of pragmatics, and that the same should hold for experimental pragmatics. But as a matter of fact neither seems to be the case. By and large, pragmatic theories focus on the hearer, and studies in experimental pragmatics follow suit. It is clear why this is so: it's a lot easier to confine one's attention to the hearer than to deal with the messy back-and-forth between interlocutors. But even I fully appreciate that dialogues are hard to control, I can't help feeling that there are serious limits to what can be accomplished with truth-value-judgment tasks and questionnaire studies. Hence, I would like to see more work on dialogue.

**[4]** About five years ago, I was invited to contribute to a volume entitled *Perspectives on framing*. I submitted a paper in which I presented a pragmatic account of some of the so-called 'framing effects' studied by social psychologists and economists. Unfortunately, the editor's feelings towards my contribution were less than lukewarm, partly because he failed to understand it and partly because it was rather critical of the views of some of the luminaries in the field, especially Tversky and Kahneman. I decided to spare myself a long and dreary debate, withdrew the paper, and submitted it to *Mind and language*, where it was duly published.

Now, although *Mind and language* is a great journal, I dare say that its readership doesn't include many social psychologists and economists, who were my target audience. So why did I submit the paper to *Mind and language*? Simple: I couldn't think of a more topical journal that might be willing even to consider the paper, let alone publish it.

This heart-rending story illustrates a general point: there is a dearth of reputable journals that (a) have the expertise to assess interdisciplinary work and (b) are prepared to publish it. This holds for experimental pragmatics, too, and the problem is particularly acute for those of us who want to be taken seriously by psychologists (e.g., because they *are* psychologists). Suppose, contrary to fact, that I had a wonderful theory on presupposition, which I could present beautifully in ten pages (though not less), and which produced the most impressive experimental results (also ten pages). How many psychology journals would consider a paper like this? Answer: very few. (*Cognition* used to be an exception, but it too has converted to the experimental monoculture that is characteristic of most of its brethren.)

Ergo: since we have conferences, workshops, and networks, why can't we have a journal as well?

**[5]** Finally, I would like to suggest that we impose a temporary moratorium on theoretical and experimental studies devoted exclusively to 'some' and/or 'or'. There must be other scalars somewhere out there.

# Do children adjust their event descriptions to the needs of their addressees?

Myrto Grigoroglou & Anna Papafragou
University of Delaware

## Abstract

Do children take into account their addressees' needs in spontaneous production? Developmental evidence for speaker adjustments is mixed. Some studies show that children are often under-informative when communicating with ignorant addressees. But other studies demonstrate successes in children's ability to integrate another person's perspective. In four experiments, we asked whether children adapt their event descriptions depending on (a) the typicality of event components, and (b) the listener's visual access. We found that children's ability to use information about the listener's visual perspective to make specific adjustments to event descriptions emerged only in highly interactive contexts, in which participants collaborated towards mutually achieved goals.

## Introduction

Do children tailor their utterances to their addressees' needs? Previous research has shown both limitations and successes in children's ability to adjust their level of informativeness to their listener's knowledge.

Experimental evidence from children's spontaneous production indicates that children often produce under-informative utterances when communicating with ignorant addressees. For instance, 3-year-olds fail to adjust their utterances appropriately when addressing ignorant vs. knowledgeable interlocutors (Perner & Leekam 1986) and 5-year-olds are often underinformative when describing one of two objects in a contrast set to an ignorant addressee (Davies & Katsos 2010). In a training study, 2-, 3- and 4-year old children originally (before training) produced ambiguous utterances when addressing an ignorant partner (Matthews, Lieven &Tomasello 2007). Even school-age children (8-year-olds) were shown to produce many ambiguous utterances when asked to describe abstract figures to an imaginary addressee (Girbau 2001).

However, other evidence suggests that children are able to make adjustments based on the knowledge state of their interlocutor. Nadig and Sedivy (2002) showed that 5- to 6-year-old children used more adjectival modifiers to refer to one of two competing objects when their addressee could see both of these objects, but not when the addressee could only see one of the objects (see Bahtiyar & Küntay 2009, for similar results with Turkish children). O'Neill (1996) found that 2-year-old children were more likely to name a hidden toy or the location of the hidden toy when asking their ignorant parents for help to retrieve it. Matthews et al. (2006) demonstrated that 3- and 4-year-olds (but not 2-year-olds) can switch from pronouns to full NPs for referents their listener cannot see, but, if provided with appropriate feedback, even 2-year-olds become more informative (Matthews, Lieven & Tomasello 2007; Matthews et al. 2012).

This experimental evidence suggests that, under some circumstances, children are able to take into account the informational needs of their addressees, but more research is required to clarify exactly which factors contribute to children's successful adaptations. Most of the research so far has focused on children's nominal reference. It is unclear whether children will be able to make addressee-specific adjustments in descriptions of events, which require more complex syntactic constructions.




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 71–75.

In this study, we asked whether adults and preschoolers adapt their event descriptions depending on (a) the typicality of event components, and (b) the listener's visual access. Both factors have been argued to play a role in adults' early syntactic choices in production (Brown & Dell 1987; Lockridge & Brennan 2002).

**Experiment 1**

In Experiment 1, twenty-four 4- to 5-year-old children and twenty-four adults watched short video clips depicting different events. In half of the events, an agent performed an action using a typical instrument (e.g., watering plants with a watering can) and in the other half the agent performed the same action using an atypical instrument (e.g., watering plants with a hat). Participants were asked to describe the events to a listener (the experimenter's confederate) who either had or did not have visual access to the events.

We found that both adults and children were more likely to mention atypical than typical instruments. We also found that adults were more likely to mention instruments when the events were not visible to their interlocutor. However, in children, visual access did not affect instrument mention. Thus, adults made both typicality-based adjustments and more specific adjustments to the informational needs of their addressee. Children, however, performed only typicality-based adjustments by mentioning only the most unusual event component.

There are several reasons why children may have ignored the needs of their addressee. One possibility is that children had difficulty estimating the goals of the task. Asking children to simply describe events for a passive listener may not have provided the necessary communicative goal that would highlight the listener's specific needs. To explore this possibility, we conducted a second experiment that clarified the goals of the task.

**Experiment 2**

Experiment 2 was a modified version of Experiment 1. Twenty-four 4- to 5-year-old children and twenty-four adult participated. The only modification was that participants were asked to describe the events so that the listener would draw the events on a sketchpad based on how participants described them. We reasoned that, given the addressee's clear goal (making accurate drawings), children would feel the communicative pressure to produce more complete event descriptions compared to Experiment 1. Results replicated the findings of the previous experiment. We, thus, concluded that young children fail to mention atypical event instruments to an addressee who lacks access to the events, even if knowing about instruments has clear communicative advantages for the addressee.

**Experiment 3**

In Experiment 3, we explored a new paradigm with the goal of making the addressee's needs more prominent for children: we asked whether instrument information can be identified and used by children to unambiguously single out and describe an event within a pair of closely matched alternatives to a listener with or without visual access to the events. Within each pair, the same event was depicted with a typical vs. an atypical instrument (e.g., a woman sweeping the floor with a broom vs. a tree branch). In such contrastive contexts, adults might be expected to produce instruments regardless of typicality or visual access, but children might be more likely to distinguish between two almost identical events for the sake of an uninformed addressee.

Ninety-six children and thirty adults participated. To seek developmental changes in the ability to make use of perspective information, the child participants fell into two age groups: a group of 4-year-olds and a group of 5-year-olds.

Results showed that both 4- and 5-year-old children mentioned atypical instruments more frequently than typical instruments, but adults used both equally frequently. There was a clear developmental effect, with older children using more instrument information than younger children, but less information than adults. Visual access to the events did not affect instrument mention in any age group. To test whether this result was due to genuine limitations in children's ability to adjust to addressees' needs or to the non-interactive presence of the confederate-listener, we introduced a more interactive modification to the experimental paradigm.

**Experiment 4**

Experiment 4 was a modified version of Experiment 3 with two main modifications: (1) the addressee was no longer a passive listener but actively involved in the task and (2) the task had a clear communicative purpose (guessing game). We predicted that in a highly interactive paradigm, in which participants interact with a "real" addressee, children might be more likely to take into account the needs of their interlocutor.

Thirty-two children and thirty adults participated. The children fell into two age groups: a group of 4-year-olds and a group of 5-year-olds. Participants were asked to play a guessing game with a "naïve" addressee (a confederate introduced to participants as another participant). In this game, the participants had to help the addressee find the "right picture" by describing one of the events from the minimal pair of the pictured typical/atypical instrument events of Experiment 3. As in the previous experiment, the addressee either had or did not have visual access to the events. Contrary to the previous experiment, the addressee interacted with the participant: At the beginning of each trial, the addressee said "I can see two pictures. Which one is it? Tell me about it!". At the end of each trial, the addressee said "I hope I got it right!" and placed a sticker next to the picture that best matched the participant's description. Feedback was provided in one practice trial at the beginning of the task.

The results from this experiment showed that both children and adults mentioned typical and atypical instruments equally frequently. There was also a developmental effect, with 5-year-olds mentioning instruments more frequently than 4-year-olds, but less frequently than adults. Crucially, participants of all age groups used more instrument information when the addressee did not have visual access to the events. Comparison between Experiments 3 and 4 demonstrated that children mentioned instruments more often in Experiment 4 than in Experiment 3. Therefore, in this more interactive paradigm, children demonstrated the ability to adjust their production to the specific needs of their addressee.

**Discussion and conclusions**

We showed that adult speakers, similarly to Brown and Dell (1987) and Lockridge and Brennan (2002), performed both 'generic' adjustments (adding information about atypical, i.e., generally unpredictable instruments) and more specific adjustments to addressees' needs (mentioning instruments more often when addressees could not see the events). Children, however, often made only generic (typicality-based) adjustments. Their ability to use information about the listener's visual perspective to make specific adjustments to event descriptions emerged only in contexts where the addressees' needs were made particularly transparent.

What were the precise factors that made addressees' needs accessible to children in Experiment 4? An important difference between Experiment 4 and Experiments 1-3 was the role of the addressee. In Experiment 4, the addressee was a "real" interlocutor, who had more genuine informational needs that children could easily identify. In fact, in studies that show early successes in children's abilities to make addressee-specific adjustments, the addressees are either the children's parents (see O'Neill 1996; O' Neill & Topolovec 2001) or confederates of the experimenter with an active role in the task (i.e., they had to follow the child's instructions, e.g., Bahtiyar & Küntay 2009; Nadig & Sedivy 2002). By contrast, in studies where the addressee is either passive or imaginary (Davies & Katsos 2010; Girbau 2001), children fail to make adaptations.

A second important difference between Experiment 4 and the previous experiments is the communicative purpose of the task. Although Experiment 2 also had a clear communicative goal (the listener to make accurate drawings), that could encourage children to attend to listener's informational needs, children largely ignored them. We believe that the crucial difference between Experiments 2 and 4 was whether the goal of the exchange was mutually achieved between the interlocutors or not. In Experiment 2, it was the addressee who carried the main burden for successfully completing the task. In Experiment 4, both the speaker and the addressee had to engage in a collaborative process (guessing game) to achieve a mutually pursued goal (finding the 'right' picture). This conclusion is supported by previous findings that demonstrate children's sensitivity to other people's perspective in tasks that require collaboration between interlocutors (e.g., Matthews, Lieven & Tomasello 2010; O'Neill 1996; O' Neill & Topolovec 2001).

Our findings, together with a growing body of work investigating how children learn to communicate in more interactive situations that best resemble real-life exchanges, contribute to our understanding of the mechanisms underlying children's pragmatic capacities.

Myrto Grigoroglou
Department of Linguistics & Cognitive Science
University of Delaware
125 East Main Street, Newark, DE 19716
E-mail: mgrigor@udel.edu

Anna Papafragou
Department of Psychological and Brain Sciences
University of Delaware
105 The Green, Newark, DE 19716
Email: APapafragou@psych.udel.edu

**References**

Bahtiyar, Sevda, & Aylin C. Küntay. 2009. Integration of Communicative Partner's Visual Perspective in Patterns of Referential Requests. *Journal of Child Language* 36(3). 529–55. doi:10.1017/S0305000908009094

Brown, Paula M. & Gary S. Dell. 1987. Adapting Production to Comprehension: The Explicit Mention of Instruments. *Cognitive Psychology* 19(4). 441–72. doi:10.1016/0010-0285(87)90015-6

Davies, Catherine & Napoleon Katsos. 2010. Over-Informative Children: Production/comprehension Asymmetry or Tolerance to Pragmatic Violations? *Lingua* 120(8). 1956–72. doi:10.1016/j.lingua.2010.02.005

Girbau, Dolors. 2001. Children's Referential Communication Failure: The Ambiguity and Abbreviation of Message. *Journal of Language and Social Psychology* 20(1-2). 81–89. doi:10.1177/0261927X01020001004

Lockridge, Calion B. & Susan E. Brennan. 2002. Addressees' Needs Influence Speakers' Early Syntactic Choices. *Psychonomic Bulletin & Review* 9 (3). 550–57. doi:10.3758/BF03196312

Matthews, Danielle, Jessica Butcher, Elena Lieven & Michael Tomasello. 2012. Two- and Four-Year-Olds Learn to Adapt Referring Expressions to Context: Effects of Distracters and Feedback on Referential Communication. *Topics in Cognitive Science* 4(2). 184–210. doi:10.1111/j.1756-8765.2012.01181.x

Matthews, Danielle, Elena Lieven, Anna Theakston & Michael Tomasello. 2006. The Effect of Perceptual Availability and Prior Discourse on Young Children's Use of Referring Expressions. *Applied Psycholinguistics* 27(3). 403–22. doi:10.1017.S0142716406060334

Matthews, Danielle, Elena Lieven & Michael Tomasello. 2007. How Toddlers and Preschoolers Learn to Uniquely Identify Referents for Others: A Training Study. *Child Development* 78(6). 1744–59. doi:10.1111/j.1467-8624.2007.01098.x

Matthews, Danielle, Elena Lieven & Michael Tomasello. 2010. What's in a Manner of Speaking? Children's Sensitivity to Partner-Specific Referential Precedents. *Developmental Psychology* 46(4). 749–60. doi:10.1037/a0019657

Nadig, Aparna S. & Julie C. Sedivy. 2002. Evidence of Perspective-Taking Constraints in Children's on-Line Reference Resolution. *Psychological Science* 13(4): 329–36. doi:10.1111/j.0956-7976.2002.00460.x

O' Neill, Daniela K. & Jane C. Topolovec. 2001. Two-Year-Old Children' S Sensitivity to the Referential (in)Efficacy of Their Own Pointing Gestures. *Journal of Child Language* 28. 1–28. http://journals.cambridge.org/abstract_S0305000900004566

O'Neill, Daniela K. 1996. Two-Year-Old Children's Sensitivity to a Parent's Knowledge State When Making Requests. *Child Development* 67(2). 659. doi:10.2307/1131839.

Perner, Josef & Susan R. Leekam. 1986. Belief and Quantity: Three-Year Olds' Adaptation to Listener's Knowledge. *Journal of Child Language* 13(2). 305–15. doi:10.1017/S0305000900008072.

**For which pragmatic phenomena is Theory of Mind necessary?:**

**Taking a different perspective**

Napoleon Katsos
University of Cambridge
nk248@cam.ac.uk

Clara Andrés Roqueta
Universitat Jaume I, Castelló
candres@uji.es

Keywords: Experimental pragmatics, implicature, Theory of Mind, perspective taking, False Belief

## 1. Theory of Mind and pragmatic inference

A question in theoretical and experimental pragmatics concerns how and when to represent the speaker's intentions (beliefs and desires, also known as Theory of Mind; ToM) in models of utterance interpretation. Some theorists argue that a representation of the speaker's intentions is required in all cases of pragmatic inference (Geurts, 2010; Sperber & Wilson, 2002, in the form of a modular mind-reading process dedicated to communication) while others suggest a limited role. In the latter camp, there is debate about which pragmatic inferences fall in which class and the mechanics of the inferential process. Subtly different views include Recanati's (1993; 2004) Associative vs Inferential or Primary vs Secondary Pragmatic Processes, Levinson's (2000) Generalised vs Particularised Implicatures, and Bach's Implicitures vs Implicatures (1994). In this group of theories, a role for representing aspects of the context including the speaker's intentions is restricted to some pragmatic inferences, and even in these cases it is limited to a final 'monitoring' stage where an inference which has been drawn may be cancelled once the speaker's beliefs and intentions are taken into account.

To illustrate the point of diversion we use the well-known case of quantity implicature. In agreement with much theoretical work, there is empirical evidence that listeners (or readers) in addition to accessing the literal meaning of the utterance, also take into account the following parameters in the process of generating implicature: the implicature's relevance to discourse goals (Breheny, Katsos & Williams, 2006), information focus (Zondervan, 2010), whether the speaker's knowledge is such that they could make the stronger statement (Bergen & Grodner, 2012; Breheny, Ferguson & Katsos, 2013), politeness considerations and face-threat (Feeney & Bonnefon, 2013; Bonnefon, Feeney & Villejoubert, 2009), likelihood of the stronger statement being true (Goodman & Stuhlmüller, 2013), the kind of alternatives (e.g. numerals as well as scalar quantifiers, in the case of scalar implicature, Degen & Tanenhaus, 2015), among others.

Of course, some of these considerations may but need not be assigned to a level of speaker's intentions. For example, relevance to discourse and information focus may be modelled at the

discourse- rather than the interlocutor- level, while likelihood of the stronger statement being true may be computed based on the listener's own expectations based on encyclopaedic knowledge rather than the speaker's expectations. However, other parameters do seem to necessitate representing a speaker, esp face-threat and epistemic state, with the experimental evidence on epistemic state strongly speaking for the need to represent individual speakers' knowledge state in specific communicative situations (Bergen & Grodner, 2012; Breheny, Ferguson & Katsos, 2013).

While the body of evidence for the factors that affect processing is quite rich for scalar implicature, similar evidence is emerging for other phenomena, or is already in hand for phenomena that are potentially relevant but not traditionally considered within the purview of pragmatics, such as word-learning (see Ambridge & Lieven, 2011). How would the theories mentioned in the introduction align with this kind of evidence? Proposals by Geurts (2010) and Sperber & Wilson (1986/1995) are fully compatible and would suggest that these and related considerations are taken into consideration from the very beginning of the process that generate implicatures. Other theorists though might argue that representing speakers' intentions is necessary for only some kinds of inferences, and/or assign their role to some secondary stage which cancels the inference.

A common feature of all these theories is that the role of ToM in pragmatic inference is discussed in terms of types of phenomena, e.g. whether ToM is or isn't necessary for scalar implicature or for word learning. This is the case not just for theorists who approach the question from linguistic or philosophical perspectives, but also for developmentalists and clinical linguists (O'Neill, 2012; Perkins, 2007).

**2. Operationalisations of Theory of Mind and inter-relation with pragmatic inferences**

Generally, at the empirical psycholinguistic level, ToM is operationalised in several ways, including success in False-Belief tasks (*FB* e.g. in Andrés-Roqueta & Katsos, submitted; Antoniou et al., submitted; Happé, 1993; Norbury, 2004), pro-social behaviour questionnaires (Nieuwland et al, 2010; Spotorno et al., 2014) or activation of neural regions and circuits independently shown to be engaged in ToM reasoning (Spotorno et al., 2012). It is not clear that anyone of these measures is fully representative of the abilities that are required when representing the speaker's perspective in pragmatic inferences (Cummings, 2015), or that they measure the same aspects of ToM and caution should be taken.

Let us now turn to empirical findings about ToM and pragmatics. We use metaphor, irony, and quantity implicature as case-studies, though word-learning (assigning referents to labels) and many other phenomena could serve to outline the debate. Findings for metaphor are mixed. Early studies report a positive correlation between success in FB and interpretation of metaphor in typically- and atypically-developing children (e.g. Happé, 1993). Subsequent work that factored in competence with vocabulary and morphosyntax as well as success in FB reports that only vocabulary and morphopsyntax are a significant predictor of success (Norbury 2004;

Andrés-Roqueta & Katsos, submitted ; see also Gernsbacher & Pripas-Kapit, 2012). The findings are mixed for quantity implicature too, with Nieuwland et al (2012) reporting a correlation between neural patterns of detection of violation of quantity and scores in the communication subscale of the Autism Quotient in neurotypical adults. Surian et al (1996) also report a positive correlation between quantity implicatures and success in FB in typically- and atypically-developing children. However, Antoniou et al, submitted, report no correlation between scalar implicature computation and any subscale of the Autism Quotient in neurotypical adults, while Andrés-Roqueta & Katsos (submitted) report no relation between scalar implicature and FB tasks in typically and non-typically developing children once vocabulary and morphosyntax are factored in the regressions. Irony on the other hand is consistently and positively associated with ToM, with correlational evidence between irony and FB tasks (author & author, submitted; Filippova & Astington, 2008; Happé, 1993) in typically- and typically-developing children, correlations between irony and scores in the social skills subscale of the Autism Quotient (Spotorno et al., 2014) as well as evidence of activation of the same neural regions (Spotorno et al., 2012).

**3. A novel proposal: epistemic dissonance as a cue for engaging ToM**

These findings may suggest that a way forward towards understanding which pragmatic phenomena do or do not engage Theory-of-Mind would be through subjecting (more) types of pragmatic phenomena to (more) careful experimental scrutiny. In any case, the prevailing trend in answering the question is to think of the relation between ToM and kinds of pragmatic phenomena (e.g. scalar implicatures or metaphors as a natural class).

In this presentation we propose that this is not the only way to pose the question. Our alternative proposal is that the engagement of ToM in real-time pragmatic inferences is not related to the type of pragmatic inference itself but to the situational context in which the inference occurs. In certain communicative contexts, such as those in which scalar implicatures tend to be tested in sentence judgment tasks, there is little need to engage ToM: in the typical experiment, the relevant contextual information (e.g. how many boxes have a token inside them or how many elephants have trunks) is always and manifestly assumed to be shared between speaker and listener. In the case of irony however as tested by Happé, (1993), Andrés-Roqueta & Katsos (submitted) and many others, participants have to interpret an unexpectedly unusual utterance based on the participant's idiosyncratic beliefs about the desirability of a certain outcome, which may or may not be commonly shared. We propose that ToM is engaged in all and only those cases where there is a manifest dissonance between two knowledge states, e.g. the speaker's and the listener's or the protagonist's and an expert's.

What is the role of representing the speaker's intentions in cases where no such dissonance arises? The logical possibilities are that the speaker's intentions are not part of the inferential process or that they are, but they are supplied from the hearer's own (egocentric) view (possibly operating on the default assumption that all critical information is shared). We believe that the latter is the case, but we agree that it is difficult to tell the two possibilities apart. Our proposal

overall is that the speaker is always represented in language comprehension, but ToM is only engaged in cases where there are cues of epistemic dissonance. In other cases there will be minimal demands on ToM as all that is required is representing the speaker's own perspective as shared.

A clear prediction of this view is that in other communicative situations (e.g. cases of manifestly privileged knowledge, e.g. in Breheny et al., 2013; cases of competing goals, e.g. as in Feeney, Scrafton, Duckworth & Handley, 2004) competence with scalar implicature will be predicted by FB scores, a hypothesis we are currently testing. Equally, in certain other cases, competence with irony will not be correlated with ToM, e.g. when the expectations for an ironic utterance are high. If this view is on the right track, it has the advantage of potentially explaining the variation in experimental findings as a function of the extent to which the interlocutors' perspectives were shared or not.

There are some limited similarities between the 'epistemic dissonance as cue' view and other approaches. For example, Giora's (2002) Graded Salience Hypothesis for irony argues that frequency, conventionality, familiarity, and prototypicality rather than literality or non-literality are the key factors on whether an ironic utterance will be easy to process or not. Here, we too argue that literality or non-literality, or membership to some category of pragmatic taxonomies is not the decisive factor. However, since the GSH is focussed on one phenomenon and on processing in general, rather than specifically the role of ToM, the two views only share some broad similarities.

The situation-dependent view we advocate for needs to identify the features of a communicative situation that signal epistemic dissonance and trigger the engagement of ToM, and to embed this in recent views of Theory-of-Mind (e.g. dual process; Apperly & Butterfill, 2009) and theories of common ground (e.g. Clark & Wilkes-Gibbs, 1985). Otherwise, it is too unconstrained and fails to make disconfirmable predictions.

**REFERENCES**

Ambridge, B., & Lieven, E.V.M. (2011). *Language Acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press. Chapter 4.

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

Apperly, I.A. & Butterfill, S.A, (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*. 116(4), 953-970.

Author & Author (submitted) The contribution of ToM and structural language to different kinds of pragmatic competence in children with ASD and SLI. *Journal of Child Psychology and Psychiatry*.

Bach, K.(1994). Conversational impliciture. *Mind and Language* 9, 124-162.

Bergen, L., & Grodner, D.J. (2012). Speaker Knowledge Influences the Comprehension of Pragmatic Inferences. *Journal of Experimental Psychology: Learning Memory and Cognition* 38 (5): 1450-1460.

Bonnefon, J.F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, 112, 249-258.

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

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1-39.

Degen, J., & Tanenhaus, M.K. (2015). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cognitive Science,* doi: 10.1111/cogs.12227

Feeney, A., & Bonnefon, J. F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of Language and Social Psychology*, 32, 181-190.

Feeney, A., Scrafton, S., Duckworth, A. & Handley, S.J. (2004). The story of *some*: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology* 58:2, 121-132.

Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*,79, 126–138.

Gernsbacher, M. A., & Pripas-Kapit, S. R. (2012). Who's missing the point? A commentary on claims that autistic persons have a specific deficit in figurative language comprehension. *Metaphor and Symbol*, 27(1), 93-105.

Giora, R. (2002). "Literal vs. figurative language: Different or equal?", *Journal of Pragmatics* 34: 487-506.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

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

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

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, 324-346.

Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology*, 23(3), 383-399.

Perkins, M. (2007). *Clinical Pragmatics*. Cambridge: Cambridge University Press.

Recanati, F. 2002. Does linguistic communication rest on inference? *Mind & Language*, 17 (1&2): 105-126.

Recanati, F. 2004. *Literal Meaning*. Cambridge: Cambridge University Press.

Sperber, D. & Wilson, D. (2002) Pragmatics, modularity and mind-reading. *Mind & Language*, 17. 3-23

Spotorno, N., & Noveck I.A. (2014). When is irony effortful? *Journal of Experimental Psychology: General*. 143(4):1649-65.

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J.B., Noveck, I., (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1), 25-39.

Surian, L., Baron-Cohen, S., & Van der Lely, H. (1996). Are children with autism deaf to Gricean maxims? *Cognitive Neuropsychiatry*, 1(1), 55–71.

# Language processing in shared task settings: How a partner influences spoken word production

**Anna K. Kuhlen (anna.kuhlen@hu-berlin.de)**
Department of Psychology, Humboldt Universität zu Berlin
Rudower Chaussee 18, 12489 Berlin, Germany

**Rasha Abdel Rahman (rasha.abdel.rahman@hu-berlin.de)**
Department of Psychology, Humboldt Universität zu Berlin
Rudower Chaussee 18, 12489 Berlin, Germany

**Abstract**

Acting jointly with a partner is different from acting alone. In this study we investigate whether speaking with a partner is different from speaking alone. Drawing upon a well-established speech production effect we investigate the degree of cumulative semantic interference experienced when naming a sequence of pictures together with a partner. Pictures of semantically related objects are either named by participants only, or also by their partner. Naming latencies increased with each additional within-category member, confirming cumulative semantic interference. Importantly, naming latencies increased more sharply when additional category members were named by the partner. A follow-up study suggests that this effect is not due to receiving additional auditory input. Instead, the mere belief of a partner naming the picture appears sufficient to elicit lexical processes comparable to naming the picture oneself. Our results speak for a profound and lasting effect of having a partner on the own speech production system.

**Keywords:** Speech production, semantic interference, joint action

## Introduction

Many pragmatic phenomena are fundamentally embedded in social interaction (Levinson, 1983). Yet, comparatively is known about how language is processed within a social interaction and how this may differ from language processing isolated from social context.

One characteristic of language use in conversational settings it that conversational partners alternate, often in quick succession, between speaking and listening (Clark, 1992; Pickering & Garrod, 2004). While one speaks, the other anticipates what is likely to be said and formulates the own response (Pickering & Garrod, 2007; Bögels et al., 2015). The two processes, attending to the partner's speech, and preparing one's own speech, are coordinated and are likely to influence each other.

In this study we investigate how a simple linguistic speech production task, naming pictures, may be influenced by the speech production of another individual. Studies investigating the cognitive processes underlying cooperation and social interaction more generally have shown that the task of one partner can influence the task of the other partner. For instance, when two partners perform complementary tasks in a shared setting, individual actors experience interference from the other person's task requirements (Sebanz et al., 2003). One explanation for this has been that the task of the partner is co-represented (Atmaca, Sebanz, & Knoblich, 2011; Sebanz, Knoblich, & Prinz, 2005)

This may also apply to speaking: In a recent study by Baus and colleagues (Baus et al., 2014), two participants took turns naming pictures of objects with high or low word frequency. Electroencephalographic recordings (EEG) during those trials in which the partner (but not the participant) had to name the object showed distinct signatures of electrophysiological activity in response to word frequency (that were less pronounced when nobody named the object). This suggests that participants engage in lexical processes not only when naming the object themselves, but also when the partner is naming the object. Yet it is still an open question whether such a simulation of the partner's speech production also affects the own speech production system.

To address this question we investigate cumulative semantic interference, a well-documented effect in single subject settings that is characterized by linearly increasing naming latencies when speakers name several semantically related objects in close succession (e.g., Belke, 2013; Costa, Strijkers, Martin & Thierry, 2009; Howard, Nickels, Coltheart, & Cole-Virtue, 2006; Navarrete, Mahon, & Caramazza, 2010).

One explanation for this effect is increased competition on the level of semantically related lexical entries: When naming a picture the depicted object elicits the activation of its concept, and in turn the activation of the corresponding lexical entry. At the same time semantically related concepts and their lexical representations are also activated. Thus, the target lexical entry needs to be selected among several co-activated, semantically related competing entries (e.g., Abdel Rahman & Melinger, 2009; Dell, 1986; Levelt et al., 1999; Roelofs, 1992). Once the target lexical entry has been selected the link between the entry and its concept is strengthened. When subsequently a different, semantically related object needs to be named, the prior named concept is co-activated along with the link to its lexical entry. This imposes strong competition, causing the new target lexical entry to be selected later. As the number of strong competitors increases, interference between semantically related items also increases (Howard et al., 2006).


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 82–86.

In a shared task setting in which two partners take turns naming objects, we hypothesize that cumulative semantic interference can be elicited not only by naming objects oneself, but also by objects that are named by a partner. This would suggest that the partner's naming of the object elicits lexical representations similar to the ones elicited when naming the object oneself. Specifically, we assume that knowing that the partner names a particular object will strengthen the link between concept and lexical entry also in the other partner even without articulating the object name. This would provide evidence that the speech production requirements of a task partner are not only co-represented but also exert a lasting influence on one's own speech production.

### Present Study

In the present study participants successively named objects, some of which were semantically related (e.g., several types of birds), together with a partner (an experimental confederate). Within some semantic categories, half of the exemplars were named by the partner; within other semantic categories half of the exemplars were named by neither partner nor participant. Thus, in both conditions participants named in close succession an equal number of semantically related objects; what differed was whether, interspersed, additional objects of the same category were named by the partner, or whether they were presented visually but were not articulated by anyone.

We predict that having a partner will lead participants to co-represent their partner's task (Baus et al., 2014), hence activating the lexical representation of the object named by the partner in a fashion similar to naming it oneself. We therefore expect a steeper increase in naming latencies (i.e. increased lexical competition) for those categories co-named with a partner compared to those categories named by the participant only.

## Methods

### Participants

Twenty-six native speakers of German (7 male, 19 female) between the ages 19-34 participated in the experiment. Participants gave informed consent and were compensated with €8 per hour or received credit towards their curriculum requirements. Two participants (1 male, 1 female) had to be excluded due to technical failure.

### Materials

Three hundred and twenty colored pictures (photographs) of man-made or natural objects were collected. The objects mapped onto 32 different semantic categories (e.g., birds, beverages, flowers). Each category held 10 exemplars. Hundred and twenty additional objects served as filler items, which were unrelated to the categories underlying the target items. All pictures were scaled to 3.5 cm x 3.5 cm and had a homogenous grey background.

### Design

Pictures were collated in a stimulus list, which was created for each participant individually with the following constraints: The order in which exemplars of one category occurred was randomly selected (by the program "Mix", van Casteren & Davis, 2006), and they were separated randomly by a minimum of two and a maximum of six unrelated items (separating items could be filler items or items belonging to a different category). To avoid a conceptual merging of two or more related categories (e.g., categories fish and birds merging to the superordinate category animal) related categories never overlapped within a list.

Half of the exemplars of a given category were assigned to "Participant Go" trials (participant names object). Under the Joint Naming condition, the other half of exemplars was assigned to "Partner Go" trials (partner names object); under the Single Naming condition, the other half was assigned to "Joint No Go" trials (nobody names object). The assignment of trial type was random with the two exceptions: Participant Go trials were separated by maximally three Joint No Go or Partner Go trials. The first and the last presented exemplar of a category were always assigned to Participant Go trials. All filler items were Participant Go trials.

The assignment of categories to naming condition (Joint vs. Single Naming) was balanced across participants.

### Procedure

Prior to the experiment, all pictures (unsorted) and their written names were presented to participants on paper. Participants had approximately 5 min to study the pictures and familiarize themselves with their corresponding name.

In the main experimental session, participants and their partner sat next to each other in front of the computer screen. One picture was presented at a time. A colored frame around the picture indicated who was to name the object: the participant, the partner, or nobody. Participants and partner were instructed to name as fast and accurately as possible those objects coded in their assigned color. In all other trials they were told to do nothing. The corresponding color codes were assigned randomly at the beginning of each experiment.

Trials began with a fixation cross of 0.5s. The picture was then presented until a response was given or for a maximum of 2s. A blank screen of 1.5s followed each picture presentation and then the next trial followed. Naming latency (reaction time) was recorded with the help of a voice-key from the onset of the picture presentation. During the experiment, the experimenter coded any failure of the voice-key (onset too early or too late) as well as erroneous trials (object named wrongly or by the wrong person).

### Data analyses

Naming latency was analyzed for those trials in which participants named the picture (Participant Go). Data obtained from filler trials were excluded. Furthermore excluded were trials with a naming latency that deviated two

standard deviations from an individual's mean value as well as pictures that were named incorrectly.

A 5 (ordinal position) x 2 (naming condition) repeated measures analysis of variance (ANOVA) with participants ($F_1$) and categories ($F_2$) as random variables tested for the hypothesized interaction between ordinal position and naming condition. In addition a dependent t-Test compared the two experimental conditions based on the slopes of the regression line between naming latency and ordinal position.

## Results

Replicating the cumulative semantic interference effect, naming latencies increased linearly with each ordinal position within a semantically related category, $F_1(4, 92)= 11.36$, p= .00, $\eta^2= .33$; $F_2(4, 124)= 15.14$, p= .00, $\eta^2= .33$. There was no main effect of naming condition, $F_1(1, 23)= 1.30$, p= .27, $\eta^2= .05$; $F_2(1, 31)= 1.27$, p= .27, $\eta^2= .04$. Crucially, though, there was the predicted interaction between naming condition and ordinal position, $F_1(4, 92)= 3.02$, p= .02, $\eta^2= .12$; $F_2(4, 124)= 2.80$, p= .03, $\eta^2= .08$, indicating that naming latencies increased more steadily in the Joint Naming condition than in the Single Naming condition, see *Figure 1*. Moreover, the slope of the regression line between naming latency and ordinal position was significantly steeper in the Joint Naming compared to the Single Naming condition, t(23) = 2.90, p = .01.

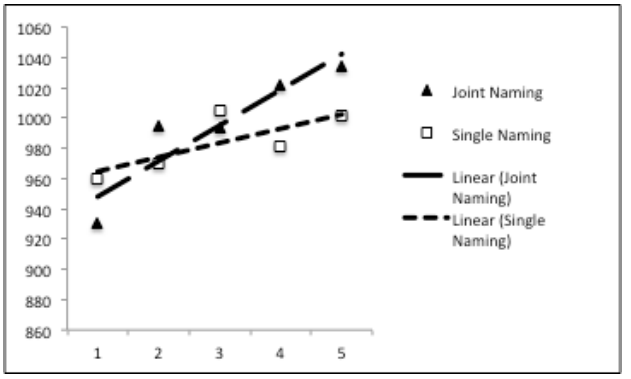



**Figure 1:** Mean naming latency broken down by ordinal position and naming condition. A linear trend visualizes the cumulative increase in naming latencies.

## Discussion

Speakers show increased picture naming latencies with each additional member within a semantic category they name in a sequence of pictures (e.g., Belke, 2013; Costa, Strijkers, Martin & Thierry, 2009; Howard, Nickels, Coltheart, & Cole-Virtue, 2006; Navarrete, Mahon, & Caramazza, 2010). In this study we show that naming latency not only increases in response to participants' own prior naming of within-category pictures, but also in response to witnessing their task partner naming the pictures: Naming latencies increased more steadily for those categories in which the partner named half of the category members compared to those categories in which additional members were presented just visually (but named by neither partner nor participant). Note that in both conditions speakers themselves named an equal number of within-category pictures. The steeper increase in categories co-named with a partner therefore results from those pictures named by the partner. More specifically, our results suggest that pictures named by the partner elicit in participants lexical processes comparable to naming the picture themselves. This is consistent with previous work showing that participants engage in lexicalization processes when task partners name pictures (Baus et al., 2014).

Going beyond previous work, our study shows that witnessing the partner name pictures has lasting effects on the participants' own speech production. According to a common understanding of the mechanisms underlying semantic inhibition effects, naming a picture strengthens the link between concept and lexical entry, and hence increases lexical competition when naming another semantically related picture (Howard et al., 2006). Accordingly, the pattern of our results suggests that merely overhearing a partner naming the picture can also strengthen the link between concept and lexical entry. As a result the degree of lexical competition experienced by a speaker can be manipulated by another person's prior lexical choices.

But how "social" is our effect? Our study cannot determine whether participants represented the partner's behavior as contribution to a socially shared task. Instead, increased semantic inhibition may have been elicited solely by *hearing* the name of additional pictures (compared to only *seeing* them). In other words, our effect may have nothing to with having a task partner, but rather with receiving auditory input (that happens to be generated by a partner). To test this possibility a follow-up experiment investigates the development of semantic interference under conditions in which participants only believe partners to name semantically related pictures (but cannot hear them).

Data collection for this experiment is currently ongoing. So far, 18 of 24 participants have been recruited. Two volunteer participants are invited to our lab at one time. They are introduced to each other as task partners. After receiving instructions identical to our first experiment, participants are told they will be performing the task of jointly naming pictures spatially isolated from each other. They are then seated in two separate cabins where they undergo procedures identical to the first experiment with the exception that their task partner is not present. Hence, during those trials in which the partner is required to name the object, participants do not receive any auditory feedback from their partner's performance. Instead, a computer simulates the partner's naming latency (which can be inferred by how quickly the picture disappears from the screen) based on average reaction times recorded during the first experiment. Thus, the two participants perform the task completely independently from each other. Only their belief of performing the task together with their remote partner defines the shared task setting.

As in our first experiment, preliminary results show that naming latencies increase with the number of within-category pictures named, $F_1(4, 60)= 10.45$, $p= .00$, $\eta^2= .41$; $F_2(4, 124)= 9.49$, $p= .00$, $\eta^2= .23$. Different from our first experiment naming latencies appear to be higher overall when categories are presumably named together with a partner than when they are presumably named alone; at least when calculating over participants, $F_1(1, 15)= 7.65$, $p= .01$, $\eta^2= .34$, but not over categories $F_2(1, 31)= .32$, $p= .57$, $\eta^2= .01$. Importantly, naming latencies increased more steeply when categories were presumably named together with a partner, resulting in an interaction effect, $F_1(4, 60)= 2.75$, $p= .04$, $\eta^2= .16$; $F_2(4, 124)= 1.87$, $p= .12$, $\eta^2= .06$. *Figure 2* visualizes our findings.

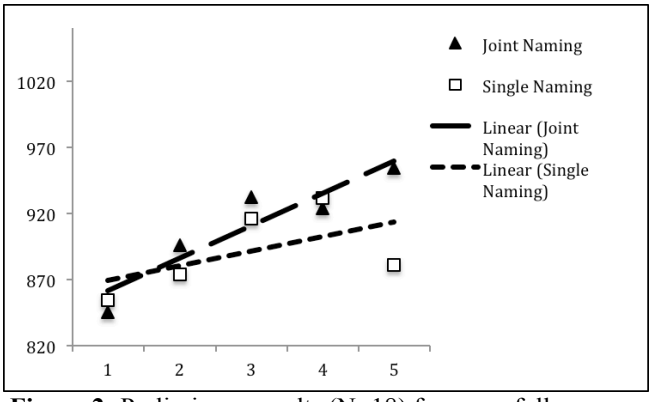



**Figure 2:** Preliminary results (N=18) from our follow-up experiment excluding auditory feedback from the partner. Mean naming latency broken down by ordinal position and naming condition. A linear trend visualizes the cumulative increase in naming latencies.

The preliminary data from this follow-up study speak against the possibility that auditory input alone is the driving factor behind the increased semantic inhibition experienced for categories co-named with a partner. Instead, our data suggest that the mere belief that a partner is naming an object is sufficient to trigger in participants lexical processes.

Together our two studies show that acting together with a partner can change profoundly how participants respond to a task. Even within a minimal social context, in which collaboration is not required, speaking together with another person is different from speaking alone. Having a partner, or even the mere belief of having a partner, leads to co-representing the partner's task. When such a representation involves simulating lexical retrieval, speaking jointly with a partner can have lasting effects on one's own speech production processes. With our study we contribute to the growing body of literature showing how in social settings, speech production and speech comprehension become tightly interwoven and influence each other.

## Acknowledgments

This work was funded by DFG Grant AB277/4-2. We thank our colleagues for helpful discussions, especially Sebastian Rose, and our student assistants for their support in collecting the data. We would like to acknowledge support by grant DFG,

## References

Abdel Rahman, R., & Melinger, A. (2011). The dynamic microstructure of speech production: semantic interference built on the fly. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37(1),* 149–161. http://doi.org/10.1037/a0021208

Atmaca, S., Sebanz, N., & Knoblich, G. (2011). The joint flanker effect: Sharing tasks with real and imagined co-actors. *Experimental Brain Research*, *211(3–4),* 371–385.

Baus, C., Sebanz, N., de la Fuente, V., Branzi, F. M., Martin, C., & Costa, A. (2014). On predicting others' words: Electrophysiological evidence of prediction in speech production. *Cognition, 133(2),* 395–407. http://doi.org/10.1016/j.cognition.2014.07.006

Belke, E. (2013). Long-lasting inhibitory semantic context effects on object naming are necessarily conceptually mediated: Implications for models of lexical-semantic encoding. *Journal of Memory and Language, 69(3),* 228–256. http://doi.org/10.1016/j.jml.2013.05.008

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5,* 12881. doi:10.1038/srep12881.

Clark, H. H. (1992). *Arenas of language use.* Chicago, IL: University of Chicago Press.

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences of the United States of America, 106*(50), 21442-21446. doi: 10.1073/pnas.0908921106

Dell, G. S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review, 93*(3), 283-321. doi: Doi 10.1037//0033-295x.93.3.283.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: experimental and computational studies. *Cognition, 100(3),* 464–482. http://doi.org/10.1016/j.cognition.2005.02.006

Levelt, W.J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and Brain Sciences, 22*(1), 1-38.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Navarrete, E., Mahon, B. Z., & Caramazza, A. (2010). The cumulative semantic cost does not reflect lexical selection by competition. *Acta Psychologica, 134*(3), 279-289. doi: 10.1016/j.actpsy.2010.02.009

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 167–226.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11(3),* 105–110.

Roelofs, A. (1992). A Spreading-Activation Theory of Lemma Retrieval in Speaking. *Cognition, 42*(1-3), 107-142.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition, 88(3),* B11–B21.

Sebanz, N., Knoblich, G., & Prinz, W. (2005). How two share a task: Corepresenting stimulus-response mappings. *Journal of Experimental Psychology. Human Perception and Performance, 31(6),* 1234–1246.

# The effect of context on generic and quantificational statements

Lazaridou-Chatzigoga[1,2], Dimitra, Katsos[1], Napoleon and Stockall[2], Linnaea
[1]Department of Theoretical and Applied Linguistics, University of Cambridge
[2]Department of Linguistics, School of Languages, Linguistics and Film, Queen Mary, University of London

## 1. Introduction

In this paper, and in the broader research program it forms part of, we are investigating the similarities and differences between ways of expressing generalisations in natural language. Quantificational generalisations, as in (1), are expressed in quantitative, statistical terms, while generic generalisations, as in (2), make general claims about kinds of entities and refer to a property that is characteristic of the kind in question, but not necessarily statistically prevalent.

(1) Some lions live in cages.
(2) Lions have manes.

Such generalisations have long been addressed by the formal semantics approach to genericity, within which genericity is viewed as a species of quantification, akin to quantificational adverbs such as 'typically' (see Krifka et al. 1995 and Mari et al. 2013 for discussion and further references therein).

In contrast to the quantificational analysis of generics, a growing body of experimental and developmental work on the topic proposes that genericity is categorically different from (and significantly simpler than) quantification (Leslie 2007, Gelman 2010). This latter hypothesis, called the *Generics-as-Default* view (*GaD* view henceforth) postulates that generics are a default and innate mode of thinking. This idea is linked to the view of cognition that assumes two different systems, made popular by Kahneman and Frederick (2002), which includes a distinction between System 1, a fast, automatic, effortless lower-level system and System 2, a slower, more effortful higher-level rule-governed system.

Within the quantificational semantics approach, the critical challenges have been to understand and model the ways in which generic generalisations are licensed, and to account for the fact that generic generalisations can be made using a wide range of different grammatical means, both within, as illustrated for English in (3), and across languages (see Behrens 2000 for typological comparisons and discussion), but no language has a unique, unambiguous marker of genericity equivalent to a quantifier or determiner.[1]

(3)
a. John drinks coffee.

[1] Our research program also investigates the cross-linguistic aspect of genericity in order to systematically compare the interpretation of generic and quantified statements across languages with distinct generic morpho-syntax. Thus, we investigate Greek adult comprehension of generics. In a language like Greek a definite plural NP is by far the most frequent in generic statements (Marmaridou-Protopapa, 1984), which furthermore is ambiguous between a generic and a specific reading (see Ionin et al. 2013 for a similar situation in Spanish):

i. I tighris ehun righes.
the tigers have.PL stripes
'Tigers have stripes/The tigers have stripes.'


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 87–94.

b. My brother drinks coffee.
c. A teacher drinks coffee.
d. Every teacher drinks coffee.
e. Coffee is tasty.

Determining which properties or attributes can be generically predicated has proven very challenging. Generic generalisations can range from exceptionless statements such as *triangles have three sides*, or *the walrus is a mammal*, through what Leslie et al. (2011) call 'majority characteristic' statements such as *dogs have four legs or stop signs are red*, which are true of the overwhelming majority of instances, with only a few exceptional individuals, through 'minority characteristic' statements like *ducks lay eggs* or *lions have manes*, which all involve primary or secondary sexual characteristics of animals, and are thus only true of less than 50% of individuals, to 'striking' generalisations like *sharks attack swimmers*, or *mosquitos carry malaria*, which are true of only a tiny fraction of individuals, but involve properties which are noteworthy in some way.

Not only is statistical prevalence not necessary to licence generic generalisation, it is not sufficient either. Statements like *books are paperbacks* or *Canadians are right-handed* can be true of 80% or more of individuals, and yet are not typically judged as true, and thus somehow fail as generic statements.

In experiments currently running, we are investigating the extent to which strikingness determines whether young children will extend a generically predicated property of a novel entity (such as *Borps love to cheat at games*). See also Prasada and Dillingham (2006) for discussion of some of the potential differences between subtypes of generic.

In the GaD approach, because generic generalisations are understood to be a basic, pre-linguistic mode of thinking, some of the specific challenges for the quantificational analysis are avoided. For instance, according to this view, there is no overt generic operator in any known language because generics are the unmarked case. Only effortful, non-default quantificational generalisations require overt linguistic exponence on this view. It is not, obvious, however, that this approach fares any better in offering a principled explanation of why *Italians are good skiers* is typically judged as true (and thus an acceptable generic statement), while *Canadians are right-handed* is not.

In Lazaridou-Chatzigoga et al. (2015), we juxtapose these two lines of research and in doing so we highlight some of the significant challenges for each approach. We argue, for instance, that the formal semantics models do not offer any clear explanation for the robust child language findings that generic utterances and generic interpretations are prevalent in children as young as 2 years old, despite not being associated with any overt morpho-syntactic marker in any known language. On the other hand, we also argue that the evidence for the GaD proposal is significantly weakened by a lack of cross-linguistic comparison, or serious engagement with the formal semantics of quantification and specificity.

In this paper we address the effect of context on generic and universally quantified generalizations (UQGs). The Generic-Overgeneralisation effect, 'GOG' effect, is "the tendency to overgeneralize the truth of a generic to the truth of the corresponding universal statement" (Leslie et al. 2011:17). The GOG effect has been used to support the GaD view (Leslie 2007, 2008, Gelman 2010). An alternative explanation, based on domain restriction (DR), is that people interpret *all ducks lay eggs* as a claim only about the relevant restricted set of female fertile ducks (see

Carlson 1999, Stanley and Szabò 2000, Greenberg 2007). Leslie et al. ruled out DR as an alternative explanation, but we argue that their results are challenged by the following observations: the contexts they use to induce specific/individual interpretations do not make salient the exceptions that would make the universal quantification over individuals interpretation untrue and are not enough to make DR to only the relevant (potentially egg laying) ducks impossible. We argue that alternative explanations for the GOG effect have not been ruled out. To do so, we show that the GOG effect can be replicated while manipulating different levels of contextual information preceding the critical utterance (contradictory, supporting or neutral). Our experiments tackle the relevance of DR for these data, as DR is routinely invoked in quantification (Heim 1991) and listeners are known to be charitable (Grice 1975). By using different levels of context and universal quantifiers with different sensitivity to DR (*all, all the, every*), we show how an explanation based on DR could work.

## 2. The GOG effect

Leslie et al. (2011) use the GOG effect to refer to "the tendency to overgeneralize the truth of a generic to the truth of the corresponding universal statement" (Leslie et al. 2011:17), while Hollander et al. (2002) and Leslie and Gelman (2012, experiment 4) report this effect in children, but it is also evident is adults.

The first detailed investigation of the GOG effect is found in Leslie et al. (2011). In their experiment 1, participants had to perform a truth value judgement (TVJ) task on sentences that were presented in one of three forms: generic, universal (*all*), or existential (*some*). The statements involved different kinds of properties as discussed above (§2.3): quasi-definitional (*triangles have three sides*), majority characteristic (*tigers have stripes*), minority characteristic (*ducks lay eggs*), majority non-characteristic (*cars have radios*), striking (*pit bulls maul children*), and false generalizations (*Canadians are right-handed*). The authors report experimental evidence that adults sometimes judge universal statements as true, despite knowing that they are truth-conditionally false. For example, participants judged a quantified statement like *all tigers have stripes* as true, even though they know it is false given that there are albino tigers. The authors claim that the participants made this 'error' because they relied on the corresponding generic statement (*tigers have stripes*), which is true. They find that the GOG effect occurs in more than half the trials when the statement involves characteristic properties: 78% for majority characteristic and 51% for minority characteristic statements.

Leslie et al. consider some alternative explanations before concluding that the GOG effect is the most suitable interpretation of their results: a) subkind interpretation, according to which people interpret *all ducks lay eggs* as 'all kinds of ducks lay eggs' and thus *all ducks lay eggs* is true under this interpretation, b) ignorance of the facts, according to which people actually think that all ducks (both male and female) lay eggs and c) domain restriction, according to which people interpret *all ducks lay eggs* as a claim only about the relevant restricted set of female fertile ducks (as per Carlson 1999, Greenberg 2007, discussed above).

The authors discarded the first explanation through a paraphrase task, which asked participants to provide paraphrases of the statements they had just read (their experiment 2b). Subtypes were almost never referred to in the paraphrases, which the authors take to mean this kind of interpretation is not readily available to participants, and thus can't explain the GOG effect, though this is hardly a knock down argument.

The second explanation was ruled out on the basis of a knowledge test that showed that people knew the relevant biological facts (their experiment 3).

Let us focus here on the third possible explanation, which they addressed in experiment 2a. In order to check for the possibility of domain restriction in the sense of Stanley and Szabó (2000), as discussed above, they provided the participants with population estimates for the kind in question in the following form:

(4) ‘‘Suppose the following is true: there are 431 million ducks in the world. Do you agree with the following: all ducks lay eggs.’

This information was supposed to prime quantification over every individual duck in the world, and thereby make it difficult/impossible to interpret *all* as restricted to only the ducks that are presupposed by *lay eggs*. If acceptance of *all ducks lay eggs* in the first experiment was driven by contextual quantifier domain restriction, the authors predict that it would disappear in the context of population information of the kind above. Nevertheless, the authors report that the GOG effect still occurred on a substantial portion of trials, with an acceptance rate for *all* statements at 55% for majority characteristic statements and 30% for minority characteristic statements, which is less than when the statements appeared with no preceding context, but is still a high percentage. The authors thus conclude that domain restriction cannot be the sole explanation for the GOG effect.

## 3. The current experiments

In our work we show that the rate of participants’ ‘GOG’-like behaviour can be altered by manipulating different levels of contextual information preceding the critical utterance (neutral, contradictory or supporting). We address the above issues by using online measures, different levels of context and three universal quantifiers with different sensitivity to DR (all, all the, each).

We illustrate here the three levels of context for a ‘majority characteristic’ statement like *tigers have stripes* that appears in either the generic condition (*tigers have stripes*) or in one of the universal quantifier conditions (*all tigers have stripes, all the tigers have stripes, each tiger has stripes*) following one of the three levels of context, as below:

(5)

a) neutral context:
Linton Zoo is home to three tigers, Tibor, Baginda and Kaytlin, whose playful games visitors love to watch and photograph.

b) contradictory context:
Linton Zoo is home to three tigers, Tibor, Baginda and Kaytlin, whose fur is all white due to a recessive gene that controls coat color.

c) supportive context:
Linton Zoo is home to three tigers, Tibor, Baginda and Kaytlin, whose black and orange coats visitors love to photograph.

Following the received view on generics (Krifka et al. 1995), the interpretation of generics should not be affected by contextual narrowing, so participants should accept generics as true in all levels of context. On our understanding, the GaD view would not make different predictions on this point. Based on the pilot study we discussed

above, we hypothesize that the level of context might affect generics, especially when the context makes the exceptions salient. We predict that we will find some effect of the context only in the case of contradictory context. The effect might be manifest either in the TVJ or in reaction times or both. We expect that if we find any errors in the acceptance of true generics these will mostly depend on the type of context. For example, we expect that participants might decide that a statement such as *tigers have stripes* is false only after being exposed to a context where some stripeless tigers have been made salient. Alternatively, if participants judge correctly that the statement is true, they might take longer to come to that judgement. Both would be indications that participants entertain the dependency of generic sentences on context.

Our predictions critically diverge from the GaD view with respect to the different universal quantifiers, since we predict variation in different contexts because there are a range of different reasons for participants to offer a 'true' response. We expect that participants' responses are not biased towards a generic interpretation, but are rather dependent on the sensitivity of the quantifier to DR, kind level interpretation and prototypical interpretation. Thus, whereas the GaD view predicts an equal percentage of 'true' responses for all quantifiers in all levels of context, we hypothesize that the acceptance of universally quantified statements will be dependent on the level of context they were paired with and on the sensitivity of the quantifier to DR.

More specifically, domain restriction is more likely if the universally quantified statement used does not require linking with a set under discussion, as is the case with all, compared to each and all the, which do (Partee, 1995). Given that all so easily lends itself to a domain restricted interpretation, it is an unfortunate choice for a universal quantifier to test the predictions of the GaD view.

Given that *all* is compatible with a contextually restricted interpretation, but does not enforce it, *all*-statements might be judged as 'false' after a neutral or supporting context and this acceptance might be even lower after a contradictory context. For a neutral context, we predict that a participant could respond 'true' either because they have generated a kind level interpretation or because they have generated a prototypical interpretation, while acceptance after a contradictory statement is only licensed under a kind level interpretation, since the individual level quantification is unavailable for this pairing of context-quantifier. More possible routes to a 'true' response should be associated with higher acceptance rates for neutral than for the contradictory contexts.

By contrast, for *all the*- and *each*-statements, we predict very low acceptance, if any, after a contradictory context, given that these quantifiers are necessarily contextually restricted (or D-linked in the sense of Pesetsky, 1987), whereas acceptance after a neutral or supportive context will be higher. In the contradictory context, participants are provided with some individuals that are exceptions to what the subsequent statement predicates. Given that *all the* and *each* have to be interpreted as referring back to the individuals introduced in the discourse, the participants are expected to interpret them as contextually restricted, yielding a 'false' response appropriate. For a neutral context, we predict that a participant would respond 'true' given that they have no reason to doubt that the normal situation holds, according to which, for instance the tigers mentioned in the context have the normal characteristics of tigers, namely, they have stripes. For a supporting context, we predict even higher acceptance rates than for the neutral context, given that the participants have to necessarily link the tigers mentioned in the statement to the tigers mentioned in the

context, for which they have supporting evidence of the relevant characteristic, on top of their encyclopaedic knowledge that tells them that this the case.

By using this careful manipulation of context, we aim to show that people interpret generic and universally quantified statements relying on their knowledge of the semantics and pragmatics of generic sentences of the quantifiers involved and on their availability for a contextually restricted interpretation.

Analyses of response patterns shows that, as predicted, there are main effects of both context type and determine type and of their interaction. There is a difference between quantifiers, with GEN being judged as true more often than the others. However, even GEN is sensitive to the preceding context to some extent (fewer true judgements in the contradictory context compared to the other two contexts). The difference between the contradictory context and the other two contexts is bigger for the other quantifiers. Moreover, within the UQGs, the difference seems to be bigger for those quantifiers that require DR because of their semantics (*all the, each*) than the one that allows but does not require DR to the relevant subset *(all*). These results suggest that context plays an important role in the interpretation of any kind of generalisation, in ways that can be understood by appealing to DR. In Figure 1 we summarize the results per context and determiner type:

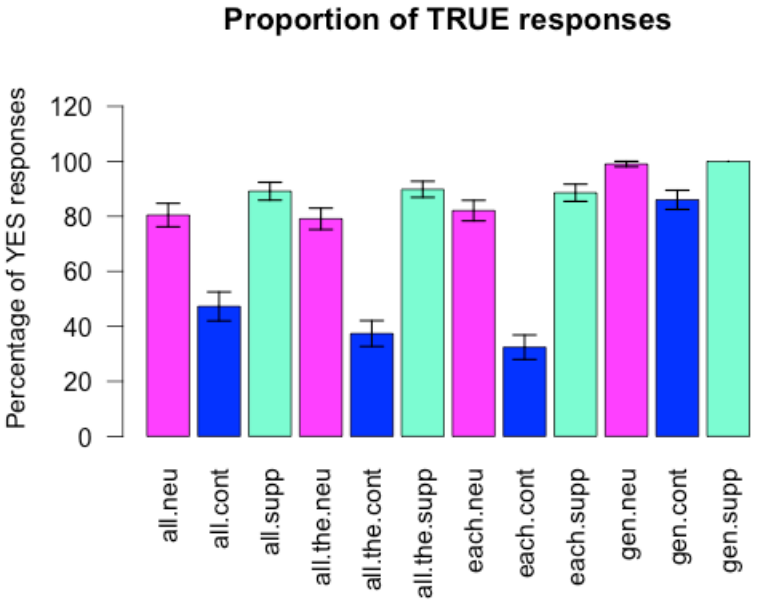



Figure 1. Proportion of TRUE responses

Furthermore, besides offline ratings, reaction times were collected in our studies, given that online measures are likely to be sensitive to depth of processing for inferential processes such as DR (particularly restricting to otherwise dispreferred subsets), which is a critical prediction. Based on previous studies (Lazaridou-Chatzigoga and Stockall 2013), we predict an RT effect on generics only when the exceptions are salient. This is in line with recent work (Sterken 2015) that claims that generics display some context sensitivity, but contrary to the received view that takes generics to strictly resist contextual restriction (Krifka et al. 1995).

## 4. Conclusion

We put forward a DR-based explanation of the GOG effect and present off-line and on-line data that support it. The general thrust of this work is that, rather than being under the influence of a default bias, children and adults are simply sensitive to the subtle interplay of quantifier semantics and pragmatics on the one hand, and context on the other. This approach has the advantage of accounting for data without postulating ad-hoc mechanisms such as GOG just for generics.

## References

Behrens, L. (2000). Typological Parameters of Genericity. Arbeitspapier 37 (Neue Folge). Institut für Sprachwissenschaft, Universität zu Köln.

Carlson, G. (1999). Evaluating Generics. In P. Lasersohn (Ed.), *Illinois Studies in the Linguistic Sciences*, 29:1, 1-11.

von Fintel, K. (1994). Restrictions of quantifier domains. PhD diss., University of Massachussets, Amherst.

Gelman, S.A. (2010). "Generics as a window onto young children's concepts". In F. J. Pelletier (Ed.), Kinds, things, and stuff: The cognitive side of generics and mass terms. New directions in cognitive science, 100-123. New York: Oxford University Press.

Greenberg, Y. (2007). Exceptions To Generics: Where Vagueness, Context Dependence And Modality Interact. *Journal of Semantics* 24(2), 131-167.

Grice, H.P. (1975). "Method in Philosophical Psychology: From the Banal to the Bizarre", *Proceedings and Addresses of the American Philosophical Association* (1975), 23-53.

Heim, I. (1991). Artikel und Definitheit. In A. von Stechow and D. Wunderlich (eds.), Semantik: Ein internationales Handbuch der zeitgenössischen Forschung, 487-535. Berlin: de Gruyter.

Hollander, M.A., Gelman, S.A., and Star, J. (2002). Children's Interpretation of Generic Noun Phrases. *Developmental Psychology* 36(6), 883-894.

Ionin, T., Montrul, S. and Santos, H. (2011). An experimental investigation of the expression of genericity in English, Spanish and Brazilian Portuguese. *Lingua,* 121*,* 963-985.

Kahneman, D. and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.) Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press. New York, 49-81.

Krifka, M., Pelletier, F., Carlson, G., ter Meulen, A., Chierchia, G. and Link, G. (1995). Genericity: An Introduction. In G. Carlson and F. J. Pelletier (Eds.) The Generic Book. Chicago. Chicago University Press, 1-125.

Lazaridou-Chatzigoga, D. and Stockall, L. (2013). Genericity, exceptions and domain restriction: experimental evidence from comparison with universals. In *Proceedings of Sinn und Bedeutung* 17, E. Chemla, V.Homer and G. Winterstein (eds.), ENS Paris, 325-343.

Lazaridou-Chatzigoga, D., Katsos, N., and Stockall, L. (2015) Genericity is easy? Formal and experimental perspectives. RATIO Special Issue: Investigating Meaning: Experimental Approaches, edited by Nat Hansen and Emma Borg. Volume 28, Issue 4, pages 470–494.

Leslie, S.J. (2007). Generics and the structure of the mind. *Philosophical Perspectives 21*, 375-405.

Leslie, S. J. (2008). Generics: Cognition and acquisition. The Philosophical Review, 117(1), 1–49.

Leslie, S. J. and S. Gelman. (2012). Quantified statements are recalled as generics: evidence from preschool children and adults. Cognitive Psychology 64(3), 186-214.

Leslie, S.J., Khemlani, S. and Glucksberg, S. (2011). All Ducks Lay Eggs: The Generic Overgeneralization Effect. *Journal of Memory and Language 65*(1), 15-31.

Mari, A., Beyssade, C. and del Prete, F. (2013). Genericity. Oxford: OUP.

Marmaridou-Protopapa, S. (1984). The study of reference, attribution and genericness in the context of English and their grammaticalization in M. Greek noun phrases. Unpublished Ph.D. thesis. Darwin College. Cambridge.

Partee, B. (1995). Quantificational structures and compositionality. In E. Bach, E. Jelinek, A. Kratzer and B. Partee (eds.). Quantification in Natural Languages. Dordrecht: Kluwer, 541–601.

Pesetsky, D. (1987). Wh-in-Situ: Movement and Unselective Binding. In The Representation of (In)definiteness, Eric J. Reuland & Alice G. B. ter Meulen, eds. MIT Press: Cambridge, Mass.

Prasada S., and Dillingham, E.M. (2006). Principled and statistical connections in common sense conception. *Cognition 1*, 73-112.

Stanley, J. and Szabó, Z.G. (2000). On Quantifier Domain Restriction. *Mind and Language* 15 (2-3), 219-61.

Sterken, R. (2015). Leslie on Generics. *Philosophical Studies*, 172(9), 2493-2512.

# Pragmatic Inference In Infancy

Olivier Mascaro,
*Department of Cognitive Science,*
*Central European University, Budapest*

Dan Sperber
*Department of Cognitive Science,*
*Central European University, Budapest*

## Abstract

In this paper, we argue for extending experimental pragmatics to the study of the inferential capacities that make, in the first place, language acquisition possible in infancy.

Previous research has established that infants are highly sensitive to ostensive cues, i.e. cues that convey an intention to communicate. This paper will present recent experimental data suggesting that being ostensively addressed does more than just attract infants' attention or intensify their processing of information. Rather, ostensive communication triggers a genuine process of pragmatic interpretation in infants and young children.

Second, we probe the strength and developmental changes of young children's priors about communicated information, with a focus on expectations of reliability. Altogether, experimental data suggest that young humans (i) interpret communicated information pragmatically (ii) rely on priors about the nature of communicated information.

We conclude by advocating the addition of an early developmental psychology dimension to the experimental study of pragmatics: both infancy research and pragmatics stand to gain from cross-fertilization.

 In: Fabienne Salfner and Uli Sauerland (eds.),
*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 95–102.

In this paper, we argue for extending experimental pragmatics to the study of the inferential capacities that make language acquisition and use possible in the first place, from infancy on. We look at the kind of expectation that infants have about communicated information, and how these expectations guide infants' interpretation of communicative behaviour - in short, we focus on pragmatic inferences. First, we will argue that infants have expectations about communicated information, and use these expectations to draw pragmatic inferences. These priors may be different in sophistication but not in kind from the expectations that, according to various pragmatic theories, guide the interpretation of linguistic information. Second, we will illustrate how the nature and development of pragmatic inference can be studied from infancy on, focusing here on expectation of reliability that attach to assertions.

**1 Prerequisites for pragmatic inference in infancy**

For long time, there were reasons to doubt that infants may engage in pragmatic inference. For example, Gricean and neo-Gricean approaches to pragmatics imply that hearers' interpretation process relies on one or several of the following ingredients: a sensitivity to speakers' rationality when pursuing their communicative goals, some capacity to represent mental states (such as beliefs) and some evaluation of communicated information (e.g. of its relevance, its accuracy, its informativeness). For a long time, it was not clear whether any of these prerequisites where within grasp of young children - let alone of infants. However, recent research results give reasons to reconsider this picture. These data provide evidence of the presence of the building blocks of pragmatic inference at a very young age.

First, a large number of studies indicate that infants can attribute beliefs and intentions to others. Studies of representations of physical actions have revealed that infants assume that agents are rational, and use these expectations to interpret others' goals and actions (Gergely & Csibra, 2003; Skerry, Carey & Spelke, 2013). Studies of early mindreading indicate that infants have some sensitivity to the content of beliefs, and some capacity to represent them (Kovács, Endress & Téglás, 2010; for reviews see Perner & Roessler, 2012; Baillargeon, Scott & Bian, 2016). Moreover, recent evidence indicate that infants recognize intentions to act on others' beliefs (Scott, Richman & Baillargeon, 2015). Thus, the basic mindreading building blocks of pragmatic inferences seem to be present from infancy on.

Second, studies of infants' and children's selective learning indicate that young humans may possess capacities to evaluate the value of communicated information. These studies indicate that young children are surprisingly discriminant learners, who prefer to learn from benevolent and competent informants (e.g. Koenig & Harris, 2005; Mascaro & Sperber, 2009). From infancy on, young humans are sensitive to the accuracy of cues (Begus & Southgate, 2012; Koenig & Echols, 2003; Thummelsthammer, Wu, Sobel & Kirkham, 2014), or to the amount of information that they can learn from a stimulus (e.g. Kidd, Piatandosi & Aslin, 2012; Stahl & Feigenson, 2015). All this research suggest that infants possess some incipient capacities to evaluate the quality of information.

In short, current evidence suggests that main components of pragmatic reasoning are in place from the first year of life on: a capacity to represent beliefs, a capacity to interpret others' actions, and a sensitivity to general properties of information, such as accuracy or informativeness.

All this makes it possible that young children - and perhaps infants - might form expectations about communicated information, and may use these expectations to interpret what speakers mean. Moreover, as we shall see, there is indeed evidence of some pragmatic inference during infancy.

## 2 Evidence for pragmatic reasoning in infancy

Several studies indicate that infants' interpretation of an action changes when it is accompanied by ostensive cues (such as raised eyebrows or infant directed speech). To take one example, in a study by K. Egyed, I. Király and G. Gergely (2013), 18-month-olds were familiarized with a demonstrator who expressed joy when looking at a unfamiliar object (positive emotion), and expressed disgust towards a second unfamiliar object (negative emotion). In the "communicative condition", the demonstrator addressed the infant ostensively (greeting the infant, calling her name, making eye contact). In the "non-communicative condition", the demonstrator displayed the same emotions, but acted as if she was alone, never looking at or talking to the infant before or after displaying emotions.

In the test, phase, a second experimenter (different from the demonstrator) asked the participant to give her one of the object. Infants were more likely to give the second experimenter the object towards which a positive emotion was expressed in the communicative condition, than in the non-communicative condition. This first result suggests that infants' interpretation of the emotion was modified by the presence of ostensive cues. When ostensive cues were present, infants assumed that the positive or negative attitudes expressed by the demonstrator could be generalized to other people (perhaps assuming that the demonstrator expressed that one object was 'good', or positive, and that the other object was 'bad' or negative) (see also Gergely, Egyed & Király, 2007). When ostensive cues were absent, however, infants were more likely to interpret the attitudes of the demonstrator as person-specific (perhaps assuming that the demonstrator liked one object, and disliked the other).

This interpretation was confirmed by a third condition, in which infants were familiarized with a non-ostensive demonstrator who emoted positive and negative emotion towards novel objects. In the test phase, the same experimenter asked participants to give her one object. This time, infants preferentially gave the experimenter that object towards which she emoted positively in the past. Therefore, even in the non-ostensive condition, infants learned something from the demonstration. Crucially, infants' interpretation of the scene (and thus what they learned from it) differed depending on whether ostensive or non-ostensive cues were present.

These results suggest that being ostensively addressed do more than just attract infants' attention or intensify their processing of information (see also Yoon, Johnson & Csibra, 2008). Rather, ostensive communication triggers a genuine process of pragmatic interpretation in infants and young children. The type of information that infants extract from ostensive communication - e.g. objects' valence (Egyed, Király & Gergely, 2013), functions (Hernik & Csibra, 2015), kinds (Futó et al., 2010), objects' typical location (Topál et al., 2008) or intentions (Martin, Onishi & Vouloumanos, 2014; Vouloumanos, Martin & Onishi, 2014; Vouloumanos, Onishi & Pogue, 2012)- is of conceptual nature, and can only be accessed through inferential processes.

The inferences triggered by ostensive cues are too diverse to be pre-coded in a semantic fashion. Rather, it appears that the recognition of a communicative intention triggers very general expectations about the nature of communicated information. So far, current evidence does not allow to specify exactly what is the exact nature of these expectations. Data may be explained by assuming that infants expect speakers to convey generalizable knowledge (Csibra & Gergely, 2009). Data are also compatible with the view that infants apply other kinds of priors, for example a Cooperative Principle, or expectations of relevance (Grice, 1975; Sperber & Wilson, 1986). Importantly regardless of the exact nature of infants' expectation, current evidence suggests that pragmatic inference is present from a very early age. These discovery call for the study of the nature and early ontogeny of humans' expectations about communicated information. In the next section, we aim at offering a proof of concept for such investigation, focusing on the expectation that what is asserted tends to be reliable.

## 3 Proof of concept: Charting the early development of expectations about communicated information

The assumption that audiences expect what is asserted to be reliable is a widely shared assumption in pragmatic theories. For example, the Gricean maxim of quality asks speakers to "try to make your contribution one that is true" (Grice, 1975). According to relevance theory, the expectation that drives pragmatic interpretation is one of relevance and not directly one of truthfulness. In the case of an assertion, however, the relevance of the information conveyed depends on its truthfulness (Sperber & Wilson, 1995; Wilson & Sperber, 2012). In short, while pragmatic theories differ on whether there is an expectation of literal truthfulness, they all take for granted that speakers' assertions should be interpreted as if they were reliable.

In collaboration with A. Kovács, one of us (O. Mascaro) probed the strength and developmental changes of this expectation of reliability. This investigation had two goals. First, it probed the emergence of the expectation that what is asserted is reliable, and evaluated how this expectation changes in communication during infancy. Second, it aimed at testing how infants' trust shapes their interpretation of communicated information. In principle, one way to expect communicated information to be reliable is simply to maximize agreement between one's interpretation of speakers' meanings and what one thinks (Davidson, 1984). This view predicts that young children should systematically reinterpret what others' communicate when it seems to conflict with what their own beliefs. Alternatively, young children may expect communication to convey novel and reliable information, and may therefore be more disposed to revise their own beliefs on the basis of what is communicated to them.

### 3.1 The development of infants' expectation of reliability

In a first study, participants had to find a reward hidden in one of two containers. To evaluate the strength of infants' expectation of reliability, we pitted communication against a pre-existing belief: Infants saw in which container the reward was located, but an informant pointed to indicate that the reward was in another container. Using this procedure, we established that infants have a strong disposition expect that what is asserted to be reliable. In Study 1, we found that 15- and 24-month-old infants were more likely to follow pointing than to trust their past perception,

even when the informant had a false belief about the location of the toy. Infants' reliance on communication was insensitive to their own certainty about the toy's location, and to the falsity of informants' belief. Moreover, reliance on communication increased with age. Fifteen-month-olds did show some sensitivity to the accuracy of pointing: They followed pointing more when it was the only source of information than when its meaning conflicted with their memory. Moreover, they decreased their trust in pointing after experiencing being misled by the informant. By contrast, 24-month-olds' pointing following remained at ceiling in all situations.

In two additional studies, we established that infants' disposition to follow pointing was genuine trust in communication, i.e. a disposition to treat communicated information as reliable. We found that infants' trust could not be reduced to mere desire to comply with the experimenter, or to an imperative interpretation of pointing. In Study 2, when the toy was hidden in transparent buckets, infants had no difficulty ignoring the misleading pointing. Moreover, infants' trust could not be reduced to an inability to reject pointing: In Study 3, infants were more likely not to accept the pointing when they possessed more evidence to pit against pointing. Altogether, these data suggest that infants expect pointing to be highly reliable.

### 3.2 The triggers of infants' expectation of reliability

In subsequent studies, we investigated the sources of infants' expectation of reliability. In principle, infants' expectation of reliability may have been acquired through repeated experience that a specific communicative action (pointing) was reliable in the past. Alternatively, infants' expectation of reliability may be attached to communication in general, and be triggered by the recognition of an intention to communicate, even when this intention is fulfilled by means of novel communicative signals.

To judge between these two hypotheses, we tested infants in a set up comparable to one of the previous studies: participants had to discover the location of a toy hidden under one of two buckets. However, instead of using a familiar communicative means, the informant indicated the location of the toy with a novel communicative action (by placing a marker placed on top of the baited bucket). In the crucial test situation, participants saw under which bucket the toy was located, and the bucket was placed on top of the other bucket. In this test situation, participants followed the marker rather than their past perception as soon as they were able to interpret the marker. Moreover, this strong trust increased during the second year of life (Study 4). It was found even when infants had a limited exposure to the novel communicative cue (Study 5). By contrast, infants' trust disappeared when the cue was presented in a non-ostensive manner. In that case, toddlers preferred to follow their past perception rather than the marker (Study 6). Altogether, these results suggest that infants' expectation of reliability may be triggered by the presence of ostensive signals, and may extend to novel forms of communication.

These studies show that infants go beyond Davidsonian charity (maximizing agreement between their own belief and what the speaker means, cf. Davidson, 1984). Rather, they assume that communication convey reliable and novel information. From the age of 15 months, infants recognize central features of assertions, in particular their "assertoric force" (Frege, 1918; Geach, 1965), i.e. acknowledge that assertions do not just convey proposition, but are an invitation to

believe those propositions. Second, the fact that infants accepted communicated information even when it conflicted with their past perception indicates that infants attach a presumption of reliability to what is asserted (Williams, 1966; Dummett, 1981). This presumption of reliability was not derived from what infants assumed about the knowledge of the communicator, since infants trusted even informants who had a false belief about the location of the toy. Rather, it appeared that the communicative action itself triggered a disposition to accept what was communicated. Third, we observe that infants' expectation of reliability increases during the second year of life, therefore suggesting that the priors supporting pragmatic inference may change during early ontogeny.

### 3.3 How expectation of reliability guide children's interpretation

In an additional series of study, we tested how a strong expectation of reliability might contribute to shape young children's interpretation of novel communicative cues, at a later developmental age (Mascaro & Sperber, in preparation). When an informant's communicative behavior has misled you, you might either withdraw your trust in the informant or *revise your interpretation* of the communicator's behavior. We designed a study that compares these two hypotheses. In this experiment, adapted from Couillard and Woodward (1999), preschoolers had to find a reward hidden in one of two boxes. A "mean" puppet informant repeatedly indicated the empty location, by placing a marker on the empty box. Over the course of 10 trials, children quickly learned to avoid the box on which the unfamiliar marker was placed and indicated the other one (a conceptual replication of Couillard & Woodward, 1999; Jaswal et al., 2010). Crucially, at the end of the experiment, children participated in the same game. However, in this case, the informant was presented as a "nice" informant. In this last phase of the experiment, participants still avoided the box on which the marker was placed.

These results are inconsistent with the hypothesis that training had caused children to distrust the "mean" informant, and consistent with the hypothesis that it had caused them to reinterpret the unfamiliar marker as an honest signal for the location of the empty box. When signals have no predetermined interpretation, young preschoolers can without difficulties reinterpret them so as to fit their expectation that communication is honest and helpful. These results do not, on the other hand, support the hypothesis that children's gullibility is based on trust in familiar signals such as speech or pointing (Jaswal et al., 2010). This result confirms that children's trust is elicited by attributing a communicative intention to an agent rather than by recognizing a familiar communicative device with an established meaning.

## 4 Conclusion

Altogether these data suggest that the addition of an early developmental psychology dimension to the empirical study of pragmatics is both feasible, and important. Humans appear to have general expectations about the nature of communicated information from infancy on, and use these expectations to interpret what others mean. These expectation also appear to undergo some changes during early infancy. Therefore, the experimental study of infants' communicative behaviour may provide crucial information about the building blocks of pragmatic inference, and their early emergence. Both infancy research and experimental pragmatics stand to gain from cross-fertilization.

**References**


Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology, 67*. retrieved from http://internal.psychology.illinois.edu/infantlab/articles/baillargeon_scott_bian_2016.pdf

Begus, K., & Southgate, V. (2012). Infant pointing serves an interrogative function. *Developmental Science*, *15*(5), 611-617.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148-153.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. New York: Oxford University Press.

Davidson, D. (1984). Truth and interpretation. *Clarendon: New York*.

Dummett, M. (1981). *Frege: Philosophy of language* (2nd ed.), Cambridge, Mass.: Harvard University Press.

Egyed, K., Király, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological science*, *24*(7), 1348-1353.

Frege, G. (1918/1965). The Thought: A logical inquiry. *Mind, 65*, 289-311.

Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative function demonstration induces kind-based artifact representation in preverbal infants. *Cognition*, *117*(1), 1-8.

Geach, P. T. (1965). Assertion. *The Philosophical Review*, *74*(4), 449-465.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287-292.

Gergely, G., Egyed, K., & Király, I. (2007). On pedagogy. *Developmental science*, *10*(1), 139-146.

Grice, H. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41-58.

Hernik, M., & Csibra, G. (2015). Infants learn enduring functions of novel tools from action demonstrations. *Journal of experimental child psychology*, *130*, 176-192.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399-e36399.

Koenig, M. A., & Echols, C. H. (2003). Infants' understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition*, *87*(3), 179-208.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development, 76*, 1261–1277.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830-1834.

Martin, A., Onishi, K. H., & Vouloumanos, A. (2012). Understanding the abstract role of speech in communication at 12 months. *Cognition*, *123*(1), 50-60.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance toward deception. *Cognition, 112*, 367–380.

Mascaro, O., & Sperber, D. (in prep.). Dumb gullibility or smart truting? Young children can reinterpret misleading signals as honest.

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in cognitive sciences*, *16*(10), 519-525.

Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive psychology*, *82*, 32-56.

Skerry, A. E., Carey, S. E., & Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proceedings of the National Academy of Sciences*, *110*(46), 18728-18733.

Sperber D. & Wilson D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Sperber, D. & Wilson, D. (1995). Postface to the second edition of Relevance: Communication and cognition. p. 255-279.

Topál, J., Gergely, G., Miklósi, Á., Erdőhegyi, Á., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, *321*(5897), 1831-1834.

Tummeltshammer, K. S., Wu, R., Sobel, D. M., & Kirkham, N. Z. (2014). Infants track the reliability of potential informants. *Psychological science*, *25*(9), 1730-1738.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science, 348*(6230), 91-94.

Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6−month−olds understand that speech can communicate?. *Developmental science*, *17*(6), 872-879.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*(32), 12933-12937.

Williams, Bernard, 1966, "Consistency and realism", *Proceedings of the Aristotelian Society* (Supplementary Volume), 40: 1-22. Reprinted in Williams 1973: 187–206.

Wilson D. & Sperber, D. (2012) Meaning and Relevance. Cambridge Cambridge UP.

Yoon, J. M., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences*, *105*(36), 13690-13695.

# On investigating intention in experimental pragmatics

Ira Noveck
*L2C2-Lyon*

Grice's distinction between *sentence meaning* (the properties of a sentence assigned to it by the grammar) and *speaker's meaning* (what the speaker intended to communicate by uttering a sentence) is at the core of experimental pragmatics. Critically, he viewed the retrieval of sentence meaning from an utterance as a matter of decoding (discovering the semantic properties that the grammar pairs to its acoustic form) and he viewed the retrieval of the speaker's meaning as a special kind of intention-reading that provides for "non-natural meaning." While certain features of Grice's (1989) program have been challenged (e.g. through dubious take-downs of the so-called Standard Pragmatic Model or through the proposal of grammatical operators), these are secondary to the fact that his general approach has remained seminal.

Grice is also the basis of a different sort of divide in Experimental Pragmatics. There are those who have focused on, or exploited, the maxims in order to create linguistic rules in the code and there are others who have focused on the role of intentions in forming *non-natural* meaning. One can view nearly every debate in experimental pragmatics as an opposition between those who incorporate maxims to some degree versus those who deny that this is central and who, instead, attribute a major role of utterance-interpretation to intention-reading.

In retrospect, intention-reading has not received nearly as much attention in experimental pragmatics as, say, the hypothesized step-by-step accounts of scalar inferences. So, I will begin by describing how Grice envisioned the communication of non-natural meaning by way of example. From there, I will describe three areas of pragmatic research where intention-reading is (or would be) crucial for advancing experimental pragmatics.

## 1 Ostensive-inferential communication

Let's consider a scenario, one in which I would like to tell my colleague, whose office is facing mine and whose door is wide open, that I would like him at this moment to keep his voice down while on the phone. This will require several steps. A first step consists of *the speaker intending to change something in the addressee's mental state*. This much is easy. I am the speaker in this scenario and I have content that I would like to communicate to my neighbor (*keep your voice down please*). A second step consists of *the addressee recognizing that the speaker has an intention to change his mental states*. In this exchange, this means that my colleague needs to recognize that I have the intention to let him know that


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 103–109.

I would like him to keep his voice down. Now, this is the tricky part; how can I go about that? One option would be to close my door (even forcefully), with the intention that he will detect my message, the one that signals that I am disturbed by his loud talking. However, even if his conversation suddenly ended or quieted down after I closed my door, we would be reluctant to say that I communicated to my colleague that I wanted him to know that I wanted him to keep his voice down. According to Grice, a critical third step would be necessary to guarantee communication: *the speaker showing the addressee that he intended to get the message across*. In my office scenario, it could be an ostensive signal, such as a conspicuous shrug-and-grimace directed at my colleague that accompanies my forceful door-closing or it could be something even more direct, such as "you are too loud; please keep your voice down." Without this third step, information transmission does not count as proper communication. It is this step that makes any Gricean proposal *ostensive-inferential* because the speaker is demonstrating to the addressee that there is an intention to gather from the speaker. As Scott-Phillips (2014) writes, this third step is crucial:

> ..it is the heart of Grice's account of meaning: to mean something, I should intend that my audience believes it, *and they should believe it at least in part because they recognize that this was my very intention*. This is the meaning of (Gricean, non-natural) meaning. It is the reason why ostensive-inferential communication can be glossed as *intentionally overt* communication.

Importantly, Grice showed that linguistic communication, while partly code-based, cannot be reduced to a mere encoding-decoding process; it involves the attribution of mental states to the speaker too. The centrality of the inferential process to understanding utterances is the critical insight that distinguishes modern accounts of pragmatics from its predecessors and distinguishes pragmat*ists* from nearly every other professional linguist. Along the way, we see how signaling becomes integral to successful communication.

Grice's account is also the inspiration for Sperber & Wilson's (1986/1995) Relevance Theory, which reduces Grice's three-part intention to two: a *communicative* intention (indicating *that* the speaker wants to communicate) and an *informative* intention (indicating *what* the speaker wants to communicate). By speaking, both of these are considered to be made ostensive and the addressee's reaction is to determine, by inference, what the informative intention is. Relevance Theory also provides general procedures that describe how effort and produced effects are critical to determining how a listener works out the utterance's intended meaning.

When it comes to experimental pragmatics, work is concerned with the participant-addressee and the way that addressee processes an utterance (an out-of-the-blue sentence or a line in an exchange or text). Typically, the addressee is required to determine what the speaker had in mind and, depending on the task, decide whether the speaker agrees with it or needs to adjust his understanding of the speaker's intention. In a classic scalar task, for example, part of a participant's job is to determine what the speaker's intended meaning is and, as part of that process, come up with a true or false judgment. In other words, addressees or participants evaluate the speaker's utterance as a way to ascertain what they think is the speaker's intention. Along the way, participants are pragmatically enriching what is encoded in the utterance.

The point is that the hypothesizing about the speaker's intention is crucial to participants, even if it seems opaque (in an experimental setting). However, considering how participants go about that is central to better understanding pragmatic phenomena. Here, I turn to three areas in experimental pragmatics that a) show *how* intention-reading can affect processing, b) reveal *when* intention reading most-likely affects processing in experimental settings, and that; c) reveal *what* is developing as more mature pragmatic processing becomes evident.

## 2 Reference

If one assumes that pragmatic enrichments are dependent on forming common ground with a speaker, it follows that even a basic pragmatic inference should be affected by a small feature of the speaker-addressee interaction (as, say, whether or not a listener considers himself allied with an anonymous speaker). On the other hand, if the coded meaning of the utterance were sufficient to communicate, it should not matter who gives an instruction.

Tiffany Morisseau and I recently completed a set of studies in which participants needed to follow some basic instructions about what to click on a screen (e.g. *click on the flower*) as they search among four objects. The critical case was one where the request, such as *click on the wet dog* potentially prompted a contrastive inference (a search for a non-wet dog). Now, some people consider it non-controversial (Sedivy, 2003) to assume that the instruction encourages participants to determine whether there is another dog, but based on developmental evidence I am not entirely sure the contrastive inference arises consistently even among adults (Kronmuller et al., 2014).

In half of the critical conditions, a dry dog was visible among the remaining three possibilities. In the other half, the dry dog was present but less visible (behind a grayish, filtering square) so that it called for a more invested search. The question we asked concerned the role of the messenger, the speaker. While instructions from two speakers were identical and always felicitous, in half the con-

ditions the search request came from the speaker's (political) ingroup and half came from the speaker's outgroup.

With the expectation that a contrast item is going to be sought out, at least by some participants, the question is to what degree. So, one question going in was, is the contrast object consistently and automatically sought out (as measured by looks at the hidden object)? Another question was, does group membership affect the search for the contrast object? Results show that the hidden item was more likely to be tracked when it was expected to contain a contrast object as opposed to nothing relevant and that this effect was enhanced in the Outgroup condition. Thus, the social affiliation between a speaker and a listener influences the extent to which pragmatic information is readily accepted.

## 3 Distinguishing spontaneous pragmatic interpretations from repeated uses

One of the consistent findings from the experimental pragmatics literature is that interpretations that call on pragmatic enrichments are effortful compared to readings without those enrichments. The support for this claim is quite robust (starting with Noveck & Posada, 2003; Bott & Noveck, 2004; Breheny et al., 2007, Huang & Snedeker, 2006, DeNeys & Schaeken, 2007 and so on). Yet, there have been a couple of studies showing that one can get participants to produce an enriched interpretation and with no apparent extra effort (e.g. Grodner et al., 2010). In that work, Grodner et al. went to some methodological extremes to facilitate the enriched interpretation of *some* (by transforming the presentation of the quantifier, by limiting the set of tested terms to three that highlight contrasts, or just by having many items). The question that this interesting work raises, then, is why the exception?

I propose that the answer lies in the intelligence of the participants and the timing over the course of an experiment. Participants are strategic players who adopt an approach to a task and part of their effort is to determine the intention of the speaker/experimenter. When there are clearly delineated options, a participant can determine the paradigm's (or experimenter's) preferred interpretation. Moreover, it could take just a couple of early trials to figure it out.

Figuring out the paradigm's preference is not free. While it might take longer to carry out an initial "pragmatic" reading compared to control items, participants can keep the same enriched interpretation over the course of an experiment without being forced to re-start the inferential process each time. This assumes, and I think correctly, that participants value consistency too, which is why participants tend to divide into two types of responders on categorization tasks (e.g. logical responders and pragmatic responders) or why seven-year-old children can be trained to answer with enrichments one week and then revert to child-like

behavior a week later without training (Guasti et al., 2007). Point is that adopting a pragmatic interpretation does not mean that "pragmatic" participants need to remain slow throughout a task. A participant does not need to start from scratch with each new experimental trial (as if it were a nonce case).

A participant's adoption of an interpretation for the sake of an experiment is reminiscent of Clark's *conceptual pacts*, in which conversational partners create a novel designation in an exchange and maintain it. There is thus a distinction to be made between one-off interpretations (that are comparable to the first encounter in an experimental task) and the maintenance of these interpretations for the remainder of a "conversation" (the experimental session). This distinction motivates the introduction of early-late analyses in our processing studies (see Spotorno & Noveck, 2014, on irony; Grodner, personal communication). It also explains why Tiffany Morisseau and I recently reanalyzed Bott and Noveck, 2004 (Experiment 3); we found that pragmatic effects attenuated over the three blocks of the experiment.

Once one considers the role of short-term conventions within experimental sessions, the question can then be extended to encounters that go beyond experimental chambers. How does a conversational convention of one exchange end up being a permanent part of a given language (and here I mean English, German, French etc)? That is, what role does pragmatics play in grammaticalization, a process in which historical changes observed in languages are unidirectional? There appears to be an overlap between historical linguistics and experimental pragmatics that deserves our attention.

A historical linguist's question is, how do individuals manage to build conventions that render communication expressively powerful. An experimental pragmatist's question is, how are we able to say and understand an endless number of new utterances. For example, both historical linguists and experimental pragmatists are concerned with metaphors. The difference between the two is that historical linguists are interested in the way certain metaphors make their way permanently into language, while we are interested in the way an original metaphor is processed for the first time (or at best over the course of a few trials, see Nieuwland and van Berkum, 2006). The potential for this intersecting line of research is enormous and, I should add, historical linguists are keen in pursuing this (see Grossman & Noveck, 2015). Some work along these joint lines has begun to emerge since a workshop (Historical Linguistics meets Experimental Pragmatics) took place in May 2014 (see Clark, 2015).

## 4 What brain networks account for pragmatic development?

In the case of deductive inference, it has been shown that by the time children are eight years of age they are able to carry out fundamental logical inferences with

relative ease but that they are still not adult-like. In one classic study (often cited in the scalar literature), Paris (1975) showed that 8-year-olds are generally competent at evaluating disjunctive, conjunctive, and conditional statements in light of provided evidence, but he also showed that there is some room for improvement as children get older. For example, when shown two pictures -- one of a monkey and another of a boy -- along with the sentence *There is a monkey or there is a ball*, only 50% of 8-year-olds say true while 90% of 17-year-olds do. One could make the case that children's increasingly adult-like performance is due generally to improved inferential abilities. When it comes to pragmatics, research has shown that children are more likely to make pragmatically valid inferences as they get older, too (Noveck, 2001, Pouscoulous et al., Papafragou & Musolino, 2002; Huang & Snedeker, 2009). This provides a different picture of growing inferential abilities among children because it appears to show that an initial logical competence increasingly gives way to pragmatic factors. But it leaves open the question as to what is developing exactly when participants appear increasingly "pragmatic" with age.

Over the last few years, my colleagues and I have carried out several neuroimaging studies to investigate pragmatic processing. In one, we determined that classic Theory of Mind regions are implicated in irony (Spotorno et al., 2012). Another study, Prado et al. (2015) introduced a reading task that presents an utterance (for example, "Amy realizes: 'That's my mom who's entering!'") that can be either a logical conclusion (because the protagonist had considered that it was either her father, sister or mother and then eliminated the first two from consideration) or an underdetermined statement that calls on pragmatic inference (because she had considered the three but had eliminated only one possibility). We called the first type *Fully-Deductive* and the second *Implicit-Premise.*

It is through a developmental version of the latter study that we realized we can pin down what develops exactly. We reasoned that if the increased ability to make pragmatic inferences in natural discourse is supported by the development of the mindreading brain network, activity in brain regions known to support mindreading such as the Temporal-Parietal Junction (TPJ), PreCuneus (PC) and Medial Prefrontal Cortex (mPFC) ought to increase with age in *Implicit-Premise* stories (compared to *Fully-Deductive* stories). Alternatively, if the increased ability to make pragmatic inferences in natural discourse is supported by the development of logical-inference mechanisms, it is the activity in the fronto-parietal regions associated with logical inference-making (Prado et al., 2015) that should increase with age in these comparisons. This ongoing research appears to show that age-related changes of activity are associated with the logical inference-making network, while there were no age-related changes in the mindreading network.

## 5 Conclusions

To summarize, understanding a speaker's intention is central to pragmatic interpretation and much can be learned by considering how an intention interacts with pragmatic processes. I do not know whether the study of intentions in utterance processing is a trend in experimental pragmatics (I'm not sure what constitutes a trend). I do know that one cannot do pragmatics without recognizing it.

## 6 References


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.

Breheny, R., Katsos, N., Williams, J. (2006). 'Are scalar implicatures generated by default?', *Cognition, 100*, 434-463.

Brennan S.E., & Clark H.H. (1996). 'Conceptual pacts and lexical choice in conversation', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482-1493.

Clark, Billy (2015) Relevance theory and language change. *Lingua* .

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

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

Grodner, D., 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*, 42-55.

Grossman, E. & Noveck, I. A. (2015). What can historical linguistics and experimental pragmatics offer each other? *Linguistic Vanguard*.

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). 'Why Children and Adults Sometimes (But Not Always) Compute Implicatures', *Language and Cognitive Processes*, *20*, 667-696.

Huang, Y., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension, *Developmental Psychology*, *45*(6), 1723-1739.

Krönmuller, E., Morisseau, T. & Noveck, I. A. (2014). Show me the pragmatic contribution: A developmental investigation of referential communication. *Journal of Child Language*.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*,1098–1111.

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

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

Papafragou, A., & Musolino, J. (2003). 'Scalar implicatures: experiments at the semantics-pragmatics interface', Cognition, 86(3), 253-282.

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

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

Prado, J., Spotorno, N., Koun, E., Hewitt, E., Van Der Henst, J.B., Sperber, D., & Noveck, I.A. (2015). Neural interaction between logical reasoning and pragmatic processing in narrative discourse. *Journal of Cognitive Neuroscience*. 27, 692-704.

Scott-Phillips, T. (2014). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan.

Sedivy, J. C. (2003). 'Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations', *J. of Psycholinguistic research, 32*, 3-23.

Sperber, D. & Wilson, D. (1985/1996). Relevance. Oxford: Basil Blackwell.

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J.-B., & Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, *63*(1), 25–39.

# You surely know what I mean.
# Theory of Mind and Non-Literal Language Comprehension

Francesca Panzeri & Francesca Foppolo
*University of Milan-Bicocca*

## 1 Introduction

A full understanding of language requires not only the recognition of the literal meaning of sentences, but also of the communicative intention of the speaker. In some cases, the latter clashes with the former, as in metaphors and irony (e.g. uttering "Your room is a battlefield" or "Your room is extremely clean" to comment on a very messy room).

Several authors linked the comprehension of non-literal language to Theory of Mind (ToM) abilities. Sullivan et al. (2005) argue that only typically developing (TD) children who pass 2nd order ToM tasks can distinguish jokes (and irony) from lies; and Happé (1993) tested children with autism spectrum disorders (ASD) and TD 5 year-olds with different levels of ToM, and claimed that 1st order ToM is sufficient for metaphor understanding, whereas irony comprehension calls for 2nd order ToM. Nevertheless, the link between non-literal language comprehension and ToM abilities has been questioned. In the first place, even if TD 5 year-olds understand metaphors grounded on physical- or action-resemblance (Keil, 1986; Vosniadou et al., 1984; Winner et al., 1980), a full understanding of metaphors is achieved only after age 11 (Billow, 1975), and there is a clear developmental trend (Cometa & Eson, 1978; Gentner, 1988), whereas 1st order ToM is reached at 4 years of age. Moreover, Norbury (2005) tested children with communication impairments (ASD and/or language impaired) and found that semantic abilities were a better predictor of metaphor comprehension than ToM abilities; and Szücs (2013) found that, in TD pre-schoolers, language, but not ToM, abilities could predict metaphor understanding, whereas irony comprehension was influenced by chronological age. And, in general, it is well known that the performance on ToM tasks is highly dependent on language abilities (Happé, 1995 and Astington & Jenkins, 1999).

## 2 The study

To further investigate the cognitive abilities that are involved in non-literal understanding and disentangle the factors at play in different kinds of non-literal language, we compared the understanding of metaphors and irony in a typical and atypical population, i.e. children with conventional hearing aids, whose linguistic and ToM abilities are delayed with respect to their TD peers (Woolfe et al., 2002; Peterson & Siegal 1999; Peterson, 2004).

*Participants*

We tested 22 Italian deaf children (12 female and 10 male) aged 8 to 11 (8;1-11;8; MA 9;7) early diagnosed with a hearing loss (41 to 70 db), with conventional hearing aids, and an exclusively oral education (HA-group). Their scores on language tests (PPVT for the receptive lexicon, and the TCGB for grammatical comprehension) matched 6 year-old TD children. A group of 24 TD children attending 1st grade at


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 110–114.

primary school (6 years of age) served as controls.

*Materials and Procedure*

Children were administered two ToM tasks: the Smarties Test, which tests 1st order ToM abilities, and Laura & Gino test, a test for 2nd order ToM abilities, adapted for children with hearing difficulties, that contains a question for 1st order ToM and a question for 2nd order ToM.

We also administered two novel tests for metaphor and for irony comprehension. The test for Metaphor was modelled after Norbury (2005) and required the completion of a total of 15 sentences: 5 Metaphors, 5 Similes and 5 Literal sentences. An example for the metaphor condition is given below. Please note that besides the target ("an earthquake"), there was always a competitor ("a waiter"), i.e., a term that could in principle be predicated of the subject, but that in the given context was irrelevant, and two distractors ("a bicycle" and "a Thursday"):

(MET) Carla leaves a mess wherever she goes. (She) is really …

| | | | |
|---|---|---|---|
| an earthquake | a waiter | a bicycle | a Thursday |

The test for Irony comprehension comprised a total of 8 short stories, followed by a remark, that was ironical in 4 stories, and literal in the other 4. Four stories presented a negative context, and the other 4 stories had a positive context. The interaction of the irony-literal and of the negative-positive context resulted in 4 conditions, schematized below:

| | **Ironic remark** | **Literal remark** |
|---|---|---|
| **Negative context** | Irony_NegContext_PosRemark<br>Tag: IrNP | Literal_NegContext_NegRemark<br>Tag: LitN |
| **Positive context** | Irony_PosiContext_NegRemark<br>Tag: IrPN | Literal_PosContext_PosRemark<br>Tag: LitP |

**Table 1**: Conditions in the test for Irony comprehension

Examples of the two types of Ironic remarks are given below.

(IrNP) Chiara is helping her mother in making a cake. Mum asks her to stir the ingredients, but Chiara let the bowl fall, and the dough ends up on the table and on the floor.
Then mum says to Chiara: **You really did a great job!**

(IrPN) Daniela tells Lucia to put in the new bookshelves all the books, more than a thousand. At the end of the day, Daniela passes by, and she sees that Lucia finished with all the books.
Then Daniela says to Lucia: **You did nothing at all!**

*Results*

As for the ToM tasks, the HA-children' scores were not significantly different from the TD children (p=.963, n.s). Since in both groups some children passed only the Smarties task, but not the 1st order question in Laura & Gino task (whereas all the children who passed 1st order question in Laura & Gino also passed the Smarties task, and all the children who reached 2nd order question in Laura & Gino correctly answered all the other questions), we decided to consider two groups of 1st order ToM. We plot children's distribution in Fig. 1.

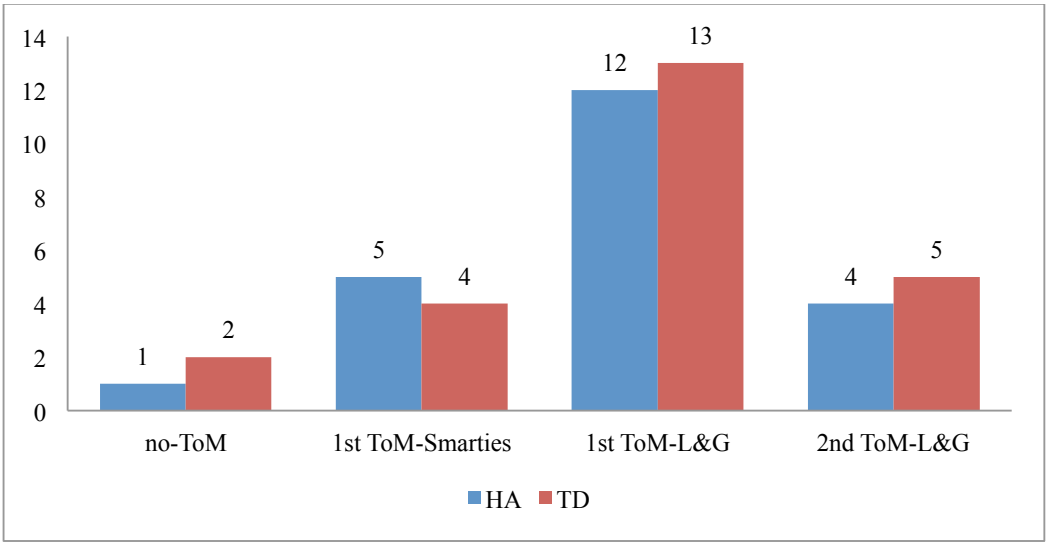



**Figure 1.** Distribution of children (HA-group=blue bars; TD-group=red bars) with respect to their performance in the ToM tasks: no-ToM= children who did not pass any ToM tasks; 1st ToM-Smarties= children who passed Smarties task; 1st ToM-L&G children who passed the 1st order question in Laura & Gino task; 2nd ToM-L&G= children who passed 2nd order question in Laura & Gino task.

As for the Metaphor task, the results for the Metaphor condition are plotted in Fig. 2.

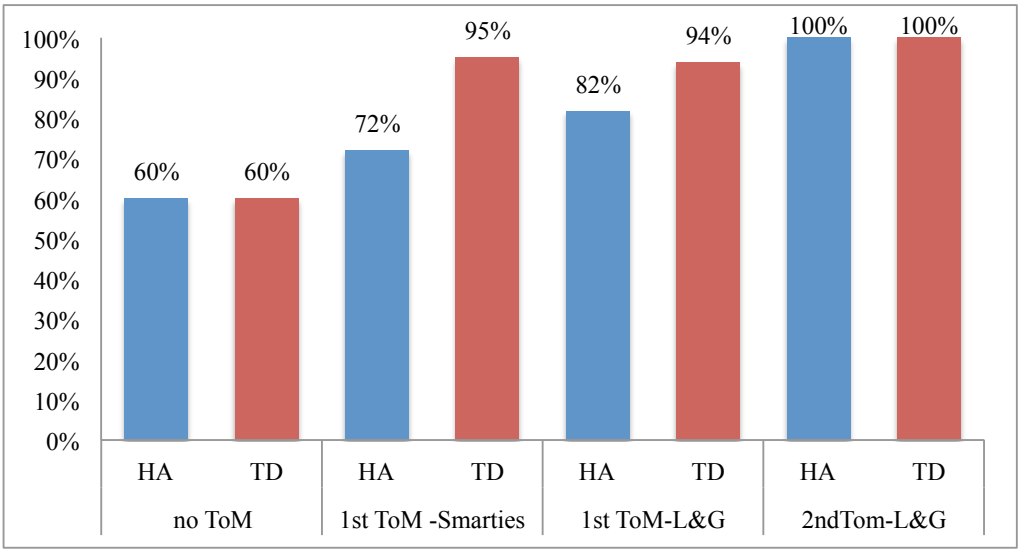



**Fig. 2.** Accuracy in the Metaphor condition, by group and level of ToM, as defined in Fig. 1.

A one-way anova revealed a difference between the HA- and the TD-group (p=.013). In the TD-group, only the two children who did not pass any level of Tom (the no ToM group) differ from the other groups. In the HA-group, since there was only one child who failed all ToM tasks, he was excluded from further analyses. The 2nd ToM group (100% accuracy) differs from 1st ToM-Smarties (72%, p=.006) and from 1st ToM-L&G (82%, p=.042).

As for the Irony task, the accuracy in the Literal condition was at ceiling for all groups. We considered the two Irony conditions (IrNP vs IrPN) and plotted them

separately in Fig. 3.

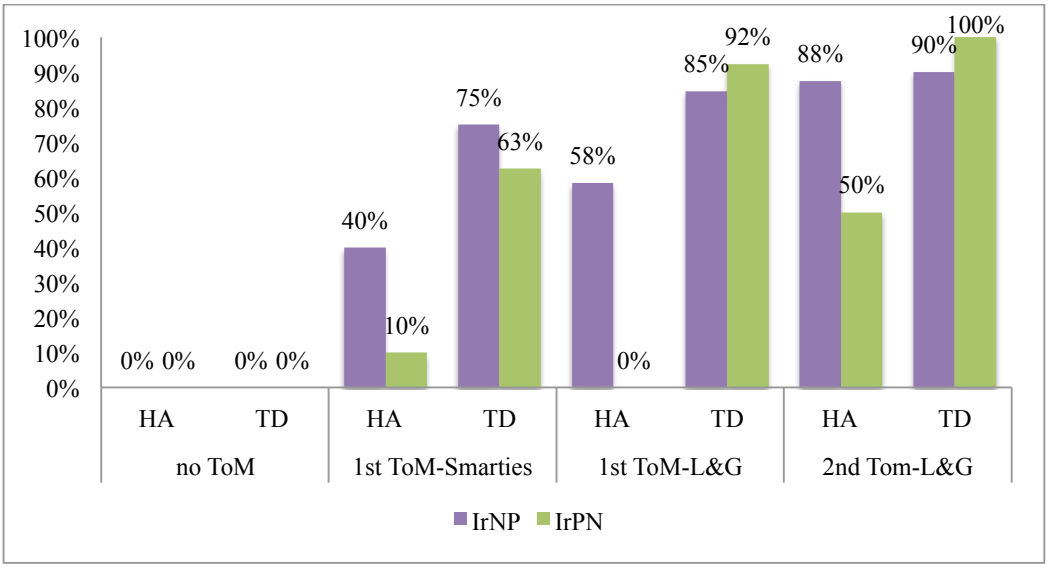



**Fig. 3.** Accuracy in the Irony task, for Irony in a negative context with a positive remark (IrNP, purple bars), and for Irony in a positive context with a negative remark (IrPN, green bars), by group and level of ToM.

In the TD group, the no-ToM children differ from the other groups in both conditions. In the IrNP condition, there is no difference in the other groups. In the IrPN condition, the 1st ToM-Smarties group differ from 1st ToM-L&G group (63% vs 92%, p=.44) and from 2nd ToM group (63% vs 100%, p=.027). In the HA-group, the child with no levels of ToM was excluded from analysis. In the IrNP condition, there is only a difference from 1st ToM-Smarties and 2nd ToM-L&G (40% vs 88%, p =.051). In the IrPN condition, the HA-group accuracy is extremely low (0% in the no-ToM and 1st ToM-L&G; 10% in the 1st ToM-Smarties, and 50% in the 2nd ToM-L&G), and 1st ToM-Smarties differs from the other groups (p=0.22 vs 1st ToM-L&G; p=.001 vs 2nd ToM-L&G).

*Discussion*

The results show that HA-children experience serious problems in the comprehension of non-literal language, and their difficulties seem to be more severe than their TD peers with analogous levels of ToM.

In the Metaphor task, as in Happé (1993), for TD children 1st order ToM abilities are sufficient for metaphor comprehension. But this is not the case for HA children: the accuracy on Metaphors is tightly linked to ToM abilities (1st order ToM differs from 2nd order ToM). At ceiling performance is only reached by the 2nd ToM group.

In the Irony task, despite their at ceiling performance on Literal controls, the No-ToM children (1 HA and 2 TD) fail to recognize all 4 ironical remarks (and interpret them literally). Ironical negative remarks in positive contexts (IrPN) are extremely hard for HA children (but not for TD). For ironical positive remarks in negative contexts (IrNP, the only ones tested by Happé), we did not find differences in the TD ToM groups (accuracy >75% from 1st ToM-Smarties). In the HA ToM groups there is a continuum in the accuracy scores : 40% (1st ToM-Smarties)-58% (1st ToM-L&G)-88% (2nd ToM-L&G).

These results suggest (contra Happé, 1993) that passing 1st order ToM is not sufficient for Metaphor understanding in the HA group; and passing 2nd order ToM is

not necessary for Irony understanding in the TD group, while in the HA group 2nd order ToM is not sufficient for a full understanding of Irony (69% overall accuracy in both conditions). Thus, we did not find a clear relation between ToM abilities and metaphor and irony understanding.

We are currently exploring the hypothesis that linguistic competence might constitute a better predictor for figurative language comprehension (as Norbury 2005 and Szücs 2013 claimed for metaphor understanding), testing a novel group of younger children, matched one-by-one to the HA group for linguistic age.

**References**

Astington, Janet Wilde & Jennifer M. Jenkins. 1999. A longitudinal study of the relation between language and theory-of-mind development. *Developmental psychology* 35(5). 1311-1320.

Billow, Richard M. 1975. A cognitive developmental study of metaphor comprehension. *Developmental psychology* 11.4. 415-423.

Cometa, Michael S., & Morris E. Eson. 1978. Logical operations and metaphor interpretation: A Piagetian model. *Child Development* 49(3). 649-659.

Gentner, Dedre. 1988. Metaphor as structure mapping: The relational shift. *Child development* 59(1). 47-59.

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

Happé, Francesca GE. 1995. The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child development* 66(3). 843-855.

Keil, Frank C. 1986. Conceptual domains and the acquisition of metaphor. *Cognitive Development* 1(1). 73-96.

Norbury, Courtenay Frazier. 2005. The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology* 23(3). 383-399.

Peterson, Candida C. 2004. Theory‐of‐mind development in oral deaf children with cochlear implants or conventional hearing aids. *Journal of child psychology and psychiatry* 45(6). 1096-1106.

Peterson, Candida C., & Michael Siegal. 1999. Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science* 10(2). 126-129.

Sullivan, Kate, Ellen Winner, & Natalie Hopfield 1995. How children tell a lie from a joke: The role of second-order mental state attributions. *British Journal of Developmental Psychology* 13(2). 191-204.

Szücs, Marta 2013. The role of Theory of Mind, age, and reception of grammar in metaphor and irony comprehension of preschool children, in Surányi, B. and Turi, G. (eds.) *Proceedings of the Third Central European Conference in Linguistics for Postgraduate Students*, Pázmány Péter Catholic University, Budapest.

Vosniadou, Stella, Andrew Ortony, Ralph E. Reynolds & Paul T. Wilson 1984. Sources of difficulty in the young child's understanding of metaphorical language. *Child Development* 55(4). 1588-1606.

Winner, Ellen, Matthew Engel, & Howard Gardner 1980. Misunderstanding metaphor: What's the problem? *Journal of Experimental Child Psychology* 30(1). 22-32.

Woolfe, Tyron, Stephen C. Want, & Michael Siegal 2002. Signposts to development: Theory of mind in deaf children. *Child development* 73(3). 768-778.

# What exactly do you mean? ERP evidence on the impact of explicit cueing on language comprehension

**Stefanie Regel (stefanie.regel@hu-berlin.de)**
Humboldt University Berlin, Department of Neurocognitive Psychology,
Rudower Chaussee 18, Berlin, 12489 Germany

**Thomas C. Gunter (gunter@cbs.mpg.de)**
Max Planck Institute for Human Cognitive and Brain Sciences, Department of Neuropsychology,
Stephanstr. 1a, Leipzig, 04103 Germany

**Abstract**

In written communication, emotion icons and pragmatic cues express speakers' emotions and attitudes. Such cueing of meanings becomes even more important for implied interpretations (e.g., irony) that cannot be derived from verbal information, but need to be inferred from contextual information and pragmatic knowledge. The sentence '*That's fantastic*' achieves either a literal or an ironic meaning depending on the context of this utterance. In two experiments using event-related brain potentials (ERPs), we examined the effects of explicit cueing (i.e., by quotation marks) on the processing of irony in comparison to literal language. In Experiment 1, cueing was ambiguous by presenting both literal and ironic sentences either with quotations (i.e., cued), or without quotations (i.e., uncued). ERPs revealed a P200 and P600 for irony relative to literal language. An impact of cueing on irony comprehension was not seen. In Experiment 2, the processing of cued and uncued irony was investigated separately in two experimental blocks. While for uncued irony a P200-P600 pattern was replicated, for cued irony an early sustained positivity emerged. The findings suggest that social cueing affects the comprehension of irony during initial and later stages of processing, but seems to depend on the cues' reliability.

**Keywords:** ERPs; P600; N400; language comprehension; pragmatics; figurative language; irony; social cues

## Introduction

In face-to-face communication a speaker's emotions and intentions can be expressed by a variety of social cues (e.g., prosody, or gestures). For written communication, however, such cues are not available. Instead, emoticons (emotion icons) and other types of pragmatic cues are used in order to convey intended meanings (see e.g., Dresner & Herring, 2010). By means of graphic signs representing objects or facial expressions, and punctuation marks (e.g., quotation marks) that accompany textual messages information for particular sentence interpretations is given. Such cueing of meanings becomes even more important when using figurative language (e.g., irony) whose meaning cannot simply be derived from verbal information. Implied sentence meanings need to be inferred from contextual information and by means of pragmatic knowledge. For instance, depending on the context the sentence "*Well done*" either expresses praise for doing something good (literal interpretation), or criticism for some failure (ironic interpretation). In contrast to literal language, irony conveys implied meanings that go beyond the literal meaning and transfer speakers' emotions and attitudes. The comprehension of figurative language has been described in different neurocognitive approaches. According to the standard pragmatic view (derived from the work of Grice (1975)), the literal meaning of a figurative sentence is initially accessed and results in semantic integration difficulty when integrating with contextual information. Inferential processes are predicted in order to derive the contextually appropriate figurative meaning. In contrast, the direct access model (Gibbs, 2002) assumes that lexico-semantic and contextual information interacts from early on allowing an immediate access of the context relevant meanings, and, thus, a direct understanding of figurative meanings. A number of ERP studies on the comprehension of irony have shown that literal and figurative language comprehension diverges as early as 200 msec (Regel, Gunter, & Coulson, 2010; Regel, Gunter, & Friederici, 2011; Spotorno, Cheylus, Van Der Henst, & Noveck, 2013). ERPs for irony reliably revealed a P200-P600 pattern relative to literal language suggesting initial semantic analysis processes followed by later pragmatic reanalysis of figurative meanings. A P200 has been reported for semantic information processing on the sentence level (Federmeier, Mai, & Kutas, 2005), and may be a refection of semantic association based on prior context supporting a non-literal interpretation of ironic sentences. The P600 in response to irony may index pragmatic reanalysis associated with the derivation of appropriate figurative sentence meanings (Regel, Meyer, & Gunter, 2014; Spotorno et al., 2013). Most interestingly, an N400 effect was not obtained in those studies confirming that a semantic integration difficulty did not arise during the processing of irony. The amplitude of N400 is associated with lexico-semantic processes (for review see Kutas & Federmeier, 2011).

The present study addresses the question of how and when explicit social cueing affects language comprehension. In two ERP experiments applying quotation marks the time-course of an impact of social cueing during comprehension of literal and ironic language was scrutinized. Quotation marks applied to single words and phrases are minimal pragmatic markers that signal an alternative interpretation of the text extracts put in quotations (Gutzmann & Stei, 2011).

 In: Fabienne Salfner and Uli Sauerland (eds.),
*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 115–120.
All papers in this collection are unedited prepublication DRAFTs!

In Experiment 1, critical words for particular sentence interpretations were either cued, or uncued (i.e., for both irony and literal language). Cues were presented with the critical word ensuring that the different types of information (i.e., linguistic and pragmatic information) appeared at the same point in time. In Experiment 2, those cues were only presented in irony to assess effects of unambiguous cueing.

If explicit social cueing has an impact on language comprehension, we hypothesize different ERP patterns for cued and uncued sentences. In case those cues provide additional information for certain sentence interpretations and are immediately incorporated in irony comprehension, an earlier onset of P600 in absence of P200 is predicted for cued than for uncued irony (Experiment 1 and 2). In presence of cues an extended semantic analysis reflected by the P200 would be redundant. Because in Experiment 1 cues are ambiguous in their function, the processing of literal language might cause more difficulty in generating appropriate sentence interpretations by showing an enhanced P600 associated with pragmatic reanalysis for cued than for uncued literal sentences. Based on the different neurocognitive approaches on figurative language comprehension the following hypotheses are derived: According to the standard pragmatic view (Grice, 1975), contextual information including social cues is initially encapsulated and affects figurative language comprehension only during later processing stages. Thus, irony compared to literal language should cause semantic processing difficulty during integration of literal meanings indicated by enhanced N400, followed by later inferential processes for derivation of appropriate meanings indicated by enhanced P600 (Experiment 1 and 2). According to the direct access view (Gibbs, 2002), however, social cueing provides a strong contextual support for potential ironic interpretations that should result in a direct comprehension of cued irony thereby making pragmatic reanalysis processes unnecessary by absence of enhanced P600 for irony in relation to literal language (Experiment 2).

## Experiment 1

Explicit social cues in form of quotation marks were applied to both literal and figurative language (i.e., irony) in order to investigate the impact of extra-linguistic information on language comprehension.

### Participants

Forty native German-speaking students (20 female, mean age 24.9 years (standard deviation (SD) 3.20)) from the University of Leipzig took part.

### Methods

**Stimuli** 120 target sentences that achieved either an ironic meaning when uttered in response to a slightly disappointing event (e.g., a cup of coffee fell over), or a literal meaning when uttered in response to a pleasant event (e.g., someone enjoyed a book) served as stimuli. Both meanings depended on the foregoing contextual information (i.e., three to four context sentences describing an everyday situation). For each target sentence two types of contexts were created, thus resulting in a total of 240 short stories. The target sentence final word contained the critical information for potential meanings. For the cueing condition, in half of the target sentences critical words included quotation marks. Thus, for 50% of the items each were cued (e.g., *That's "fantastic"*) and uncued (e.g., *That's fantastic*). By means of cueing respective meanings were favored for irony, but hindered for literal language. Target sentence meanings were pretested for semantic expectancy by means of a sentence completion task (Taylor, 1953), which revealed an average cloze probability of about 94% (SD 8.15). Ironic meanings were less expected (about 5%) than literal ones (paired t-test on items $t(119) = 28.25$, $p < .001$). In an additional pretest, the interpretation of intended sentence meanings was assured by means of a rating on the sentences' ironic and literal meaning (significant difference in meaning (paired t test on items $t(119) = 2187.9$, $p < .001$)). Experimental factors *pragmatics* (ironic/literal) and *cueing* (cued/uncued) were fully crossed yielding four experimental conditions. Stimuli (context and target sentences) were presented visually. For experimental presentation, the 120 items were equally divided into four lists in a pseudorandomized order (i.e., 30 items each) to avoid repetition of target sentences. Each participant saw only one list.

**Procedure** During the experimental session (about 50 minutes), participants were seated in a sound-attenuated cabin in front of a monitor (at a distance of about 100 cm). A trial started with the visual presentation of the context sentences in one block of three to four lines on the monitor in a self-paced reading mode (automatic continuation after 20 sec). After presentation of a fixation cross for 200 msec, target sentences were presented word-by-word in Rapid Serial Visual Presentation (RSVP) mode with 300 msec per word and 200 msec in between. After sentence offset and a blank screen for 1500 msec, the experimental task had to be completed (i.e., content question judgment). In this task, a test statement outlining prior contextual information had to be judged with a *yes* or *no* response (response time of max. 6000 msec). The inter-stimulus interval was 1000 msec.

**Date acquisition and analysis** Behavioral data contained the judgments on the experimental task, and were analyzed in a repeated-measure ANOVA with the two-leveled factors *pragmatics* (ironic/literal) and *cueing* (cued/uncued). The electroencephalogram (EEG) was recorded continuously from 52 Ag/AgCl electrodes[1] mounted in an elastic cap (Electro Cap International). In order to control for eye movement artifacts, the bipolar horizontal and vertical electrooculogram (EOG) placed on the outer canthus of

[1] Fp1, Fpz, Fp2, Af7, Af3, Afz, Af4, Af8, F7, F5, F3, Fz, F4, F6, F8, Ft7, Fc5, Fc3, Fcz, Fc4, Fc6, Ft8, T7, C5, C3, Cz, C4, C6, T8, Tp7, Cp5, Cp3, Cpz, Cp4, Cp6, Tp8, P7, P5, P3, Pz, P4, P6, P8, Po7, Po3, Poz, Po4, Po8, O1, Oz, O1, and left mastoid.

each eye was recorded. The sampling rate was 250 Hz. EEG recordings were referenced online to the left mastoid, and re-referenced offline to the average of the left and right mastoid. Electrode impedances were kept below 5 kΩ. For the ERP analysis, epochs of EEG data were averaged for the critical word for each electrode position for each experimental condition in the period of -200 msec to 1000 msec relative to stimulus onset of the critical word. Only correctly answered trials free from any artifacts (approx. 11% rejections due to ocular or movement artifacts (EOG rejection +/-40 μV)) entered the analysis. For statistical analysis of the ERP data, three time windows were calculated: *200–300 ms*ec (P200), *300–500 msec* (N400), and *500–900 msec* (P600). ERPs were quantified using multivariate analyses of variance (MANOVA) to prevent violations of sphericity (Vasey & Thayer, 1987). For all time windows, an overall MANOVA including the factors *pragmatics* (ironic/literal) and *cueing* (cued/uncued) was conducted. For distributional ERP analysis, the topographic factors *Anterior/posterior* (2) and *hemisphere* (left/right) were defined and clustered into four different *Regions of Interest* (ROIs) each containing eight electrodes[2]. Midline electrode positions (Fz, Fcz, Cz, Cpz, Pz, and Poz) were analyzed separately. Whenever the overall analysis showed interactions between the experimental and topographic factors, further analyses within particular ROIs were carried out. All effects revealing a significance level of $p<0.05$, and of $p<0.1$ for marginal significance are reported.

## Results

**Behavioral data** For the judgments of the comprehension task, a mean accuracy rate of 95.3% (SD 3.80) was obtained indicating that participants performed excellently. The statistical analysis showed an interaction of pragmatics with cueing ($F(1,39)=7.59$, $p<0.01$). Follow-up analyses for each sentence meaning separately showed an effect of cueing ($F(1,39)=4.27$, $p<0.05$) for literal sentences only. Participants performed slightly better when literal sentences were uncued (mean accuracy rate 96.4% (SD 4.02)) than cued (mean accuracy rate 94.9% (SD 5.06)).

**ERP data** Grand average ERPs showed an enhanced P600, preceded by a P200 in response to irony (Figure 1). An N400 component related to irony was not seen. In response to cueing positivity with latency onset at 200 msec was evoked (Figure 2).

The overall analysis of the *200-300 msec* time window revealed a marginally significant interaction of pragmatics and anterior/posterior ($F(1,39)=3.05$, $p<0.1$). Subanalyses for anterior and posterior electrode sites separately showed an anterior effect of pragmatics ($F(1,39)=4.48$, $p<0.05$) suggesting the emergence of a P200 for irony relative to literal sentence. Analysis of the midline electrodes showed main effects of pragmatics ($F(1,39)=3.02$, $p<0.1$).

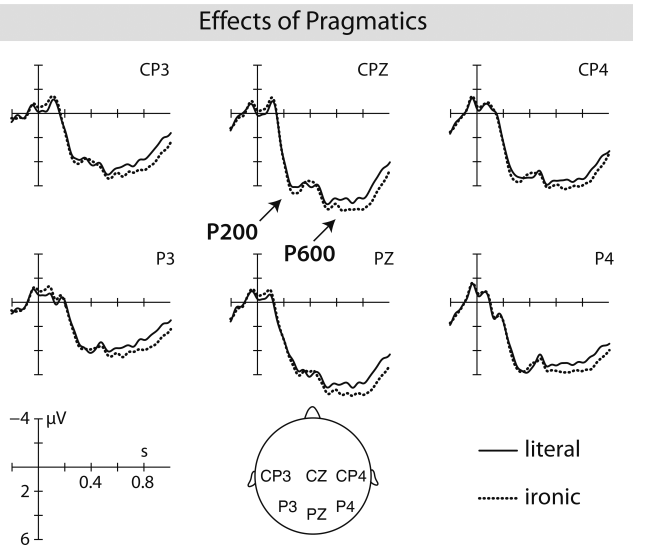



Figure 1: Grand average ERPs evoked by literal (solid line) and ironic (dotted line) sentences.

In the overall analysis, an effect of cueing ($F(1,39)=14.61$, $p<0.001$), and an interaction of cueing with anterior/posterior ($F(1,39)=11.74$, $p<0.005$) were also obtained. Resolving this interaction revealed effects of cueing for both anterior ($F(1,39)=21.30$, $p<0.001$) and posterior sites ($F(1,39)=5.07$, $p<0.05$) suggesting the presence of an early positivity in response to cueing. In the statistical analysis of the midline electrodes this positive ERP effect was confirmed by showing an effect of cueing ($F(1,39)=11.70$, $p<0.01$).

In the *300-500 msec* time window, the overall analysis revealed a main effect of cueing ($F(1,39)=18.21$, $p<0.001$), as well as interactions of cueing with the topographic factor anterior/posterior ($F(1,39)=6.61$, $p<0.05$). Separate analyses for anterior and posterior electrode sites showed effects of cueing for anterior ($F(1,39)=10.47$, $p<0.01$) as well as posterior sites ($F(1,39)=23.83$, $p<0.001$) confirming a positivity for cued relative to uncued sentences. Analysis of the midline electrodes affirms this positive effect by showing an effect of cueing ($F(1,39)=13.62$, $p<0.001$).

In the main analysis of the *500-900 msec* time window, an effect of pragmatics ($F(1,39)=6.03$, $p<0.05$) and an interaction of pragmatics with anterior/posterior and ROI ($F(1,39)=4.78$, $p<0.05$) were present. The resolution of this interaction showed effects of pragmatics for both the left ($F(1,39)=3.21$, $p<0.1$) and right posterior ROIs ($F(1,39)=4.32$, $p<0.05$), which confirms a P600 component for irony compared to literal sentences. Further, in the main analysis an effect of cueing ($F(1,39)=5.82$, $p<0.05$), and an interaction of cueing with anterior/posterior ($F(1,39)=12.96$, $p<0.01$) were observed. Resolving this interaction by anterior/posterior showed effects of cueing for anterior electrode sites only ($F(1,39)=13.75$, $p<0.001$). Analysis of

[2] ROIs were left anterior (F5, F3, Ft7, Fc5, Fc3, T7, C5, C3), left posterior (Tp7, Cp5, Cp3, P7, P5, P3, Po7, Po3), right anterior (F4, F6, Fc4, Fc6, Ft8, C4, C6, T8), and right posterior (Cp4, Cp6, Tp8, P4, P6, P8, Po4, Po8).

the midline electrodes revealed main effects of pragmatics ($F(1,39)=6.14$, $p<0.05$) and cueing ($F(1,39)=7.32$, $p<0.01$). Interactions of pragmatics with cueing were not seen in this time windows ($F(1,39)<0.85$, n.s.).

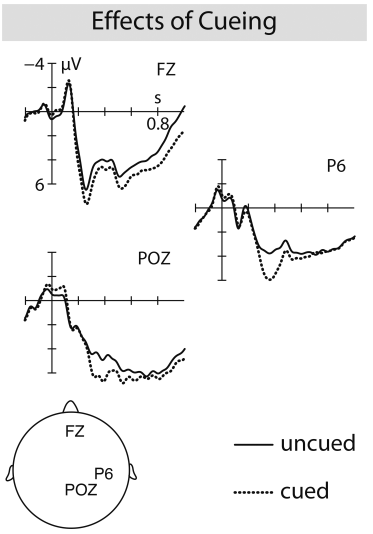



Figure 2: Grand average ERPs seen in response to cueing.

## Discussion

This experiment investigated the impact of explicit social cueing on language processing. Sentence final words containing critical information for potential literal or ironic sentence interpretations were either cued by quotation marks, or uncued. While for irony such additional cueing of meanings facilitates the recognition and interpretation of figurative meanings, for literal language those cues rather hinder contextually appropriate interpretations as implied by behavioral data for literal sentences. ERPs revealed a replication of the P200-P600 pattern for irony as seen previously (Regel et al., 2010; Regel et al., 2011; Regel et al., 2014). Comprehending ironic sentences may initially involve semantic association processes (Federmeier et al., 2005), and pragmatic reanalysis during later stages of processing (see e.g., Regel et al., 2014). An impact of cueing, however, on the processing of sentence meanings was not seen. Social cueing was applied to both irony and literal language, which apparently caused an uncertainty in meaning for both types of languages. The findings suggest that explicit cueing that is anomalously employed as in the case of literal language (i.e., overall 25% of all items), may override potential effects of cueing on language processing. Still, main effects of cueing were obtained suggesting that this extra-linguistic information has been processed. For cued in relation to uncued sentences a positivity with a latency onset of around 200 msec emerged. Independently of the type of language (i.e., ironic or literal language), cued sentences engaged enhanced processing during initial stages of processing. On basis of Experiment 1 an impact of cueing on language processing was approved, however, whether such an effect is facilitatory for figurative language comprehension cannot be concluded from the data.

# Experiment 2

In Experiment 2, explicit social cueing is employed such that the cues serve as clear hint for figurative interpretations, and allow a comparison of the processing of cued and uncued irony (i.e., by means of two experimental blocks). In the first experimental block, both irony and literal language were uncued. In the second experimental block succeeding the first one, irony was cued by quotation marks, whereas literal language remained uncued in order to scrutinize potential facilitatory effects of cueing.

## Participants

Forty native German-speaking students (20 female, mean age 23.5 (SD 2.30)) participated. All of them were right handed and had normal or corrected-to-normal vision.

## Methods

**Stimuli and procedure** The same stimuli and procedure were applied as in Experiment 1. Stimuli were presented in two separate experimental blocks. In the first block, both literal and ironic sentence were uncued, whereas in the second block merely irony was cued by quotation marks. Each block contained a total of 60 items (i.e., 30 ironic and 30 literal sentences). Experimental factors were pragmatics (ironic/literal) and blockwise cueing (cued/uncued).

**Data acquisition and analysis** The acquisition and analysis of data was identical to Experiment 1. About 13% of all trials had to be rejected due to ocular or movement artifacts.

## Results

**Behavioral data** Mean accuracy rate of 96.2% (SD 3.01) was found implying that participants performed at ceiling. The statistical analysis showed a significant effect of cueing ($F(1,39)=4.56$, $p<0.04$) indicating that participants performed slightly better for the uncued (mean accuracy 96.7% (SD 1.12)) than the cued items (mean accuracy 95.5% (SD 1.22)).

**ERP data** Grand average ERPs for uncued (Figure 3, line A) and cued irony (Figure 3, line B) revealed processing differences emerging already from 200 msec on. Relative to literal sentence for uncued irony a P200-P600 pattern was elicited, whereas for cued irony an early sustained positivity was evoked.

The overall analyses of the *200-300 msec* time window revealed effects of pragmatics ($F(1,39)=16.08$, $p<0.001$) and cueing ($F(1,39)=8.60$, $p<0.01$). An interaction between pragmatics, cueing and anterior/posterior ($F(1,39)=3.40$, $p<0.1$) was marginally significant. Resolving this interaction by cueing revealed for uncued irony an anterior effect of pragmatics ($F(1,39)=3.56$, $p<0.1$) indicating an irony-related P200 compared to literal language. For cued irony effects of

pragmatics were obtained for both anterior ($F(1,39)=23.67$, $p<0.01$) and posterior electrode sites ($F(1,39)=8.28$, $p<0.01$) suggesting an early positivity. Analysis of the midline electrodes confirms the presence of an early positivity for irony. Separate analysis of an interaction of pragmatics with cueing ($F(1,39)=7.05$, $p<0.01$) showed effects of pragmatics for cued irony ($F(1,39)=15.78$, $p<0.001$).

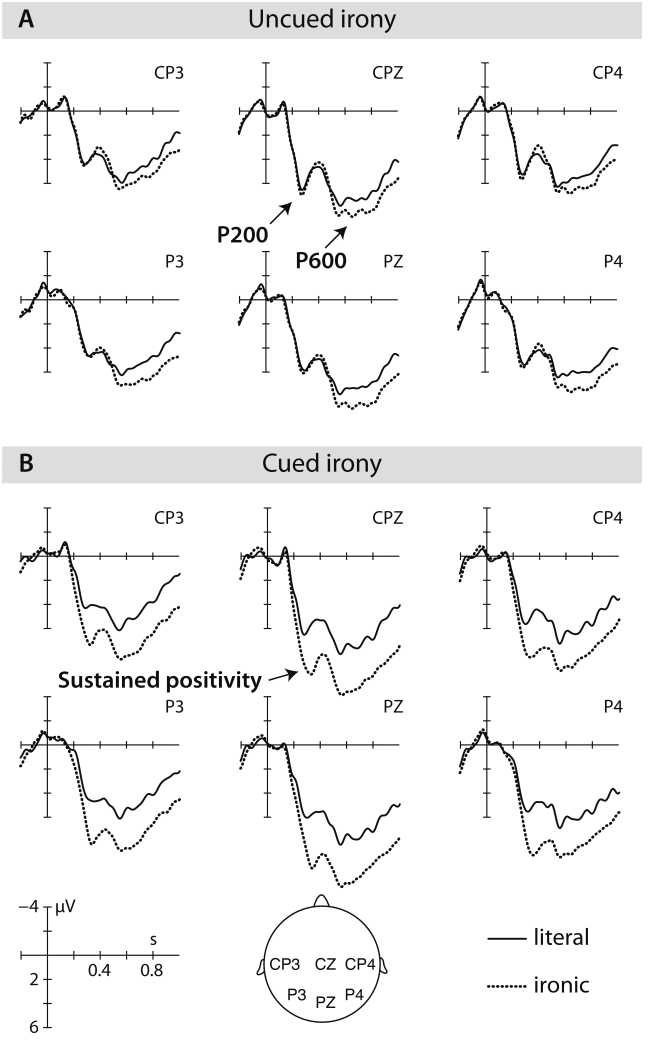



Figure 3: Grand average ERPs for uncued (line A) and uncued (line B) irony in comparison to literal language.

In the *300-500 msec* time window, the overall analysis revealed effects of pragmatics ($F(1,39)=14.31$, $p<0.001$) and cueing ($F(1,39)=44.01$, $p<0.0001$), as well as an interaction of pragmatics, cueing and anterior/posterior ($F(1,39)=14.69$, $p<0.001$). The resolution of this interaction by cueing showed an interaction of pragmatics with anterior/posterior ($F(1,39)=23.01$, $p<0.0001$) for cued irony only. In subanalysis for both sites effects of pragmatics were present for anterior ($F(1,39)=10.54$, $p<0.005$) and posterior sites ($F(1,39)=32.41$, $p<0.005$) indicating that the positivity in response to cued irony is sustained. In the statistical analysis of the midline electrodes main effects of pragmatics ($F(1,39)=14.76$, $p<0.001$) and cueing ($F(1,39)=37.94$, $p<0.0001$), and an interaction of pragmatics and cueing ($F(1,39)=17.12$, $p<0.0001$) were observed. The resolution of this interaction confirms the sustained positivity for cued irony by an effect of pragmatics ($F(1,39)=23.94$, $p<0.0001$).

In the time window of *500-900 msec*, main effects of pragmatics ($F(1,39)=30.84$, $p<0.0001$) and cueing ($F(1,39)=4.82$, $p<0.05$) were found. A four-way interaction of pragmatics, cueing and both topographic factors was also significant ($F(1,39)=3.07$, $p<0.1$). In separate analyses for cued and uncued irony, effects of pragmatics were seen for uncued irony ($F(1,39)=3.92$, $p<0.05$) suggesting the emergence of P600, and for cued irony ($F(1,39)=25.34$, $p<0.05$) confirming the sustained positivity. Statistical analysis of the midline electrodes further substantiate both findings by showing significant effects of pragmatics ($F(1,39)=36.99$, $p<0.0001$) and cueing ($F(1,39)=5.71$, $p<0.05$), and an interaction of both factors ($F(1,39)=4.76$, $p<0.05$). In separate statistical analysis, effects of pragmatics were significant for cued ($F(1,39)=5.46$, $p<0.05$) as well as uncued irony ($F(1,39)=30.30$, $p<0.0001$).

## Discussion

In Experiment 2, the impact of explicit social cueing for comprehending figurative language has been investigated by presenting irony either cued by quotation marks (as in the second block), or uncued (as in the first block). While for uncued irony the construction of sentence meanings solely relied on contextual information, for cued irony additional extra-linguistic information pointing to potential figurative interpretations was provided. ERPs in response to uncued irony revealed a P200-P600 pattern in comparison to literal language, which further substantiates previous findings (e.g., Regel et al., 2011; Regel et al., 2014). The comprehension of irony seems to involve initial semantic association (P200) and pragmatic reanalysis processes (P600). Evidence for semantic integration difficulty reflected by enhanced N400 amplitude, however, was not found. The findings suggest that literal meanings of ironic sentences do not need to be fully accessed for deriving contextually appropriate sentence meanings. With regard to an impact of cueing, the present data imply the emergence of a sustained positivity in response to cued irony, which had a latency onset of around 200 msec post stimulus presentation. In case extra-linguistic information is provided for particular sentences meanings, the processing of cued irony diverged from literal language beginning with initial stages of processing. While in Experiment 1 this type of cueing appeared to be unreliable in its function of pointing to potential meanings (i.e., by occurrence with both ironic and literal languages), in Experiment 2 those cues were clearly related to figurative interpretations by being merely related to irony. The present findings suggest that extra information entailed by quotation marks was effective in cueing ironic meanings, and resulted in a different ERP pattern as seen for uncued irony (i.e., P200-P600). This sustained positivity may be interpreted as an early starting

P600. In case sufficient cues for particular sentence interpretations are provided, processes of pragmatic reanalysis allowing a derivation of ironic meanings are initiated already during initial stages of processing. Thus, explicit social cueing seems to have a facilitatory effect at least for the recognition of figurative language.

## General discussion

In two experiments, the impact of explicit social cueing on language comprehension was investigated. ERPs analyzed at the critical word revealed that cueing by quotation marks has an immediate effect on the processing of irony. In Experiment 2, cued irony evoked an early sustained positivity compared to literal language. This positive ERP response resembled an early starting P600 suggesting that pragmatic reanalysis processes have been initiated simultaneously with lexical access (i.e., around 200 msec post stimulus onset). In presence of explicit cues potential figurative meanings are favored thereby enabling an immediate derivation of intended ironic meanings. Moreover, an impact of social cueing on language processing depended on the unambiguity of its function. In Experiment 1, both the processing of literal and ironic language was unaffected by cueing implying that in case of ambiguous cueing (i.e., by being simultaneously conform to linguistic norms for irony, but non-conform for literal language) this kind of information was neither integrated during initial, nor later stages of processing. For irony compared to literal language, a P200-P600 pattern was replicated as seen previously (Regel et al., 2014; Spotorno et al., 2013). This finding implies that contextual strength, which varied by the occurrence of cues, modulates figurative meaning processing. In absence of additional cues for particular sentence interpretations, the processing of irony engages initially semantic association processes, and later pragmatic reanalysis in order to derive contextually appropriate sentence meanings. In presence of additional cues, however, pragmatic reanalysis seems to be initiated already during lexical access. With regard to neurocognitive approaches on figurative language processing, the present findings partially accord with the assumptions of the standard pragmatic view (Grice, 1975). By observation of P600 in response to irony, inferential processes allowing a derivation of intended sentence meanings may have been engaged during later processing stages. Still, an irony-related N400 was not obtained suggesting the absence of semantic integration difficulty. Contextual information already seems to affect initial processing stages as indicated by P200 seen for irony. While this view can be confirmed for the presumed later processing stages, no evidence was found for an initial literal access of ironic sentences. Since irony evoked early and late ERP effects, a direct comprehension of figurative language was not supported as assumed by the direct access view (Gibbs, 2002). Nonetheless, contextual information plays a crucial role for comprehension of intended sentence meanings.

## Acknowledgments

We are grateful to Angela D. Friederici for motivating discussions on the ERP results of both experiments. We also thank Cornelia Schmidt and Christiane Klein for their assistance with the acquisition of the EEG data.

## References

Dresner, E., & Herring, S. C. (2010). Functions of the Nonverbal in CMC: Emoticons and Illocutionary Force. *Communication Theory, 20*(3), 249-268.

Federmeier, K. D., Mai, H., & Kutas, M. (2005). Both sides get the point: Hemispheric sensitivities to sentential constraint. *Memory & Cognition, 33*(5), 871-886.

Gibbs, R. W. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics, 34*(4), 457-486.

Grice, P. H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts: Syntax and semantics* (pp. 41-58). New York: Academic Press.

Gutzmann, D., & Stei, E. (2011). How quotation marks what people do with words. *Journal of Pragmatics, 43*(10), 2650-2663.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychophysiology, 62*, 621-647.

Regel, S., Gunter, T. C., & Coulson, S. (2010). The communicative style of a speaker can affect language comprehension? ERP evidence from the comprehension of irony. *Brain Research, 1311*, 121-135.

Regel, S., Gunter, T. C., & Friederici, A. D. (2011). Isn't it ironic? An electrophysiological exploration of figurative language processing. *Journal of Cognitive Neuroscience, 23*(2), 277-293.

Regel, S., Meyer, L., & Gunter, T. C. (2014). Distinguishing Neurocognitive Processes Reflected by P600 Effects: Evidence from ERPs and Neural Oscillations. *Plos One, 9*(5), e96840.

Spotorno, N., Cheylus, A., Van Der Henst, J.-B., & Noveck, I. A. (2013). What's behind a P600? Integration Operations during Irony Processing. *Plos One, 8*, 1-10.

Taylor, W. L. (1953). "Cloze Procedure": a new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology, 24*, 479-486.

# Off-record indirectness: In theory and in practice

Jessica Soltys and Napoleon Katsos
*Department of Theoretical and Applied Linguistics, University of Cambridge*

## 1 Introduction

Indirect speech acts are instances in which an illocutionary act, such a request or offer, "is performed indirectly by way of performing another" (Searle 1975: p.60). This paper focuses on off-record indirectness (henceforth ORI), a strategy through which a speaker intentionally conveys to the addressee two related meanings – a literal, direct meaning and an indirect meaning, the latter of which arises via implicature in accordance with contextual clues (Grice 1975). ORI utterances are non-conventionalised, with no restrictions on either propositional content or linguistic form (Blum-Kulka 1987). ORI speech acts are pragmatically ambiguous, as "it is not possible to attribute only one clear communicative intention to the act" (Brown & Levinson 1987: p. 211) and, as such, are plausibly deniable.

Because of this ambiguity, there is a risk that the hearer may be unable to recognise the intended meaning. Likewise, the complexity of the act may require greater planning on behalf of the speaker. In what situations, then, do speakers prefer ORI over more explicit and straightforward options? Several theoretical accounts have been proposed to address this question, each of which attributes different motivations to ORI and assumes different functional, contextual, or interpersonal conditions under which ORI is ideal. This paper focuses on two accounts, Politeness Theory (PT, Brown & Levinson 1987) and the Strategic Speaker approach (SS, Lee & Pinker 2010; Pinker 2007; Pinker, Nowak, & Lee 2008).

Under PT, ORI is considered the most polite option from amongst a range of strategies aimed at balancing the speaker's desire to express the intended communicative act with his commitment to respecting the addressee's negative face wants. The decision to use ORI is based on the speaker's assessment of three sociological factors – the differences in power (P) and social distance (D) between the interlocutors and the degree of imposition placed on the hearer (R), with the likelihood of ORI use increasing as the settings of P, D, and R, collectively known as 'face threat', increase. A request for favour is a prototypical case of polite ORI, though under PT, ORI can be used to redress any speech act so long as the settings of P, D, and R dictate maximum levels of politeness.

In SS, ORI is a game-theoretic means of balancing the risks and rewards – legal, financial, social, and/or emotional – associated with potentially conflictual acts such as bribes and sexual propositions; acts which, according to Pinker and colleagues, are not governed by politeness. While in PT, both interlocutors abide by social and communicative norms to ensure that face is respected and Gricean maxims are observed, only the latter can be assumed in SS, as the speaker and hearer may have competing interpersonal aims. The speaker's interest is, primarily, self-serving – to achieve the best possible outcome and to deny an (inappropriate) implicated meaning when socially necessary – and is not influenced by P, D, and/or R or other politeness concerns.


*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 121–127.

This paper summarises three related empirical studies, each of which used offline questionnaires to explore one aspect of the use of ORI by native English speakers. Findings are considered both individually and comparatively with regards to the competing predictions associated with PT and SS, with the overall aim of determining which account(s) best explains the use of ORI.

## 2 Experiment 1

Experiment 1 is a qualitative experiment about trends in the use of ORI, as informed by participants' self-reports of their intuitions about, and experiences with, indirect speech. There were 94 participants (64 female), aged 18 to 73 years with a mean age of 38. All lived in the US and were recruited on MTurk.

Participants were asked to recount up to three situations in which they used ORI. They were instructed to reproduce the off-record utterance, describe the context in which it was used, and explain why they chose to speak indirectly. They were shown a brief introduction to ORI, stating "In some situations, we prefer to say things indirectly. We may use hints or innuendo to imply what we want instead of stating it explicitly." Five versions of the questionnaire were disseminated – a neutral version which included no additional explanatory or contextual information and two versions each which included illustrative examples of ORI used in SS situations (a bribe and an invitation for a date) and PT situations (requests between colleagues and friends).

In total, 85 responses were analysed: 38 in the PT scenarios, 40 in the SS scenarios, and 8 in the neutral version. Responses were coded in accordance with the theory they best fit – PT or SS. Responses were classed as 'multiple interpretations possible' if they reflected both theories and 'other' if they described uses of ORI not related to PT or SS. PT responses frequently cited motivations associated with displays of negative politeness, as in example (1) while SS responses described the participants' own use of bribes and propositions. Table 1 shows the mean responses in each theoretical category.

(1) My babysitter cancelled on me last minute and needed someone to watch kids so I could go out to dinner at a new restaurant / "I'm so bummed I have to cancel these dinner reservations that we have had for 2 months." / I didn't want to inconvenience my friend.

**Table 1: Participants' reported use of ORI**

| | PT | SS | Multiple | Other |
|---|---|---|---|---|
| **PT scenarios** | 0.55 | 0 | 0.29 | 0.16 |
| **SS scenarios** | 0.50 | 0.13 | 0.15 | 0.22 |
| **All versions** | 0.51 | 0.07 | 0.22 | 0.20 |

The use of ORI was most frequently attributed PT, with 51% of the responses across all five versions of the questionnaire. There were no significant differences in PT uses between the PT and SS scenarios ($t(37) = 0.83$, $p = 0.41$). SS uses were not mentioned in either the PT scenarios or the neutral version. The difference in SS uses between the PT and SS scenarios was statistically significant ($t(37) = -2.37$, $p < 0.05$). Across all versions, there was a statistically significant difference between the frequency of PT-based and SS-based uses of ORI ($t(84) = 6.41$, $p < 0.001$).

The findings suggest that PT-based uses of ORI are the most accessible to native English speakers, characterising both their patterns of use and their intuitions about motivation. Participants' prototypical understanding of ORI is defined by PT principles, as evidenced by the high frequency of PT-based uses in both the neutral scenario and the scenarios biasing SS uses. SS uses, on the hand, were less readily accessible, cited by participants only when prompted by SS examples. SS uses, then, are unlikely to mould native English speakers' prototypical understanding of ORI.

## 3 Experiment 2

Experiment 2 is inspired by Lee & Pinker's 2010 multiple choice study on the use of ORI in bribes, sexual propositions, and favours. The former two are prototypical SS scenarios, presenting the speaker with potential risks and rewards – legal/financial and social/emotional, respectively. The latter, a favour, is a scenario for which PT, but not SS, predicts ORI under suitably high settings of P, D, and/or R.

In Lee & Pinker 2010, participants viewed one bribe, one proposition, and eight versions of a favour, with manipulations to the latter depicting binary changes to the settings of P, D, and R. They selected from five response types – blunt on-record, negatively polite, positively polite, somewhat indirect, and very indirect, the latter two of which were analysed as a single ORI category. Participants opted for ORI in 58% of bribes, 91% of the propositions, and 21% to 34% of the eight favours. The negatively polite on-record option was the most popular choice in the favours. Scalar ratings of P, D, and R were collected for each scenario. In the two SS scenarios, the ratings of P, D, and R were either comparable to or lower than the respective ratings in the most face-threatening version of the favour. Lee & Pinker conclude, then, that the use of ORI is not a function of face-threat. They suggest that the SS acts are categorically distinct from the PT-based favour in three ways: the frequency with which ORI is used; the effect of P, D, and R; and the role of politeness. With regards to the latter point, they state that the two SS acts "may be highly indirect, but they are unlikely to be clad in the kinds of constructions that protect the hearer's face in other requests, such as Please, Do you think you might', or 'I'm sorry to have to ask but'" (Lee & Pinker 2010: p. 786).

Since no manipulations were applied to the bribes or propositions, one cannot determine whether it is the act itself, or the particular settings of the P, D, and R employed in the sole example of the act, that influence the participants' use of ORI. Further analysis would determine whether qualitative changes to the settings of P, D, and R significantly affect the frequency with which participants opt for ORI utterances in each of the SS acts and, further, would reveal whether there are conditions under which participants prefer blunt on-record, positively polite, or negatively polite utterances for bribes and propositions. The use of multiple choice is restrictive as well, as one cannot determine whether the ORI options were chosen with higher frequency than other utterances due to a true preference for ORI or a dispreference for the remaining options which was not related to the level of directness but to the specific wording used.

Experiment 2 departs from Lee & Pinker in two ways in order to address the aforementioned criticisms. Firstly, speech act data is elicited through an open-ended format, a condition which allows both quantitative analysis (frequency of use and correlations with P, D, and R) and qualitative analysis (the participants' choice of politeness markers and other content-based and discoursal features). Secondly, binary manipulations to P, D, and R are applied to all three speech act types – a bribe between a speeding driver and

a traffic officer, a sexual proposition between co-workers, and a favour amongst colleagues. Ratings of P, D, and R are collected in each scenario using a seven-point Likert scale in order to measure the effect of P, D, and R on the use of ORI in the bribes and the propositions

A total of 110 participants (77 females) were recruited from the University of Cambridge; 35 did the bribe scenario, 34 did the favour, and 41 did the proposition. The participants ranged in age from 18 to 63, with a mean age of 23 years and all were university students and native English speakers.

Open-ended responses were coded using a binary scheme – a response was either ORI or not ORI. Responses were deemed ORI if multiple interpretations were available; in others words, if the speech act was pragmatically ambiguous and, as a result, plausibly deniable. The ORI responses in examples (2), (3), and (4) were provided by participants in the bribe, proposition, and favour, respectively.

(2) Officer, I'm really sorry. I didn't mean to go so fast. Perhaps we can reach some sort of arrangement.

(3) Fancy coming back to mine for a drink?

(4) I'm really struggling to complete this report. Do you know of anyone who may be able to help me with it?

Table 2 shows the mean rates of ORI per speech act, as well as the correlations between the participants' use of ORI and their ratings of P, D, and R. The favour is excluded from the correlations due to low ORI use.

**Table 2: Mean rates of ORI and correlations with P, D, R**

| | ORI | P | D | R |
|---|---|---|---|---|
| **Bribe** | 0.78 | $r = 0.19, p < 0.01$ | $r = -0.10, p = 0.11$ | $r = -0.05, p = 0.46$ |
| **Proposition** | 0.92 | $r = 0.08, p = 0.17$ | $r = 0.05, p = 0.35$ | $r = 0.10, p = 0.08$ |
| **Favour** | 0.03 | N/A | N/A | N/A |

As predicted by the SS account, ORI was used frequently in the two SS-based scenarios – the bribe and the sexual proposition – and rarely in the favour. There were statistically significant differences in the use of ORI between the proposition and the bribe ($t(263) = 4.747$, $p < 0.001$), the proposition and the favour ($t(261) = 41.942$, $p < 0.001$), and the bribe and the favour ($t(261) = 27.567$, $p < 0.001$). The data suggest a categorical distinction between speech act types – bribes and propositions on one the one hand, where ORI is a frequent choice (notwithstanding the significant difference in the rate of ORI between the two) and favours on the other, where ORI is almost non-existent.

In the bribe scenario, there was one statistically significant correlation in support of PT: as the rating of P increased, reflecting a greater difference in relative power between the interlocutors, the use of ORI likewise increased. There were no other statistically significant correlations between the use of ORI and the ratings of P, D, or R in the SS scenarios, suggesting that these factors alone are not a reliable predictor of ORI.

The ORI data were further analysed for the use of both negative and positive politeness markers, including deference and solidarity-based address terms, hedging, polite pessimism, and the use of conventionally polite forms such as 'would/could you…' as shown in the ORI bribe in example (5).

(5) Sir, there is no possibility that you might be able to help me, is there?

Politeness markers were frequently used within ORI utterances: 92% of the bribes were coded as both ORI and polite, as were 66% of the propositions. The difference between the two acts was statistically significant ($t(204) = 5.95$, $p < 0.001$). These data challenge the SS account, namely the assertion that "[…] politeness and indirectness do not reside on the same scale but are rather distinct mechanisms elicited by different types of social encounters" (Lee & Pinker 2010: 787). The data show that these mechanisms can indeed be used within a single encounter, albeit for different ends. While ORI affords plausible deniability, politeness markers allow the interlocutors to reap (at least ostensibly) the social benefits of facework.

## 4 Experiment 3

Experiment 3 is a qualitative study on the strategic motivations ascribed to the use ORI. It was completed by 20 native English speakers (13 females), all of whom were recruited from the University of Cambridge student body and ranged in age from 22 to 46 years, with a mean age of 29.

The study comprised a judgement task in which participants were asked to infer the reasons for which a quoted speaker opted for ORI in scenarios depicting bribes, propositions, and favours. Participant viewed two full versions each of the three speech acts – representing the –P –D +R and –P –D –R settings, respectively and, after completing a scalar rating task (not reported here), were shown an ORI utterance attributed to the depicted speaker. The utterances were taken from the data collected in Experiment 2 (examples (2), (3), and (4)) and represented prototypical uses of ORI. Participants were asked: "Based on your experience, why do you think that [speaker's name] has decided to phrase the [speech act] in this way?"

The responses were coded following the criteria used in Experiment 1, with the participants' responses categorised as reflecting 'PT only', SS only', 'multiple interpretations,' and 'other'. The aim of the study was to determine whether participants identified similar motivations to the ones described by PT and SS and, further, whether these motivations underlie the categorical distinction between PT and SS based uses of ORI. Examples (6) and (7) represent data coded as PT and SS and were used in response to the favour and bribe scenarios, respectively. Table 3 shows the mean response by speech act type.

(6) This avoids the obtrusiveness of asking directly for the favour, which could be seen as an affront as it puts the other person under pressure […]

(7) Because it is intentionally vague. If the officer retorts 'are you trying to bribe me?' (either out of alarm or hesitancy) [the speaker] can reply that that is not what she intended by her statement at all […]

**Table 3: Motivations for ORI**

| | PT only | SS only | Multiple | Other |
|---|---|---|---|---|
| **Bribe** | 0 | 0.70 | 0.25 | 0.05 |
| **Proposition** | 0.05 | 0.65 | 0.15 | 0.15 |
| **Favour** | 0.60 | 0.05 | 0.15 | 0.20 |

There were statistically significant differences between the PT motivations in the favour and both the bribe ($t(19) = 5.34$, $p < 0.001$) and the proposition ($t(19) = 4.82$, $p < 0.001$) and in the SS motivations in the favour and both the bribe ($t(19) = -4.95$, $p < 0.001$) and the proposition ($t(19) = -4.49$, $p < 0.001$).

In all cases, there was a clear and significant preference for one motivation over the other. As predicted, PT motivations were cited in response to use of ORI in the favour scenario while in both the bribe and the proposition, participants mentioned motivations related to the SS. The data support the categorical distinction found in Experiment 2. Bribes and propositions differ from favours not only with regards to the frequency with which ORI is used, but also in terms of the motivations for its use. The use of ORI, then, is not limited to a single motivation, but to several aims, in accordance with the context and speech act type.

## 5 Discussion

When Experiments 1 and 2 are viewed concurrently, an unexpected discrepancy between intuition and practice arises. In Experiment 1, participants rarely recounted SS-based uses of ORI, even when prompted by prototypical SS examples. In Experiment 2, however, participants frequently produced ORI utterances, of their own accord, in the two scenarios for which SS predicts ORI – the bribe and the sexual proposition. Similarly, PT-based uses of ORI predominated in Experiment 1, yet were virtually absent from the favour in Experiment 2.

On the surface, the findings challenge both theoretical accounts, as neither predicts both the intuition and the practice of the participants. SS fails to impact the participants' intuition, either in terms of the game theoretic logic that motivates the use of ORI or the conflictual scenarios upon which the account is focused. With regards to practice, both theories fall short. Although, as predicted by SS, ORI is used frequently in the both the bribes and the propositions, the widespread and strategic use of politeness markers within these utterances explicitly contradicts the assumptions of the account. The near absence of ORI in the favours, even in the scenarios for which P, D, and/or R are set to comparably high levels, speaks strongly against the principles of PT.

When reconsidered, however, the data provide support for both PT and SS. Experiment 1 confirms that PT-based uses of ORI are both familiar and accessible to native English speakers, shaping their prototypical understanding of ORI. Within cultures oriented to negative politeness, a preference that has been attributed to users of both American and British English (Blum-Kulka 1987, Sifianou 1992), polite uses of ORI may be considered desirable by both speakers and hearers, further encouraging self-reports. In Experiment 2, each of the settings of P, D, and R may have been evaluated at a level below the threshold necessary for ORI. This conclusion is supported by the high frequency of negatively polite on-record responses, a strategy ranked by PT as reflecting a level of politeness below ORI. SS-based uses of ORI, on the other hand, are limited in scope and less likely to be experienced on a regular basis. The social stigma attached to acts such as bribes and propositions may hamper the participants' willingness to self-report, a conclusion that supports Pinker and colleagues' assertion that these speech acts are categorically distinct from other acts due to both their potentially conflictual nature and to the tangible and intangible consequences associated with them. As predicted by SS, participants draw on ORI when faced with these acts under experimental conditions, regardless of intuition or first-hand experience.

Experiment 3 lends additional support to both PT and SS and, by extension, highlights the versatility of ORI. Participants differentiated between the speech acts they viewed – the bribe and the sexual proposition on the one hand and the favour on the other – by attributing distinct strategic and interpersonal motivations to the use of ORI in each case. The motivations described by the participants aligned with the predictions of SS for the bribe and proposition and with PT for the favour. The data show that there is no singular purpose for the use of ORI, but rather, ORI is used strategically in different situations with recognisably different aims.

The wide-spread and unexpected use of negative and positive politeness within ORI bribes and propositions presents a unique challenge to SS. While both the frequency of use of ORI in Experiment 2 and the strategic motivations attributed to the use of ORI in Experiment 3 strongly support the predictions of the account, the addition of politeness contradicts Pinker and colleagues' assumptions.

In light of these findings, we propose an updated version of Pinker and colleagues' account, henceforth referred to as the 'Moderate Strategic Speaker approach' (Moderate SS). This account incorporates the facets of Pinker and colleagues' theory that were upheld by the experimental findings – namely, the game-theoretic logic and the lack of sensitivity to P, D, and R – with the face-based politeness principles associated with Brown & Levinson's PT.

Under the Moderate SS, ORI and politeness are separate phenomena and, with regards to high-risk acts such as bribe and propositions, the use of each is motivated by different factors. ORI is motivated by the strategic interests outlined in SS – the desire to balance risks and rewards in the speaker's favour – and is not influenced by the settings of P, D, and R. The use of negative and positive politeness is motived by facework. The choice of politeness markers may be influenced by the speaker's assessment of P, D, and R. ORI and politeness can be used concurrently within a single exchange.

While each theory has merit, neither is all-encompassing. PT and SS may independently motivate the use of ORI under conditions well-suited to each, while in other strategic contexts, motivations related to both PT and SS, respectively may each contribute to an aspect of the speaker's choice of ORI.

**References**

Blum-Kulka, Shoshana, 'Indirectness and politeness in requests: Same or different?', *Journal of Pragmatics*, 11.2 (1987), 131–146.

Brown, Penelope and Stephen C. Levinson, *Politeness: Some universals in language use* (Cambridge: CUP, 1987).

Grice, H. Paul, 'Logic and conversation,' in *Syntax and Semantics, Volume 3: Speech Acts*, ed. by Jerry L. Morgan and Peter Cole (New York: Academic Press), pp. 41–58.

Lee, James J. and Steven Pinker, 'Rationales for indirect speech: The theory of the Strategic Speaker', *Psychological Review*, 117.3 (2010), 785–807.

Pinker, Steven, 'The evolutionary social psychology of off-record indirect speech acts', *Intercultural Pragmatics*, 4.4 (2007), 437–461.

Pinker, Steven, Martin A. Nowak, and James J. Lee, 'The logic of indirect speech', *Proceedings of the National Academy of Sciences*, 105.3 (2008), 833–838.

Searle, John R., 'Indirect speech acts' in *Syntax and Semantics, Volume 3: Speech Acts*, ed. by Jerry L. Morgan and Peter Cole (New York: Academic Press), pp. 59–82.

Sifianou, Maria, *Politeness phenomena in England and Greece* (Oxford: Clarendon Press, 1992).

# What would a compositional hearer do? - controlling for prior expectations in visual world timecourse studies

Chao Sun
*University College London*

Richard Breheny
*University College London*

**Abstract** Studies on the timecourse of pragmatic enrichment from *some* to *not all* have found conflicting evidence. Here we present research showing that participants in such studies use an association between set size and quantifier to predict the target. This and other factors make the interpretation of previous visual world data problematic. We argue that the best measure that controls for extraneous factors driving anticipatory looks is looks to a 'residue set'.

**Keywords:** Scalar enrichment, Prior expectation, Eye-tracking

## 1 Introduction

Visual-World studies on the timecourse of processing pragmatically enriched *some* compared to *all* have found conflicting evidence (Degen & Tanenhaus 2015; Grodner et al. 2010; Huang & Snedeker 2009). The use of visual-world methods in experimental pragmatic research exploits the fact that, at a given point in the linguistic input, eye-gaze data reveals that participants anticipate upcoming content based on incremental processing of the stimuli up to that point. However, factors that affect anticipatory eye movements may go beyond computation of the composition of (pragmatically enriched) semantic content. For example, Kamide et al. 2003 shows that priors based on world knowledge affect anticipatory looks. To date no visual-world research that we know of has established that prior expectations about what visual information to expect given a quantifier expression affect eye gaze. Here we present research showing that participants have prior expectations about the relative set size of the target in the display given the determiner (*all*/*some*). They use such expectation between set size and quantifier to predict the target. We propose that considering the effects of such expectation on the anticipatory looks to the target, visual search for a complement set is a better indicator to explore whether the pragmatically enriched meaning of *some* is available in the same timecourse as the semantic interpretation of *all*.

The point of departure for these studies is Huang & Snedeker (2009). In their visual-world eye-tracking studies, participants viewed a display with four quadrants.




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 128–136.

Each quadrant contained a character, and in total there were two boys and two girls. Two sets of objects (e.g., socks and soccer balls) were distributed among these characters. In the critical trials, one set of four objects (e.g. socks) was evenly divided between a boy and a girl, while the other set of three objects (e.g., soccer balls) was distributed as a total set to the other girl. Participants were instructed, *Point to the girl that has [Det] of the socks/soccer balls*, where [Det] is one of 'some', 'all', 'two', 'three'. The results showed that participants were slower to disambiguate reference in *some* compared to *all*, *two* and *three*. Moreover, the referential disambiguation in *all* was as quick as *three* trials which both were observed from 400ms after the quantifier onset. The authors attributed the delay for *some* items to the fact that some pragmatic process is required to anticipate the referent of the definite expression, while *all* and *number* items do not require such a process. Contradictory results were reported in Grodner et al. (2010) who found that convergence on the target for utterances with *some* was as fast as utterances with *all*. The effects of the pragmatic *some* were observed 200-400ms after the onset of the quantifier. The key difference between two studies is that there was no number instruction in Grodner et al. (2010)'s study. They suggested the delay in Huang & Snedeker (2009) was due to the presence of number items which may reduce the felicity of definite descriptions using the less specific *some*. By contrast, Huang, Hahn & Snedeker (2011) argued that the presence of number items blocks pre-coding of regions of the visual display prior to the onset of the quantificational determiner.

Notwithstanding the effect of the presence of number items in a session, we conjecture that there is another factor at play in these studies. In the papers mentioned above, the *all* target was a character that had a larger collection of objects than the *some* target. We conjecture that participants have a relatively 'low-level' expectation that a person with *all* of something will possess a relatively large set of objects. It is also possible that participants have an expectation that a person with *some* of something will possess a relatively small set. It is unknown whether the expectation for the *all* referent to be the big set and the *some* referent to be the small set is equally strong. If the expectation between set size and quantifier affects the processing of quantificational NPs differently, then the delay in *some* can be the result of prior expectation rather than the pragmatic processing. Degen & Tanenhaus (2015) have noticed this factor and they controlled for the association between quantifier use and set size by counterbalancing the target set size. *Some* and *all* were both used for big (e.g.4 gumballs) and small set (e.g. 3 gumballs). Therefore participants could not associate an absolute set size with quantifier use. Their *number* present experiment replicated the results in Huang & Snedeker (2009) that there was a delayed response in small-set *some* items compared to big-set *all* items. Degen & Tanenhaus (2015) suggested the delay is due to the availability of context-specific alternatives, "in particular, the availability of number terms as potential alternatives with which to

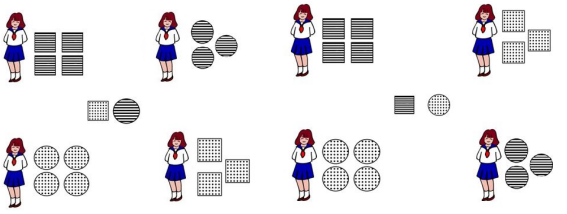
describe small sets of objects decreases the speed with which *some* is processed". However, they found a quantifier-by-target size interaction, the relative delay in generating implicature of *some* only happened when the target size was big but not when it was small. Such an interaction seems surprising in that there was no timing difference between *some* and *all* when set size was small but in the presence of numbers.

Given that lower-level expectations may influence gaze independently of compositional processing, we designed our study so that participants could inspect a 'residue set' in order to verify that the target is correct. To date no visual-world research that we know has investigated the anticipatory looks to the residue set. Such anticipatory looks is a product of the verification process of the quantificational NPs. To illustrate, consider Huang & Snedeker's study where the instructions were *Point to the girl that has [Det] of the socks/soccer balls*. Identifying the referent of the description involves verifying the relative clause containing a quantificational NP against the sets of objects associated with the characters. In the case of number items, it is sufficient to inspect only the cardinality of the sets in each quadrant only. Whereas for quantifiers, if *some* is unenriched, anticipating the referent does not require checking the complement set; to establish whether a character is a girl with *all*, or a girl with *some but not all*, it is necessary to check the residue set. Therefore, looks to the residue set reflect the incremental processing of the quantificational NP and it is relatively unaffected by the prior expectation discussed above.

## 2 The Current Study

The current study investigates the timecourse of pragmatic enrichment of *some* in a visual world eyetracking paradigm similar to previous research. In particular, we are interested in examining whether the expectation between set size and quantifier affects the processing of quantificational NP. Adapting from Huang & Snedeker (2009), each trial began with a display in which four agents surround four groups of objects. Participants heard a description of the types of objects in the middle. Then the objects were distributed to four identical agents. There were always two agents that have all of one kind of object and the other two have some but not all. The residue remained in the centre. Figure 1 are examples of critical displays. Participants were given an instruction of the form *Click on the girl with [Det] of the [modifier] [shape]*. [Det] is one of *some*, *all*, *three*, *four*, [modifier] is one of *dotted*, *stripy*, *checked* and [shape] is one of *circle*, *square*, *triangle*.

We predict there to be lower bias to the residue set for numbers than *all*. We also predict that, to the extent that *some* is enriched in the same timecourse as *all*, there should be also lower looks to the residue target in number items than *some* items. In addition, to examine whether the expectation between quantifier use and target size



(a) big *all*-small *some* (b) small *all*-big *some*

**Figure 1** Example visual scene

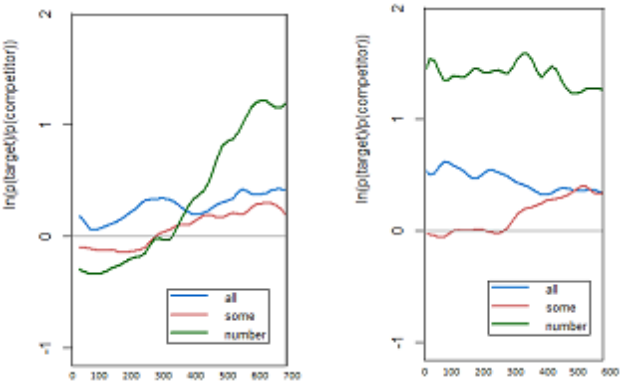
affects the anticipatory looks to main target, half of the trials follow the typicality expectation (big *all*-small *some*, Figure 1a) and the other half violate the expectation (small *all*-big *some*, Figure 1b). We predict the target bias is greater in big-set *all* and small-set *some* compared to big-set *some* and small-set *all* respectively.

### Materials, participants, procedure

The experiment employed 3x2 design, with determiner type (all/some/number) and target set size (big/small) as repeated-measure factors. 36 experimental displays were paired with an audio instruction in one of three conditions. One version of each item was assigned to one of three presentation lists, each display only appeared once in each list. Each list contained 36 experimental items, 12 in each of the three conditions. In addition, there were 18 fillers in each list. The audio instructions were cross-spliced in order to avoiding co-articulation information in favour of any condition. The average duration for determiner window is 719ms, and the average duration for modifier window is 632ms.

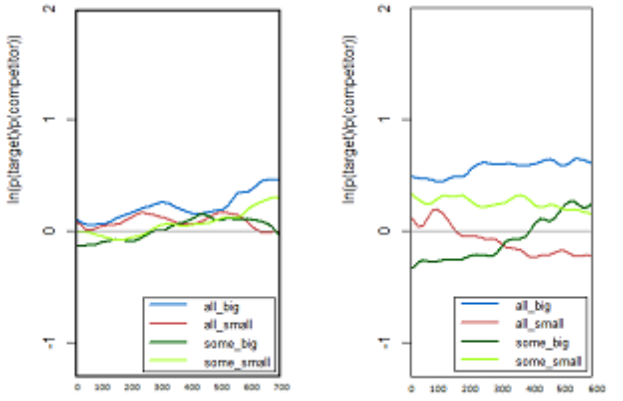
36 native English speakers participated the study. The experiment was conducted using E-Prime software and a Tobii TX300 eye-tracker. Fixations were sampled every 17ms. Participants were randomly assigned to one of the presentation list, and 54 trials were presented in random order. For each trial, participants viewed the initial display and listened to an audio description of the type of objects contained in the middle of the display. 6 seconds after the onset of the description, the next display appeared. Participants had 2.5 seconds to preview the display. After the preview, they were instructed to click on certain image on the screen. There were six practice trials in the beginning. The whole experiment lasted approximately 25 minutes.

**Results**




(a) determiner window (b) modifier window

**Figure 2** Log ratio of percentage of looks to target over competitor for experimental conditions in both windows




(a) determiner window (b) modifier window

**Figure 3** Log ratio of percentage of looks to target over competitor by quantifier for two set sizes in both windows

We excluded trials where there was no response (3.4%), and trials with the wrong response (2%). To plot the anticipatory looks to the target compared to the competitor,

we calculated natural log ratio of percentage of looks to the target over competitor as a function of time. For the determiner window, both agents with correct amount of the objects were targets and the other two agents were competitors. For the modifier window, the target was the agent of the description and the competitor was the agent with objects of the same pattern. Given the change of regions of interest between determiner window and modifier window, we visualised the data for two time windows separately. Note that for all plots and data analysis, word regions have been offset by 200ms. Eye movements and auditory input have been resynchronized according to individual word onsets (Altmann & Kamide 2009).

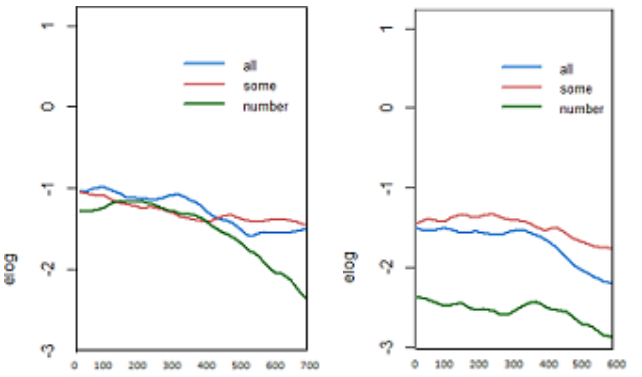
For statistical analysis, due to the eye-movement based dependencies, the data was aggregated over 50 ms time slices (3 frames of eye data) and over all of the trials in a given condition for each participant. The natural log ratio of percentage of looks to the target over competitor *(ln(P(T)/P(C))* was calculated for each bin as dependent measure for the first analysis. The empirical logit for each bin was calculated as dependent measure for the second analysis. We carried out separate statistical analysis for each time window by fitting mixed effects linear regression models using R (version 3.1.2, R Core Team) with lme4 package (Bates et al. 2015) and lmerTest package (Kuznetsova et al. 2013). The two levels of target size (small/big) were coded using contrasts (-.5 vs. .5), the three levels of determiner type (all/some/number) were coded using helmert coding. When the complex interaction were significant, we broke down the data into smaller dataset.

For the first analysis, in each time window, we fitted a model to predict log ratio from fixed effects of determiner type (all/some/numbers), target set size (big/small), a continuous time variable and their interactions. The model contained maximal by-subject random effects structure supported by the data.

Figure 2 plotted the log ratios over each 17ms sample for *all*, *some*, and *number*. During the determiner window (Figure 2a), as time increased, the bias towards the target was formed faster in numbers than quantifiers (*all*: $\beta$= -0.24, SE= 0.74, p<.001; *some*: $\beta$=-0.20, SE=0.74, p=.003). Meanwhile *All* and *some* behaved the same. On average log ratios did not differ between *all* and *some* ($\beta$=-0.02, SE= 0.23, p=0.50), and as time increased, the bias to the target did not develop differently ($\beta$=0.02, SE=0.35, p=.06). During the modifier window (Figure 2b), log ratios were significantly higher for numbers than for quantifier conditions (*all*: $\beta$=-0.26, SE=0.22, p<.001; *some*: $\beta$=-0.42, SE=0.22, p<.001). In the modifier window, on average log ratio was higher for *all* than for *some* ($\beta$=-0.07, SE=0.22, p=.04), but the target bias was increased faster in *some* than *all* ($\beta$=0.04, SE=0.32, p<.001).

Figure 3 plotted the log ratios by quantifier type for big and small target size. For determiner window (Figure 3a), there was a three way interaction between quantifier, target size and time ($\beta$=-0.04, SE=0.71, p <.001). Analysis of the effect of target size on each level of quantifiers showed that after hearing *all*, the target bias

increased faster when the target set was big ($\beta$=0.08, SE=0.35, p<.001); whereas after hearing *some*, the target bias increased faster when the target set was small ($\beta$=-0.06, SE=0.38, p=.008). For modifier window (Figure 3b) there was a significant quantifier by target set size interaction ($\beta$=0.04, SE=0.39, p <.001). When the target set size was small, the target bias was greater after *some* than after *all* ($\beta$=-0.16, SE=0.32, p=.037); when the target set size was big, the bias was greater after *all* than after *some* ($\beta$=0.35, SE=0.27, p<.001). This results contrasts with Degen& Tanenhaus 's finding that when the target set was small, looks to the target did not differ between *some* and *all*. Our results revealed that participants expect *all* referent to be a big set, and *some* referent to be a small set.




(a) determiner window (b) modifier window

**Figure 4** Looks to the residue set for *all*, *some*, and *number*

For the second analysis, to compare looks to the residue set across three conditions, we fitted a model to predict the empirical logit from fixed effects of determiner type (all/some/numbers), a continuous time variable and their interactions. The model contained maximal by-subject random effects structure supported by the data.

Figure 4 plotted the empirical logit over each 17ms sample for *all*, *some*, and *number*. For determiner window (Figure 4a), there were significantly less looks to the residue set after hearing numbers than *all* ($\beta$=0.08, SE=0.16, p=.046). There was a marginally significant difference also between numbers and *some* ($\beta$=0.06, SE=0.15, p=.09), while looks to the residue set decreased faster in numbers than both *all* ($\beta$=0.05, SE=0.40, p=.049) and *some* ($\beta$=0.08, SE=0.55, p=.01). For modifier window (Figure 4b), there were significantly less looks to the residue set in numbers than *all* and *some* (*all*: $\beta$=0.22, SE=0.14, p<.001; *some*: $\beta$=0.25, SE=0.15, p<.001). These results provided further evidence that verification processes of the

quantificational NP differ between numbers and non-numbers. Among quantifiers, as shown in Figure 4a, in determiner window, overall looks to the residue set did not differ between *some* and *all* ($\beta$=-0.02, SE=0.15, p=.58), and as time increased, looks to the residue set decreased regardless of the quantifiers ($\beta$= 0.03, SE=0.42, p=0.17). In the modifier window, we found overall looks to the residue set did not differ between *some* and *all* ($\beta$=0.03, SE=0.13, p=.28).

### Follow up

To determine if strength of association between *all* and big sets is the same as that between *some* and small sets, we asked a separate group of participants for their judgments on which image fits better with the statement, e.g. *Mary has all/some of her sister's roses/lollies*, using a slider scale with images differing only in set size located on two ends (Figure 5). Results confirmed those in our main study: a big set preference in *all* trials, small in *some* trials (p's < 0.01). More importantly, the strength of association between *all* and big sets was significantly stronger than that between *some* and small sets (t(42)=2.78, p=.008). Then the observed delay found in previous comparisons between big-set *all* and small-set *some* is compatible with another possible explanation. That the delay in *some* is due to the difference in prior expectation rather than the pragmatic processing.

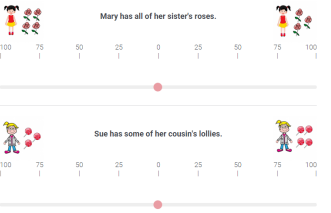



**Figure 5** sample items in follow up study

## 3 General discussion and conclusion

In our study, we found evidence that the anticipatory looks to the target could reflect the prior expectation about the relative set size given a quantifier expression. The aim of previous research is to see if participants use the enriched interpretation of *some* in compositional processes in the same timecourse as they use the plain meaning of *all*. We suggest that the best measure that controls for extraneous factors driving anticipatory looks is looks to a residue set. We obtained evidence that the general looking pattern to the residue set did not differ between enriched *some* and *all*, which suggested that the pragmatic enrichment is fast.

## References

Altmann, Gerry T.M. & Yuki Kamide. 2009. Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition* 111(1). 55 – 71. http://dx.doi.org/http://dx.doi.org/10.1016/j.cognition.2008.12.005. http://www.sciencedirect.com/science/article/pii/S0010027708002904.

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

Degen, Judith & Michael K. Tanenhaus. 2015. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science* n/a–n/a. http://dx.doi.org/10.1111/cogs.12227. http://dx.doi.org/10.1111/cogs.12227.

Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary & Michael K. Tanenhaus. 2010. "some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116(1). 42 – 55. http://dx.doi.org/http://dx.doi.org/10.1016/j.cognition.2010.03.014. http://www.sciencedirect.com/science/article/pii/S0010027710000788.

Huang, Yi Ting & Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology* 58(3). 376 – 415. http://dx.doi.org/http://dx.doi.org/10.1016/j.cogpsych.2008.09.001.

Kamide, Yuki, Gerry T.M Altmann & Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language* 49(1). 133 – 156. http://dx.doi.org/http://dx.doi.org/10.1016/S0749-596X(03)00023-8. http://www.sciencedirect.com/science/article/pii/S0749596X03000238.

Kuznetsova, Alexandra, Brockhoff Per Bruun & B.C. Rune Haubo. 2013. *lmertest: Tests for random and fixed effects for linear mixed effect. r package version 2.0-3.*

Chao Sun
Chandler House, 2 Wakefield Street
London
chao.sun.13@ucl.ac.uk

Richard Breheny
Chandler House, 2 Wakefield Street
London
r.breheny@ucl.ac.uk

# Homogeneity and enrichability affect scalar processing

Ye Tian
*Université Paris Diderot*

Chao Sun
*University College London*

Richard Breheny
*University College London*

**Abstract** Following the recent investigation by Van Tiel et al. (2014), we investigated other sources of variation in the rates of scalar inferences across scalar terms. Here we present research showing that scale homogeneity and frequency of local enrichment account for a significant proportion of the scalar diversity previously reported.

**Keywords:** Scalar enrichment, Scales

## 1 Introduction

Saying (1) *John ate some of the cookies* often implies that *John did not eat all of the cookies*. The *not all* interpretation derived from the literal meaning of the scalar expression *some* is known as a scalar inference. Recent experimental studies investigated the rates at which scalar expressions of different lexical categories generate scalar inferences (Doran et al. 2009; Van Tiel et al. 2014). These studies suggest that different scalar expressions give rise to scalar inferences (SIs) at different rates.

van Tiel et al. (2014) provided the latest evidence for the variability. They employed an inference paradigm to test participants' interpretation of statements containing scalar expressions. Several classes of scalar expressions were examined including quantifiers (e.g. <*all, some*>), modals (<*certainly, possibly*>), adjectives (<*beautiful, pretty*>) and verbs (<*dislike, loathe*>). Figure 1 is an example of a critical item (exp.2).

Participants read a statement uttered by a character. Then they were asked whether or not the speaker implied the negation of the stronger statement in which scalar expression was replaced by its stronger scale mate. For example, whether *the student is intelligent* implies *the student is not brilliant*. Participants gave yes/no judgements. Yes responses indicated that the scalar expression gives rise to an upper-bounded inference, whereas No responses indicated the lack of inference. The results showed that there was significant variation at the derivation rates of scalar




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 137–145.

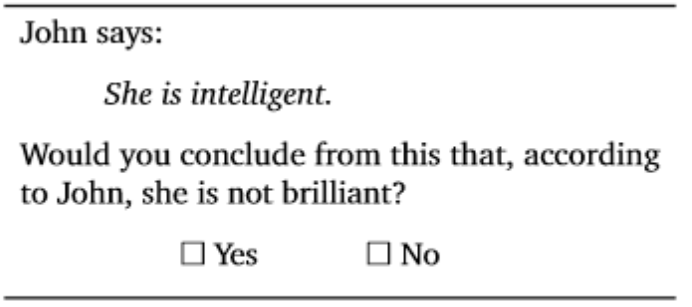



*Figure 1: Sample item used in Experiment 1.*

**Figure 1** an example of critical item in van Tiel *et al.* (2015)

inferences across different supposedly scalar expressions, ranging from 4% to 100%. Quantifiers and modal expressions generated scalar inferences more frequently than adjectives and verbs. Moreover, while quantifiers and modal expressions consistently gave rise to scalar inferences, there was much greater variability within adjectives and verbs. The authors explained the variability in terms of word class and semantic distance. Word class correlated with the derivation rates. The greater the semantic distance within each pair of scalar expressions, the higher the rate of inference derivation. However, semantic distance could only account for a relatively small amount of the variation (20%) found in the inference task, leaving a large amount of variation unexplained.

We propose two additional factors that affect the derivation rates of scalar inferences in van Tiel et al. (2014): homogeneity and frequency of local enrichment. The first factor, scale homogeneity, measures how frequently the stronger term is interpreted to be entailing the weaker term. "Scales" with low homogeneity are those whose "scalar terms" have meanings that often differ in dimensions other than quantity – that is, they are not good "scales". It is possible that "scales" in van Tiel et al. (2014) differ in homogeneity.

Under the classic Gricean account, scalar implicatures are derived by reasoning about more informative alternatives on the same scale, where the informativeness can be defined in terms of entailment relation. Considering (1) as an illustration, <all, some> scale is involved in reasoning about more informative alternatives, in this case, (1) is less informative than *John ate all the cookies*. Assuming the speaker is cooperative and following the conversational maxims, saying (1) instead of the more informative alternative implies that the quantity maxim is in tension with the

other maxims, usually the maxim of quality i.e. that you do not say what you believe to be false. Hence saying (1) implicates that the negation of the more informative alternative, that is. John did not eat all of the cookies. Under this account, an important assumption is that the alternatives must be on the same entailment scale as the given term. When this assumption is not met, no scalar implicature can be drawn. However, if this is the case, how come the rates of scalar implicatures based on different scales cover a wide range rather than being just ones and zeros? This is because word meanings often, if not always, depend on context. Scales can differ in how frequently their terms are used to communicate homogeneous senses. We hypothesize that other things being equal, the more frequently they are used homogeneously, the higher the rate of scalar implicature derivation. On one extreme, numerals should always satisfy the entailment relation. On the other extreme, there are words that almost always differ in dimensions other than scale. For example, the pair of "king" and "emperor". We suggest that entailment relation is more consistently met by quantifiers (e.g. *all/ some*) and modals (*certain/ possible*) than adjectives (*brilliant/ intelligent*). This variation should partially explain the scalar variability reported in van Tiel et al. (2014).

The second factor, frequency of local enrichment, measures how frequently a scalar term is enriched during utterance comprehension.Carston (02) suggests that if the literal meaning of a lexical item is not sufficiently relevant for the context, listeners will enrich the literal meaning to yield an optionally relevant interpretation. For example, if a speaker says "the cinema is some distance from the restaurant", the literal meaning of "some" does not make the utterance relevant enough and has to be enriched. This pragmatic process targets a particular lexical item and strengthens the concept it encodes (Carston 02; Sperber 2007), and is an alternative way to the above mentioned Gricean mechanism for inference derivation. Our conjecture is that scalar expressions differ in the relative frequency of local enrichments and the enrichability of scalar expressions, and this variability correlates with the scalar inferences derivation rates. For quantifiers (e.g. *some*) and modal expressions (e.g.*possible*), the unenriched meaning is underspecified (i.e. *at least some*, *at least possible*). In order to achieve an optimally relevant interpretation, these expressions frequently get enriched locally to have an upper bound meaning (i.e. *some but not all*, *possible but not certain*). Therefore in the inference task, quantifiers and modal expressions produced higher rates of scalar inferences. Whereas for adjectives (e.g. *intelligent*) and verbs (e.g. *start*), the unenriched meaning are on average more specific, hence compared to quantifiers and modals, these expressions are less frequently enriched and this may explain why they give rise to lower rates of scalar inferences.

## 2 The Current Study

The current study aims to investigate the variability of scalar inferences and test empirically whether homogeneity and frequency of local enrichment can account for some of the variation. In experiment 1, we replicated van Tiel *et al*.'s inference task (exp.2) using the same scalar expressions to obtain the derivation rate for each scale. Next, in experiment 2, we measured the degree of homogeneity for each scale <X, Y> investigated in experiment 1 using an offline naturalness rating task. A separate group of participants rated the naturalness for sentences of the form *X but not Y*, e.g. 2a-2c:

- 2a. The student is brilliant but not intelligent. <brilliant, intelligent>
- 2b. The weather is hot but not warm. <hot, warm>
- 2c. The dancer finished but she did not start. <finish, start>

Given that *but* has a denial-of-expectation or contrastive conventional implicature, an *X but not Y* sentence is felicitous to the extent that X can be construed to not strictly entail Y but normally or often would imply Y, or be expect to co-occur with Y. A scale with high homogeneity is one where the stronger term is interpreted to entail the weaker term. Entailment relations require that if X entails Y, whenever X holds, Y must hold. Thus these *X but not Y* sentences should be very unnatural if the contrasting predicates X and Y are on the same entailment scale. So if naturalness rating for 'but' sentences is low, it suggests a higher degree of homogeneity for the given scale; whereas if the rating is high, then the degree of homogeneity is relatively low. We predicted that the naturalness rating for scalar expressions in 'but' task should negatively correlate with the inference task results.

In experiment 3, we measured the frequency of local enrichment for each scale. Another separate group of participants rated the naturalness for sentences of the form *X so not Y*, e.g. 3a-3c.

- 3a. The student is brilliant so not intelligent. <brilliant, intelligent>
- 3b. The weather is hot so not warm. <hot, warm>
- 3c. The dancer finished so she did not start. <finish, start>

The discourse function of *so* contrasts with that of *but* in a number of ways (Blakemore, 2002). *So* implies that the second segment follows in some way from the first. So, while *X but not Y* suggests that one might expect Y, given X, *X so not Y* suggests that one might expect not Y, given X. Thus, *X so not Y* sentences should be more coherent to the extent that the weaker scalar expression can be locally enriched

to have an upper bound meaning. For example, to understand 3b, *warm* must be enriched to mean *just warm*. Notice that this process has to be local enrichment rather than Gricean reasoning because the weaker term is in the scope of negation. The negation of an un-enriched weaker term is in fact more informative than the negation of an enriched weaker term (*not at least warm* is more informative than *not warm-but-not-hot*). In our task, if the naturalness rating for 'so' sentences is low, it suggests a lower frequency of enrichments for the scalar expression; whereas if the rating is high, then the frequency of enrichments is relatively high. We predicted that the naturalness rating for scalar expressions in the 'so' task should positively correlate with the inference task results.

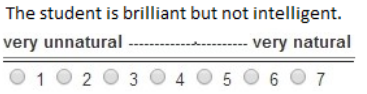
**Experiment 1 Replicating van Tiel et al.'s inference task**

We tested the derivation rates of scalar expressions investigated in van Tiel et al. (2014) via a similar inference paradigm. Instead of providing a yes/no response, participants rated on a 0-100 scale to indicate to what extent they could infer from the speaker's statement that the speaker do not believe the stronger alternative. There were 43 experimental items. 7 filler items were created guided by the design discussed in van Tiel et al. (2014). 32 native English speakers participated the experiment.

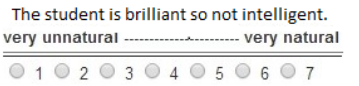
We carried out one-way ANOVAs with the ratings on the inference task as the dependent variable and lexical categories (word class) as the independent variable. There was a main effect of lexical categories ($F1(3, 93) = 32.20$, $p < .001$; $F2(3,39) = 9.46$, $p < .001$). Pairwise comparisons showed that the ratings of scalar inference for adjectives (M= 35.32) were significantly lower than for quantifiers (M=74.67, p<.001) and modals (M=62.79, p<.001). Meanwhile, ratings for verbs (M= 35.25) were significantly lower than for quantifiers (p<.001) and modals (p<.001). Neither the difference between adjectives and verb (p=.985) nor the difference between quantifiers and modals (p=.053) was significant. Our results replicated the general response pattern in van Tiel *et al.*'s inference task.

**Experiment 2 'But' task**

Figure 2(a) is an example item. We used the scales < X, Y> investigated in experiment 1 to construct sentences for the 'but' task. The experiment sentences were of the form *X but not Y*, where according to van Tiel et al. (2014), X and Y are a pair of scalar terms and X is stronger than Y. There were 43 experimental sentences and 7 fillers. Fillers contained clearly felicitous (e.g. *The banker is rich but not happy*) and clearly infelicitous items (e.g. *The man left the party but he never came*).




(a) experiment 2




(b) experiment 3

**Figure 2** Example items

20 Participants rated how natural these constructions are on a 1 (very unnatural) -7 (very natural) scale.

We carried out one-way ANOVAs with the naturalness ratings of 'but' task as dependent variable and lexical categories (word class) as independent variable to examine whether the degree of scales homogeneity differs across lexical categories. There was a main effect of lexical categories ($F1(3, 57) = 15.19$, $p < .001$; $F2(3,39) = 4.56$, $p = .008$). Pairwise comparisons showed that the naturalness ratings for adjectives ($M= 2.78$) were significantly higher than for quantifiers ($M=1.67$, $p<.001$), modals ($M=1.98$, $p<.001$) and verbs ($M= 2.25$, $p<.01$). Ratings for verbs ($M= 2.25$) were significantly higher than for quantifiers ($p<.003$). But neither the difference between quantifiers and modals ($p=.053$) nor the difference between modal and verb ($p=.068$) was significant. These results suggested that entailment relation is more consistent among quantifiers and modals than adjectives. More importantly, we found a significant negative correlation (figure 3) between the rating of the 'but' task and the inference task ($r=-.341$, $p=.025$). Moreover, a linear regression analysis was conducted to predict inference task results from the rating of 'but', the result showed that the 'but' task accounted for a significant amount of the variance of the inference data ($R^2=.12$, $F(1,41)= 5.40$, $p=.025$). These results confirmed our prediction and provided evidence that the degree of scale homogeneity partially explains the variations.

### Experiment 3 'So' task

Figure 2(b) is an example item. We used the same scales as in experiment 1 to construct sentences for 'so' task. The experiment sentences were of the form *X so not Y*, where X is stronger than Y. There were 43 experimental sentences and 7 fillers. Fillers contained clearly felicitous (e.g. *The cup is red so not blue*) and clearly infelicitous items (e.g. *The banker is rich so not happy*). 20 Participants rated how natural these constructions are on a 1 (very unnatural) -7 (very natural) scale.

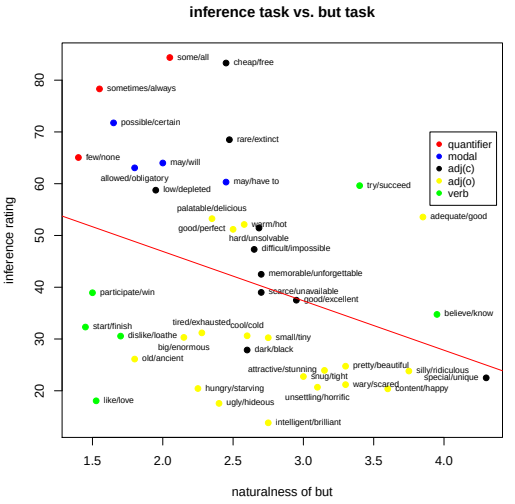



**Figure 3** Negative correlation between the absence of homogeneity and inference rate

One-way ANOVAs with the naturalness ratings of the 'so' task as the dependent variable and lexical categories (word class) as the independent variable showed that there was a main effect of lexical categories ($F1(3, 57) = 52.37$, $p < .001$; $F2(3,39) = 9.40$, $p < .001$). Pairwise comparisons showed that the naturalness ratings for adjectives (M= 2.90) were significantly lower than for quantifiers (M=5.05, p<.001). Ratings for verbs (M= 1.88) were significantly lower than quantifiers (p<.001), modals (M=3.14, p<.001) and adjectives (p<.001). Quantifiers yielded significant higher rating than models (p<.001). The difference between modals and adjectives was not significant (p=.22). These results indicated that quantifiers are most likely to be enriched locally, followed by modals. Adjectives and verbs are less likely to be enriched locally, in fact verbs have the lowest frequency of local enrichment. Furthermore, we found a significant positive correlation (figure 4) between the rating of the 'so' task and the inference task (r=.417, p=.005).

Combining data from experiment 2 and 3, we found that the ratings of the 'but' and the 'so' tasks did not correlate (r= .063, p= .69), suggesting that they each account for a distinct part of the variability. A multiple regression analysis was conducted to predict inference task results from the rating of 'but' and 'so' task. The results indicated that together, 'but' and 'so' task accounted for a significant amount of the variance of the inference data ($R^2$=.30, F(2,40)= 9.77, p<.01).

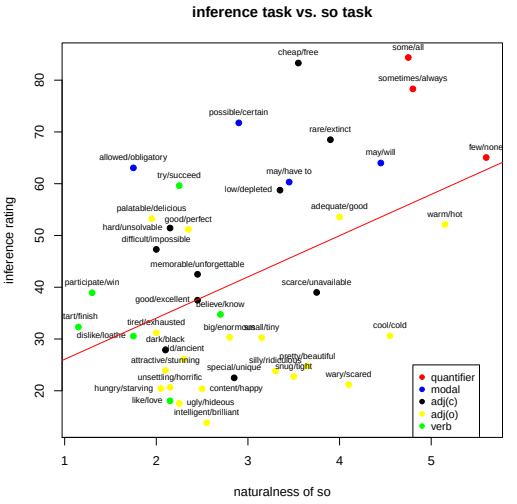



**Figure 4** Positive correlation between the frequency of local enrichment and the inference rate

## 3 General discussion and conclusion

Following the recent investigation by van Tiel et al. (2014), we investigated other sources of variation in the rates of scalar implicatures across scalar terms. In experiment 1, we replicated the results of van Tiel *et al.*'s inference task that different scalar expressions generate scalar inferences (SIs) at different rates. We suggest that the variability observed in experiment 1 can be partially explained by two new factors: scale homogeneity and frequency of local enrichment. Scale homogeneity measures how frequently the stronger term is interpreted to be entailing the weaker term. We propose that the more frequently this entailment relation is met, the higher the rate of scalar implicature derivation. This was tested in experiment 2 using sentences such as *John ate all but not some of the cookies*. The naturalness of these sentences negatively reflects the degree of scale homogeneity. We found that the degree of homogeneity varied across different scales, and the naturalness of the 'but' sentences correlated negatively with the inference task results. The second factor, frequency of local enrichment measures how frequently a scalar term is enriched during utterance comprehension. We propose that the more frequently the weaker term is enriched, the higher the rate of scalar implicature derivation. This was tested in experiment 3 using sentences such as *John ate all so not some of the cookies*. The naturalness of these sentences positively reflects the frequency of local enrichment. We found that the frequency of local enrichment were different among scalar expressions. Some scalar expressions investigated in experiment 1

are more frequently enriched to have an upper bound meaning. The frequency of enrichment positively correlated with the inference task results. Combining data from experiment 2 and 3, we found that the ratings of the 'but' sentences and 'so' sentences do not correlate, suggesting that the two experiments measured different aspects of the variation. We conclude that scale homogeneity and frequency of local enrichment account for a significant proportion of the scalar diversity previously reported. Both factors capture variability separate from the mechanism of genuine scalar implicature derivation. It suggests that the mechanism of genuine SI derivation does not contain as much diversity as previously suggested.

## References

Carston, Robyn. 02. *Thoughts and utterances: The pragmatics of explicit communication*. John Wiley & Sons.

Doran, Ryan, Rachel E. Baker, Yaron McNabb, Meredith Larson & Gregory Ward. 2009. On the non-unified nature of scalar implicature: An empirical investigation. *International Review of Pragmatics* 1(2). 211 – 248. http://dx.doi.org/10.1163/187730909X12538045489854.

Sperber, Ira Noveck Dan. 2007. The why and how of experimental pragmatics: The case of 'scalar inferences'. In Noel Burton-Roberts (ed.), *Advances in pragmatics*, Palgrave.

Van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2014. Scalar diversity. *Journal of Semantics* http://dx.doi.org/10.1093/jos/ffu017. http://jos.oxfordjournals.org/content/early/2014/12/22/jos.ffu017.abstract.

Ye Tian
Salle 751, Olympe de Gouges, Rue Albert Einstein
Paris
tiany.03@gmail.com

Chao Sun
Chandler House, 2 Wakefield Street
London
chao.sun.13@ucl.ac.uk

Richard Breheny
Chandler House, 2 Wakefield Street
London
r.breheny@ucl.ac.uk

# Processing pragmatic inferences

Bob van Tiel

## 1 Introduction

One assumption that permeates much of the literature is that pragmatic inferences can be categorised into a number of well-defined classes. Each of these classes is associated with a set of criteria that a proposition has to fulfill to be categorised as a bona fide member of that class: presuppositions must project, conversational implicatures must be cancellable, calculable, and nondetachable, conventional implicatures must be lexicalised and truth-conditionally inert, et cetera.

Insofar as this assumption is on the right track, one might wonder whether these theoretical distinctions have some kind of correspondence in behavioural data. In other words, one might wonder whether different kinds of pragmatic inferences are associated with differences in response patterns and reaction times. In order to evaluate this *correspondence hypothesis*, I compared eight types of pragmatic inferences within a verification task. This contribution is based on previous work I did in collaboration with Walter Schaeken (van Tiel & Schaeken 2015).

## 2 Eight kinds of pragmatic inferences

In this verification task, participants read sentences that were followed by pictures. It was their task to indicate whether the sentence was an appropriate description of the subsequent picture. In the target condition, the sentence was followed by a picture that falsified an inference that might have been triggered by the sentence. I assume that participants who compute the inference judge the sentence 'false' and participants who do not judge the sentence 'true'. In addition, each sentence type was presented in control situations in which it was unambiguously true or false. Figure 1 shows for each inference type an example sentence, its hypothesised pragmatic inference, and a target situation in which the sentence is true on its unenriched interpretation but in which the hypothesised inference is false.

In total, we tested eight types of pragmatic inferences. First, three types of scalar inferences; one based on the ⟨or, and⟩ scale and two based on the ⟨some, all⟩ scale. Second, three types of inferences that are often explained as varieties of quantity implicature: distributivity inferences, conditional perfection, and exhaustivity in 'it'-clefts. A more extensive discussion of these inferences can be found in van Tiel & Schaeken (2015). Third, the uniqueness presupposition associated with definite




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 146–152.

descriptions. Lastly, a syntactically ambiguous sentence for which the stronger surface scope reading entails the weaker inverse scope reading.

If the correspondence hypothesis is correct, it is expected that the theoretical boundaries between these four groups of inference types will be visible in truth judgements, verification times, or perhaps even both.

## 3 The experiment

### 3.1 Participants

40 students at the Université Libre de Bruxelles, all native speakers of French, participated in the experiment for financial compensation (33 females, mean age: 21, range: 18-28).

### 3.2 Materials

The experiment consisted of 62 items and included eight types of sentences corresponding to the eight types of pragmatic inferences. For each inference type, three kinds of situations were constructed: two control situations and one target situation. In the first control situation, the sentence was unambiguously true; in the second control situation it was unambiguously false; in the target situation its truth value depended on whether the pragmatic inference was derived. See Figure 1 for examples of target situations.

Target situations occurred four times for each inference type; control situations twice. The order of the items was randomized for each participant.[1]

### 3.3 Procedure

On each trial, the target sentence was displayed first. Participants were instructed to press the space bar as soon as they had read and understood the sentence. Thereupon, the sentence disappeared and was replaced by a picture. Participants had to decide as quickly as possible whether the sentence was true or false as a description of the depicted situation, and had to register their decision by pressing one of two keys. Thereupon, the picture disappeared and was replaced by the message '(Press the space bar to continue.)'. Upon pressing the space bar, the next trial commenced.

1 I included four other types of pragmatic inferences. For various reasons, these had to be excluded from the analysis.

**Scalar inference 'some'**
Some of the shapes are red.
⇝ Not all of the shapes are red.

**Scalar inference 'some not'**
Some of the shapes are not red.
⇝ Some of the shapes are red.

**Scalar inference 'or'**
The square or the circle is red.
⇝ Not both of them are red.

**Distributivity inference**
Each of the shapes is red or green.
⇝ Not all of the shapes are red.

**Conditional perfection**
Each shape is red if it is a circle.
⇝ Not all of the shapes are red.

**Exhaustivity in 'it'-clefts**
It is the circle that is red.
⇝ Only the circle is red.

**Uniqueness presupposition**
The circle is not red.
⇝ There is only one circle.

**Syntactic ambiguity**
All of the shapes are not red.
⇝ None of the shapes are red.

**Figure 1** Target sentences followed by their pragmatic inferences and a situation that falsifies the inference but verifies the unenriched interpretation of the target sentence.

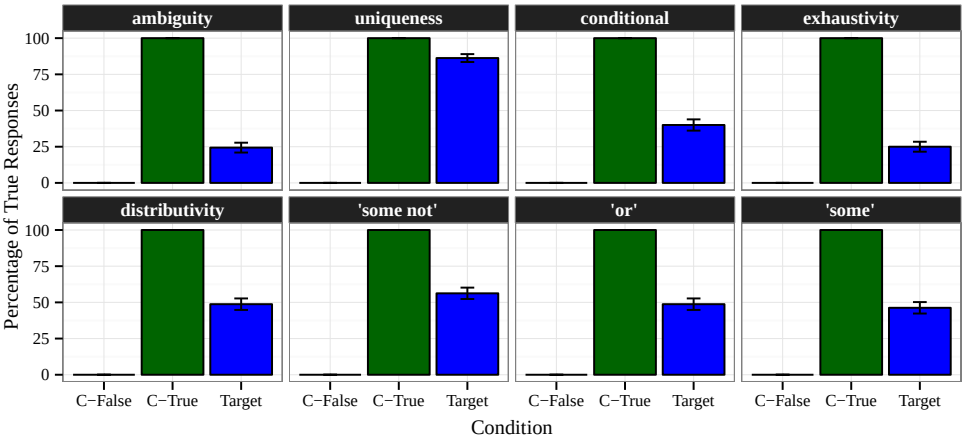



**Figure 2** Percentages of 'true' responses for each type of pragmatic inference. Error bars represent standard errors.

### 3.4 Data treatment

All participants and trials were included in the analysis. Decision times were logarithmised to reduce the skewness of their distribution.

### 3.5 Choice proportions

The percentages of 'true' responses for each inference type are summarised in Figure 2. For scalar inferences (46% 'true' responses for 'some', 49% for 'or', and 56% for 'some not'), conditional perfection (40%), and distributivity inferences (49%) 'true' responses were roughly as frequent as 'false' responses. Participants in the case of exhaustivity in 'it'-clefts (25%) and syntactic ambiguities (24%) had a pronounced preference for 'false' responses. Conversely, in the case of uniqueness presuppositions (86%), there was a pronounced preference for 'true' responses.

In general, error rates were quite low, with one exception: in the case of 'some not', many participants judged the sentence false in a situation in which it was true on both its literal and enriched reading. I do not have a straightforward explanation for this anomaly. Perhaps participants considered the sentence too unspecific to be considered an appropriate description of the situation.

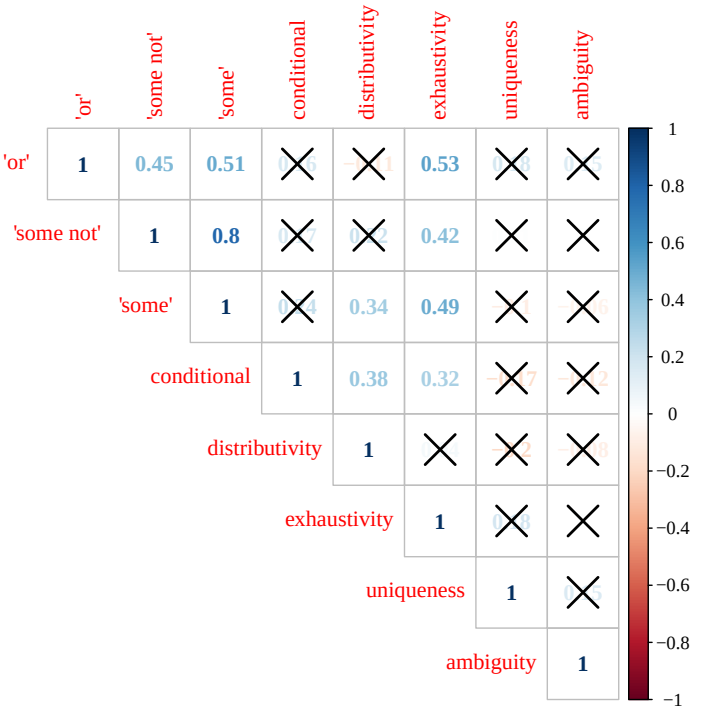



**Figure 3** Product-moment correlation coefficients for each pair of inference types. Correlations for which $p > .05$ are crossed out.

### 3.6 Correlations

One might wonder whether participants were consistent across different inference types, i.e. whether participants gave a comparable number of 'false' responses for each inference type. To test this, we calculated the product-moment correlation between the number of 'false' responses for each pair of inference types. The results of this analysis are summarised in Figure 3.

There were significant correlations between each pair of scalar inferences, between exhaustivity in 'it'-clefts and each kind of scalar inferences, between exhaustivity in 'it'-clefts and conditional perfection, between conditional perfection and distributivity inferences, and between distributivity inference and scalar inferences associated with 'some'. None of the other correlation coefficients were significant.

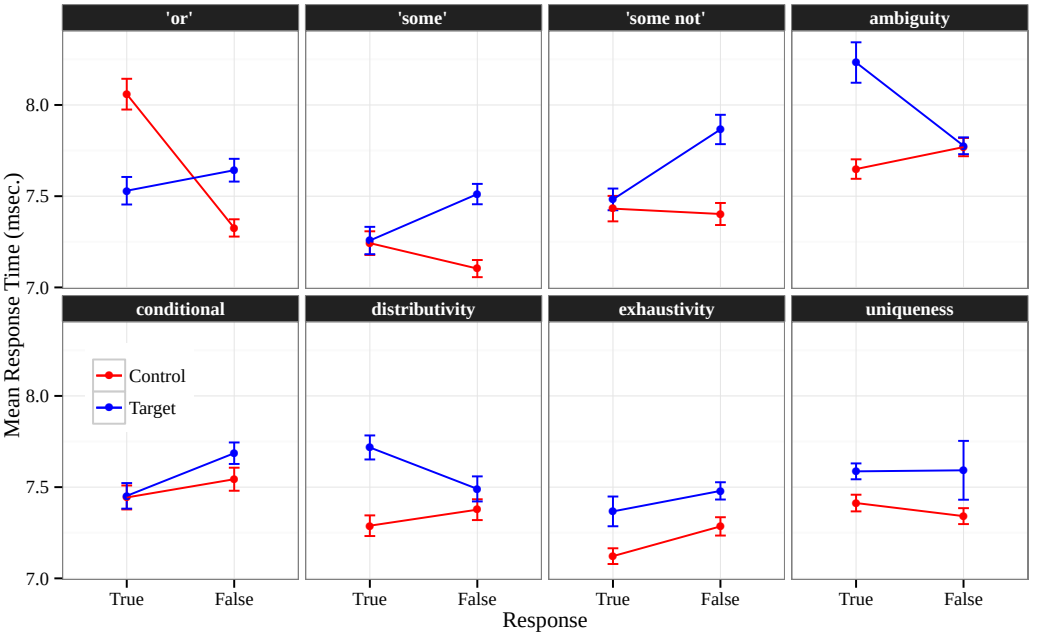



**Figure 4** Mean logarithmised decision times for each type of pragmatic inference. Error bars represent standard errors.

### 3.7 Decision times

Decision times refer to the time from the onset of the picture to the button press indicating the truth judgement. For the analysis of the decision times, all error trials were removed. Mean logarithmised decision times are summarised in Figure 4.

Mixed models were constructed predicting logarithmised decision times based on condition (target or control) and response ('true' or 'false'), including random intercepts for participants and items. Degrees of freedom and corresponding $p$-values were calculated based on Satterthwaite's procedure as implemented in the `lmertest` package. I assume that an interaction between condition and response indicates that the computation of a pragmatic inference facilitated or delayed decision times.

The interaction between condition and response was significant for each kind of scalar inference: 'or' ($\beta$ = -0.77, $SE$ = 0.26, $t$ = -3.00, $p$ = .022), 'some' ($\beta$ = -0.34, $SE$ = 0.11, $t$ = -3.00, $p$ = 0.03), and 'some not' ($\beta$ = -0.45, $SE$ = 0.12, $t$ = -3.6, $p < .001$). It was also significant, but in the opposite direction, for distributivity inferences ($\beta$ = 0.31, $SE$ = 0.11, $t$ = 2.76, $p$ = .006) and ambiguous sentences ($\beta$

= 0.45, $SE$ = 0.11, $t$ = 4.09, $p < .001$). It was not significant for any of the other inference types (all $p$'s $> .2$).

In summary, the computation of scalar inferences was associated with a processing cost. Calculating distributivity inferences and arriving at the stronger surface scope reading of the ambiguous sentence facilitated decision times. There were no effects on decision times for the computation of conditional perfection, exhaustivity in 'it'-clefts, and the uniqueness presupposition of definite descriptions.

## 4 Discussion

The correspondence hypothesis predicts that theoretical distinctions between different types of pragmatic inferences are reflected in experimental data, in casu the results of a verification task. This hypothesis was partly confirmed. Scalar inferences behaved relatively homogenously: participants were highly consistent across different kinds of scalar inferences and their computation was associated with a delay in decision times, unlike other kinds of pragmatic inferences.

To some extent, the three varieties of quantity implicature also patterned together, as can be seen in the correlation plot in Figure 3. However, this pattern was clearly less strong and was not associated with a particular processing signature. The uniqueness presupposition and syntactically ambiguous sentence behaved differently from the three types of quantity implicature in terms of choice proportions and consistency but not in terms of verification times.

Hence, although experimental data can provide an insight into the provenance of a type of pragmatic inference, there does not seem to be a specific litmus test to arrive at a decisive verdict. Rather, data from different measures should be taken into consideration. In order to learn more about what measures are relevant for which types of pragmatic inference, however, it will be necessary to extend the scope to different kinds of presuppositions and ambiguities.

## References

van Tiel, Bob & Walter Schaeken. 2015. Processing conversational implicatures: alternatives and counterfactual reasoning. To appear in: *Cognitive Science*.

# Obligatory and optional focus association in sentence processing

Barbara Tomaszewicz
*University of Cologne*
*University of Wrocław*

Roumyana Pancheva
*University of Southern California*

**Abstract** Formal semantic research on focus has identified two types of focus association: *obligatory* and *optional*. Obligatory focus association is taken to be encoded in the lexical semantics of focus sensitive expressions, whereas optional/free association is a result of the contextual setting of the restrictor of a quantificational operator (Beaver and Clark 2008). We conducted three experiments comparing the processing of focus structures with obligatorily associating *only* and *even* (Rooth 1985, Tancredi 1990, Krifka 1992, a.o) and optionally associating *many* (Herburger 1997) and *most* (Heim 1999). The results indicate that both obligatory and optional associators create a processing bias for narrow focus.

## 1 Two types of focus association

A focus associator is an expression whose contribution to the meaning of a sentence depends on the position of sentence focus. In (1) and (2) the exclusive particle *only* and the adverb *always* associate with focus in different positions (realized as prosodic prominence) which results in different truth conditions for the (a) and (b) versions.

(1) a. John *only* bought [MAry]$_F$ a cake.
'John bought no one else but Mary a cake.'
b. John *only* bought Mary a [CAKE]$_F$.
'John bought Mary nothing else but a cake.'

(2) a. In St. Petersburg, officers *always* escorted [balleRInas]$_F$. (Rooth 1992, 1996)
'Whenever officers escorted somebody, they escorted ballerinas.'
b. In St. Petersburg, [Officers]$_F$ *always* escorted ballerinas.
'Whenever ballerinas were escorted by somebody, they were escorted by officers.'

Quantificational operators such as *only* and *always* carry an (implicit) domain argument and when these operators *associate with focus* the value of the domain variable is set with respect to the focus structure of the sentence (Rooth 1992, 1996, von Fintel 1994). The focus structure of a sentence introduces a presupposition that a set of relevant alternatives is retrievable from the context – these alternatives contribute to the restriction of the domain of quantification. In (1a) the focal presupposition is that there is a set of individuals for whom John bought a cake, and the value of the domain variable of *only* is the set of alternative propositions of the form *John bought x a cake*. In (1b) *only* quantifies over propositions *John bought Mary f(x)* because focus contributes alternatives to the property *f* that holds of the individual *x* that John bought for Mary. In (2) the restriction of *always* is the set of times whose value varies with the focus alternatives as indicated by the paraphrases.

The mechanism of focus association is based on the fact that the focus structure of a sentence is a result of discourse congruence. Focus is licensed if it is congruent with the current question under discussion, (3a) vs. (3b), that is either explicit or implicit (salient in the current discourse).

(3) Who did John buy a cake?
   a. John bought [MAry]$_F$ a cake.
   b. #John bought Mary a [CAKE]$_F$.

The focus effects on quantifier domains are a result of the anaphoric dependence of the quantifier's domain variable on the same background context that licenses focus (or more specifically, Rooth's (1992) focus interpretation operator ~).[1] This approach predicts that irrespective of focus, the context can constrain the value of the domain variable. Beaver and Clark (2008) argue that this is true for quantificational adverbs such as *always* but not for the exclusive *only*. Consider the following diagnostic: given the context in (4) and the focus on the verb 'salutes', *always* can associate with the elided verb in the *because*-clause, (4a), but *only* cannot, (4b). In (4a) the restriction of *always* is the set of times at which Sandy is at a ceremony and this set is established purely on the basis of the discourse context. *Only* cannot associate with a semantic focus for its restriction to be contextually set, it requires a focus that is phonologically realized.

(4) Context: At the ceremony, some soldiers salute and others fire a round in the air. Some do both. What about Kim and Sandy?
   a. Kim always [salutes]$_F$ because Sandy *always* does.
      (can mean: 'Kim salutes at every ceremony because Sandy salutes at every ceremony.')
   b. *Kim only [salutes]$_F$ because Sandy *only* does.
      (cannot mean: 'Kim salutes (and does nothing else) because Sandy salutes (and does nothing else).')

(Beaver and Clark 2008: 178)

Furthermore, when a phonological focus is present, *only* must associate with it, but *always* does not have to. The focal presupposition in (5) is that there is a set of alternatives to exams which Mary managed to complete, but in the setting of the domain of *always* this presupposition can be ignored, and instead the presupposition of the verb 'manage' (that Mary took exams) can be used, (5a). The resolution of the domain with respect to the focal presupposition results in the reading in (5b) which is dispreferred. In the same context, *only* must associate with the focus on 'exams', (6).

(5) Mary *always* managed to complete her [exams]$_F$.
   a. 'Whenever Mary took exams, she completed them.'
   b. ?'Whenever Mary completed something, it was invariably an exam.'

(6) Mary *only* managed to complete her [exams]$_F$.
   a. *'What Mary did when taking exams was complete them and do nothing else.'
   b. 'What Mary completed was an exam and nothing else.'

(Beaver and Clark, 2008: 204)

Since the pragmatic theory of focus association always allows the domain variable to be contextually resolved, for cases where operators obligatorily associate with the phonological focus, focus association needs to be lexically encoded (Rooth 1992). Beaver and Clark (2008) argue that focus association is not a uniform phenomenon resulting from discourse congruence, but that *obligatory focus association* is specified in the lexical semantics of operators such as exclusives (*only*, *merely*), additives (*too, also*), scalar additives (*even*) or intensives (*really, truly*). *Optional* or *free focus association* is a result of the contextual setting of the restrictor of operators such as quantificational

[1] The *presuppositional theory of focus* of Rooth (1992, 1996) contrasts with the *structured meanings* approach (Stechow 1991, Krifka 1991 a.o.), where focus results in the restructuring of a proposition which can be directly accessed by an operator.

adverbs (*always*, *usually*), quantificational determiners (*every, many*) or the superlative morpheme *-est*.

The distinction between the two types of focus association predicts that in the absence of a discourse antecedent that licenses narrow focus, optional associators can be restricted by purely pragmatic relevance, but obligatory associators will require the interpretation of a sentence constituent as narrow focus (licensed by implicit context). What are the empirical consequences of this distinction? How do the two types of focus associators affect expectations in online sentence processing? Obligatory associators can be predicted to create an expectation for the presence of focus in their scope during incremental semantic processing, but what happens in the case of optional associators? In the next section we outline the prior-experimental results showing that o*nly* facilitates the processing of focus structure as containing a narrow focus, and in Section 3, we present our three experiments comparing obligatory and optional associators.

## 2 *Only* and focus processing

In a silent reading study, Stolterfoht et al. (2007) compared the effects of the presence/absence of *only* on the integration of focus in the following context and established an ERP signature for 'focus structural revision' (positivity at 350–1100 ms). Their experiment design is based on the assumption that when a sentence is read without a preceding context, it receives a wide focus reading, i.e. the entire sentence is in focus as all of it is new information, (7). This assumption is supported by experimental evidence from Birch & Clifton (1995), Bader & Meng (1999), Stolterfoht & Bader (2004).

(7) [John bought cakes]$_F$.
(8) John *only* bought [cakes]$_F$.

When only is added to the sentence as in (8), during silent reading narrow focus is assigned to *cakes* for two reasons: (i) *only* requires a focus associate, and (ii) the constituent *cakes* is the default location for the nuclear stress. Stolterfoht et al. (2007) hypothesized that the wide focus reading assigned to sentences silently read out of context will have to be revised if the following context requires a narrow focus interpretation in the preceding discourse. When (9a) is read without preceding context, the first conjunct receives a wide focus interpretation (as marked in red), and when the processor encounters the ellipsis remnant in the second conjunct (in blue), it must revise the focus structure of the first conjunct from wide to matching narrow focus (marked in green), in order to license ellipsis. The presence of *only* in (9b) requires narrow focus on its associate (in green), which is congruous with the ellipsis remnant. The Accusative case marking guarantees that any differences in ellipsis resolution can be attributed to the processing of focus structure.

(9)a. [Am Dienstag hat der Direktor [den Schüler]$_F$ getadelt]$_F$, und nicht [den Lehrer]$_F$.
On Tuesday has the principal.$_{Nom}$ the pupil.$_{Acc}$ criticized and not the teacher.$_{Acc}$
b. Am Dienstag hat der Direktor *nur* [den Schüler]$_F$ getadelt, und nicht [den Lehrer]$_F$.
On Tuesday has the principal.$_{Nom}$ *only* the pupil.$_{Acc}$ criticized and not the teacher.$_{Acc}$

The reanalysis hypothesized for (9a) was correlated with an ERP signature, a bilateral sustained positivity (350–1100 ms) at the ellipsis remnant, which was absent in contexts like (9b), where, as hypothesized, only induced the narrow focus interpretation. Thus, Stolterfoht et al. interpreted the positivity as an ERP signature for 'focus structural revision', required by the revision from a wide to narrow focus reading.

Carlson (2013) hypothesized that the facilitative effect of *only* that was found in the Stoltefoht et al. ERP study should also be reflected in reading times. She found the effect with English materials parallel to the German materials of Stoltefoht et al., which she attributed to only generating an

expectation for focus alternatives. In the absence of case marking in English *only* played a disambiguating role, indicating the antecedent for the contrastive ellipsis, therefore the increased reading times at the ellipsis site in the absence of *only* may not have been associated with just the revision of focus structure.

## 3 *Only* and *even* vs. *most* and *many*

The use of Polish allowed us for a direct comparison between obligatory associating *only* and *even* (Rooth 1985, Tancredi 1990, Krifka 1992, a.o) and optionally associating *many* (Herburger 1997) and *most* (Heim 1999) for two reasons: (i) the case marking indicates that 'sculptors' is the syntactic associate in (10)-(13), therefore any differences in ellipsis resolution can be attributed to the processing of focus structure; (ii) in Polish the focus on 'sculptors' in (13) yields a superlative reading that is unavailable in English or German (as indicated by the translation). Pancheva and Tomaszewicz (2012) propose that this reading arises via focus association.

(10) [Fotografowie ucałowali [rzeźbiarzy]$_F$ na powitanie]$_F$, a nie [malarzy]$_F$
photographers.$_{Nom}$ kissed sculptors.$_{Acc}$ forgreeting and not painters.$_{Acc}$
'Photographers kissed sculptors for greeting, and not painters.'

(11) Fotografowie ucałowali ***tylko*** [rzeźbiarzy]$_F$ na powitanie, a nie [malarzy]$_F$
photographers.$_{Nom}$ kissed **only** sculptors.$_{Acc}$ for greeting and not painters.$_{Acc}$
'Photographers kissed only sculptors for greeting, and not painters.'

(12) Fotografowie ucałowali ***najwięcej*** [rzeźbiarzy]$_F$ na powitanie, a nie [malarzy]$_F$
photographers.$_{Nom}$ kissed **most** sculptors.$_{Gen}$ for greeting and not painters.$_{Gen}$
'Photographers kissed more sculptors for greeting than anybody else, and not painters.'

(13) Fotografowie ucałowali ***wielu*** [rzeźbiarzy]$_F$ na powitanie, a nie [malarzy]$_F$
photographers.$_{Nom}$ kissed **many** sculptors.$_{Gen}$ for greeting and not painters.$_{Gen}$
'Photographers kissed many sculptors for greeting, and not painters.'

### *3.1 Experiment I*

In Experiment I, we tested whether *only* creates a bias towards a contrast set (Sedivy 2002) detectable with offline measures. The contrastive function of *only* alone could account for the facilitation of ellipsis resolution in prior studies. Participants ($n$=36) were asked to fill in the gap in the position of 'painters' in (10)-(13). We expected that participants would be biased towards a continuation related to the most recently read phrase (the adverbial). We found this effect in the Baseline and *many* conditions but not with *only* and *most*, (14). This suggests that focus structural information has an effect on the interpretation of contrastive ellipsis.

(14) *Experiment I. Proportion of responses for adverbial and object continuations*

| | Baseline (10) | *Only* (11) | *Most* (12) | *Many* (14) | Effect of Sentence Type ANOVAs & Planned Comparisons |
|---|---|---|---|---|---|
| Adverbial Continu-ation | 0.523 | 0.125 | 0.421 | 0.463 | F1(3,105)=32.25, $p$<.0001; F2(3,69)= 22.64, $p$<.0001; Baseline vs. *Most* F1(1,35) = 7.07, $p$=.01; F2(1,23) = 3.32, $p$=.08; Baseline vs. *Only* F1(1,35) = 64.14, $p$<.0001; F2(1,23) = 81.79, $p$<.0001; |
| Object Continu-ation | 0.384 | 0.832 | 0.509 | 0.444 | F1(3,105)= 39.54, $p$<.0001; F2(3,69)= 27.2, $p$<.0001; Baseline vs. *Most* F1(1,35) = 7.48, $p$=.01; F2(1,23) = 6.8, $p$=.02; Baseline vs. *Only* F1(1,35) = 72.84, $p$<.0001; F2(1,23) = 109.85, $p$<.0001; |

### *3.2 Experiment II*

In a moving-window self-paced reading task ($n$=36) (run in Linger by Doug Rohde), we tested whether both *only* and *most* would result in shorter RTs on the ellipsis site than in the Baseline and *many* conditions. The conditions and materials were identical as in Experiment I; with 24 experimental items and 36 fillers. The presence of the adverbial mitigated against any recency effects where the last constituent of the clause creates an expectation for the upcoming discourse. A phrase consisted either of a single word, of a preposition and a noun, or of all the words making up a connective. After each sentence a comprehension question requiring a 'yes' or a 'no' response was displayed.

In a self-paced reading task longer reading times for a particular region compared across the conditions are interpreted as reflecting a higher level of processing difficulty. We predicted a difficulty in the processing of the ellipsis in the baseline condition that should be manifested in longer reading times of the ellipsis remnant than in the *only* and *most* conditions. *Many* could either pattern with the baseline (given the results of Experiment I), but as an optional associator, it could also pattern with *most*. The Polish counterpart of *many* that we used, *wielu*, allows for the proportional, (15a), and the cardinal reading, (15b); the latter results from association with focus (Herburger 1997). Additionally, *wielu* carries the masculine gender morpheme *-u*, which makes it different from the non-agreeing *only* and *most* (we only used masculine nouns in the experimental items).

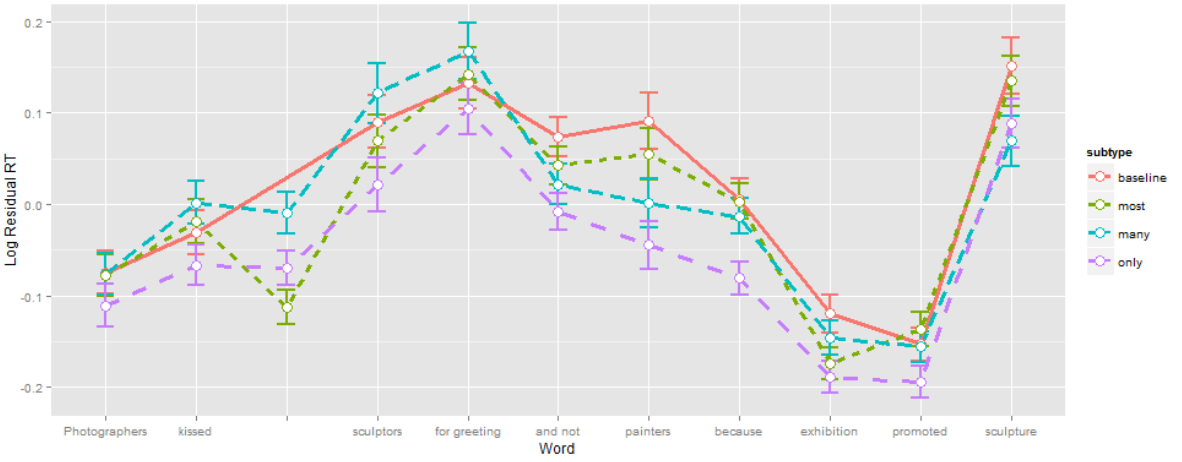
(15) a. Photographers kissed *many* sculptors for greeting.
'Photographers kissed a large proportion of sculptors for greeting.'
b. Photographers kissed *many* [sculptors]$_F$ for greeting.
'Of those kissed by photographers for greeting sculptors were a large group.'

We, therefore, predicted the following contrasts: (i) *only* is faster than the baseline; (ii) *most* is faster than the baseline; (iii) *many* is possibly faster than the baseline. We did not have specific predictions for direct comparison between *only, most, many*. The key goal was to demonstrate that there is no slow-down in the *most* (and *many*) conditions, as opposed to the baseline, which can be attributed to focus if the same experiment shows a facilitative effect of focus with *only* vs. baseline.

The comparison of average log-transformed residual reading times at each word position[2] shows that while there are no significant differences in RTs in the first clause (as visualized by the overlapping error bars in the plot in (16)), at the conjunction 'and not' and at the ellipsis site 'painters' both *only* and *many* conditions are significantly faster than the baseline, (17). *Only* remains significantly faster than the baseline on the first and second word of the spillover region ('because exhibition promoted sculpture'). *Most* becomes significantly faster than the baseline at the word 'exhibition', (17).

[2] We reported the results of Experiment II at the AMLaP 2015 and LCQ2015 conferences (Tomaszewicz et al. 2015a,b) using residual reading times obtained from trimmed raw RTs. At the ellipsis site 'painters' *only* and *many* were numerically close (-12.9 and -11.6, vs. *most* 18.1 and the baseline 77.1), but the difference was not significant. *Many* and *most* both significantly differed from the baseline only at the second word of the spillover region, 'exhibition'. Subsequently, we determined that without log transformation the assumption of homogeneity of variances was violated (see Vasishth, Chen, Li, and Guo, 2013, for the discussion of the importance of appropriately transforming reading time data).

(16) *Experiment II. Average Log-Transformed Residual Reading Times*




(17) *Experiment II. Significance tests. ANOVAs and Planned Comparisons.*

| | |
|---|---|
| 'and not' | F1(3, 105) = 2.695, $p$ = **.05** by subjects; F2(3, 69) = 4.149, $p$ = **.009** by items<br>*Only* vs. Baseline F1(1,35)= 10.032, $p$ = **.003**; F2(1,23)= 18.655, $p$ < **.001**<br>*Many* vs. Baseline F1(1,35)= 2.445, $p$ = .127; F2(1,23)= 5.78, $p$ = **.025** |
| 'painters' | F1(3, 105) = 5.229, $p$ = **.002** by subjects; F2(3, 69) = 6.083, $p$ < **.001** by items<br>*Only* vs. Baseline F1(1,35)=9.63, $p$ = **.002**; F2(1,23)= 23.482, $p$ < **.001**<br>*Many* vs. Baseline F1(1,35)=8.97, $p$ = **.004**; F2(1,23)= 5.25, $p$ = **.032** |
| 'because' | F1(3, 105) = 5.784, $p$ = **.001** by subjects; F2(3, 69) = 3.996, $p$ = **.011** by items<br>*Only* vs. Baseline F1(1,35)= 12.749, $p$ = **.001**; F2(1,23)= 8.16, $p$ = **.009** |
| 'exhibition' | F1(3, 105) = 3.555, $p$ = **.016** by subjects; F2(3, 69) = 4.035, $p$ = **.01** by items<br>*Only* vs. Baseline F1(1,35)=11.735, $p$ = **.002**; F2(1,23)= 8.485, $p$ = **.008**<br>*Most* vs. Baseline F1(1,35)= 4.777, $p$ = **.036**; F2(1,23)= 6.778, $p$ = **.016** |

Our three predictions have been partially met. *Only* facilitates processing relative to the baseline, and there is a facilitatory effect emerging at a later stage for *most*. *Many* patterns with *only* at the ellipsis site, which indicates that unlike in the baseline condition, ellipsis resolution is easy. This effect could be attributed to either focus association (indicating that the reading in (15b) is more dominant that the reading in (15a)) or to the presence of gender morphology creating an expectation for a masculine syntactic associate (unlike with *most* or *only*). Despite the unclear result for *many*, we have some evidence that the difference between *only* and *most* could be attributed to the difference between obligatory and free/optional focus association. In both cases, an expectation for narrow focus is created, but it is less strong with optional associators which allow different locations for narrow focus.

On the other hand, the fact that the faciliatory effect with *most* is seen later than the ellipsis site means that we do not have direct evidence that the speed-up with *most* vs. the baseline is directly attributable to ellipsis resolution. Furthermore, the early facilitatory effect of *only* can be linked to either obligatory focus association, or the contrast needs of *only* (the exclusive assertion may be creating an expectation for the explicit mention of the alternative set). We designed Experiment III to test this alternative explanation, as well as the question whether the observed differences reflect obligatory and optional/free focus association.

### *3.3 Experiment III*

In Experiment III ($n$=36), we used exactly the same materials and fillers as in Experiments I and II, but we replaced *many*, (13), with *even*, (18), which introduces a presupposition of the existence of an

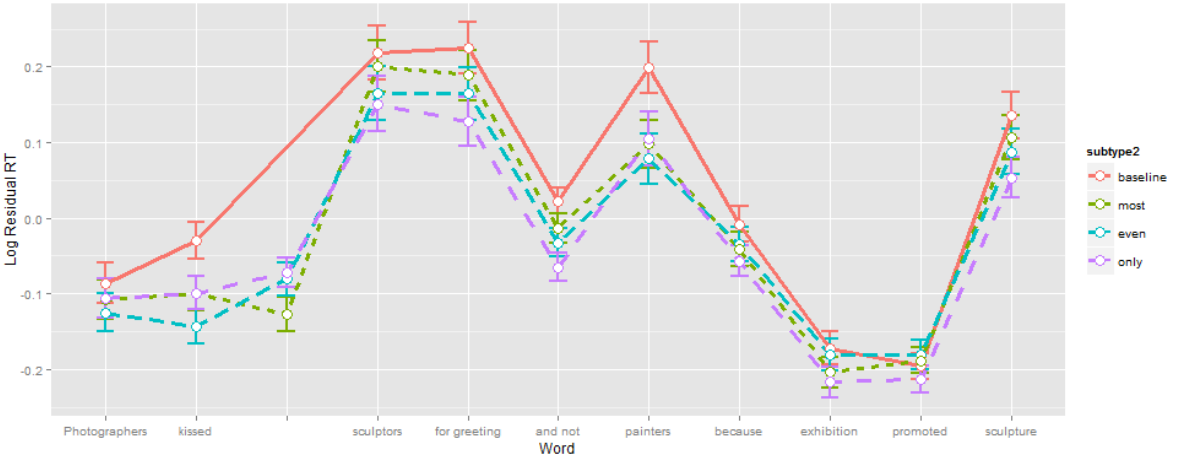
alternative set (and a scalar ordering between its members). Importantly, '*even x and not y*' is less frequent in discourse than *'only x and not y'*.

(18) Fotografowie ucałowali ***nawet*** [rzeźbiarzy]$_F$ na powitanie, a nie [malarzy]$_F$
photographers.$_{Nom}$ kissed **even** sculptors.$_{Acc}$ for greeting and not painters.$_{Acc}$
'Photographers kissed even sculptors for greeting, and not painters.'

We predicted that if Experiment II indicates differences between obligatory and optional associators, *only* and *even* should pattern together, as they both involve obligatory focus association, and *most* should pattern differently. If, on the other hand, the facilitatory effect with *only* in Experiment II comes from its semantics (the assertion of the exclusion of alternatives), then the *only* condition should pattern differently from *even*.

The results confirm that *even* and *only* create expectations that have an early facilitating effect in processing – a significant difference between the conditions appears at the conjunction 'and not', (19)-(20). *Even* and *only* are significantly faster than the baseline at 'and not' and 'painters', and so is *most* at 'painters', (20).

(19) *Experiment III. Average Log-Transformed Residual Reading Times*




(20) *Experiment III. Significance tests. ANOVAs and Planned Comparisons.*

| | |
|---|---|
| 'and not' | F1(3, 105) = 4.286, $p$ = **.006** by subjects; F2(3, 69) = 3.624, $p$ = **.017** by items<br>*Only* vs. Baseline F1(1,35)= 13.71, $p$ < **.001**; F2(1,23)= 6.924, $p$ = **.015**<br>*Even* vs. Baseline F1(1,35)= 6.055, $p$ = **.019**; F2(1,23)= 4.712, $p$ = **.04** |
| 'painters' | F1(3, 105) = 3.196, $p$ = **.026** by subjects; F2(3, 69) = 2.681, $p$ = **.054** by items<br>*Only* vs. Baseline F1(1,35)= 3.696, $p$ = **.06**; F2(1,23)= 4.577, $p$ = **.043**<br>*Even* vs. Baseline F1(1,35)= 7.13, $p$ = **.011**; F2(1,23)= 6.373, $p$ = **.019**<br>*Most* vs. Baseline F1(1,35)= 5.075, $p$ = **.03**; F2(1,23)= 4.769, $p$ = **.039** |
| 'because' | F1(3, 105) = 1.112, $p$ = .348 by subjects; F2(3, 69) = .915, $p$ = .438 by items |

The results confirm that *even* and *only*, which obligatorily associate with focus, create expectations that have an early facilitating effect in the processing of ellipsis. Although with *even* the reference to focus alternatives is part of the presupposition and with *only* it is part of the assertion, both associators in the same way create an expectation for narrow focus.

*Most* similarly facilitates processing relative to the baseline, but the effects are seen slightly later, on the ellipsis site and not on the conjunct. The results are supportive of the proposal that *most* optionally associates with focus.

## 4 Conclusion

Expressions that obligatorily associate with focus can be predicted to create an expectation for the presence of narrow focus in their scope during incremental semantic processing. A sentence with an optional focus associator, on the other hand, does not have to contain a narrow focus to be interpretable. Our experiments show, however, that both obligatory and optional associators create a processing bias for narrow focus indicating that optional focus association is not on par with contextual domain restriction of quantifiers (von Fintel 1994), i.e. the processing of a set of focus alternatives is lexically triggered and not merely the result of the fact that the restrictor variable tends to be resolved to focus alternatives that are contextually salient. The results indicate that both *obligatory* vs. *optional* association with focus, and effects having to do with the status of focus alternatives play a role in the online processing of focus structure. Obligatory focus associating expressions like *only* and *even* behave alike in some respects that set them apart from an optionally focus associating expression like *most.*

**Acknowledgements**

This work has been supported by the NSF DDIG Award #1430803 Doctoral Dissertation Research. We would like to thank Joanna Błaszczak, Anna Czypionka, Jakub Dotlačil, Elsi Kaiser, Dorota Klimek-Jankowska, Colin Philips, Petra Schumacher, as well as the audiences at AMLaP2015, LCQ2015, Linguistischen Arbeitskreis Köln and Experimental Linguistics Talks Utrecht for helpful comments and discussion.

**References**

Bader, Markus and Michael Meng (1999). Subject-object ambiguities in German embedded clauses: an across-the-board comparison. *Journal of Psycholinguistic Research*, 28, 121–143.

Birch, Stacy and Charles Jr. Clifton (1995). Focus, accent, and argument structure: effects on language comprehension. *Language and Speech*, 38(4), 365–391.

Beaver, David and Brady Clark (2008). *Sense and Sensitivity: How Focus Determines Meaning*. Oxford: Wiley-Blackwell.

Carlson, Katy (2013). The Role of Only in Contrasts In and Out of Context, *Discourse Processes*, 50:4, 249-275.

von Fintel, Kai (1994). *Restrictions on Quantifier Domains*. Ph.D. Dissertation, University of Massachusetts, Amherst.

Heim, Irene (1999). Notes on superlatives. Ms., MIT.

Herburger, Elena (1997). Focus and Weak Noun Phrases. *Natural Language Semantics* 5, 53–78.

Krifka, Manfred (1992). A Framework for Focus–Sensitive Quantification. *Proceedings of SALT 2*, 215-236.

Pancheva, Roumyana and Barbara Tomaszewicz (2012). Cross-linguistic Differences in Superlative Movement out of Nominal Phrases. *Proceedings of WCCFL XXX*, 292-302.

Rooth, Mats (1985). *A Theory of Focus Interpretation*. University of Massachusetts, Amherst: Doctoral Dissertation.

Rooth, Mats (1992). A Theory of Focus Interpretation. *Natural Language Semantics* 1: 75–116.

Rooth, Mats (1996). On the Interface Principles for Intonational Focus. *Proceedings of SALT 6*, 202–26.

Sedivy, Julie C. (2002). Invoking discourse-based contrast sets and resolving syntactic ambiguities. Journal of Memory and Language, 46, 341–370.

von Stechow, Arnim (1991). Current Issues in the Theory of Focus. In *Semantik*, ed. Arnim von Stechow and DieterWunderlich, 804–25. Berlin:Walter de Gruyter.

Stolterfoht, Britta, Markus Bader (2004). Focus structure and the processing of word order variations in German. In A. Steube (ed.), *Information Structure: Theoretical and Empirical Aspects*. 259–275.

Stolterfoht, Britta, Angela D. Friederici, Kai Alter, Anita Steube (2007). Processing focus structure and implicit prosody during reading: Differential ERP effects. *Cognition*, 104, 565–590.

Tancredi, Christopher D. (1990). *Not only* even*, but even* only. Cambridge, MA: MIT Press.

Tomaszewicz, Barbara M., Joanna Błaszczak, Roumyana Pancheva (2015a). Focus association revealed in reading times. Poster presented at the *Architectures and Mechanisms for Language Processing* (AMLaP) conference, University of Malta, 3-5 September 2015.

Tomaszewicz, Barbara M., Joanna Błaszczak, Roumyana Pancheva (2015b). Focus association revealed in reading times. Poster presented at the *Workshop on Linguistic and Cognitive Aspects of Quantification* (LCQ2015), Budapest, 16-17 October 2015.

Vasishth, Shravan, Zhong Chen, Qiang Li and Gueilan Guo (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE, 8.*

# Are false implicatures lies? An experimental investigation

Benjamin Weissman and Marina Terkourafi
University of Illinois at Urbana-Champaign

Meibauer (2014: 103, 125) proposed a definition of lying according to which an assertion *p* that carries an implicature *q* counts as a lie if *p* is true but *q* is false, for both particularized and generalized conversational implicatures. This stands in contrast to a majority of definitions of lying, which tend not to count false implicatures as lies. Saul (2012: 37, 118) regards false implicatures as instances of "mere misleading", and Fallis (2009: 40) also states explicitly that false implicatures are not lies.

Meibauer's definition can be rather complex to assess empirically, not least because of the wide variety of kinds of conversational implicatures. In addition to the well-known distinction between generalized and particularized conversational implicatures (GCIs and PCIs) (Grice 1975), Neo-Gricean accounts have distinguished between different types of GCIs. For instance, Levinson (2000) proposes three heuristics: Q ("What isn't said, isn't" applying to scalars and clausals—meaning that if a weaker expression is used, this implies that the stronger one is not the case); I ("What is expressed simply is stereotypically exemplified"); and M ("Marked message indicates marked situation"), which give rise to conversational implicatures by default but may still be canceled in special circumstances. Recently, Doran et al. (2012) set out to test Levinson's proposal experimentally, attempting to determine where people draw the line between what is said and what is implicated for different types of GCIs. The researchers found that linguistic expressions realizing the three types of GCIs proposed by Levinson (Q, I, and M) do not pattern neatly into three blocks but rather exhibit significant degrees of permutation in terms of how readily respondents are prepared to incorporate the resulting GCIs into what is said.

With these two foundations in mind, the research questions for the current study were formulated as follows:

(RQ1) —Do people consider false implicatures to be lies?

(RQ2) —Do implicatures of the same kind, theoretically, pattern together in terms of people's natural lie judgments?

The current study aimed to find out whether, and to what extent, people consider different types of implicatures to be lies by asking participants to judge if a speaker who licenses a false implicature in a given context has lied. Eleven examples of GCIs (adapted from Doran et al. [2012]—four Q-based, four I-based, three M-based) were tested alongside four examples of PCIs (one each, based on Quality, Quantity, Relation, Manner). First, a norming study was carried out to confirm that target utterances did indeed give rise to the intended implicatures. Fourteen subjects were asked to rate the

 In: Fabienne Salfner and Uli Sauerland (eds.), *Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 162–169.

likelihood of implicature *q* arising given an assertion *p* in a generic context. Implicatures scoring consistently over 2.5 (on a 1–4 Likert scale) were used in the main experiment.

The experimental methodology followed that of Arico & Fallis (2013) and Coleman & Kay (1981). Target utterances were alternately embedded in three types of contexts, strongly biasing their interpretation as either a "straightforward truth" (henceforth: T condition) or a "straightforward lie" (L), or generating a "false implicature" (X). Sixty native speakers of English, naïve to the theoretical underpinnings of the experiment, were recruited through Amazon MTurk and surveyed on Qualtrics. Participants were asked to judge on a 1–7 Likert scale whether the speaker in the story has lied. On this scale, 1 indicated "absolutely not a lie"; 4 indicated "neither a lie nor not a lie"; and 7 indicated "absolutely a lie". Since our goal was to tap into the participants' own naïve intuitions as native speakers of English, we provided no definition of lying on which they were supposed to base their responses. Each subject saw 15 stories in total, five in each condition, and no subject saw the same target utterance in more than one condition. Three lists were arranged and counterbalanced so that each condition of each story was ultimately seen by 20 participants.

An example of one story in each of the three conditions is given below in (1)-(3). In each case, the participants were asked to answer the question in (4).

(1) **T**: Rumors have spread about an incident in the art studio yesterday. Alex was in the studio all day and saw Sarah, frustrated with a project, pick up a hammer and use it to smash a statue to bits. The following day, Alex talks about the incident.

Mark: I heard Sarah had a meltdown in the art studio yesterday! What happened?
**Alex: Yes. In a fit of rage, Sarah picked up a hammer and broke a statue.**

(2) **L**: Rumors have spread about an incident in the art studio yesterday. Alex was in the studio all day and saw Sarah, frustrated with a project, fling a paintbrush across the room, breaking a window. Later that night in the studio, Alex accidentally tipped over a statue, causing it to smash all over the floor. The following day, Alex talks about Sarah's incident.

Mark: I heard Sarah had a meltdown in the art studio yesterday! What happened?
**Alex: Yes. In a fit of rage, Sarah picked up a hammer and broke a statue.**

(3) **X**: Rumors have spread about an incident in the art studio yesterday. Alex was in the studio all day and saw Sarah, frustrated with a project, pick up a hammer and then walk over to a statue and kick it over with her foot, causing it to smash all over the floor. The following day, Alex talks about the incident.

Mark: I heard Sarah had a meltdown in the art studio yesterday! What happened?
**Alex: Yes. In a fit of rage, Sarah picked up a hammer and broke a statue.**

(4) Did Alex lie? (1–7 Likert scale given.)

A linear mixed-effects model run on the data using R (R Core Team 2015) indicated that in the target false implicature condition (X), 11 of the 15 stories were rated significantly lower than the numerical "lie" cutoff (a 4 on the 7-point Likert scale) with >95% confidence, and only two were rated significantly higher. Moreover, GCIs based on the same Levinsonian heuristic did not pattern together. Figure 1 below displays the mean rating for each of the 15 stories in this condition.

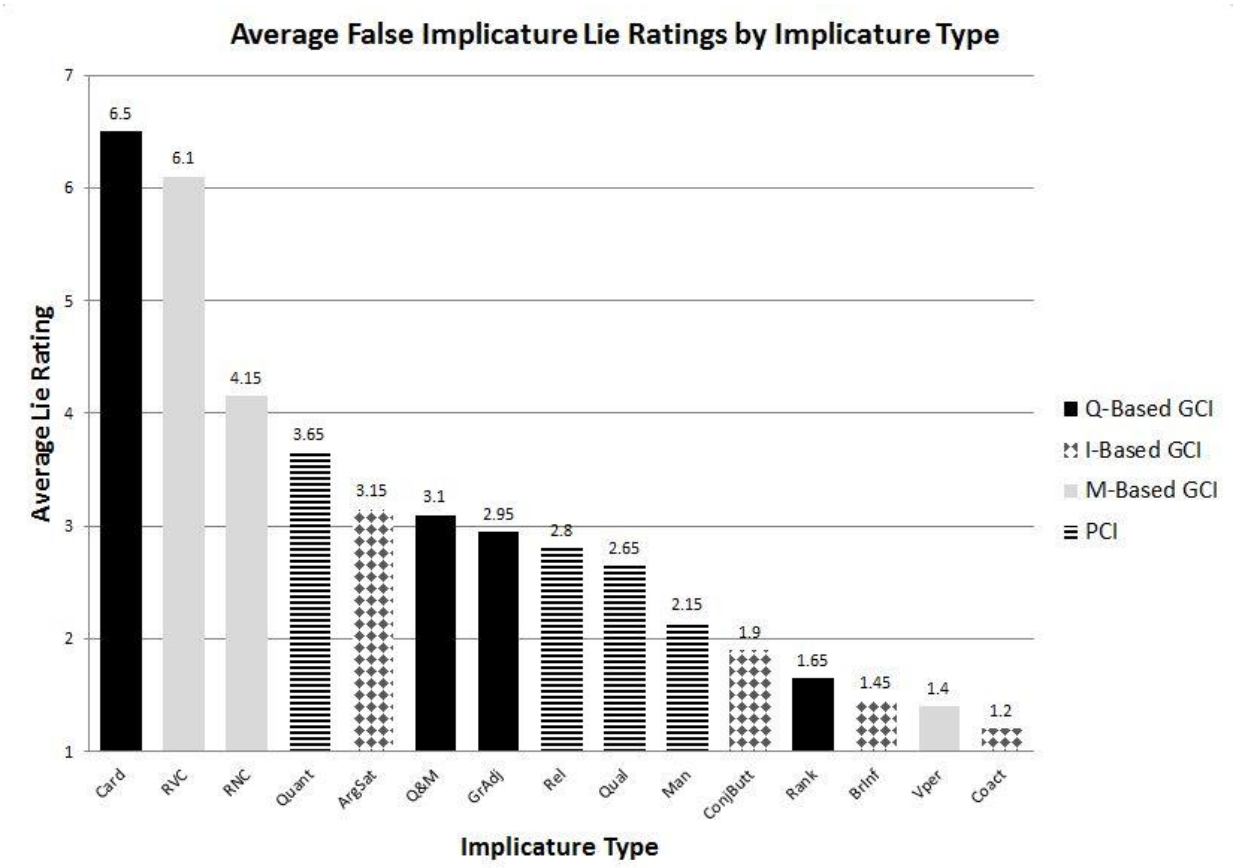



**Figure 1** —Average lie ratings (1–7) in "False Implicature" condition for each type of implicature tested; 4 is the cutoff line between "lie" and "not a lie".

Not only were mean ratings mostly below the lie cutoff on the 7-point Likert scale, this result is also upheld if the Likert scale gradient is transformed into a simple binary "lie" vs. "not lie" choice. Figure 2 below shows the percentage of people that regarded each story to be a lie, regardless of how strongly they felt this to be the case. Any rating from 5–7 would be considered a lie of some sort; these responses are tallied in the chart below.

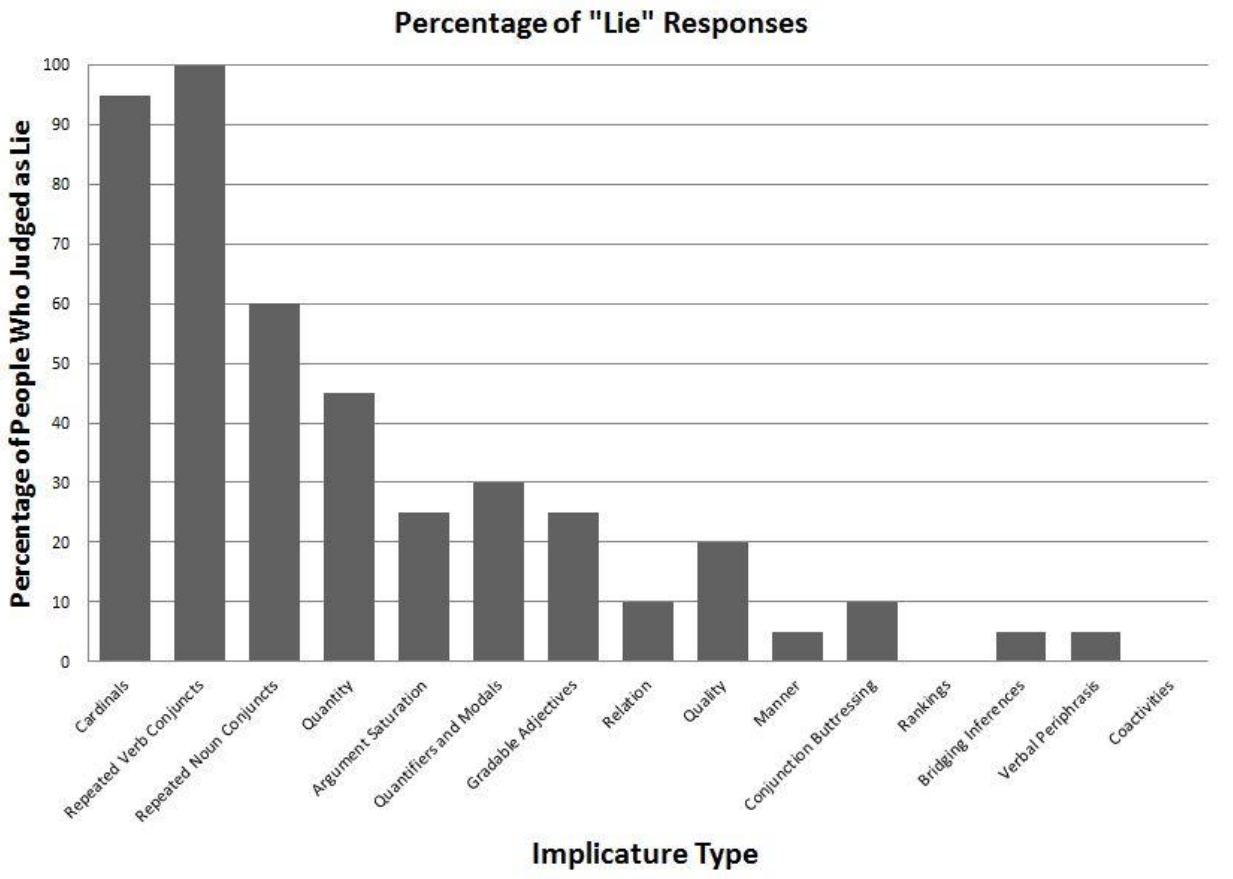



**Figure 2** —Percentage of respondents who judged each false implicature story to count as a lie (rating >4 on Likert scale).

Cardinals and Repeated Verb Conjuncts, the two stories rated significantly higher than the lie cutoff, were also considered to be lies by a vast majority of the participants. The Cardinals story in the false implicature condition is given below in (5).

(5) Jenn made a half dozen cupcakes in the morning and left them out to cool on the counter in her apartment. Her roommate, Molly, ate all of them. Jenn comes back in the evening to find that the cupcakes are gone, so she asks Molly about them.

Jenn: Did you eat my cupcakes?
**Molly: Yes, sorry, they were so good! I ate three.**

Cardinal numbers, though considered to generate scalar implicatures by many Neo-Griceans, such as Levinson (2000: 87-90), have also been demonstrated perhaps not to carry these implicatures at all. In this experiment, the cardinal term was almost unilaterally regarded as having exact semantics and not carrying a scalar implicature (i.e., "three" means *exactly three*, and not *at least three*). This finding joins recent

experimental work (e.g., Huang et al. 2013) in lending support to hypotheses like that of Carston (1998), which regard cardinal numbers as having exact semantics.

Repeated Verb Conjuncts, the other category that was rated significantly higher than the lie cutoff, with consensus, is not without controversy itself. The RVC story is given in (6) below.

**(6)** Last night, Liza and Kelsey went to a bar for a night out. Liza was a responsible designated driver, and had only one drink. Kelsey, feeling bad, did not drink a lot, and had only two beverages throughout the night. The following day, Liza talks about the night.

Abby: How was the bar last night? Were you and Kelsey drunk?
**Liza: I drove, so I wasn't, but Kelsey drank and drank.**

It is possible that these cases of Repeated Verb Conjuncts have in fact been grammaticalized as constructions in the language (cf. Goldberg 1995). Reduplication is a common construction in many languages around the world; repeating a word or a morpheme is often used to represent increased intensity or repetition (cf. Moravcsik 1978) as in Estonian (Erelt 2008) and various pidgins and creoles (Bakker & Parkvall 2005). According to Levinson, these conjuncts in English relate to the M-Heuristic ("Marked message indicates marked situation"), but if it is the case that "drank and drank" is a grammaticalized construction that automatically means "drank an excessive amount" or "drank more than is expected/normal", then this would actually be closer to Levinson's I-Heuristic ("What is expressed simply is stereotypically exemplified"). This would still make it a generalized conversational implicature but not of the same type.

The results of this experiment indicate that the unenriched literal meaning of an utterance is what people tend to consider when judging if an utterance is a lie or not, a result wholly in line with the theory proposed by Saul (2012), who argues that a lie must be part of what is said and not simply part of what is communicated/implicated. In order for a statement like "I won the lottery and bought a house" to consistently be judged not a lie when the house was in fact bought first ("conjunction buttressing" story; mean rating = 1.9, SD = 1.34, only 10% judged it a lie in X condition, significantly different from L condition rating with >95% confidence), the GCI that the events happened in the order they are described cannot possibly be what is being judged, even though this GCI is clearly generated. In the norming study, this utterance was consistently judged to include the implicature that the events happened in this order (7/7 participants indicated that the utterance means that the events were more likely to happen in this order). This discrepancy suggests that when asked if the speaker has lied, people are judging the unenriched literal meaning of the utterance—it is true that I won the lottery and it is true that I bought a house. Since both of these are, in fact, true, the utterance is judged not to be a lie. If the implicature

is what was being considered in these lie judgments, the ratings would be higher (toward the lie end of the scale).

Given this finding, the results of this experiment can be used as an alternate avenue to test the question raised and tested by Doran et al. (2012) about what counts as part of what is said. Comparing theirs and our own experiments, it is evident that the order of GCI types follows to some extent their findings (see Figure 3).

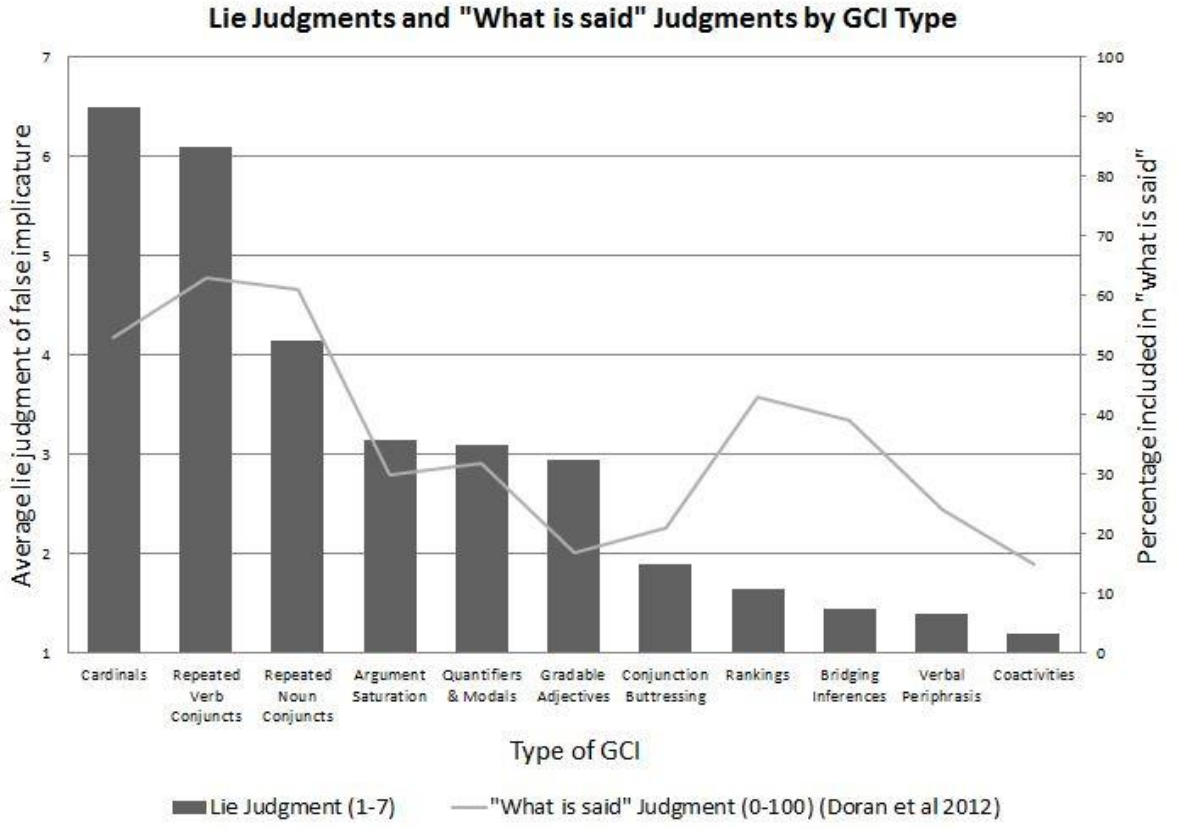



**Figure 3** —A comparison of results from the current experiment (columns) to results from Doran et al. (2012). Columns indicate mean rating (1–7) of false implicatures as lies for each GCI type; line indicates percentage of inclusion as part of "what is said" for each GCI type as found by Doran et al. (2012) in their *Literal Lucy* condition.

The data reported from Doran et al. (2012) are only from one of the three conditions they tested—the condition in which participants were asked to make a truth value judgment from the perspective of an extremely literal-minded character named *Literal Lucy*. The other two conditions were *literal*, in which participants were asked to make the truth value judgment literally, and *baseline*, in which no specific instruction was given as to how they should make their judgments. The baseline condition is most similar to the task in the current experiment. Though the full set of baseline data is neither reported in the paper nor available now, the authors do mention in a footnote that "the relative ordering of frequency of incorporation for the different GCI types was

preserved: in neither of the other two conditions was the rank order of a GCI type more than two positions away from its rank ordering in the *Literal Lucy* condition" (Doran et al. 2012: 143, fn. 15). This statement leaves open the possibility that results from the baseline condition, the condition most similar to the current experiment, may look even more like the results presented here. The one data point from the baseline condition that is given is the average response to the Cardinals stories, which rose to 92% incorporation in that condition. That would place it at almost exactly where the current experiment's Cardinals story's ratings landed. Minor reordering of the categories in the baseline could assuage some of the relative discrepancies between the two sets of results, but on the whole, even here, they do follow the same general pattern.

Doran et al. only tested GCIs because the incorporation of PCIs in what is said is not a debated issue. In order to fully test Meibauer's proposed definition, however, it was necessary to include PCI examples in the stimuli. The Quantity-based false implicature scored rather highly, the fourth-highest-rated story of all. Even this story, however, cannot be viewed as being completely a lie—only 45% of respondents scored it as a lie and its average rating (3.65) was slightly less than the lie cutoff of 4. With the other three PCI types scoring lower than that, it can be deduced that most people considered false PCIs not to count as lies, in contrast to Meibauer's proposed definition.

Our results suggest that the amount of variation among types of implicatures makes a universal statement regarding the treatment of all implicatures as lies (or not) unlikely. Two types of implicatures were uniformly and strongly rated as lies; however, several other types were uniformly and strongly rated as *not* lies. Of the 15 types of implicatures tested, overall, more types were taken not to be lies than to be genuine lies, indicating that Meibauer's (2014) proposed definition, while theoretically grounded, may fall short of the folk definition of lying that most native English speakers use. In addition, our methodology provides an additional avenue to test the claim of Doran et al. (2012) that a different classification of GCIs from the three-pronged one proposed by Levinson (2000) is warranted by the experimental facts.

The next stage of this research will include more examples of each of the 11 Doran et al. (2012) implicature types and the four PCIs. It is evident from this experiment and others that not all implicatures behave alike. With that in mind, further testing these categories by including more examples of each type is a necessary step to draw more-robust conclusions about the behavior of different implicature types. Our current results suggest that most false implicatures are not considered to be lies, but further testing will enable us to refine and further clarify this suggestion.

Email for correspondence: bpweiss2@illinois.edu

**References**

Arico, Adam and Fallis, Don 2013. Lies, damned lies, and statistics: An empirical investigation of the concept of lying. *Philosophical Psychology.* 26(6), 790-816. http://dx.doi.org/10.1080/09515089.2012.725977

Bakker, Peter and Mikael Parkvall. 2005. Reduplication in pidgins and creoles. In *Studies on Reduplication*, Bernhard Hurch (ed.), 511-532. Berlin: Mouton De Gruyter.

Carston, Robyn. 1998. Informativeness, relevance, and scalar implicature. In *Relevance theory: applications and implications*, Robyn Carston & Seiji Uchida (eds.), 179-236. Amsterdam: John Benjamins.

Coleman, Linda and Paul Kay. 1981. Prototype semantics: The English verb *lie*. *Language* 57 (1), 26-44. http://dx.doi.org/10.1353/lan.1981.0002

Doran, Ryan; Ward, Gregory; Larson, Meredith; McNabb, Yaron; and Baker, Rachel E. 2012. A novel experimental paradigm for distinguishing between what is said and what is implicated." *Language* 88(1), 124-154. http://dx.doi.org/10.1353/lan.2012.0008

Erelt, Mati. 2008. Intensifying reduplication in Estonian. *Linguistica Uralica* 44 (4), 268-277. http://dx.doi.org/ 10.3176/lu.2008.4.02

Fallis, Don. 2009. What is lying? *The Journal of Philosophy* 106 (1), 29-56. http://www.jstor.org/stable/20620149

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Grice, H. Paul. 1975. Logic and conversation. *Syntax & Semantics Vol. 3,* Cole and Morgan eds., 41-58. New York: Academic Press.

Huang, Yi Ting; Spelke, Elizabeth; and Snedeker, Jesse. 2013. What exactly do numbers mean? *Language Learning and Development* 9, 105-129. http://dx.doi.org/10.1080/15475441.2012.658731

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature.* Cambridge, MA: MIT Press.

Meibauer, Jörg. 2014. *Lying at the semantics-pragmatics interface*. Berlin: Mouton De Gruyter.

Moravcsik, Edith. 1978. "Reduplicative Constructions," in Joseph H. Greenberg, (ed.), *Universals of Human Language,* Vol. 3: Word Structure, Stanford: Stanford University Press, 297-334

R Core Team 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Saul, Jennifer. 2012. *Lying, misleading, and what is said*. Oxford: Oxford University Press.

# In a manner of speaking: an empirical investigation of Manner Implicatures *

Elspeth Wilson
*University of Cambridge*

Napoleon Katsos
*University of Cambridge*

**Keywords:** Experimental pragmatics, implicature, manner

## 1 Introduction

Grice proposed that not only the content of what is said but also the *form* can give rise to implicatures (Grice 1975): Manner inferences derive from observing or violating the maxim 'be perspicuous'. In this study we address the questions: (i) do Manner implicatures pattern like Quantity inferences – are they similarly robust and commonly derived by hearers? and (ii) do Manner implicatures, like Quantity, exhibit varying degrees of conventionalization? In addition, we evaluate the kind of experimental paradigms that are suitable for investigating manner implicatures.

## 2 Context

Although the submaxims of Manner ('avoid obscurity of expression', 'avoid ambiguity', 'be brief', and 'be orderly') are diverse, for Grice the defining feature is the relation "not to what is said but, rather, to *how* what is said is to be said" (Grice 1975: 46). Some post-Gricean theorists have subsumed Manner under other principles (Horn 1984), while others have maintained it, albeit in modified form (Levinson 2000, Franke 2009). This latter approach seems justified on theoretical grounds at least, because Manner implicatures can be distinguished from Quantity and Relevance precisely because it is the linguistic form, not the content, which triggers an inference.

Levinson (2000) argues that marked forms are used by speakers as a 'shortcut' to a marked meaning (M-forms give rise to M-implicatures), whereas the unmarked I-form produces a stereotypical interpretation (I-implicature). We adopt Levinson's terminology here, because it allows us to draw a potentially useful distinction between the two forms and corresponding implicatures, while acknowledging that

* We thank the audience at the Experimental Semantics & Pragmatics workshop at University of Cambridge, June 2015, for their comments, and Dimitris Alikaniotis for technical advice. Elspeth Wilson is funded by an ESRC studentship.




*Pre-proceedings of 'Trends in Experimental Pragmatics'*, XPRAG.de, Berlin, Germany, pp. 170–176.

I-implicatures are not included as prototypical implicatures in many typologies. Besides Levinson's Generalised Conversational Implicatures, there are of course Particularised Conversational Manner Implicatures, which are not tied to systematic alternations between forms expressing similar meanings, although the distinction is subject to much debate (e.g., Degen 2015) – see Table 1 for examples.

Theoretically, there are interesting parallels with Quantity: both M- and Quantity implicatures involve reasoning with reference to an alternative that was not said – either an alternative that is 'stronger' in meaning, or more lengthy. Both also include more or less conventionalised instances (traditionally, GCIs and PCIs). To date, however, there has been almost no empirical research on Manner.

| | **Context** | **M-ending** | **I-ending** | **M-implicature** | **I-implicature** |
|---|---|---|---|---|---|
| **GCI** | Nick and Dan were watching a history programme, but Nick fell asleep. Afterwards he asked, "What happened at the end?" Dan replied, | "The invaders caused the villagers to die." | "The invaders killed the villagers." | The invaders killed the villagers but indirectly, by introducing disease. | The invaders killed the villagers directly. |
| **PCI** | Jamie got home from his grandmother's. His mum asked, "Did she give you a drink?" He answered, | "She put a teabag into a cup and poured over boiling water." | "She gave me a cup of tea." | She gave me a cup of tea, only it didn't taste like tea. | She gave me a normal cup of tea. |

**Table 1** Example of experimental items

## 3 Experiments

Experiments were conducted via Prolific Academic, a UK-based crowd-sourcing platform for research. Participants were British English speakers born in UK, which was important as linguistic form is sensitive to variety, possibly even at the dialectal level. Pretests established scenarios consisting of short dialogues, which were rated as equally natural with both a 'marked' ending (the M-implicated meaning) and an 'unmarked' ending (the I-implicated meaning). There were 8 items with marked forms with lengthy paraphrases (PCIs), and 8 GCIs with lexical alternatives (modals, causatives, negation of contradictories).[1]

### 3.1 Experiment 1

#### 3.1.1 Procedure

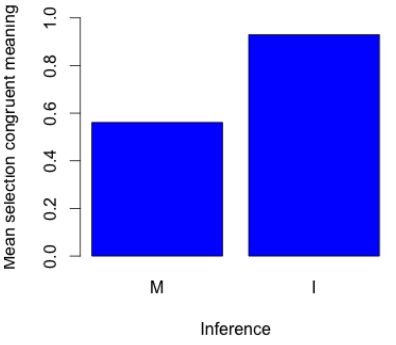
Participants ($N = 39$) in two groups were shown a series of scenarios, each of which ended with an utterance with an M-form or I-form, with its text in red. Below,

1 For the full stimuli set, see elspethwilson.uk/research/resources/

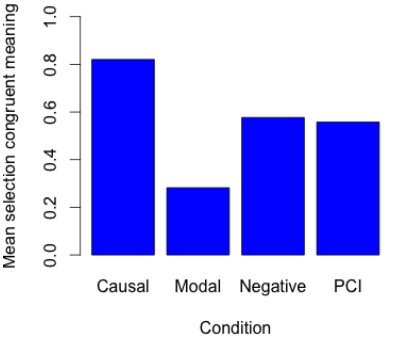
they read the M-implicated and I-implicated meanings, and were asked to select the sentence that was most similar in meaning to the part of the scenario in red. Scenarios were presented in randomised order, and each participant saw each item in only one condition. If participants are sensitive to M-implicatures, we predicted higher selection rate of M- rather than I-implicated meanings for M-forms, and vice versa.

### 3.1.2 Results

The results (Fig. 1) show a significantly higher rate of selection of I-implicated meanings for I-forms (92%) than M-implicated meanings for M-forms (56%), Wilcoxon signed-rank by-item: $W = 2$, $p < 0.01$, $r = -0.98$. There is no difference between GCI and PCI items for either M- or I-forms (M GCI 56%, M PCI 56%, I GCI 94%, I PCI 91%), Wilcoxon signed rank: $W = 11$, $p = 0.92$, $r = -.04$; $W = 17.5$, $p = 0.55$, $r = -.21$. Numerically, the results for different types of M-form suggest different strengths of cues across subtypes (Fig. 1b), with causatives seemingly the strongest cue to a marked meaning (82%) and modals the weakest (28%). Binary choice tasks can, however, mask subtle differences in participants' responses to the stimuli; Experiment 2 uses a rating scale, which might show more sensitivity to pragmatic interpretations.




(a) Experiment 1: Mean selection of I-implicated meaning for I-form, and M-implicated meaning for M-form




(b) Experiment 1: Mean selection of M-implicated meaning for M subtypes

**Figure 1** Experiment 1 results

### 3.2 Experiment 2

#### 3.2.1 Procedure

In a $2 \times 2$ (form $\times$ meaning) design, participants ($N = 80$) were asked to rate on a scale of 1 to 7 how similar a sentence with an I- or M- implicated meaning was to the part of the scenario in red (I- or M-form), based on Degen's (2015) paradigm for Scalar Implicatures. Four groups of participants saw each item in one condition. We predicted that ratings would be ordered: unmarked form/unmarked meaning > marked form/marked meaning > marked form/unmarked meaning > unmarked form/marked meaning. The inference to the stereotypical from the I-form is arguably closest in form and meaning to the literal meaning of the utterance, and so rated highest; the marked form introduces information not implicated by an unmarked form, so is rated the lowest.

#### 3.2.2 Results

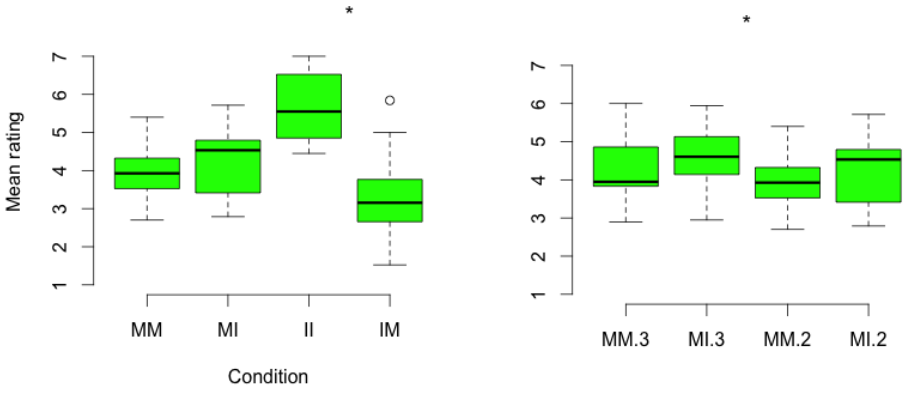
The results (Fig. 2a) show a significant effect of condition (by-item analysis, Friedman test $\chi^2(3) = 33.8$, $p < 0.01$). Planned pairwise comparisons show that the congruent and incongruent conditions (MM and MI ) are not significantly different (Wilcoxon signed-rank by-item: $W = 45$ , $p = .25$, $r = -0.29$), while II and IM are (Wilcoxon signed-rank by-item: $W = 136$, $p < .01$, $r = -1.04$). Planned pairwise comparison of GCI and PCI subtypes for each condition again found no significant differences (Wilcoxon signed-rank – MM: $W = 26$, $p = 0.26$; MI: $W = 24$, $p = 0.46$; II: $W = 32$, $p = 0.055$; IM: W = 18.5, $p = 0.94$).

#### 3.2.3 Discussion

Experiments 1 and 2 show that M-implicatures do not seem to be as robust as, for example, the most robust Scalar Implicatures, in either a binary choice or rating paradigm: in Experiment 1, the M-implicated meaning is selected for the M-form only around half the time, and in Experiment 2, there is no significant difference between the ratings for M congruent and incongruent conditions (seeing the M form and M-implicated meaning or I-implicated meaning). In contrast, in picture-selection tasks, adults tend to be at ceiling for 'pragmatic' interpretation of SIs with 'some' and ad hoc Quantity implicatures (Katsos & Bishop 2011, Stiller et al. 2015), while in a rating task, Degen (2015) finds some utterances with 'some' whose implicated meaning 'some but not all' is consistently given a high rating towards ceiling on a 1-to-7 Likert scale.

Two mitigating explanations are conceivable: firstly, in Experiment 2, the presence of II and IM items rated at the top and bottom of the scale could leave partici-

pants with less space in the middle of the scale to distinguish MM and MI conditions. Experiment 3 was run to address this concern. Secondly, M-Implicatures could differ from, say, Scalar Implicatures in that the implicated meaning does not have such a direct relation in form or meaning to what is said (e.g., 'some' > some but not all). This could mean that participants did infer an M-implicature, but that the explicitly stated M-implicated meaning in the tasks did not match their inferences. Furthermore, the I-implicated meaning is included in the M-implicated meaning, which may mean it is chosen more often or rated more highly than the M-implicated meaning. Experiment 4 was carried out to ascertain the variety in participants' inferred meanings.




(a) Experiment 2: Mean rating of implicated meaning per condition

(b) Experiment 3: Mean rating of implicated meaning for MM and MI conditions, Exps 2 and 3

**Figure 2** Experiment 2 and 3 results

### 3.3 Experiment 3

This was the same as Experiment 2, but with only MM and MI conditions ($N = 38$). A by-item analysis of the results (Fig. 2b) using Friedman's ANOVA and planned pairwise comparisons indicates that the only significant difference between MM and MI conditions in Experiments 2 and 3 was between Experiment 2 MI and Experiment 3 MM (Wilcoxon signed-rank, $W = 112$, $p = 0.02$, $r = -0.58$). This suggests the $2 \times 2$ design of Experiment 2 did not affect the distribution of participants' ratings.

### 3.4 Experiment 4

Participants (N = 41) were asked to write what they thought the speaker of the utterance in red meant. Participants saw both M- and I-forms (but only one condition per item), and, in the instructions, were given two example interpretations, one for an I-form and one for an M-form.

Results were coded as 'M-implicature' or 'no M-implicature', with 39% M-implicature responses for M-forms, and no difference between GCI and PCI types (40% and 38% respectively). Qualitatively, a wide variation in interpretations of M-forms was observed, even given the context of the short dialogues. Sometimes interpretations were even in opposite directions – see Table 2 for examples.

This suggests that one factor leading to low rates of M-implicatures in binary choice and ratings tasks, compared to I-implicatures, is that, given minimal background context in the experiment, participants arrive at varying M-implicated interpretations, which do not necessarily match the one explicitly stated. However, even with free response, M-implicated meanings constituted fewer than half of all responses to M-forms, suggesting that, at least in written mode, M-implicatures are not as robust as other inferences, such as the most robust examples of Scalar Implicature (cf.Van Tiel et al. 2014).

| **Item (critical utterance)** | **M inference** | **No M inference** |
|---|---|---|
| "I'm not unpleased by what you've done." | That the feature he has written is satisfactory, but that it is not amazing. | I'm pleased. |
| | I am quite please by what you have done. | I'm pleased with your work. |
| | His boss is not overly impressed but thinks his work is acceptable. | Mick's boss was happy with his feature. |
| "She put a teabag into a cup and poured over boiling water." | Jamie's grandmother makes tea so badly it is just boiling water over a teabag. | Jamie explained how his grandmother made a cup of tea. |
| | Jamie might like his tea with milk and sugar, neither of which his grandma has. | Jamie's Grandmother made a cup of tea for Jamie. |
| | She made little effort. She couldn't be bothered. | She made me tea. |
| "The invaders caused the villagers to die." | At the end of the history programme some villagers died as a result of the actions of some invaders. | The invaders killed the villagers. |
| | The invaders killed the villagers indirectly. They all died in the end. | They killed them. |

**Table 2** Examples of responses in Experiment 4

## 4 Conclusion

Speakers may not be as sensitive to Manner triggers as theories such as Levinson (2000) would suggest, although there is evidence that speakers sometimes derive inferences suggested in the literature. Some of these findings may be particularly due to the way binary choice and ratings contexts interact with Manner. Further research could investigate whether, for example, prosodic cues in speech affect the reliability of manner inferences, and whether other experimental paradigms, such as those measuring sensitivity to maxim violation, might reveal a difference between proposed GCI and PCI implicatures. Finally, as the results of Experiment 1 suggested that different lexical triggers within the GCI category may be more or less reliable cues, future research could also investigate these subtypes with a sample size and design suitable for statistical analysis, to assess whether M-implicatures constitute a homogenous class.

## References

Degen, Judith. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics* 8(11). 1–55.

Franke, Michael. 2009. *Signal to act: Game theory in pragmatics*: University of Amsterdam PhD Thesis.

Grice, H. Paul. 1975. Logic and conversation. In Robert Stainton (ed.), *Perspectives in the philosophy of language*, 41–58. Broadview Press.

Horn, Laurence. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schriffen (ed.), *Meaning, form, and use in context: Linguistic applications*, 11–42. Georgetown University Press.

Katsos, N. & D.V.M. Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition* 120(1). 67–81.

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

Stiller, Alex J., Noah D. Goodman & Michael C. Frank. 2015. Ad-hoc implicatures in preschool children. *Language, Learning & Development* 176–190.

Van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina & Bart Geurts. 2014. Scalar diversity. *Journal of Semantics* 1–39.

Elspeth Wilson
Department of Theoretical and Applied Linguistics
9 West Road
Cambridge
ep321@cam.ac.uk

Dr Napoleon Katsos
Department of Theoretical and Applied Linguistics
9 West Road
Cambridge
nk248@cam.ac.uk